Computational Biology meets Data Science

Gabriela Cohen Freue (University of British Columbia),
Robert Gentleman (Harvard Medical School),
Maribel Hernandez-Rosales (Center for Research and Advanced Studies),
Andrew McDavid (Ozette Technologies)

May 7, 2023 – May 12, 2023

1 Overview of the Field

Recent advances in high-throughput technologies have opened new venues for researchers and scientists to dig deeper into biology and understand underlying causal mechanisms of diseases and conditions. Computational biology, which fuses components from mathematics, computer science and statistics, has generated a wide variety of promising tools to explore, visualize and extract knowledge from these complex data. However, this diversity has also presented challenges, both practical and methodological. For practitioners, accessing and choosing the appropriate tool to respond to sophisticated biological questions remains a challenge. For developers, too often tools are developed without comparison or context of related methods.

These shortfalls have been exacerbated by the increasing scale and dimensionality of biological data. For example, with the recent development of single-cell genomic and proteomic technologies, scientists can now study the function of individual cells, which can unmask important underlying mechanisms from complex heterogeneous samples. In turn, new methodologies are required to analyze different types of data and address new pertinent challenges (for example heavily zero-inflated distributions in single cell RNAseq data. While it is undoubtedly true that computational biology has grown considerably in the last decade, the field continues to evolve and respond to new emerging challenges.

With comparable momentum, data science is becoming an independent discipline. While there is not a clear consensus about the definition of this new discipline, it is unquestionable that data science is providing an overarching framework for innovative and computational efficient solutions for the collection, management, processing, analysis, visualization, and interpretation of complex data. From generative to predictive models, mathematical to algorithmic optimization, static to interactive visualization tools, local to cloud computing and super-computing centers, data science is contributing with creative solutions in data-intensive research and sparking current advancement in science.

Overall, we observe that the hybridization of data science principles to computational biology has already led to important advances, thus increasing the interaction between these disciplines can help address remaining and new emerging challenges. Our workshop has explored different areas of impact of data science on computational biology with special focus on: *i) the eagerness to adopt predictive and unsupervised models from machine learning for the analysis and integration of multi-omics data, ii) the readiness to share resources from large-scale data sets to reliable software, and iii) the consensus on the relevance of reproducible and transparent workflows*.

2 Recent Developments and Open Problems

The impact of data science on computational biology has become evident in many areas of study. In particular, we highlight three main areas below, which this workshop has focused on:

i. Eagerness to adopt predictive and unsupervised models from machine learning for the analysis and integration of multi-omics data

In recent years, we have evidenced the generation of terabytes to petabytes of biological data and a wider and more diverse characterization of complex samples. Undoubtedly, these big biomedical data offer an unprecedented opportunity to improve our understanding of compound biological processes, as well as our success in predicting patient survival outcome and treatment response. However, there is a risk for false discoveries and unsupported conclusions if analyzed with inappropriate bioinformatic and computational methods.

- **Supervised methods from machine learning for biological problems.** Uses of supervised techniques in machine learning have traditionally included predicting risk of disease by exploiting case-control datasets that included genomic sequence or transcriptomics data, such as The Cancer Genome Atlas project or Wellcome Trust Case Control Consortium. Increasingly, machine and deep learning techniques have been proposed to attack fundamental questions in biology. A key challenge is providing interpretable predictions that inform on the mechanisms or laws that underlie the predictions. The workshop gathered experts in various machine learning and statistical areas to share what works, what does not, and why.

- **Clustering and dimension reduction for genomic data.** The Human Cell Atlas and related efforts are attempting to use single cell RNA sequencing to characterize previously cryptic subpopulations of cells in a variety of human tissues. These subpopulations are primarily defined by applying clustering approaches to the gene expression profiles of individual cells. Because of the difficulty of defining a ground truth, and the rapid adoption and evolution of novel genomic assays, an incredible diversity of methods have been proposed. By a recent count of tools and publications on the database scrna-tools.org, over 260 methods have been devised to perform unsupervised clustering on single cell RNAseq but only a few attempts have been made to synthesize and benchmark these tools. Given the rate that scientific hypotheses are being made on the basis of unsupervised clustering, it seems important to expand these efforts. A central topic of this workshop was to review and discuss up-to-date methods to mitigate the problem of unsupported conclusions in Science.

- **Data integration.** Researchers can now simultaneously gather or generate different -omics data to learn about basic biology, identify drug targets, or find biomarkers for various predictive models. For example, in cancer research large sets of tumors and preclinical cancer models have been characterized at the epigenomic, transcriptomic and proteomic levels. The vast majority of existing research has inspected each genomic data in isolation and the development of integrative analysis across multiple platforms to simultaneously study multi-omics data has gained some traction, but largely is still in its early stages. Innovative statistical and computational methods including integrative clustering, classification, graphical models, pathway/network analysis are needed to reveal plausible biological mechanisms and interactions and simultaneously study multiple aspects of diseases. One of the goals of the workshop was for researchers that are actively pursuing this field to get an opportunity to discuss the relative merits and limitations of different approaches to integration, and to take full advantage of the growing data sources.

ii. Readiness to share resources from large-scale data sets to reliable software

Large-scale, public databases are being generated by international collaborative projects, including the cancer genome atlas (TCGA) project, the international cancer genome consortium (ICGC), the genotype-tissue expression (GTex) project, the international nucleotide sequence database (INSD) collaboration, the trans-omics for precision medicine (TOPMed) program, and the human proteome project (HPP), among others. Smaller-scale projects are also generating unique data, which are being deposited in public databases such as the Gene Expression Omnibus (GEO) and Array-Express databases. While this fertile material may empower the creation, scrutiny and validation of biomarker discoveries and predictive models, data are usually archived as raw sequencing reads or have not been processed with unifying standards or pipelines. Thus, subsequent computational analyses integrating various of these data resources remains a significant challenge.

A hallmark of data science is its ability to solve large scale industrial problems by leveraging well-designed software APIs, automated unit and integration testing among other techniques. Computational biology has increasingly recognized the importance of these principles by using semantic versioning of open-source software and containerization, thus increasing the transparency of the analytical pipelines. This workshop has been enriched by the expertise of bioinformaticians and data science experts working in this domain.

iii. Consensus on the relevance of reproducible and transparent workflows

In recent years, major concerns have been raised on the reproducibility and replicability of scientific conclusions as well as the overuse (and misuse) of the "statistical significance" paradigm. The extreme complexity of genomic data in medical research increases the difficulty of reproducible research substantially. Thus, the development and application of quality control practices and standards for implementing, preserving, and disseminating the analysis of genome-scale data has become imperative to extract meaningful and supported conclusions from the rich available datasets.

The solutions to the "reproducibility crisis" range from contributing with open software tools to top-down prescriptions relating to the release and format for raw data, code, and results. The computational biology community has responded to this demand to ensure the transparency and reproducibility of genomic research. For example, the Massive Analysis and Quality Control (MAQC) Society was founded to "communicate, promote, and advance reproducible science principles and quality control for analysis of the massive data generated from high-throughput technologies". In addition, the integration of data science tools in analytical workflows, such as public databases, RMarkdown (rmarkdown.rstudio.com) and Jupyter (jupyter.org) notebooks and public GitHub repositories and is supporting the generation of reproducible analyses and re-use of code scripts. Despite ongoing improvements, opportunities still exist to improve research standards and to advocate the notion and practice of reproducible and transparent research in the interdisciplinary field of statistical and computational genomics. The workshop explored best practices in this area.

3 Presentation Highlights

Invited and online speakers discussed topics relating to the workshop's foci. The workshop opened with an extensive overview of novel technologies and approaches that revolutionize our understanding of cell biology in health and disease from a complex hierarchy of different -omics. Presentations not only demonstrated the rapid advances in single-cell and spatial -omics technologies but also highlighted the computational challenges and developments to interrogate the data generated by these new methods.

Dr. Robert Gentleman: *Spatial transcriptomics*

Dr. Gentleman gave an overview of different activities at the Center for Computational Biomedicine. An overarching goal of the Center is to develop new tools to support high throughput reproducible analyses of the complex and rich data generated by novel technologies. In particular, Dr. Gentleman talked about different milestones and challenges in spatial transcriptomics technologies, including `MERFISH`, `Codex` and `CyCIF`. The spatial resolution of single-cell transcriptomics provides crucial information for therapeutic design. For example, with `MERFISH` we can not only simultaneously measure hundreds to thousands of genes but also know their spatial distribution within individual cells with 3D resolution. Dr. Gentleman highlighted the importance of recovering this information, in particular in complex diseases with strong genetic components. However, key challenges associated with image processing, analysis and segmentation can jeopardize the potential benefits of these cutting-edge technologies. A fair amount of supporting infrastructure has been developed, including `SpatialExperiment` in Bioconductor, Shiny and Posit Connect approaches for visualization, and other Python interfaces. Yet the size of the images that need to be displayed requires the use of sufficiently powerful servers to operate.

Dr. Gentleman also talked about the Center's contribution to gene ontologies and the development of tools that help scientists organize what is known about particular sets or collections of entities. In particular, he presented The Human BioMolecular Atlas Program (HuBMAP) ontology, an open platform to map healthy cells in the human body to better understand the relation between cells, tissue organization and function that can affect human health. To ease the interrogation of such a rich platform, Dr. Gentleman's group had developed a computational tool called HuBMAPy to interact with HuBMAP and to program ontology queries. Other relevant tools presented include `OpenGWAS` and `text2term` to search, aggregate, and compare data.

Dr. Celia Greenwood: *Modelling covariate and SNP influences on DNA methylation data*
Dr. Greenwood, presenting remoting, described an effort to identify genomic methylation sites (from bulk sequencing) that correlate with anti-citrullinated protein antibodies (ACPA), which is a biomarker for rheumatoid arthritis, while adjusting for covariates and confounding variables, such as cell type composition, genetic background, and exposures. Methylation was measured by bisulfite sequencing, which gives count data to compare among groups of samples with high and low ACPA. This effort required tackling numerous data challenges, including measurement error, zero inflation problems, over-dispersion, and high-dimensionality in the methylation status. To obtain a realistic pattern of dispersion, a quasi-binomial with random effects has been proposed to add two terms to model overdispersion, a multiplicative and an additive term. Developing this model required iteratively examining and interpreting a variety of model diagnostics to ensure the modeling assumptions accurately reflected the properties of the data. Dr. Greenwood lab developed appropriate algorithms and packages to implement their proposed model. Simulation and case studies results were presented to demonstrate the contribution of their work.

Dr. Maria Chikina: *Mixture deconvolution: a new perspective on an old problem*
Dr. Chikina provided an overview of cell-type deconvolution methods, including score based methods, regression based methods as well as reference free methods, to look at certain cell types from a mixture in a sample when bulk assays are used. Importantly, Dr. Chikina discussed how single cell datasets can be used to construct bulk data with control of cell proportions to obtain benchmark sets. Experiments demonstrate that the current simulation pipelines are unrealistic and do not accurately account for inter-subject heterogeneity in cell type expression states, which leads to optimistic bias when benchmarking methods. Dr. Chikina's group proposed alternative simulation strategies of bulk samples mixing different samples to account for the heterogeneity observed in real data. As the simulation strategies become more realistic, the correlation between predicted and true values drops. Results show that `BayesPrims` is a top performer in more realistic simulation strategies but computationally expensive. Dr. Chikina proposed computational optimizations by derandomizing `BayesPrism` and developed a method called `InstaPrism`.

Dr. Sandrine Dudoit: *Learning from Data in Single-Cell Transcriptomics*
Dr. Dudoit, presenting remoting, surveyed statistical and bioinformatic methods developed by her group for the main steps of a workflow to analyze single cell RNA sequencing data. These include `EDASeq` for summaries and visualizations, `RUVSeq` and `scone` for normalization and pre-processing, `scPCA` for dimensionality reduction, `ZINB-WaVE` to analyze the data characterized by zero inflation and over-dispersion while adjusting for known and unknown factors of unwanted variation, and two clustering methods: `RSEC` and `Dune` to generate and merge clusters maximizing their concordance. Dr. Dudoit presented `Slingshot`, a method developed by her group that provides a flexible and robust framework for inferring cell lineages and pseudotimes from cell clusters. These methods were used in a study that examined stem cell differentiation in mouse olfactory epithelium (OEp63) data. Noteworthy, three lineages were identified and analyzed by trajectory-based differential expression method `tradeSeq` to gain insight into the biological processes underlying differentiations. Results can potentially be applied in the prevention and treatment of neural tissue damage and degeneration, e.g. Alzheimer.

Dr. Gerald Quon: *DNA sequence models for predicting single cell expression*
Dr. Quon highlighted a challenge that characterizes gene network inference problems from RNA sequencing data. In such tasks, the number of available samples is typically much smaller than the number of genes in a particular genome. In order to improve the effective number of samples used to perform network inference, Dr. Quon group proposes a novel deep learning strategy in which multiple networks are inferred simultaneously for related conditions, and through multitasking, samples and covariance matrices are shared across networks. In this talk, he demonstrated the efficacy of the proposed approach on a series of scRNA-seq datasets collected by fibroblast cells that were reprogrammed into patient-specific iPSC lines, and then

subsequently differentiated into terminal neurons and endoderm cells. Results suggest that their multitask network inference strategy can yield novel insights into how gene network structure varies across conditions.

Dr. Fabien Plisson: *Auditing for structural bias in machine-learning peptide design*
In this presentation, Dr. Plisson discussed the current challenges and solutions to develop robust models for the design of antimicrobial peptides (AMPs). AMPs have become important in therapeutic avenues against antibiotic-resistant infections. However, some major limitations prevent their translation into clinical settings, including their low metabolic stability, poor oral bioavailability and high toxicity. Dr. Plisson showed that peptides' Sequence-Structure-Function relationships are complex and exploring peptide-fitness landscapes is costly and therefore limited. He described the experimental Design-Make-Test-Analyze cycle that leads to a multi-objective optimization to design peptides, and explained how machine learning algorithms can be used to develop a more cost-effective computational-protein design to generate optimal AMPs. Dr. Plisson also discussed the computational challenges faced in developing fair machine-learning models for the discovery and design of safe AMPs, including data scarcity, class imbalance, taxonomic bias, and predictive limitations. Various studies from his group were focused on benchmarking protein prediction methods and documenting sampling and structural biases in AMP predictive models.

