Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Array Platform Modeling and Analysis (A)

- **Li-Xuan Qin**
  Associate Member of Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
- **Shuangge Ma**
  Associate Professor of Biostatistics, Yale University, New Haven, CT, USA.
- **Yen-Tsung Huang**
  Assistant Professor of Epidemiology, Brown University, Providence, RI, USA.
- **Hui Zhang**
  Assistant Member of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA.
- **Hongyuan Cao**
  Assistant Professor of Statistics, University of Missouri-Columbia, Columbia, MO, USA.

Supplement Aims and Scope

Cancer informatics represents a hybrid discipline encompassing the fields of oncology, computer science, bioinformatics, statistics, computational biology, genomics, proteomics, metabolomics, pharmacology, and quantitative epidemiology. The common bond or challenge that unifies the various disciplines is the need to bring order to the massive amounts of data generated by researchers and clinicians attempting to find the underlying causes and effective means of treating cancer.

The future cancer informatician will need to be well-versed in each of these fields and have the appropriate background to leverage the computational, clinical, and basic science resources necessary to understand their data and separate signal from noise. Knowledge of and the communication among these specialty disciplines, acting in unison, will be the key to success as we strive to find answers underlying the complex and often puzzling diseases known as cancer.

This supplement is focused on array platform modeling and analysis, and article topics may include:

- Reverse Phase Protein Arrays (RPPA)
- Single Nucleotide Polymorphism Arrays
- RNA Arrays
- Surface Adjustment and Tissue Array Profiling
- Normalization Methodology
- Multiplicative Spatial Effects
- Multiple Small Scale Variation Tools
- Insertions, Deletions, Microsatellites and Non-Polymorphic Variants
- mRNA Transcripts
- Physical Mapping
- Functional Analysis
- Multi-Dimensional Association Studies
- Evolutionary Analysis
- RefSNP Attribubutes
- Mendelian Inheritance
- Estimating Smooth Surface from Positive Controls
- Generalized Additive Modelling of Micro-Array Data
- Analysis of Spatial Artifacts
- Quantitative Intensity Modulation
- Molecular/Proteomic Profiling
- Reproducibility Metrics
- Transcript and Protein Expression
- Analysis of Signaling Pathways
- dbSNP and JSNP Database Search Tools
- HapMap
- Promiscuous Protein in Silico
- Geometric Scoring Criteria
- Mean Signal Intensity Ratio

The year of 2015 marks the 20th anniversary of the microarray technology. Since the publication of the first microarray paper in 1995, this high-throughput technology has revolutionized our ability to study the molecular features (such as RNA expression levels, DNA copy number changes, and epigenetic regulations) of tumors, and has greatly advanced our understanding of cancer and its prevention and treatment. We take the opportunity of this special issue to celebrate the past achievements in the modeling and analysis of microarray data and to highlight some recent methodologic development in this area.

The availability of high-throughput microarray data, characterized by their large scale and complex underlying structure, has led statisticians to develop highly innovative analytic methods for detecting potential drug targets and prognostic factors for cancer. Major advances in statistical methodologies have been made on issues such as data normalization, multiple comparison adjustment, and high-dimensional variable selection and classification. Nevertheless, there still remain analytic challenges that require development of better statistical methodologies in order to more fully reap the rich information that resides in microarray data.

In this special issue, we have put together a set of articles by leading researchers in the field of microarray data analysis. These articles present novel statistical approaches to a variety of current challenges for microarray data analysis. Some of these articles are summarized below by their intended type of molecular data.

