Libertas Academica
FREEDOM TO RESEARCH

Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

- **Kellie J. Archer**
  Professor, Department of Biostatistics, Director of the Massey Cancer Center Biostatistics Shared Resource, Virginia Commonwealth University, Richmond, VA, USA.
- **Kevin Dobbin**
  Associate Professor of Biostatistics, University of Georgia, Athens, GA, USA.
- **Swati Biswas**
  Associate Professor, Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX, USA.
- **Roger S. Day**
  Associate Professor of Biomedical Informatics, Associate Professor of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA.
- **David C. Wheeler**
  Assistant Professor, Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA.
- **Hao Wu**
  Assistant Professor, Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, 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.

Articles focus on computer simulation, bioinformatics, and statistical analysis of cancer data and processes and may include:

- Multi-dimensional Simulation Models of Tumour Response
- Simulating Tumour Growth Dynamics
- Spatio-Temporal Simulation Models
- Parametric Validation of Simulation Models
- Simulation of Dynamic Phenomena in Cancer using Highly Specialized Algorithms
- Hyper-High Performance and Biocomplexity Systems Modelling of Cancer
- Sequence Alignment Methods for DNA-Seq, RNA-Seq, miRNA-Seq, CHiP-Seq Experiments
- Spectra Analysis
- Generic Visualization Tools
- Array-Comparative Genomic Hybridization Visualization
- Statistical Methods for Next Generation Sequencing Data
- Predictive Modeling for High-Dimensional Data
- Robust Feature Selection
- Integration and Analysis of Big Biomedical Data
- Meta-Data Imaging
- Equivalent Cross-Relaxation Imaging
- Mathematical Modeling and Image Enhancement of MRI Cancer Data
- Rapid Imaging Analysis of PET Cancer Scans

Cancer is a widely-recognized public health burden and a complex disease. High-dimensional "Omics" research – research that uses high dimensional genomic, proteomic, and/or metabolomic data – has often been motivated by cancer applications[1]. Important lessons have been learned in processing and analyzing such data. Further, additional sophisticated technologies in medical imaging, geographic information systems, and environmental exposure monitoring also give rise to high-dimensional data. Yet, as quickly as methods adapt to the new technologies, it seems that biotechnological developments produce even higher dimensional and more complex datasets – making bioinformatics more and more critical to the processing and understanding of patient data. But, in the rush of data processing we must not lose sight of biostatistics principles so that data are correctly interpreted, signal separated from noise, and new discoveries efficiently translated into clinical practice without unnecessary delays or inefficiencies. This supplement includes articles by various researchers in statistics, biostatistics and bioinformatics who are developing methods for addressing a wide spectrum of scientific questions related to big data. Some of these main areas are summarized below.

As mentioned, the challenges facing today in cancer informatics are multifaceted lying at the interface of bioinformatics, genetic epidemiology, epigenetics, and risk prediction, and several authors contributed in these areas. In particular, some papers focused on the approaches for detecting gene-gene and gene-environment interactions. Talluri and Shete considered an information theory approach for detecting epistasis, compared it with a standard logistic regression approach, and demonstrated utility by applying it to head and neck cancer data. Liu and Xuan constructed a mutual SNP association network (SAN) based on information from various sources of gene interaction such as protein-protein interaction and gene co-expression, with the goal that SAN reflects the real functional associations between genomic loci. Zhang and Biswas focused on detecting interactions of rare haplotypes with environmental covariates with an enhanced version of Logistic Bayesian LASSO and used it to uncover a rare haplotype interaction with smoking for lung cancer. Zang et al. considered lung cancer and explored the association of its sub-types with CDKN3 gene expression, and found that higher expression of CDKN3 was associated with poorer survival outcomes for lung adenocarcinoma but not for squamous cell carcinoma. Sun and Li developed a pipeline to identify hemi-methylation, that is, DNA methylation occurring in one strand only, and characterized different patterns in the entire genome, and applied the method to breast cancer cell lines. Mazzola et al. reviewed many recent enhanced capabilities of the widely used breast cancer genetic risk prediction model BRCAPRO that estimates the probability that a counselee carries mutations of BRCA1 or BRCA2 genes given her family history.

Other papers focused on extensions and applications of penalized or regularized models, such as the LASSO. Zemmour et al. took an important look back at centrally important breast cancer datasets. These datasets had previously been used to develop bioinformatic diagnostic tools that are currently in use. But these tools were developed when we knew less about the statistics of this type of data than we do today. They applied regularized methods that are becoming the gold standard in this area. They found that shorter and arguably more robust gene lists with equivalent prediction value are the result. They also found that the overlap of the gene lists from different modern tools is much higher than with the previous tools, suggesting that the previous poor overlap may have caused unnecessary concerns. Other articles pertained to extending penalized methods for discrete response modeling. For example, ordinal responses have an inherent ordering, but the distance between the responses cannot be quantitatively measured. Examples of ordinal variables include tumor grade (T0 to T4), degree of spread to the lymph nodes (N0 to N3), and cancer stage (I to IV), all of which are commonly reported in cancer research studies. The ordinal generalized monotone incremental forward stagewise method (GMIFS) was previously described for fitting ordinal response models in the presence of a high-dimensional covariate space. Ferber and Archer demonstrated how the GMIFS algorithm, via a forward continuation ratio model, could be used to fit a discrete survival time response model, applicable when survival is recorded on an ordinal scale such as short-, intermediate-, and long-term survival. Gentry et al extended the ordinal GMIFS algorithm to enable inclusion of covariates that should not be penalized in the model fitting process, and applied the method to predict stage of breast cancer using methylation of high-throughput CpG sites as penalized predictors with important clinical covariates coerced into the model. Makowski and Archer extended the GMIFS method to the Poisson regression setting for modeling a count response in high-dimensional covariate spaces, and applied the method for predicting micronuclei frequency using gene expression features as predictors.

In order to understand the gene regulatory mechanism in cancer, Wei P et al. proposed a novel statistical framework to integrate diverse types of genomic data from The Cancer Genome Atlas (TCGA) to decipher expression regulation. Wei Y et al. tackled a similar problem and proposed joint models for integrative analysis of multiple types of omics data. The results from the analysis shed light on personalized medicine. Statistical methods for differential expression analysis in both microarray and RNA-seq have been well-developed, however, the method for equivalent expression analysis is still seriously lacking. Cui et al. proposed a novel statistical method to test equivalent expressions among multiple treatment groups. Zhou et al. compared the traditional approach of univariate modeling to linear mixed effects modeling when assessing the relationship between psychoneurological symptoms (anxiety, depression, and stress) and methylation of CpG sites. O'Donnell et al. examined the exciting and emerging area of immunology signatures in cancer. Their goal was to

investigate the potential for development of an early detection and diagnostic tool that can be used on a simple blood draw. Their intriguing approach uses a peptide microarray and a time-frequency analysis. Zhao et al. reported on the development of informatics for a precision oncology – personalized medicine – clinical trial. The molecular profiling-based assignment of cancer therapy trial (MPACT) uses genomic profiling in novel ways to determine patient treatment. It is one of the first in what is likely to be a long line of such trials, and they describe how complex molecular characterization of the patients can be folded into clinical trial randomization and quality control. They show that careful development of bioinformatics tools can facilitate optimal interactions among diverse multidisciplinary teams composed of clinicians, molecular oncologists, statisticians and bioinformaticians.