Dr. Ingo Ruczinski: *Detecting And Quantifying Antibody Reactivity In PhIP-Seq Data*
Dr. Ruczinski presented a novel analytical pipeline designed for the analysis of PhIP-Seq data. PhIP-Seq enables high throughput characterization of antibody response to thousands of target antigens. In his presentation, Dr. Ruczinski presented the basics of the technology and discussed differences to similar data types such as RNA-Seq. For example, the inclusion of negative controls (beads-only samples) in the data needs to be addressed in the analysis, peptides of low counts are of biological interest, and read frequencies follow a Beta distribution. Based on these differences Dr. Ruczinski' group proposed to improve existing methods designed for RNA-Seq data analysis, like `edgeR`, to better address the characteristics observed in PhIP-Seq data. Dr. Ruczinski presented `BEER` (Bayesian Enrichment Estimation in R), a software package developed specifically for PhIP-Seq data. `BEER` estimates a hierarchical model to generate posterior probabilities for peptide reactivity and to quantify the magnitude. Results from a simulation and data analysis showed that although it takes longer to run, `BEER` is more sensitive to weakly reactive peptides with fewer false positives.

Dr. Richard Bonneau: *Accelerating anybody optimization and de novo design with AI/ML*
Dr. Bonneau gave an overview of the activities at Prescient Design and their transition from foundation till becoming a part of Genentech. He presented their solution to the complex problem of computational protein design. Combining a sequence denoising autoencoder with a length predictor and a function classifier their deep manifold sampler can generate diverse sequences of variable length with desired functions. Dr. Bonneau also presented their contribution to protein function prediction using `DeepFRI`. He discussed how protein language models and protein structures can be used to extract sequence features and how this work inspired them to build a design method. In December 2021, Prescient Design delivered the first set of proof-of-concept antibody sequences. Dr. Bonneau presented their most recent contributions to improve antibody generative models with response to sequence diversification, property optimization and de novo sequence generation, including LMS, a multi-property optimization via compositional energy functions, `EquiFold` for structure prediction in high-throughput applications, and `AbDiffuser` for de novo antibody generation.

Tuesday talks were mainly focused on the presentation of novel algorithms proposed for the analysis of complex -omics datasets. A brief description of each talk follows.

Dr. Katya Rodriguez-Vazquez: *Bio-inspired Algorithms in Biology*
Dr. Rodriguez-Vazquez presented on evolutionary and bio-inspired algorithms, which are search and optimization techniques designed to emulate natural processes. Notable examples include Genetic Algorithms (GA) and Genetic Programming (GP), both utilizing natural selection concepts to evolve a population of potential solutions for problem-solving. These algorithms employ genetic operators such as recombination and mutation to create improved solution populations. Ant Colony Optimization, another bio-inspired algorithm, mimics the foraging behavior of ant colonies to generate optimal paths from the nest to a food source. Bio-inspired algorithms prove useful in addressing various biological problems, including clustering, classi-

fication, modelling, and gene expression analysis. Dr. Rodriguez's talk delved into the application of these algorithms to solve diverse biological challenges, showcasing their versatility and efficacy in optimization contexts.

Dr. Anthony-Alexander Christidis: *Robust Multi-Model Subset Selection*
Dr. Christidis introduced a method called Robust Multi-Model Subset Selection (`RMSS`), to learn a robust ensemble composed of a small number of sparse, diverse and robust models to accurately predict a continuous response, e.g., the concentration of a compound. The degree to which the models are sparse, diverse and resistant to outlying observation is driven directly by the data. Dr. Christidis discussed some of the statistical properties of the proposed method and described a tailored computing algorithm to learn the ensembles by leveraging recent developments in optimization. Results from simulation studies demonstrate the competitive advantage of the proposed method over state-of-the-art sparse and robust methods.

Kat Clark: *Trimming Outliers in Matrix-Variate Normal Mixtures using the OCLUST Algorithm*
In this talk, Kat Clark presented an extension of the `OCLUST` algorithm to matrix-variate normal mixtures. The original version of the `OCLUST` algorithm trims outliers iteratively in multivariate normal mixtures. Leveraging the fact that Mahalanobis squared distances are chi-squared distributed (or scaled beta-distributed when using sample parameter estimates) for multivariate normal data, suspected outliers are removed one-by-one until the subset log-likelihoods conform to the specified distribution. Using the matrix-variate normal analogue of Mahalanobis squared distance, they show that the log-likelihoods approximate a shifted chi-squared mixture distribution. This distribution was employed to detect likely outliers in matrix-variate normal mixtures as well as to predict the proportion of outlying points.

Dr. Jason Xu: *Learning hierarchical covariance structure from multiple studies via subspace factor analysis*
Dr. Xu discussed limitations of existing factor analysis and hierarchical extensions of factor analysis methods when applied to high-dimensional data in which the same set of variables are often collected under different conditions. Dr. Xu's group proposed a class of Subspace Factor Analysis (SUFA) models, which characterize variation across groups at the level of a lower-dimensional subspace. They proved that the proposed class of SUFA models lead to identifiability of the shared versus group-specific components of the covariance, and study their posterior contraction properties. Taking a Bayesian approach, these contributions are developed alongside efficient posterior computational algorithms. Their sampler fully integrates out latent variables, is easily parallelizable, and has complexity that does not depend on sample size. In his talk, Dr. Xu illustrated the methods through application to integration of multiple gene expression datasets relevant to immunology.