- **RNA Expression:** Normalization has been shown to be an essential preprocessing step for RNA microarray data and useful methods have been developed to normalize mRNA data. Zhou et al assess the use of quantile normalization for microRNAs, a class of small RNAs that play a regulatory role in a cell, in relation to two other data preprocessing steps – log transformation and probeset summarization. Much of the microarray data analysis has defined 'interesting' genes as those with differential expression between two groups. An alternative and less explored definition is genes with outlying expression among samples. Ghosh et al explore the use of C(α) test for detecting outlier expression and develop a bivariate extension to this test to accommodate data from two platforms on the same sample set. For data collected in a longitudinal study, analysis using multivariate adaptive splines allows flexible modeling of the trajectory. Duan et al demonstrate the use of this method using a breast cancer data to characterize genes that have age-varying expression.
- **DNA Copy Number:** Copy number changes, due to insertions, deletions, or inversions of DNA, are a major source of genomic alterations in tumors. Li et al introduce a novel method based on hidden Markov models to detect copy number variants using SNP array data and demonstrate its use in a breast cancer dataset. This method first estimates copy number for each SNP in a single array and then standardizes the estimated copy number among multiple arrays. The relationship of copy number changes to clinical outcomes can be better understood in the context of biological pathways. Huang et al propose a novel statistical method to analyze copy number alterations of a pre-defined gene set (for example, a biological pathway) and use this method to investigate how cigarette smoking may affect the expression profile of 1814 biological pathways in non-small cell lung cancer.
- **DNA Epigenetic Regulation:** The transcriptional potential of DNA molecules can be regulated by epigenetic mechanisms in a cell such as DNA methylation and histone modification, without changes to underlying DNA sequences. The role of epigenetic regulation in carcinogenesis has been increasingly appreciated. Houseman et al examine the use of DNA methylation data to build phylogenetic classification for normal breast tissues and breast tumors using three publicly available breast cancer datasets and suggest a close relationship in epigenetic states between normal breast cells and breast cancer cells.
- **Integrative Analysis of Multiple Data Types:** Data are becoming increasingly available for multiple types of molecules on the same set of samples in the private and public domains such as the Gene Expression Omnibus (GEO) and the Cancer Genome Atlas (TCGA). When multi-platform genomic data are analyzed in an integrative framework, information can be enriched across data types and the regulatory relationships between molecules can be studied. Dellinger et al develop an innovative integrative method based on gene pathways using a graph-based learning algorithm to derive a classifier for clinical outcome. They apply this method to derive classifiers of tumor stage based on RNA expression levels, DNA copy number changes, and DNA methylation profiles, using data from TCGA.

The past two decades have seen considerable improvements in our ability to molecularly characterize the cancer genome and to quantitatively understand the genetic causes of the disease. Statistical thinking and methodologies have played an important role in leveraging the wealth of genomic data collected on microarrays and other more recent profiling technologies such as next generation sequencing. They will continue to be an integral part bridging genomics data and clinical practice in the personalized medicine era.

Lead Guest Editor **Dr Li-Xuan Qin**

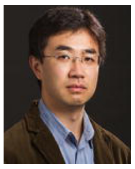
Dr Li-Xuan Qin is an Associate Member of Biostatistics at Memorial Sloan Kettering Cancer Center. She completed her PhD at the University of Washington. Her current work focuses on the statistical analysis of high-dimensional data for translational cancer research. Dr Qin is the author or co-author of many published papers, and holds several NIH grants as PI or co-investigator. She has been invited to give presentations in national/international conferences and academic departments.

qinl@mskcc.org
http://www.mskcc.org/research/epidemiology-biostatistics/biostatistics/staff/li-xuan-qin

Guest Editors

DR SHUANGGE MA

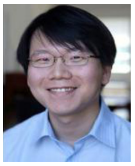
Dr Shuangge Ma is an Associate Professor of Biostatistics at Yale University. He completed his PhD at the University Of Wisconsin and did post-doctoral research at the University of Washington. His current research projects include the development of new statistical methodologies for complex data, and the study of epidemiology and pathogenesis of multiple cancers. Dr Ma is the author or co-author of many published papers and has presented at many conferences, and is an elected member of the International Statistical Institute and the American Statistical Association.



shuangge.ma@yale.edu
http://publichealth.yale.edu/people/shuangge_ma.profile

DR YEN-TSUNG HUANG

Dr Yen-Tsung Huang is an Assistant Professor of Epidemiology at Brown University. He completed his ScD at Harvard University. His research focuses on the incorporation of new biological discoveries into statistical methodologies for a better understanding of cancer genomics. Dr Huang is the author or co-author of 19 published papers and has presented at 12 conferences.



yen-tsung_huang@brown.edu
https://vivo.brown.edu/display/yh70

DR HUI ZHANG

Dr Hui Zhang is an Assistant Member of Biostatistics at St. Jude Children's Research Hospital. He completed his PhD at the University of Rochester. He now works primarily in the fields of categorical data analysis, count data in next generation sequencing, U-statistics extended nonparametric theory, and computational neuroscience. Dr Zhang is the author or co-author of 37 published papers and has been invited to present or chair sessions at multiple conferences.

Hui.Zhang@stjude.org
http://www.stjude.org/zhang_h

DR HONGYUAN CAO

Dr Hongyuan Cao is an Assistant Professor of Statistics at the University of Missouri-Columbia. She completed her PhD at the University of North Carolina, Chapel Hill, and has previously worked in the Health Studies Department at the University of Chicago. She now works primarily on high dimensional and large scale statistical inference, longitudinal data analysis, and survival analysis.

caohong@missouri.edu
http://faculty.missouri.edu/~caohong/index.shtml

SUPPLEMENT TITLE: Array Platform Modeling and Analysis (A)

CITATION: Qin et al. Array Platform Modeling and Analysis (A). *Cancer Informatics* 2014:13(S4) 91–93 doi: 10.4137/CIN.S22973

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Editorial

CORRESPONDENCE: qinl@mskcc.org

All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines.