Published in final edited form as:
Stat Modelling. 2017 August ; 17(4-5): 338–357. doi:10.1177/1471082X17728576.

Rejoinder to statistical contributions to bioinformatics: Design, modelling, structure learning and Integration

Jeffrey S Morris[1] and **Veerabhadran Baladandayuthapani**[1]

[1]Department of Biostatistics, UT MD Anderson Cancer Center, Houston, Texas, USA

Abstract

We thank the discussants for their kind comments and their insightful analysis and discussion that has substantially added to the contribution of this issue. Overall, it seems the discussants have affirmed many of our primary points, and have also raised a number of other relevant and important issues that we did not emphasize in the paper. Several common threads emerged from these discussions, including the importance of software development, appropriate dissemination, and close collaboration with biomedical scientists and technology experts in order to ensure our work is relevant and impactful. Each discussant also mentioned other areas of bioinformatics that have been impacted by statistical researchers that we did not highlight in the original paper. In response, we will first summarize discuss these general themes, and then respond to specific comments of each discussant, and finally talk about the additional areas of bioinformatics impacted by statisticians that were mentioned by the reviewers.

1 General Themes

1.1 Missed Opportunities

All of the discussants seemed to agree with our premise that, while statisticians have made some incredible impacts on molecular biology and medicine through bioinformatics, our field has missed opportunities to establish ourselves as more central leaders in the field. Micheal Newton in his review made the interesting point that in the mid to late 1990's as high-throughput genomics was beginning to take off, the already- established field of statistical genomics was well positioned to take leadership in this arena, but failed to "throw their hats into the ring." The reasons for this are not clear – were they already busy with the set of DNA-based QTL and disease mapping problems that was their focus and were not interested in branching out into this new area, did they not think that this field would grow as it eventually did, or did the technical preprocessing problems that dominated much of the early work turn them off? It is not clear why more in the statistical community did not get involved on the front lines as this field arose, but our hope is that our statistical community does not miss such opportunities in the future as new paradigm-changing technologies emerge.

Address for correspondence: Jeffrey S Morris, Department of Biostatistics, UT MD Anderson Cancer Center, 1515 Holcombe Blvd, Houston, Texas, USA. jefmorris@mdanderson.org. **Phone:** (+1) 713 563 4284. **Fax:** (+1) 713 563 4242.

In reading the discussions, we see three key practical points emerge that can help us avoid these missed opportunities for statisticians to make a stronger impact as leaders of the field: the dissemination of freely available, implementable software, an increased focus on publishing our work in subject area and practice-oriented journals that may be higher impact and more visible to the biomedical community, and a commitment to work closely with biomedical collaborators and technology experts to guide our work towards the most important problems and ensure we provide relevant input into all steps of collecting and processing the data.

1.2 Software

As emphasized by Kechris and Ghosh, and by Ma, Song, and Tseng, the availability and dissemination of implementable software is crucial for new statistical methods to make a practical impact on science. We thank them for bringing up this important point, a point with which we strongly agree but did not emphasize. Indeed, one major component of all of the examples we included of statistical impacts on bioinformatics was the development of freely available, usable software to implement the underlying methods. The development of the *bioconductor* package for *R* has strongly enabled this process, as it established *R*, the preferred software for statistical methodology developers, as the most commonly used platform for molecular biologists and provided an open-source package onto which their contributions can be added and quickly disseminated.

However, not all statistical methods developed are accompanied by usable software. One reason for this is that the culture of the statistical methodology community has perhaps undervalued the importance of providing freely available, practically usable software when new methods are published. These software should not just work, but be written in general enough terms so they can be broadly applied to new problems, should be well-documented to enable users to effectively apply and adapt them for their uses, and should be computationally efficient enough to scale up to the big data settings for which they were intended.

Besides methodology, another area of software development in which statisticians as a group have not thought deeply enough is visualization. Given the complexity and volume of high-dimensional data, the provision of computationally efficient tools to visualize these complex data is a major need of practitioners, and nowadays deeply valued and rewarded. For example, colleagues in the Department of Bioinformatics and Computational Biology at the University of Texas M.D. Anderson Cancer Center, led by Bradley Broom and John Weinstein, have developed next- generation clustered heatmaps (NG-CHM, http://bioinformatics.mdanderson.org/TCGA/NGCHMPortal/), an online tool that uses clever multi-resolution data representations to produce these useful graphical summaries for enormously sized multi-platform genomics data sets along with associated covariates. In many of these cases, it is not even feasible to load the raw data into the computer at once, but their clever data management and down-sampling strategy allows users to visualize the heatmap for the entire large data set at a lower resolution, and then zoom in to regions of interest as the software produces higher resolution maps showing more detail. This work has been utilized in most of the high impact publications emanating from the The Cancer

Genome Atlas (TCGA) initiative sponsored by the NCI, and has been highlighted in major journals including *Nature*. Statisticians are able to uncover deep insights from the complex, big data generated by emerging technologies through their innovative methods. As statisticians begin to value the development of visualization techniques, the results from their methods will become more clearly visualized and understood by their collaborators, and this will in turn increase the practical impact of their work. By partnering with other quantitative scientists including computer scientists and biomedical informaticians, statisticians can work in these interdisciplinary teams to develop outstanding and impactful tools.

Unfortunately, the reward system in the statistical community has been more aligned towards innovation and mathematical depth than the provision of implementable software or visualization tools. As a result, there are many potentially impactful methods developed by statisticians appearing in top statistical journals that are not appreciated or being used by others because of the lack of such software. This has begun to change in recent years, as more journals are requesting or even requiring the sharing of usable software with published statistical methodology papers, and some journals publishing articles on data visualization approaches. Some are even requesting or requiring the sharing of data used in the paper, which as emphasized by Houwing-Duistermaat, Uh, and Gusnanto, can also contribute to greater reproducibility and more impactful work. We need to keep this trend moving forward, and purposefully choose to value the development of good software when evaluating methodological contributions of our colleagues and trainees. As this trend continues, there is promise that the statistical community's work can make a broader and deeper impact on science.

