Mathematical methods in cancer biology, evolution and therapy

Peter Van Loo (MD Anderson Cancer Center),
Wenyi Wang (MD Anderson Cancer Center),
Ronglai Shen (Memorial Sloan-Kettering Cancer Center),
Quaid Morris (Memorial Sloan-Kettering Cancer Center)

May 14 – May 19, 2023

1 Overview

The unifying theme of the our workshop was the development of creative mathematical and statistical models for analyzing patient samples, with a particular focus on tumor evolution and immuno-oncology, two of the most timely and important foci of cancer research. There is an exciting growth of research in these foci fueled by the increasing availability of whole-genome and whole-transcriptome next-generation sequencing data in a large number of tumor samples, as well as in single-cell and cell-free DNA settings. It is now well understood that most patient samples are comprised of a mixture of cell types, including multiple genetically distinct populations of cancer cells (i.e., 'subclones') and a variety of normal cell types (e.g., normal stromal tissue cells, fibroblasts, and diverse immune cells). Effective use of these samples to extract relevant inference on tumor biology requires information from individual cell types to be deconvolved from the 'bulk' data collected. This workshop provided the opportunity to identify synergies in mathematical deconvolution techniques developed independently in different cancer subfields.

Our workshop brought together leading experts from diverse geographic regions – Canada, US, UK, Australia, Spain, Singapore and others – and a wide range of scientific backgrounds – computational biology, mathematics and statistics, computer science, and various biomedical fields – and a variety of career stages including trainees, early stage researchers and senior faculty. It provided a forum for cross-disciplinary learning, discussion, collaboration, and mentoring. We organized the talks in the workshop into five focus areas: (1) intra-tumor heterogeneity and subclonal reconstruction, (2) copy number analysis of cancer genomes, (3) signatures of mutational processes in cancer, (4) deconvolution of tumor, normal and immune transcriptomes and epigenomes, and (5) immunogenomics and immuno-oncology.

2 Background and Open Problems in the Five Focus Areas

1. **Intra-tumor heterogeneity and subclonal reconstruction algorithms.** Intra-tumor heterogeneity, i.e., the presence of a multitude of genetically and phenotypically distinct cancer cell populations in a tumor, is a key challenge in cancer medicine. These intra-cellular differences can include somatic point mutations, chromosomal aberrations and epigenetic changes. Heterogeneity introduces genetic and phenotypic diversity

that fuels ongoing tumor evolution and resistance to therapy, ultimately contributing to disease progression and metastasis.

The reconstruction of subclonal populations and their evolutionary relationships from sequencing data thus represents an important but extremely challenging deconvolution problem. Ideally subclonal reconstruction algorithms would distinguish and characterize populations of normal cells with no somatic aberrations and multiple populations of tumor cells with both shared and unique somatic changes. Multiple bespoke approaches such as non-parametric, tree-based Dirichlet processes and Hidden Markov Models have been developed, including many prominent works from our invitees. While this is a thriving field, multiple opportunities for further mathematical and statistical method development remain, including time-efficient and optimal phylogenetic inference, integration of information across multiple time points and multi-regional samples and across bulk and single-cell sequencing data, and integration across different classes of somatic variants (e.g. subclonal copy number changes and subclonal point mutations).

2. **Mathematical and statistical approaches for copy number analysis of cancer genomes.** Chromosome instability is a hallmark of cancer, and is associated with poor prognosis and therapeutic resistance. These chromosomal aberrations can range from small-scale deletions and duplications, through gains and losses of entire chromosomes, to even duplication of the entire genome. They play an important role in cancer development, and in addition are a confounder that needs to be accounted for in many genomics analyses (including intra-tumor heterogeneity, subclonal reconstruction and in immunogenomics analysis). Cancer genomics data such as SNP arrays and exome or whole-genome sequencing can be used to infer genome-wide copy number profiles of cancers. However large-scale aneuploidy and the admixture of normal cells are confounders that need to be addressed. While important strides have been made in tackling this deconvolution problem over the past decade, several mathematical challenges remain, including: (i) ambiguity in whole-genome duplication inference in 10-20% of cancers, (ii) cancer samples with high normal cell admixture, and (iii) accurate inference of subclonal copy number changes. These are issues also relevant in applying copy number analysis in clinical settings. In addition, there are several unfulfilled opportunities to infer copy number profiles from other -omics data, including transcriptomic and epigenomic.