Some articles pertained to analyzing the spatial dimension in cancer. A catchment area is the geographic area and population from which a cancer center draws patients, and defining a catchment area allows a cancer center to describe its primary patient population and assess how well it meets the needs of cancer patients. Wang and Wheeler estimated diagnosis and treatment catchment areas for the Massey Cancer Center at Virginia Commonwealth University using cancer registry and patient billing data and Bayesian hierarchical regression models. Historically, generalized additive models with bivariate smoothing functions have been applied to estimate spatial variation in cancer risk. Siangphoe and Wheeler evaluated the ability of different smoothing functions in generalized additive models to detect overall spatial variation of risk and elevated risk in diverse geographical areas using a simulation study.

Another aspect of spatial dimension in cancer is the heterogeneity within each tumor, especially in regard to changes over time. Diffusion tensor imaging provides a way to image brain tumors. To assess changes over time from treatment and tumor growth, it is necessary to register serial images against a high-quality reference of pure brain tissue. This allows comprehensive visualization of the entire tumor progression through the course of treatment. Ceschin et al provide an open source pipeline, sfDM (serial functional diffusion mapping), for registering and processing the images. They also evaluate different registration methods. Hobbs and Ng are concerned with monitoring tumor characteristics via perfusion imaging, and describe a model-based approach for inferring stability for stochastic curve estimation, a useful statistically-based technique.

Other articles described methods for assessing environmental chemical exposures. Environmental variables used in regression models to explain environmental chemical exposures or cancer outcomes are typically modeled at the same spatial scale. Grant, Gennings, and Wheeler presented four model selection algorithms that select the best spatial scale for each area-based covariate to explain variation in groundwater nitrate concentrations in Iowa. In the evaluation of cancer risk related to environmental chemical exposures, the effect of many chemicals on disease is ultimately of interest. The method of weighted quantile sum (WQS) regression attempts to overcome these problems by estimating a body burden index that identifies important chemicals in a mixture of correlated environmental chemicals. Czarnota, Gennings, and Wheeler assessed through simulation studies the accuracy of WQS regression in detecting subsets of chemicals associated with health outcomes and found that WQS regression had good sensitivity and specificity across a variety of conditions. The relationship between correlated environmental chemicals and health effects was not always constant across a study area, as exposure levels may change spatially due to various environmental factors. Czarnota, Wheeler, and Gennings assess through a simulation study the ability of geographically weighted regression (GWR) and geographically weighted lasso (GWL) to correctly identify spatially varying chemical effects for a mixture of correlated chemicals within a study area and find that GWR suffered from the reversal paradox, while GWL over-penalized the effects for the chemical most strongly related to the outcome.

A promising approach to cancer treatment leverages new understanding of the role of cancer stem cells (CSCs) and discovery of agents targeting them. Day uses tumor dynamics modeling to illustrate two special challenges in CSC-targeted treatment: how to design combination regimens that stand the best chance of success, and how to design clinical trials with time frames and endpoints sufficient to detect highly successful regimens that otherwise would be missed.

REFERENCE

1. McShane LM, Cavenagh MM, Lively TG, et al. (2013) Criteria for the use of omics-based predictors in clinical trials. Nature, 502: 317–20.

Lead Guest Editor **Dr. Kellie J. Archer**

Dr. Kellie J. Archer is a Professor in the Department of Biostatistics and Director of the Massey Cancer Center Biostatistics Shared Resource at Virginia Commonwealth University. She completed her PhD at The Ohio State University and previously worked there in support of research associated with the Cancer and Leukemia Group B (CALGB) Leukemia Correlative Sciences Committee. She now works primarily in developing innovative statistical methods and software for the analysis of high-dimensional datasets such as those arising from high-throughput genomic platforms. Dr. Archer is the author or co-author of 99 published papers, two book chapters, has 30 university seminars/professional conference presentations, holds an editorial appointment at *Progress in Transplantation* and is a Statistical Consultant for *Radiology* and the *Nature* Publishing Group.

kjarcher@vcu.edu
http://www.people.vcu.edu/~kjarcher/

Guest Editors

KEVIN DOBBIN

Dr. Kevin Dobbin is an Associate Professor of Biostatistics at the University of Georgia. He completed his PhD at the University of Minnesota and has previously worked at the National Cancer Institute. He now works primarily in high-dimensional data, experimental design, classifier development and validation, cancer biomarkers, and causal modeling. Dr. Dobbin is the author or co-author of 34 published papers and has presented at 31 conferences, and holds editorial appointments at *Biometrics*, *Scandinavian Journal of Statistics*, *Journal of the National Cancer Institute* and is a Statistical Consultant for the *Nature* Publishing Group.

dobbinke@uga.edu
http://www.dobbinuga.com

SWATI BISWAS

Dr. Swati Biswas is an Associate Professor in the Department of Mathematical Sciences at the University of Texas at Dallas. She completed her PhD at The Ohio State University, her postdoctoral training at the M D Anderson Cancer Center, and has previously worked at the University of North Texas Health Science Center. Her research interests include statistical genetics, genetic epidemiology, cancer genetics, risk prediction models, and Bayesian cancer trials. Dr. Biswas is the author or co-author of 25 peer-reviewed published papers and has presented at 17 conferences. She has several research grants from NIH and other organizations as a PI or co-investigator.

swati.biswas@utdallas.edu
http://www.utdallas.edu/~swati.biswas

ROGER S. DAY

Dr. Roger S. Day is an Associate Professor of Biomedical Informatics and Associate Professor of Biostatistics and a member of the biomedical informatics training program core faculty at the University of Pittsburgh. He completed his ScD in biostatistics at Harvard School of Public Health. At the University of Pittsburgh Cancer Institute, he developed and ran the biostatistics facility for 14 years. His current research focuses primarily on evaluating identifier mapping and filtering methods for high-throughput proteomics and expression platforms, promoting innovative and ethical design and execution of clinical trials, software architecture for comprehensive cancer modeling and validation, multi-scale modeling in cancer, strategies for overcoming drug resistance in cancer, and novel approaches to modeling pharmaceutical and biological interactions. Dr. Day is the author or co-author of 80 published peer-reviewed or proceedings papers. He is a devoted teacher and mentor, and highly active musician.

day01@pitt.edu
http://www.pittsburghartistregistry.org/accounts/view/RogerDayPanEthnicTuba

DAVID C. WHEELER

Dr. David C. Wheeler is an Assistant Professor in the Department of Biostatistics at the Virginia Commonwealth University. He completed his PhD at The Ohio State University and his MPH at Harvard University. He was previously a postdoctoral fellow at Emory University and a Cancer Prevention Fellow at the National Cancer Institute. His research areas are spatial epidemiology and cancer control and prevention with a focus on environmental and occupational risk factors. Dr. Wheeler is the author or co-author of 50 peer-reviewed publications and has presented at 39 conferences.

dcwheeler@vcu.edu
http://www.biostatistics.vcu.edu/david-c-wheeler/

HAO WU

Dr. Hao Wu is an Assistant Professor at the Department of Biostatistics and Bioinformatics at Emory University. He joined the department in 2010 after obtaining Ph.D. in Biostatistics from Johns Hopkins University. Dr. Wu's researches have been mainly focused on bioinformatics and computational biology. He is particularly interested in developing statistical methods and computational tools for interpreting large-scale genomic data from high-throughput technologies such as microarrays and second-generation sequencing. He collaborates closely with researchers working on epigenetics to characterize different types of DNA methylation and histone modifications. Dr. Wu is the author or co-author of 37 published papers in peer-reviewed journals, two book chapters, and has over 30 invited seminars at different academic institutes and professional conferences.

hao.wu@emory.edu
http://web1.sph.emory.edu/users/hwu30/

SUPPLEMENT TITLE: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

CITATION: Archer et al. Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes. *Cancer Informatics* 2015:14(S2) 247–251
doi: 10.4137/CIN.S32525

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Editorial

FUNDING: Authors disclose no funding sources.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: kjarcher@vcu.edu

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.