Dr. Clara Bermudez: *What would we discover about viruses' secrets if the computation attacks them?*
Dr. Bermudez discussed the collaborative efforts of the research community and health systems in addressing the challenges posed by COVID-19. She highlighted the role of interdisciplinary collaborations, computational biology, and data science in rapidly providing valuable information on genomic sequences, clinical data, and epidemiology associated with SARS-CoV-2. The talk emphasized the ongoing monitoring of Coronaviruses and other potentially zoonotic RNA viruses, utilizing platforms like GISAID. The presentation focused on dual aspects of RNA viruses' genomes, functioning both as Coding Sequences (CDS) and as macromolecular secondary structure engines. Overall, the presentation highlighted the importance of genomic and structural analyses, as well as the application of Artificial Intelligence in understanding and addressing emerging and reemerging viral threats.

Dr. Natasa Tagasovska: *Tailored Multiobjective Optimization and Sampling for Molecular Design*
Dr. Tagasovska gave an overview of the drug design workflow and how therapeutic antibodies are designed at Prescient Design. Dr. Tagasovska explained the benefits of using energy based models to sample new antibodies and generate realistic data. In particular, she proposed a Pareto-compositional energy-based model (`pcEBM`) that uses multiple gradient descent for sampling new designs and adheres to various constraints in optimizing distinct properties. Dr. Tagasovska also discussed the difficulty of evaluating the goodness of the candidate designs and proposed to use the hypervolume as a metric of evaluation. Results from in-silico experiments showed the benefits of the proposed methodology. Dr. Tagasovska explained the limitations of this approach for antibody design methods, in which multiple properties need to be taken into account. With

her team, they proposed an acquisition function, called `BOtied` for multi-objective Bayesian optimization, based on a cumulative distribution functions indicator and implemented with copulas, that is more suitable for molecular design.

Dr. Iuliana Ionita-Laza: *Knockoff-based statistics in genetics: opportunities and challenges*
Dr. Ionita-Laza gave an overview of methods to identify important genetic variants that are associated with phenotypes of interest in genome-wide association studies (GWAS). She first showed how Quantile Regression (QR) can be used in GWAS studies and its advantage in settings relevant to genetics, including cases of heterogeneous effects across quantiles and non-additive effects. Dr. Ionita-Laza continued her presentation with an overview of the knockoff-based inference framework, which has some appealing features, including increased power, reduced confounding due to linkage disequilibrium and population structure. She described a general knockoff-based procedure and made general remarks about some of its challenges, such as being computationally/memory intensive and not dealing well with high correlation from group knockoffs. As a solution, Dr. Ionita-Laza proposed a scalable method via sub-sampling, called the `KnockoffScreen`, for knockoff generation. Results from a simulation study and a case study based on UK biobank data to find causal genes of cholesterol demonstrate the advantage of the proposed method.

Dr. Pei Wang: *iProMix: A mixture model for studying the function of ACE2 based on bulk proteogenomic data*
Dr. Wang presented `iProMix`, a novel statistical framework to identify epithelial-cell specific associations between ACE2 and other proteins/pathways with bulk proteomic data. The identification of such interactions are crucial to examine how SARS-CoV-2 enters and perturb human cells. `iProMix` decomposes the data and models cell-type-specific conditional joint distribution of proteins through a mixture model. It improves cell-type composition estimation from prior input, and utilizes a non-parametric inference framework to account for uncertainty of cell-type proportion estimates in hypothesis tests. Simulations demonstrate that iProMix has well-controlled false discovery rates and favorable powers in non-asymptotic settings. Dr. Wang's group have applied iProMix to analyze proteomic data of 110 (tumor-adjacent) normal lung tissue samples from the Clinical Proteomic Tumor Analysis Consortium lung adenocarcinoma study, and identified interferon-response pathways as the most significant pathways associated with ACE2 protein abundances in epithelial cells. Strikingly, the association direction is sex-specific. This result casts light on the sex difference of COVID-19 incidences and outcomes, and motivates sex-specific evaluation for interferon therapies.

Dr. Amilcar Meneses: *Towards a replacement for glyphosate as a pesticide using data mining and artificial intelligence*
In this talk Dr. Meneses presented the use of the CRISP methodology to work with the QSTR model to describe the relationship between the molecular descriptors of dragon and DFT with toxicity for agrochemicals. In Mexico, the use of agrochemicals in agriculture is common. One of the most widely used agrochemicals is Glyphosate. Glyphosate is an herbicide that has been used since 1974. In 2015, the World Health Organization classified it as a probable human carcinogenic. The results of this methodology in carbamates demonstrate that two DFT descriptors and eight Dragon-CHEMID are necessary to predict the toxicity in carbamate compounds.

The main focus for Wednesday's talks was on methods developed for the integration of different -omics datasets.

Dr. Javier De Las Rivas: *Multivariate logistic regression, survival analysis and other algorithms applied to the study of genomic and clinical data from cancer patients*
Dr. De Las Rivas described the importance of addressing patient survival and risk prediction problems in clinical research. His group developed an algorithm, called ASURI, that combines a set of functions for disease survival analysis and patient risk predictions based on gene signatures. The algorithm includes functions to perform a robust feature selection for the discovery of gene markers linked to survival and identification of their associations with clinical variables or phenotypic characteristics. The tool can also be used to construct robust patient risk predictors based on gene signatures using univariate and multivariate approaches. Dr. De Las Rivas presented two relevant applications of ASURI, one related to an integration of multiple survival

studies of colorectal cancer and another one related to breast cancer studies. An R package to implement ASURI is currently under construction.

Dr. Hector Corrada Bravo: *The Single Cell Hub at Genentech: how to manage and query data for 130M cells*
Dr. Corrada Bravo explained how large scale genomic data collection can be used to understand cell states and their organization by creating single cell hubs. He presented a framework consisting of 4 stages: observe, recall, synthesize and predict. He started with the description of the first stage, “observe”, where datasets are collected, curated and harmonized using three-table representations, consisting of gene-by-cell matrix of gene expression measurements, metadata about genes and metadata about cells. Each dataset is then mapped to metadata attributes, which are mapped to proper ontologies. Dr. Corrada Bravo then described the second phase, “recall”, to find appropriate datasets in useful and meaningful ways. For example, datasets can be identified based on a particular cell type ontology. He also illustrated how to find datasets by a gene signature expression or by cell similarity. He finished his talk discussing new problems related to the “synthesize” stage of their framework and open questions related to spatial datasets and cell state organization in tissue.

Dr. Ali Yazbeck: *Integrating germline and somatic genetic profiles through machine learning to understand cancer etiology*
In this presentation, Dr. Yazbeck introduced semi-automated and automated non-coding RNA (ncRNA) curation methods designed to streamline the annotation and analysis of ncRNAs. Dr. Yazbeck first described a method he designed to address the 7SK RNA by defining the invertebrates’ secondary structure. He then presented a second method, `MIRfix`, which is a fully automated tool for correcting and curating all metazoan families of microRNAs. The tool integrates advanced algorithms for predicting secondary structures and functional domains, enhancing the understanding of ncRNA functions. Dr. Yazbeck also gave an overview of his current work in the field of genetics. He summarized his current progress in collecting the somatic mutations, processing germline variants and using artificial intelligence to find the association between germline variants and somatic mutations. His work focuses on Whole Genome Sequencing of matched pairs samples of Esophageal Adenocarcinoma patients.

Dr. Katia Aviña Padilla: *A multiomics approach to identify potential driver genes in cancer diseases*
Dr. Avina Padilla presented her work on multi-omics approaches using bioinformatics strategies to investigate driver genes in cancer. Her work explores cancer genomics, tumor microenvironment, intronless genes, complex networks, transcriptional reprogramming, and systems biology to enhance our understanding of cancer and guide future clinical strategies. In particular, her work in Glioblastoma Multiforme (GBM) uncovered transcriptional profiles aligning with the stem cell model and distinct somatic mutation patterns between young and adult patients. Furthermore, she emphasized the regulatory roles of differentially expressed human intronless genes (DE-IGs) across eight different cancer types, highlighting their relevance in protein-protein interaction networks. Her multi-omics approach offers insights into predictive molecular markers, aiding clinical decision-making in cancer management.

Thursday talks were focused on the analysis of data from Biobanks to infer characteristics of populations and their evolution. A brief description of each talk follows.

Dr. Laxmi Parida: *Topology and Logic in Life Sciences*
Dr. Parida’s talk was focused on the use of topological data analysis (TDA) in three different areas of life sciences: population genomics, metagenomics and phenomics. For the first application, she explained how topological characteristics learned from SNPs data can be used to detect admixtures, which are populations that descended from multiple populations through interbreeding, migration, mixing and so on. Using simulated and real data, Dr. Parida demonstrated that TDA and essential simplicies can be used to identify the presence of admixtures, determine their number and deduce their relative timing. In metagenomics, TDA can be used to examine the influence of the immediate environment on the phenotype of an organism. However, it is challenging to identify truly present organisms in a sample since short sequencing reads can match multiple organisms in a database due to sequence similarities. Dr. Parida showed how TDA can be used to extract information from the geometric structure of data and demonstrated the power of this approach using data

from 36 serotypes of salmonella. As a third application, she described how TDA can be used in phenomics to construct a pipeline to automatically extract candidate pathways associated with COVID-19 from clinical notes.

Dr. Jessica Dennis: *Genetic epidemiology in the era of big data, biobanks, and -omic technologies*
Dr. Dennis presented her contributions in post-GWAS studies of loneliness and functional genetic studies of placenta. Loneliness is defined as discontent with social connections and has been associated with a 38% increased risk of early mortality. She conducted a phenome-wide association study using loneliness polygenic scores and found that susceptibility to loneliness was associated with different mental health disorders but also with different circulatory system conditions, which could explain the increase in mortality. A follow-up study shows that the association between the polygenic score and risk of coronary artery disease is not attenuated after controlling for heart disease risk factors. Another study showed that this association was not driven by genetic risk factors shared between major depressive disorder and loneliness. In her work on functional interpretation of GWAS her group examined enrichment of placental mQTL in GWAS of 19 traits, including neuropsychiatric traits, immune related traits, and growth related traits. All the analyzed mQTL were enriched in GWAS results for growth- and immune-related traits in childhood, like type one diabetes, asthma, or different kinds of allergies. A colocalization study found a considerable number of colocalized CpG sites in several of these childhood traits.