1.3 Dissemination

Another aspect of our reward system in the statistical community contributing to our limited impact is the journals we value. Understandably, the statistical research community tends to most strongly value publication in the top-level statistical methodology journals. Perhaps this is appropriate, as we strive for these to include the best new methodological work our profession has to offer, and we should recognize what we as a field value as the strongest work. However, an unfortunate by-product of this emphasis on publishing in statistical journals is that we undervalue contributions to other journals, especially subject-matter journals. The publishing of new methods in subject-matter journals can be challenging in its own right, and can have even greater potential to impact a scientific field than a paper in a statistical journal. This will require a change in culture and pedagogy – but is already happening to some degree in multiple academic departments.

As the low impact factors of even the highest ranked statistical journals can attest, few researchers outside of the statistical community read or cite statistical papers. By contrast, papers in subject-area journals providing statistical guidance or introducing new statistical methods to a specific scientific field can have exceptionally high citation counts, in the hundreds or thousands, and can effect change in statistical practice within that field. Statisticians may or may not think the methods in these papers are the best, but the point remains that articles providing clear articulations of new methods in the subject area journals, speaking the language of that field and addressing what they recognize as their key quantitative challenges, are heavily read by biological and medical researchers and effect change in the corresponding fields of science. More statisticians should be publishing articles like these in order to better disseminate our methods and provide guidance on best statistical practice and thus exercising quantitative leadership in these fields.

The discussions of Kechris and Ghosh and Ma, Song, and Tseng both hit on this point. Kechris and Ghosh emphasize that statisticians tend to be slow to the game and then when they do publish, our statistical journals are so slow that by the time the work appears, it can already be obsolete. By publishing only in statistical journals and not in the high impact subject-area journals, many statisticians are forfeiting the opportunity for their methods to become more widely known and make an impact in science. If many of us in our profession continue to make this choice, we as a group will forfeit our leadership of statistical analysis practice within the scientific community and will miss opportunities to have a seat at the table of influence in science and society. In these cases, undoubtedly other quantitative scientists will step up in our place.

As emphasized by Kechris and Ghosh, there are practical steps we can take to make stronger impacts. We can publish the deeper mathematical principles and technical details in a statistical paper, but then publish more applied papers showing how a new method makes a difference in a subject area journal or we can publish a software focused paper in a bioinformatics journal. In this way, we can still publish the best methods in top statistical methodology journals yet still have our work disseminated in other venues that will reach a broader community of researchers. While it takes greater effort to learn how to write a paper that resonates with the broader audience of these subject area journals, this effort will pay off in greater impact and position our field well for the future.

1.4 Interdisciplinary collaboration

Various discussants highlighted the importance of interdisciplinary collaboration and effective communication and coordination between biomedical scientists and statisticians if we are to make a practical impact on the science. Baggerly highlighted that it is crucial for statisticians to understand the biological goals of our collaborators, and that communication is key for us to be able to incorporate biological knowledge in an intelligent way into the analysis. Houwing-Duistermaat, Uh, and Gusnanto mention their training program “Innovative Methods for Future Datasets” that seeks to educate both young statisticians and young chemists and biologists about experimental design, reproducible research, and preprocessing of noisy -omics data, and effectively engenders communications between them. Kechris and Ghosh emphasize that statisticians should directly work with biomedical investigators so they can develop methods in context and provide results that are as interpretable as possible so they resonate with biomedical scientists and are most likely to get used. In the context of mass spectrometry metabolomics and proteomics, Dowsey emphasizes the need for statisticians to be more heavily involved in the scientific process, working closely with other scientists so they can influence the design and operation of the instrumentation, and even optimize the acquisition process itself. Our own experiences, of working in premier research-oriented cancer center, attest to this fact.

Newton's discussion contains a deep and interesting discussion of this topic. He introduces the terminology "contextual features" to deal with the biological underpinnings of measurement technologies that must be understood for proper analysis. He emphasizes that statisticians tend to partition out these contextual issues in their attempt to distill the context-specific analytical problems into generic data problems, and that this practice can make it hard for the statistician to address the most important problems of the science. He characterizes bioinformatics as operating as a "three-legged stool" of context, statistics, and computational details, with the implication that all three are equally needed and work together in effective methods. We wholeheartedly agree with this well-conceived characterization, and with the insights mentioned by the discussants.

2 Response to Specific Discussants

2.1 Keith Baggerly

We thank Keith Baggerly (2017, KAB17 henceforth) for his discussion of our article. We consider Keith an international leader in reproducible research with a detailed eye and mind for the key fundamental issues often missed or glossed over by other statisticians, and with a no-nonsense approach to seek to uncover the true insights contained in these data while being extra careful to look out for spurious results. It is reassuring for us to see his agreement with our premise of the major role of statisticians in bioinformatics and his sympathy to the goal of making fuller use of the data. He also augments the discussion with several crucial points, on which we now briefly comment.

Robustness to "Garbage" Observations—KAB17 brings up the key point that the complexity and high dimensionality of these data mean that occasionally there are aberrant observations, and these can be difficult to detect. This makes *robustness* another key property we would like our methods to possess. He highlights examples in DChip and RMA microarray preprocessing as well as RNAseq analysis of gene counts. We think that robustness could be added as another fundamental principle underlying many statistical methods that have made strong contributions to bioinformatics.