3. **Deconvolution of signatures of mutational processes in cancer.** The somatic changes in cancer genomes are caused by the interplay of DNA damage and repair. From patterns of co-occurence of different types of somatic mutations across cancers, signatures of mutational processes can be identified. For example, tobacco mutagens cause C->A mutations, UV radiation causes CC->TT mutations. These signatures can help understand the causes and progression of cancer, as well as helping to identify treatment strategies that target deficient DNA damage repair unique to cancer cells. Mutational signature analysis is a type of “mixed membership” statistical problem that is often approached as a (non-negative) matrix factorization problem where each tumor’s mutational repertoire is modelled as a combination of core mutational ‘signatures’ (representing the recurrent mutational processes active across tumors), each with appropriate ‘weights’ (representing the different activities of these mutational processes in the given tumor). While different methods have been developed for point mutations leading to more than 70 signatures of mutational processes with identfied etiologies (e.g. tobacco mutagens, UV light exposure, exposure to aristolochic acid), there is yet no complete consensus on the best deconvolution approaches; and current approaches lack sensitivity, leaving many mutational processes unidentified. Also, promising approaches to identify mutational signatures of structural variants and copy number changes are starting to emerge, opening up new potential avenues of discovery and treatment optimization.

4. **Mathematical and computational approaches to deconvolve tumor, normal and immune signals from transcriptomic and epigenomic data.** The clinical analysis of tumor samples is complicated by the tumor-stroma-immune interaction. The number and types of tumor-infiltrating immune cells in these samples predicts clinical outcome, and can affect treatment decisions. Decomposing tumor samples into their constituent parts in the lab is expensive, technically challenging and time-consuming, making it difficult to discern the relevant immune signals and motivating computational approaches to integrate the estimation of cell type-specific expression profiles, and epigenetic profiles, for tumor cells, immune cells, and the tumor microenvironment. Most deconvolution methods assume that malignant tumor tissue consists of two distinct components, epithelium-derived tumor cells and surrounding stromal cells, and are thus unsuitable for char-

acterising immune populations. Other deconvolution methods for more than two compartments require list of immune cell-type-specific reference gene lists, thus restricting the resolution, robustness, and scope of these analyses. Fortunately, this is a fast-growing field, where many research groups are developing models that relax these restrictive assumptions and input data requirements. However, key questions remain unsolved, such as dissecting expression signals from individual cell types for all 20,000 genes at once. A promising direction of inquiry is using matching bulk and single cell RNAseq data to borrow strength across data types, initial results suggests that this leads to better deconvolution of signals. Nonetheless, significant opportunities remain for theoretical and methodological research in this complex and high-dimensional deconvolution problem.

5. **Immunogenomics.** The immune system plays a critical role in the body's defense against cancer, and the recent development of cancer immunotherapies has clearly demonstrated the importance of host immune cells in cancer treatment. The immune response to cancer is activated by 'neoantigens', novel, tumor-specific peptides generated by missense mutations. These neoantigens appear on the cancer cell surface bound by the patient's HLA molecules and specific T-cells identify and respond to these neoantigens by initiating an immune response. The study of this process, immunogenomics, has benefitted from new sequencing technologies and associated computational data analysis methods, including the deconvolution of the appropriate signals from bulk sequencing data. In individual cancers, immunogenomics has improved the prediction of neoantigens for prognostic purposes or to inform immunotherapeutic interventions. New methods have also been developed to study HLA sequences, to investigate changes in the T cell repertoire, to characterize the gene expression signatures of the immune cell types present in the tumor mass (as described above), and to design personalized vaccines or adoptive cell transfer (ACT) therapies. The new immunogenomic methods have the promise of being widely used in clinical applications. However, this is an emerging field and various analytical challenges remain to be resolved, including the prediction of MHC presentations, an integrative analysis of somatic mutations and the immune repertoire, and the prediction of response to immunotherapies, both in terms of tumor killing effects and autoimmune side effects.

3 Presentation Highlights

The talks over the five-day workshop covered dynamic discussions on the past lessons, current state-of-art methods and future directions in which the field will develop.