Dr. Mashaal Sohail: *Mexican Biobank advances population and medical genomics of diverse ancestries*
Dr. Sohail gave a brief overview of the Mexican Biobank (MXB) project that includes phenotypic and genotypic data that can be used for various analyses of population genetics and complex traits. Although 40,000 individuals from all 32 states in Mexico are included in the database, only 6000 individuals were genotyped, including individuals that speak indigenous languages as well as those from rural localities. Dr. Sohail demonstrated how using data from the MXB project provides new insights into the genetic histories of individuals in Mexico and dissects their complex trait architectures. In particular, using ancestry deconvolution and inference of identity-by-descent (IBD) segments, Dr. Sohail inferred ancestral population sizes across Mesoamerican regions over time, unraveling indigenous, colonial, and post-colonial demographic dynamics. They also found that several traits are better predicted using the MXB GWAS compared to the UK Biobank GWAS. Results are crucial for making precision and preventive medicine initiatives accessible worldwide.

Dr. César Díaz: *Evomining as a tool to clusterize metagenomes*
Dr. Diaz's talk focused on metagenomics, a field that studies genetic material directly from environmental samples. The primary challenge in metagenomics is accurately classifying organisms to the species level. However, Dr. Diaz addressed a related issue: given the presence of species A in a metagenomic sample, the goal is to classify the reads to its subspecies with high specificity. This is particularly crucial for organisms where only specific subspecies are pathogenic, leading to costly treatments. Dr. Diaz proposed a modification to `EvoMining`, a tool for genome mining, which identifies copies of genes with the potential to produce new natural products. This modification may offer insights into solving the subspecies classification problem.

Dr. Nelly Selem: *MicroAgrobiome - a comparative metagenomics platform applied to hunt genera in the crop microbiome*
Dr. Nelly Selem discussed the need for an integrated analysis platform for publicly available data on microbiomes associated with crops. While there is abundant data on these microbiomes, there is currently no platform for analyzing the traditional composition, co-occurrence networks, and stability of crop microbiomes across different plants. Microagrobiome aims to fill this gap and become a public reference for agriculturally relevant microbiomes, similar to the understanding of the human microbiome and its impact on health. The focus is on the Clavibacter genus, which specifically infects crops such as tomatoes, corn, chili, potatoes, wheat, and alfalfa. Over the past two years, data on the composition of microbiomes associated with these agriculturally relevant plants have been released. Dr. Selem's work proposes two main objectives: first, to build a map that combines data from metagenomic studies to establish the common microbiome among agriculturally relevant crops, and second, to use the platform to understand Clavibacter as a model for studying speciation resulting from bacterial infections caused by changes in the host.

Dr. Deisy Gysi: *Network Medicine, Disease Module (in)completeness and other tales*
In this presentation, Dr. Gysi explained how Network Medicine (NM) can enable the prioritization of drugs for treating emerging diseases. In particular, she presented the relevance of these contributions in pandemic times when the traditional approaches to develop new drugs were not feasible and repurposing clinically approved drugs that have therapeutic effect in COVID-19 patients was crucial. Using disease modules to identify close diseases in a protein-protein-interaction network can be used to select such drugs. However, there were no other diseases that were very close to COVID-19 so a treatment could not be derived from already approved therapies. Implementing three network drug repurposing methods, network proximity, network diffusion and AI prioritization, they found complementary ranked lists of effective drugs. The experimental screening of over 6K drugs provides strong evidence towards their pipeline. Including ncRNA mediated interactions in the analysis they could identify more disease modules, elucidate disease progression, and find better comorbidity detection.

4 Highlight of Activities

In this workshop, scientists from both Data Science and Computational Biology have presented their innovative solutions to address current computational and analytical challenges co-emerging with complex biological datasets and to share and join efforts and solutions to advance knowledge in Science. Unlike other conferences, our workshop has blended the knowledge arising from multiple related fields, opening the opportunity to scientists with a diverse background and areas of expertise to collaborate in the development of innovative methods for the analysis of large-scale multi-omics data. The workshop was enriched with several joint discussions, brainstorming moments, a career panel and project work presentations. In particular, we highlight below two important activities.

Industrial Experts and Leaders Panel Discussion. Industrial experts and leaders in the fields participated in a hybrid career panel. The panel was formed by Drs. Laxmi Parida (IBM Master Inventor and group leader at the Watson Research Center and Courant Institute of Mathematical Sciences, among several other consortia and projects), Natasa Tagasovska (Machine Learning Scientist at Prescient Design, Genentech, and Roche), Robert Gentleman (co-founder of R programming language and the Bioconductor project, vice president at 23andMe, and currently the director of the Center for Computational Biomedicine at Harvard Medical School), Raymond Ng (Director of the Data Science Institute at University of British Columbia, Canada).