When to Say "There's Nothing There"—Another excellent point by KAB17 is the importance of being able to identify when there is no signal strong enough in the data such that we should act. Our ability to find structure in any complex big data set along with the constant pressure on scientists to find results and publish produces papers with false positive or practically irrelevant results that can be misleading and distract scientists from discovering the key principles that are in fact driving the science.

As we mentioned in our article, one of the unique perspectives we statisticians bring to the table is a deep understanding of variability and uncertainty, which allows us to provide inferential summaries and probability statements, not just point estimates. Some quantitative scientists focus on providing a point estimate that optimizes the data, without a rigorous sense of how likely the results are to characterize the underlying true process. Our inferential procedures allow us to make probability statements that account for multiple testing, such as experimentwise error rate and FDR methods, and thus have the potential to tell us when there is "nothing there." Now, this idea of "uncertainty quantification" has grown in prominence among other quantitative scientists, but we statisticians are experts in understanding these concepts and should be the ones to assert ourselves as leaders.

Baggerly also briefly mentions the importance of considering both statistical and practical significance in order to ensure results reported are likely to be real and important. We strongly agree with this point, and put this into practice in our Bayesian modeling whereby we flag results that have high probabilities of some minimum practical effect size, allowing us to account for both statistical and practical significance in our inference.

Positive and Negative Controls—Another great point in KAB17 is the important place positive and negative controls, so-called “sanity checks”, can play in the process of trying to discover insights from our data and “separate the wheat from the chaff.” We point out that this is another area in which close communication and coordination between biomedical scientists and statisticians is key, as often statisticians themselves will not know how to choose appropriate controls, while subject area scientists know which results should be there or should not be there if the assay worked out right.

Extending Results to Integrative Analysis—KAB17 points out that these diagnostics are harder to do in the context of multi-platform integrative analysis. We agree that with more data types, there is more checking to do, raising new challenges. On the other hand, the inherent relationships among these platforms provide new structure that can be used for quality control checks, positive/negative controls, or to provide a finer focus on biologically relevant effects.

For example, in a given context certain genes may be known to be regulated by particular transcription factors, methylation, or copy number changes. Integrative analyses allow us to investigate that these relationships are indeed apparent and in the expected direction. Also, these inter-platform relationships allow us to reduce the model space by focusing on the most biologically relevant changes, thus reducing false positives and gaining power to detect significant differences. For example, if our goal is to find differentially methylated regions, rather than focusing on the entire set of 450k CpGs across the entire genome, we can first filter those CpGs whose methylation is associated with gene expression, and then only considering those. This reduces the number of multiple tests and focuses on those most likely to be biologically relevant. Cross-platform integrative modeling can effectively increase the sample size of available information and provide additional power for discoveries.

2.2 Michael A. Newton

We thank Michael Newton (MAN17) for his insightful discussion of a number of key issues in the application of statistical methods to bioinformatics. His extensive experience and unique perspective leads him to provide some deep insights and viewpoints on these issues. We comment on some of them here.

Contextual Features—As discussed in Section 1 above, Newton introduces the term “contextual features” to refer to scientific background that needs to be taken into account for effective bioinformatics research and through his three-legged-stool characterization, emphasizes that taking these contextual features into account is equally important as the statistical and computational details of the underlying algorithms. We thank him for this apt analogy, which also highlights the need for deep collaboration and effective communication with biomedical collaborators and technology experts, since most statisticians will not be experts in these contextual issues.

Transcription Factor Binding Sites and Model-Based Methods—We are glad MAN17 included an extended discussion of the work of Chip Lawrence, Jun Liu, and others to develop methods to identify transcription factor binding sites. These methods were on our short list of methods to emphasize given their sophistication and effectiveness in solving a challenging problem, but we omitted this example for brevity’s sake. We are glad Newton is the one to discuss these issues, as he pointed out a fact that may have eluded us and we would not have mentioned: that this is an example whereby model-based methods have taken the field forward in a way that the usual algorithmic-based approaches could not. This is an outstanding point and worth further consideration.

Algorithmic approaches have dominated in a vast majority of problems in bioinformatics, with model-based methods underutilized and sometimes even viewed with skepticism by many practitioners. This field was one where the use of statistical models provided key insights not contained in the data themselves, namely the key patterns in the model-estimated probabilities. He mentions this key insight has made an enormous impact in many other areas of sequence analysis. We add to this that this example also illustrates the benefits of unified modeling, learning probabilities across and within samples through a hierarchical model, to provide improved inference.

Benefits and Drawbacks of Model-based Approaches—Newton recognizes our paper as providing an “impressive review of model-based approaches,” illustrating in a broader sense that these can provide great statistical benefits and insights that might be missed by algorithmic approaches. He agrees that the potential to share information across data elements helps mitigate the “big p, small n” problem characterizing this field, and to integrate multiple data types. He mentions that models provide analysts with a “powerful language” that multi-step algorithmic approaches cannot. We believe that carefully constructed models that account for contextual features like directionality of association between genomic platforms or connection of molecules working within a given biologically established pathway, can increase power as well as yield interpretable results summarized in a way that makes it easy for biomedical scientists to understand, explain, and connect to existing knowledge. They also produce various measures of uncertainty quantification to help investigators know how strong the apparent patterns are and perform informed cost-benefit analyses to determine which discoveries are worthwhile to prioritize and follow up.

Hazard of Model Misspecification—While Newton agrees with these potential upsides of model-based approaches, he also warns about inconsistency that can result from model misspecification. We agree with him that one must be careful, but believe that robustness to assumptions is not unique to model-based approaches, that sensitivity analyses can be done to assess dependence of results on model choice, and use of flexible, adaptive models can mitigate some of the misspecification concern.

First, one of my most influential mentors was Emmanuel Parzen, one of the great pioneers in time series and kernel density estimation as well as the author of the seminal Probability Theory textbook that educated a generation of mathematical statisticians. One saying he liked to make that stuck with me and shaped my understanding of statistics was (paraphrased by me), "When you get results from a statistical procedure, you need to ask yourself what part of the answer came from the data, and what part came from the model." This principle is not unique to model-based approaches, but is true of any statistical method, as all statistical methods have underlying assumptions that must be made and can have an impact on the resulting inference. This issue is most commonly raised to criticize model-based approaches, and especially Bayesian ones where the prior is another element of the model that must be somehow specified, but it is also true of algorithmic or seemingly nonparametric approaches. There are always assumptions, and we need to test these assumptions and assess robustness of our results to them whenever possible.

We believe this is an essential practice that must be followed by careful statisticians, especially in the context of complex, high-dimensional data, since models may have to make assumptions about multiple aspects of the data and any of these assumptions may have a strong impact on results. We believe that various types of model checking should routinely be done, including graphical and statistical testing of assumptions as well as sensitivity analyses tweaking the model by varying certain assumptions about structural, distributional, and prior characteristics as well as regularization parameters, and the nassessing persistence of the key results across these conditions.

In much of our own work on models for complex, high-dimensional data, we aim to build models that, while borrowing strength across multiple data elements to gain efficiency and reduce effective dimensionality, are flexible enough to adapt to the key features of the data. This approach can mitigate some of the inherent bias of model-based approaches that is the source of Newton's concern, while still effectively discovering and accounting for structure in these data elements to yield increased efficiency over piecemeal elementwise approaches. Indeed, this is the key motivation behind the large set of works we have done on Bayesian functional mixed models for complex, high-dimensional functional data. These principles also extend to other complex modeling efforts such as integrative genomics and graph-based methods to discover associations.

Superficial or Superfluous, Needlessly Complex Models—Another insightful point made by MAN17 is that at times, model-based approaches can be "superficial or superfluous," unnecessarily complex and adding little practical benefit relative to simpler algorithmic approaches. While it is obvious by our own publication record that we like model-based approaches, we also understand this criticism and potential downside to these strategies. As we mentioned in the article, we do not believe they are the best for all circumstances, but have highlighted them since we believe their strengths are perhaps underappreciated and their drawbacks overemphasized, and wanted to attempt to motivate more researchers to exploit their benefits.

One must assess when it is worthwhile to utilize a complex modeling approach and when simpler, perhaps algorithmic approaches are sufficient. The key factor is that, like most anything, there are tradeoffs that must be considered. When utilizing a complex model or statistical method, there is an inherent “complexity cost” that is incurred. This cost takes various forms, from extra time to build the model and assess its fit, a greater level of expertise to understand the various components of the model, the greater difficulty explaining it and getting readers and reviewers to understand it, and the risk of a reviewer rejecting the paper because of a lack of understanding of the innovative model. In many cases, it is much easier to use the simplest possible method, especially if the simple method is what scientists in the given field are used to seeing. This complexity cost is real, and should be taken into account.

However, at times, advanced methods, including complex, flexible modeling frameworks, can yield significant benefits that make them worthy of use. These benefits can come in the form of improved prediction, increased power for discoveries, and reduced of false discovery rates, and are provided by the data integration, dimension reduction, and incorporation of scientific information undergirding effective unified modeling approaches. In cases whereby these benefits are realized, the upside of the modeling can justify the complexity cost. We believe that it is the bounden duty of the statistical modeler to assess the benefits of any more complex model vs. the inherent complexity cost, and should be a regular query we raise with authors of such methods in statistical journals.

Bright Future for Statistical Bioinformatics—In light of his great experience, we are encouraged by Newton’s assertion that he sees a bright future for statistical bioinformatics. He cites forensic bioinformatics as bringing attention to statisticians as having an essential role in the field (we agree), and the emergence of data science as a recognized discipline in biomedicine and other field bringing the problems of complex, high-dimensional data into the forefront of the mind of scientists. He provides examples of some emerging areas bringing significant challenges, including electronic detection systems and single cell measurements. It is the challenges raised by these emerging areas that provide statisticians with the opportunity to get heavily involved early, and establish themselves as quantitative leaders in the field.

2.3 Genevera I. Allen

We thank the Genevera Allen (2017, GIA17 henceforth) for an insightful discussion of our article, noting that our article is a “comprehensive and compelling review” and in particular, highlighting some major challenges and opportunities in nascent field of data integration. We summarize and respond to some of these comments below.

Data Integration vs. Integromics—GIA17 prefer the term “data integration” to “integromics”. Our usage of the term was mainly from a bioinformatics perspective, but we do agree that data integration spans a broader context. Modern biomedical research has generated unprecedented amounts of data. A combination of clinical, environmental and public health information, proliferation of associated genomic information, and increasingly complex digital information, from sources like imaging, electronic health records, social media, mobile health, have created unique challenges in assimilating, organizing, analyzing and interpreting such structured data. In our view, statistics can not only only maximize access to, and usability of, such data to enhance, improve, and inform decision making, but it can also translate basic research into the evidence-based healthcare decision-making process through rigorous analysis and integration of the available information.

Newton also expressed his dislike for this term "integromics." Allen suggested "data integration". We can sympathize with the over-use of the -omics suffix, and are not wed to the term integromics. However, we believe that for these types of analyses combining information across multiple modalities, integration is the best word, so perhaps "integrative analysis" could be used. Whatever the term, we believe this is among the most important areas of quantitative science, and we statisticians need to be on the front line of developing powerful methods to integrate information across multiple modalities to gain deeper knowledge of the underlying biomedical processes. We need to make sure that we do not miss the opportunity to establish ourselves as leaders in this area.

Data analyses challenges—As GIA17 also point out, this comes with major challenges from both methodological and practical perspectives. From a conceptual point of view, data integration is actually the reverse of classical meta-analyses problems in terms of the inherent scientific and biological questions being tackled – hence the need for new integrative statistical methods. These issues are further exacerbated by the facts that each view/mode of the data can be ultra-high dimensional and measured on completely different scales, leading to mixed-data types. The latter, in particular, is a very relevant problem, for which few statistical methods are currently available. From a practical perspective, this encompass a few data challenges as well.

The first is to obtain the relevant data that can contain the information to answer important scientific questions. Various national/international consortium level efforts have been made at this level which includes The Cancer Genome Atlas (TCGA, cancergenome.nih.gov) and International Cancer Genomics Consortium (ICGC, icgc.org). Many data extraction pipelines have been developed to extract formatted and analyzeable data from these portals, especially TCGA. GIA17 mentioned `TCGA2STAT` and some others in this domain include `TCGAAssembler` (Zhu et al., 2014) and a standardized data portal available publicly at http://bioinformatics.mdanderson.org/TCGA/databrowser/ – developed at our home institution MD Anderson Cancer Center. GIA17 also emphasize the importance of a collaborative team in these endeavors and it is vital to involve collaborators who are biological and/or clinical experts to focus on the most relevant scientific questions. Furthermore, it is also critical that the preprocessing steps for each of data types are fully documented to ensure reproducibility of the process.

From an analyses perspective, a major issue is existence of batch effects, since typically these data types are collected at different time points and locations – which might induce technical artifacts, not attributable to any real biology. We alluded to this in our main article and would like to re-emphasize the importance of this step, and point out with increasing types of data generated – there exists a rich opportunity for methodological developments in this area. Another issue is missing data, which currently poses a major challenge for developing coherent statistical methods, especially model-based – since the "matching" of both samples and platforms, often leads to severe loss of available complete cases. We believe this can be a very fertile ground for developing new missing data methods, especially those where the biological knowledge about the interactions within and between platforms can be brought into the methodologies.

Statistical modeling challenges—The area of integrative analyses has given rise to a plethora of opportunities for developing advanced statistical modeling frameworks for such information-rich, complex structured and high-dimensional datasets. GIA17 bring forth some recent advances in this area. For prediction, there are several methods that can handle mixed data types, mostly from machine learning or ensemble-type approaches to merge predictions from different data types. Although they are not ideally suited for high-dimensional data, one could use a prior feature selection step before fitting such models. However, coherent joint prediction methods with multi-view data is still an open area of investigation.

Moving beyond prediction, the goal of many multi-platform genomics studies is to generate new biological/scientific hypothesis from multi-view data that might have some potential clinical or translational relevance, and endeavor that could be callsed “discovery science.” Most of the recent advances in this area have been through exploratory learning methods e.g. data visualization, dimension reduction, pattern recognition, clustering, feature selection, network structure learning. There are few existing methods for integrative learning as most of these methods model in a latent space, such as lower-dimensonal summaries provided by matrix decompositions, which may not capture all the dependencies. Also, many of these tools are only available for continuous data and hence, not ideally suited to mixed data types. This might be another germane area for investigation.

Integration via mixed graphical models—GIA17 review some of their recent work on graphical models for mixed data, which is an outstanding contribution that nicely complements the developments we detailed in our main paper. Graphical models for mixed data can be constructed assuming univariate exponential family distributions and the extended to multi-view setting allowing potentially different exponential families in nodes. Alternatively, chain graphical models with Gaussian models conditional on discrete (Ising) model is another approach, but are only applicable to continuous and discrete data. GIA17’s recent work combines the concepts of chain graphical models and graphical models via exponential families to formulate mixed chain graphical models for which the variables are written as a chain of conditional distributions, each arising from potentially different exponential families, and admitting directional dependencies between the chain links such as mutation → RNA. This strategy greatly simplifies the calculations, allows a more flexible class of dependencies, and effectively incorporates biological information into the modeling through these directionalities.

We would also like highlight a recent Bayesian treatment of this problem. Recently Bhadra et al. (2017) proposed a unified Bayesian graphical modeling procedure for inferring dependence structure for mixed as well non-normal data-types using Gaussian scale mixtures. They introduce the concept of a *conditional sign independence* that captures stochastic independence in terms of signs for different variables as opposed to magnitude. This conditional sign independence metric is especially relevant to mixed-data types as it has a very intuitive interpretation. For example, in mixed- data context, it might not make sense to compare actual numeric values, e.g. mRNA, to binary data, e.g. mutation data, but rather one might be interested to evaluate if positive values indicating presence of mutation co-occurs with upregulation of some gene conditional on the rest of the variables of interest. Conversely, one might also want to investigate if two arbitrarily coded binary deleterious mutations are likely to co-occur, accounting for the effect of the rest of the variables. This makes conditional sign independence a versatile tool for establishing networks between mixed-data types.

2.4 Katerina Kechris and Debashis Ghosh

We thanks Kechris and Ghosh (KG17) for their thoughtful and deeply illuminating review of our article. The have brought out several complementary points emphasizing the importance of dissemination and interpretability. We highlight and comment on some of the issues raised in their review.