1. **Intra-tumor heterogeneity and subclonal reconstruction algorithms.** Speakers in this focus area included Sohrab Shah, Nicholas McGranahan, Tony Papenfuss, Ben Raphael, Mohammed El-Kebir, Paul Boutros, Quaid Morris, Cenk Sahinalp, Gryte Satas, Yuchao Jiang, Sitara Persad, Leah Weber, Ethan Kulman, Chay Paterson and Adam Olshen. Several speakers presented on phylogeny analysis for the reconstruction of tumor evolution from single-cell and bulk sequencing. Such study typically involves multi-region tumor sequencing or sequencing of autopsy samples where multiple tumor samples are included. Phylogeny construction is a combinatorial problem that are often computationally intractable. Parsimony is a key principle in phylogeny analysis, for example, by limiting the transformation from one state to the other with minimal number of events. Ben Raphael discussed models for cancer evolution and lineage tracing, including the constrained k-dollo phylogeny model and dynamic lineage tracing using CRISPR-induced mutations. Cenk Sahinalp focused on the inference of mutational progression history and subclonal composition of tumor samples, highlighting the CITUP and MQIP methods for constructing mutation trees. Mohammed El-Kebir presented MACHINA, a tool for reconstructing metastatic migration histories, and discussed the identification of temporally consistent co-migrations using MACH2. Quaid Morris discussed clone tree reconstruction and introduced SubMARine, a noise-free phylogeny method that improves tree construction performance, and Pairtree, a scalable and fast bulk sample reconstruction method. Leah Weber introduced Phertilizer, a method for building clonal trees from ultra-low coverage single-cell DNA sequencing data, addressing the tradeoff between depth and uniformity and assessing clone trees based on a probabilistic model and fit to observed data. Ethan Kulman presented Orchard, a method for reconstructing mutation trees from bulk DNA data using a combinatorial search approach and the Gumbel Max Trick, allowing a ten-fold increase in the size of the cancer trees that could be reconstructed. Chay Paterson discussed multi-stage clonal expansion

models for simulating pre-invasive cell populations and estimating cancer risk, with applications to colorectal adenocarcinoma and vestibular schwannoma. Paul Boutros discussed research on the prostate cancer epitranscriptome, exploring the clonality of m6A peaks, germline m6A interactions. Additionally, he presented work on the impact of host factors such as hypoxia, sex and ancestry on tumor evolution.

Several speakers presented their work on using single cell sequencing technologies to study intra-tumor heterogeneity. Gryte Satas presented ArtiCull, a feature-based classifier for removing variant calling artifacts from DLP+ single-cell WGS DNA sequencing data, and discussed its applications in improving concordance with bulk data and identifying subclones. Yuchao Jiang presented work on scRNA+ATAC multiomic single cell. Leah Weber presented work on addressing sparsity of SNV signal in ultra low coverage DNA-sequencing (0.01x). Adam Olshen presented an integrative clustering method for Dab-seq with joint DNA and cell surface protein sequencing and future directions in combining Dab-seq with other single cell sequencing modalities such as CITE-seq. Tony Papenfuss presented an ensemble learning approach to integrate structural variant calls from whole-genome sequencing. Sohrab Shah identified distinct categories of high-grade serous ovarian cancer (HGSOC) based on genomic features and developed TreeAlign for single-cell integration to understand phenotypic consequences. Sitata Persad presented SEACELL, which leverages single-cell genomics data to study biological heterogeneity, including inferring epigenetic regulation and studying gene accessibility dynamics.

2. **Mathematical and statistical approaches for copy number analysis of cancer genomes** Speakers in this focus area included Peter Van Loo, Nana Mensah, Carla Castignani, Barbara Hernando, Geoff Macintyre, Henri Schmidt, and Kristiana Grigoriadis. Peter Van Loo discussed the analysis of copy number intratumor heterogeneity, inferring clonality from point mutations, constructing phylogenetic trees from mutation clusters and DNA methylation deconvolution using CAMDAC. Nana Mensah presented CAMDAC-WGBS for allele-specific methylation analysis in metastatic cancer, including the distinction between matched normal or panel methods. Carla Castignani introduced CREDAC, a tool for copy number-based expression deconvolution analysis of cancers, which models pure tumor expression profiles form bulk tumor expression alone. Barbara Hernando introduced a framework for modelling chromosomal instability (CIN) using copy number mutational signatures, identifying 17 CIN signatures in TCGA data and exploring integration with additional covariates. Geoff Macintyre highlighted the impact of chromosomal instability (CIN) in precision medicine, correlating CIN signatures with treatment response and exploring the translational potential of CIN signatures for ovarian cancer. Henri Schmidt presented Lazac, a zero-agnostic model for copy number evolution in cancer, addressing challenges in constructing evolutionary histories and proposing a model for efficient solution of the small-parsimony problem. Kristiana Grigoriadis presented ALPACA, a tool for inferring subclonal copy number states based on SNV-derived phylogenetic trees, considering evolutionary constraints.