This diverse and inclusive panel of experts from industry and academia generated broad panel discussions that were particularly inspiring for students, trainees and young researchers, participating both in person and online, to define their career paths.

Collaborative Research Projects. In order to facilitate a close interaction between students, professors and researchers, we created 3 collaborative projects to work on throughout the time of the workshop. The theme of the projects were jointly suggested by participants prior to the workshop but finalized onsite. Throughout the first 3 days of the workshop, we reserved several time slots for the groups to meet and discuss. On the last day, a representative student within each group presented their results to the audience. This activity has promoted a fruitful interaction between students and PIs from different backgrounds, disciplines and institutions. This activity has also helped PIs to identify outstanding students seeking for new opportunities in their career. For example, Dr. Christidis, postdoctoral fellow at UBC has recently started working in a new position at the Center for Computational Biomedicine at Harvard Medical School with Dr. Gentleman. A brief description of the projects follows.

- **Influence of Question Formulation on ChatGPT Responses in Scientific Contexts.** This project aimed at assessing the impact of question formulation on ChatGPT's responses, the team focused on investigating the differences between positive and negated formulations of both general and scientific statements. Specifically, we created pairs of questions, such as "True or false, RNA is transcribed from DNA" and its negated counterpart "True or false, RNA is not transcribed from DNA" totalling

20 question pairs. Our objective was to determine whether the formulation of questions influenced ChatGPT's answers.

These question pairs were posed to three independent instances of ChatGPT, and the results revealed intriguing patterns. Notably, we observed varying responses across different threads for a significant number of questions, emphasizing the probabilistic nature of large language models (LLMs). This variability was particularly pronounced in questions that allowed for multiple unspecified interpretations, underscoring the contextual sensitivity of ChatGPT's responses.

While acknowledging the limited power of 20 questions to yield robust results, our findings hinted at a potential inclination for ChatGPT to agree with user-presented statements. This suggested a preference for responding affirmatively, with a tendency to provide 'yes' or 'true' answers rather than 'no' or 'false' answers.

In summary, this project shed light on the nuanced dynamics of question formulation and its influence on ChatGPT's responses. The observed variability and the potential propensity to agree with user-provided statements underscore the importance of understanding the probabilistic nature of LLMs when interacting with such models. These findings contribute to the broader discourse on the use and interpretation of AI language models in various contexts.

- **A perspective on the challenges to achieve fine taxonomic classification of metagenomes**

  In this project, the challenge of achieving taxonomic classification at the species level in metagenomic samples was addressed. The primary issues considered were as follows:

  Classic classifiers, dependent on databases containing genomes of previously sequenced organisms, were scrutinized for their limitation in excluding unsequenced species. False positive problems arose when reads matched with more than one organism, and the sensitivity of available classifiers to high intraspecific variability was questioned.

  Exploration of potential solutions led to the consideration of neural networks. Deep neural networks, encompassing convolutional and recurrent architectures, were explored for their capacity to capture intricate patterns within metagenomic sequences. Transfer learning, involving the utilization of pre-trained models on extensive datasets, was seen as a means to enhance classification abilities, particularly in scenarios with limited reference genomes.

  Generative models, exemplified by generative adversarial networks (GANs), were contemplated for their potential in addressing intraspecific variability and the presence of contaminants. The theoretical application of GANs involved learning to reconstruct metagenomic data, thereby identifying patterns and mitigating the impact of noise.

  The project unfolded as a comprehensive exploration of the limitations inherent in classic classifiers and the potential of neural networks, transfer learning, and generative models to redefine taxonomic classification in metagenomic samples.

- **Reading group: "Orchestrating Single Cell Analysis with Bioconductor"**

  The aim of this project was to read and discuss some key chapters of the new online book "Orchestrating Single Cell Analysis with Bioconductor". Dr. Robert Gentleman led the discussions and showed some key features of the online book. Members took some independent time to become familiar with the book content and play with some of the tools to process, analyze, visualize, and explore scRNA-seq data covered in the book. The group then discussed together different topics related to common workflows for the analysis of single-cell RNA-seq data (scRNA-seq), and tools.

A summary of the discussions and conclusions from each group were presented to the rest of the participants on the last day of the workshop.

5 Outcome of the Meeting

Overall, participants contributed and shared ideas on the design of efficient experimental designs to benchmark algorithms, data structures, statistical methods, machine and deep learning techniques for the analysis of large scale datasets, multi-omics data integration, and good practices for reproducibility / replicability of experiments.

Importantly, our workshop promoted and strengthened collaborations among a diverse pool of scientists from a variety of ethnic backgrounds, genders, cultures, locations, and career stages. This was particularly important to create a supportive community for young researchers that may lack diverse interactions at their home institutions. It also gave the chance to participants from under-represented groups to gain visibility and to contribute with valuable and unique ideas.