Other methods/technologies and common themes—Like MAN17, KG17 also highlight that statistical approaches in bioinformatics started in the early days of sequence databases in early 1970's, which encompassed likelihood-based methods for phylogenies and extreme value distributions for scoring alignments. We also thank them for pointing out several new technologies that we missed in our review. These include new sequencing technologies: Hi-C for chromosome conformation, GRO-Seq for real-time transcription, HITS-CLIP for RNA binding as well as metabolomics/proteomics using MS and proteomics with aptamer based technologies (SOMAscan). As these technologies mature and advance, statisticians are critically placed to contribute to this field (see Section 3 as well)

They also point out the constantly changing nature of high-throughput technologies, although "exciting and rewarding", present some challenges, for example requiring the updating of courses. They also agree that there are some constant themes that emerge regardless of technology, so by applying principles we have learned on older technologies forward to newer technologies, we avoid trying to "reinvent the wheel." As we pointed in our review as well, they iterated the importance of primacy of pre- processing, high-dimensional ($p >> n$) problems, and structured dependencies in the variables and data such as higher-order biological interactions.

Dissemination—KG17 mention how dissemination of statistical methods can be improved which will have increase the impact of our discipline. As we highlighted in Section 1 of this rejoinder, KG17 described a repeating pattern that partially explains the missed opportunities for impact by statisticians. In particular, our "slowness" to the process, extended publication times and publishing in statistical literature which is usually ignored by practitioners. They point out a possible solution/strategy/model whereby the core statistical methods are published in a main-stream statistical journal and the software in presented in a bioinformatics journal for greater visibility.

Tradeoff of speed and complexity—KG17 highlight that more effort in methods development on implementation and efficient computations would help our profession have a bigger impact. They highlight the need to balance statistical complexity and ease of use of accompanying software. This is especially true for the Bayesian community that needs to pay more attention to efficient computations for large- scale data sets through variational Bayes, GPU-based computing and divide/conquer Markov Chain Monte Carlo (MCMC).

Interpretability—KG17 also emphasize the importance of interpretability. They state that statisticians should work directly with biomedical investigators so they can develop methods in context and provide results as interpretable as possible. As we have found out as well, practitioners place a high value on interpretation of the results, such that results that resonate with biomedical scientists and fit into their paradigms are most likely to be recognized and built upon. These include some basic outputs like effect sizes, coefficients, predictive values and appropriate measures of significance including p-values and posterior probabilities. They bring out the importance of getting involved in experimental design stage, and how a failure of statisticians and biomedical investigators to connect can lead to bad science.

2.5 Tianzhou Ma, Chi Song and George C. Tseng

We thank Ma, Song and Tseng (MST17) for a very thorough and highly informative review of our article. MST17 have made several points, and as true statisticians even backed their points by empirical justifications, that are both very complementary as well as in-line with points we have made. We highlight some of these below.

Leadership roles of statisticians in Bioinformatics—MST17 make a very good point about how statisticians have evolved from consulting to leadership roles in Bioinformatics. They highlight this aspect through data collected on the growing number of NIH grants on which they were Principal Investigators (PI) from 2001-2015 which broadly encompasses many allied fields. They also highlight some additional subfields of bioinformatics, as defined by the *Bioinformatics* journal, that covers other aspects such as sequence analysis, phylogenetics, structural bioinformatics, data and text mining, databases and ontologies, and bioimage informatics – that we did not cover in our review. The also mined data from the journal to provide evidence that clearly demonstrates the strength of statisticians in certain key areas, especially “genetics and population analysis”, “gene expression” and “genome analysis.” These are provided in Table 2 of MST17.

Additional select contributions—MST17 also expand on other topics not included in our paper and also describe several key areas where they believe statisticians need to take a larger leadership role. They discuss highly cited papers in differential expression analyses including SAM, LIMMA, edgeR, DEseq for microarrays and RNA-seq data, to which we alluded but did not go into detail for brevity’s sake. They also cover a very interesting and pertinent area of study design and power calculations and trace the evolution of this area, especially in the context of microrrays. They summarize several influential papers in this area that are now routinely used in practice for design of these high-throughput experiments. This also highlights how careful experiment design can help scientists avoid potential pitfalls.

MST17 also cover machine learning areas in which statisticians have made a tremendous impact, in particular supervised and unsupervised learning. This area has taken off over the last couple of decades and have found wide applicability in bioinformatics. For example, several machine learning approaches such as shrunken nearest centroid, random forests, support vector machines, and multiple kernel learning are often top performers in classification problems with high-throughput data. In clustering, there are several methods that have found traction such as frequentist and Bayesian Gaussian mixture models, weighted correlation networks, gap statistics, resampling-based methods for both single and multi-platform data, including methods from the Tseng lab. As MST17 point out, there is a great deal of overlap with computer scientists and applied mathematicians in these efforts, but statisticians should continue to play an important role and strive for leadership.

MST17 also cover the important area of meta-analyses, especially horizontal meta- analyses, which involve the integration of information from several single-platform studies across multiple studies. This integration of information can increase power by combining samples across studies and enable researchers to find more subtle signals. They describe some success stories in this area and also other works that conduct meta-analyses for specific purposes such as differential expression, quality control, pathway analyses, clustering, classification, to name a few. Finally, MST17 cover some evaluative and comparative studies that are instrumental in guiding practitioners to select the best method for a given task, e.g. classification, clustering, missing value imputation, microarray processing and GWAS meta-analysis, differential expression and fusion transcript detection in RNA-seq.

Better contributions and leadership—MST17 present some poignant thoughts on how statisticians can continue to contribute and participate in the bioinformatics field to take the lead. We have already covered several of these areas in Section 1 of this rejoinder and our article including software packages and computational considerations, reproducibility of statistical methodological papers, primacy in project planning and study design, publishing in subject area journals and close collaboration with scientists.