3. **Deconvolution of signatures of mutational processes in cancer.** Speakers included Steve Rosen, Bin Zhu, Teresa Przytycka, and Caitlin Harrigan. Steve Rosen presented a hierarchical Dirichlet process for mutation signature decomposition with a prior on the shape of single base substitution (SBS) signature. Flat signatures (e.g., SBS3 vs SBS40) are hard to discriminate and additional information (indels with microhomology) is needed to refine the separation. Bin Zhu presented a non-negative matrix decomposition for mutation signature analysis of targeted sequencing with very sparse mutation counts. SBS count is biased for targeted panel (and whole-exomes) as coding regions are more enriched for certain trinucleotide changes. He presented a normalization of the SBS counts with respect to the reference genome. He presented analysis of the AACR-GENIE consortium data and simulation studies to investigate the sensitivity and specificity of detecting signatures with certain prevalence and flatness (using Shannon equitability index) of shape. He showed 100,000 samples are needed to detect flat signatures (e.g., SBS3, SBS5) with low frequency. Teresa Przytycka presented computational approaches to study mutagenic signatures, including EcoSigClust for identifying signature etiology. Caitlin Harrigan presented DAMUTA, a Bayesian method for identifying damage and misrepair mutation signatures, extracting 18 damage and 6 misrepair signatures, and assessing their association with gene mutations.

4. **Mathematical and computational approaches to deconvolve tumor, normal and immune signals from transcriptomic and epigenomic data.** Speakers included Wenyi Wang, Aaron Newman, Francesca

Petralia, Venkatraman Seshan, Carla Castignani, Kyle Coleman, and Yaoyi Dai. The Wenyi Wang lab aims to understand variations in both the transcriptome and genotype and integrate deconvolution models to define plasticity features in human tissues mathematically. They developed a new metric, total mRNA expression per haploid genome in tumor cells (TmS), and found that it can serve as a marker for worse prognosis and potentially predict response to therapy. Yaoyi Dai demonstrated that TmS is a pan-cancer prognostic marker in triple-negative breast cancer (TNBC) and correlated with TNBC subtypes, immune activation, and differential gene expression pathways. Carla Castignani introduced CREDAC, a tool for copy number-based expression deconvolution analysis of cancers, which models pure tumor expression profiles from bulk tumor expression alone. Aaron Newman discussed the challenges of profiling cell heterogeneity from bulk RNA-seq data and introduced CIBERSORTx as a method to improve accurate profiling of cell subsets. Francesca Petralia presented BayesDeBulk, a reference-free method for quantifying cell types in the tumor microenvironment using bulk data, highlighting its improved association with overall survival. Kyle Coleman introduced MISO, a multi-modal spatial omics method that integrates data from different modalities and aligns well with pathologist annotations. Venkatraman Seshan explored the application of James-Stein shrinkage estimation in principal components analysis (PCA), revealing overestimation in PCA and the potential application of shrinkage to these estimates.

5. **Immunogenomics.** Speakers included Wei Sun, Nicholas McGranahan, Vicky Yao and Ronglai Shen. Wei Sun presented his work on identifying co-occurrence of HLA allele and TCR clonal-type and predicting surface neoantigen generated by somatic mutations. Nicholas McGranahan presented work built on the TRACERx project. He discussed the role of HLA genes in tumor evolution and immune evasion. They found association of loss-of-heterozygosity (LOH) of HLA allele with elevated tumor mutation burden, and that alternate splicing of HLA molecule could change its function. Exon skipping in HLA domains is observed in 30% of lung adenocarcinomas. Ronglai Shen presented work on using Latent Dirichlet Allocation (LDA) models for flow cytometry analysis of cancer patients' blood samples to infer the immune contexture and pharmacodynamics upon immunotherapy exposure. She further presented a spatial LDA approach to delineate tumor microenvironment from multiplex imaging data. Vicky Yao explored the cancer-associated microbiome using semi-supervised NMF, identifying microbial signatures associated with survival.

4 Outcome of the Meeting

The 5-day workshop sponsored by BIRS included 38 participants attended in person and several virtual attendees with a mix of established and early-career investigators and 14 postdoc or graduate student trainees. The dynamic discussions and interactions led to several outcomes: dissemination of research results, techniques, and ideas; new connections, collaborations, and projects being formed; networking opportunities for trainees (postdocs and graduate students). The success of the workshop would be impossible to achieve without the support of BIRS.