An additional point raised by MST17 that we find to be very pertinent is the importance of making data publicly available. There are public databases like GEO for microarray data and Sequencing Read Archive (SRA) for sequencing data that serve as repositories for enhancing additional learning and replication of study results. They point out some issues that impede open data sharing, such as privacy issues that prevent sharing of raw patient-level sequencing data. Potential solutions to this problem include protected databases such as dbGaP, but these involve substantial administrative work, and establishment of data sharing standards such as MIAME (Minimum Information About a Microarray Experiment). Such protocols that describe useful and minimal information and also potentially protect patient privacy. This might be another area in which statisticians can contribute moving forward.

2.6 Jeanine J. Houwing-Duistermaat, Hae Won Uh, and Arief Gusnanto

We thank Houswing-Duistermaat, Uh, and Gusnanto (HDUG17) for their compliments and for their thoughtful discourse on glycomics and how the issues we discussed relate to this emerging field. Their discussion explores various key challenges in the field of glycomics, discusses the current state of the art relative to statistical practice, and raises open problems and issues for this field. We will respond to some of the issues they raised.

Simplistic Analytical Tools, Preprocessing, and Effect on Downstream Analyses—Like many fields with complex, high-dimensional data, HDUG17 characterize analytical tools for glycomics as hindered by a lack of knowledge of key technological and preprocessing issues, and elementwise statistical methods that ignore correlation and dependencies. This does not surprise us, and highlights opportunities for statisticians to improve the state of the art and make strong contributions to this field. They also mention that it is not so apparent how to perform filtering steps and express concerns about the effects of commonly-used preprocessing steps on downstream analysis. We also share this concern that for many technologies, preprocessing steps that seem necessary may have unforseen and unintended consequences on the data and obscure true signals that could be detected. We believe these concerns highlight the need for deep statistical involvement and coordination with subject area scientists and technology experts, to understand the technological details enough to assess which preprocessing steps are indeed necessary, and to carefully check whether final results are sensitive to the specifics of these steps.

Challenges of Graphical Models—They highlight the challenges of studying associations in glycomics given that the data are not Gaussian, but have a dependence structure induced by the normalization, and a vast majority of available methods for discovering associations through graphical models are designed for Gaussian data. The mixed graph methods introduced in the GIA17 discussion and mentioned in the rejoinder above have the potential to solve this problem. Multinomial data share the same constraint as these normalized data, and the generality of the exponential family representation undergirding these methods may provide sufficient flexibility to account for this structure.

Challenges of Integrative Analysis—In the context of glycomics, HDUG17 highlight some of the fundamental practical challenges of integrative analysis, including different scales, size, sparsity, and measurement error, are also present. They mention simple "stacking" of different types of -omics data sets often does not work well because of these issues. Some of these issues are discussed in the mixed graphical models papers highlighted by GIA17, and we also believe that utilizing known biological relationships has advantages over such stacking in that it reduces the model space and focuses on what are likely to be the most biologically plausible and relevant associations. They mention PLS strategies that can be useful to identify some types of joint structure, but is one of many tools that can shed light on these problems.

Differential Expression and Effect Estimation—HDUG17 also nicely highlight some of the fundamental contributions of statisticians to differential expression analysis and false discovery rate, pointing out some important work not discussed in as much detail in our paper. They also discuss some of their own work in which they utilize a 3-part mixture prior to identify probes that are either differentially expressed high, low or not differentially expressed. This type of modeling strategy is a useful way to approach the problem, and as insinuated by HDUG17 the ideas have the potential to be scaled up to the functional regression context to find these latent components while accounting for the internal structure of the assays through functional modeling techniques. This highlights one of the strengths of model-based approaches, that different modeling components can be integrated and assembled as modules in order to construct models for more complex settings that better account for the salient structure.

Protein Geometry—One area we did not discuss is protein folding and 3d structure detection, an important part of proteomics that has numerous unique quantitative challenges. HDUG17 discuss these problems and nicely summarize some of the statistical methods proposed in recent years to solve them. The centrality of geometry to the understanding of these data is somewhat unique among high throughput - omics data, and highlights the benefit of multi-disciplinary collaboration in solving these problems. While mathematicians can bring the geometric intuition to bear on the problem, statisticians can bring the stochastic perspective to bring uncertainty quantification into the analytical methods of this field of study.

2.7 Andrew W. Dowsey

Dowsey (AWD17) presents a detailed discussion of statistical issues in mass spectrometry proteomics and metabolomics, an important area that generates enormous structured data sets on the order of many gigabytes to even terabytes. He paints a picture of the current analytical practices, points out their limitations, and raises interesting ideas for how to improve the status quo, while connecting to the fundamental principles we raise in our article. He does so at a great level of detail and with strong insights coming from one of the international leaders in this area, and as a result his presentation poignantly touches on the points we are trying to make in this article. We summarize and comment on some of this points here.

Algorithmic, Ad Hoc, Reductionistic, Piecewise Strategies—After giving an overview of mass spectrometry and its place in biomedical science, Dowsey describes the typical MS analysis pipeline and makes the point that these existing methods miss lots of insights and that statisticians have a primary role to play in moving the science forward. He quotes our statement about how algorithmic, ad hoc, reductionistic, piecewise strategies dominate in many areas of high throughput -omics in painting the picture of the typical pipelines that extract symbolic representations and throw away the raw data at an early stage, and that follow multi-step procedure in which errors propagate, and information is lost on peaks not detected or spectral signals that overlap. These issues are especially prominent for these technologies since they were first developed for basic biology and chemistry settings, in which the studies are more controlled and the samples are more homogeneous and the data are cleaner. When applying to biomedical settings with heterogeneous samples of varying quality and complex underlying biology, these problems are exacerbated.

Unfortunately, the quantitative problems raised by these challenges have not been adequately acknowledged by the field, which Dowsey mentioned has instead chosen to focus on improved instrumentation rather than better processing algorithms. We see this general pattern at work in numerous biotechnology fields, looking to overcome limitations by developing more sensitive instruments, not improving the the data processing algorithms that are in many cases the key factors preventing effective extraction of the full information from the current instruments. Unless the quantitative algorithms are optimized, the advantages provided by the new instrumentation can largely go to waste.

AWD17 points out that the typical MS-based bioinformatic tools are multi-step, and fail to propagate uncertainty or borrow information across steps, making them inefficient and propagating errors throughout the process. He raises some issues that make stochastic modeling of peptides especially challenging, more difficult than DNA or RNA, including the lack of uniqueness of peptide signals to specific proteins or processes, and the complex missing data problems. These issues make the design of more rigorous statistical pipelines challenging, but worth the effort as there is a great deal of information lost by the current error-prone strategies.

Improved LC-MS Pipelines—Dowsey proceeds to describe some improved processing pipelines that attempt to put the statistical principles highlighted in this paper to work in order to more efficiently extract the rich information contained in these raw data, and quantify the proteomic measurements with less propagation of error. This pipeline utilizes well-chosen statistical distributions, sparsity, borrowing of strength using image registration and basis function representations, hierarchical Bayesian modeling, and functional modeling. This approach tries to perform statistically and biologically informed feature extraction while following up with functional data modeling to squeeze out any information missed by the initial feature extraction. He expressed hope that these pipelines can improve on the status quo, but also mentioned some important problems that remain to be solved, including improved protein identification, better use of parallel processing capabilities to deal with the computational intensiveness of these methods, development of tree-based models to capture the hierarchy of peptides, proteins, and pathways, and the development of better uncertainty quantification measures.

3 Other Statistical Contributions to Bioinformatics

As we disclaimed, our paper did not attempt to provide an exhaustive summary of areas in which statisticians have made a strong impact on bioinformatics. We sought to describe several key principles that we see as veins underlying many of the strong contributions statisticians have made to the field, and then highlight these areas through some examples of great impact and some other work that illustrates the principles even if not as impactful. By elucidating these principles, we hope we are able to help stimulate researchers to apply these principles and make strong contributions in other new technologies as they emerge. The various discussants have highlighted other areas of bioinformatics that have been strongly impacted by, and in some cases have been led by, statisticians, or are in need of greater statistical input.

Both Newton and Kechris and Ghosh remark that statistical approaches in bioinformatics started in the early days of sequence databases in the early 1970's, and also mention likelihood-based methods for phylogenies. These phylogenetic approaches are useful for modeling single-cell sequencing data, an emerging area mentioned by both Newton and Ma, Song and Tseng as one needing significant statistical input. With the ever-increasing efficiency of sequencing platforms, sequence-based methods now dominate the DNA, RNA and epigenetics field, and while not emphasized in our paper, there are many contributions and problems in this area. Kechris and Ghosh mention extreme value distributions for scoring alignments, and along with Ma, Song and Tseng highlight technologies studying gene regulation including Hi-C, HITS-CLIP, ChIP-chip and ChIP-seq, and eQTL analysis. Newton described some details about contributions to transcription binding site detection which serves as a nice example of a setting in which model-based methods were able to solve a problem that was not adequately solved by algorithmic approaches.

Kechris and Ghosh reference metabolomics and proteomics technologies based on mass spectrometry, and aptamer-based protemic technologies such as SOPAscan, and Dowsey provides an illuminating detailed discussion of statistical contributions and problems in LC-MS metabolomics and proteomics analysis. Houwing-Duistermaat, Uh and Gusnanto dissect various fundamental issues and statistical problems in glycomics, and mention protein folding, and important subfield of proteomics with some intricate and challenging quantitative problems.

Ma, Song and Tseng discuss a number of fundamental contributions to early work in genomics not mentioned by us. They recognize methods for differential expression analysis including SAM, LIMMA, edgeR, DEseq that are all very highly cited, as well as methods for study design and power calculations that have also been extremely impactful.

Statistical machine learning has also made fundamental contributions to bioinformatics, as statistical leaders have taken some of the principles used by machine learning computer scientists, framed them in a statistical framework, and further developed them. They provide numerous examples of methods for classification, clustering, and integrative multi-platform modeling problems at the interface of machine learning and statistics. They also mention horizontal meta-analysis methods that can increase power by combining information across studies, and also describe a number of comparative/evaluative studies in which statisticians have compared various methods, including classification and clustering tools for microarrays, missing value imputation methods for microarrays and GWAS studies, and differential expression and fusion detection methods for RNAseq data. These comparative studies tend to be highly cited and provide strong guidance to practitioners.

It is encouraging to see so many areas in which statisticians have made an impact, and also to see a clear delineation of some of the key areas currently in need of greater statistical input.

4 Conclusions

Once again, we thank the various discussants for their insightful comments that adds significantly to our discussion of statistical contributions to bioinformatics, and we thank the editors and publishers for their devotion of this issue to this topic. We believe that the quantitative problems raised by the complexity and high dimensionality of multi-platform genomics data are the most important problems in biomedical science, and lack of efficient and effective tools to solve these problems has the potential to be the bottleneck preventing the scientific community from gaining a deeper understanding of basic human biology that is needed to provide the next generation of medical strategies that can more effectively treat our most vexing medical ailments. Quantitative scientists of various ilks will certainly be involved in the front lines of science solving these problems. It is our hope that statisticians, equipped with deep understanding of variability and uncertainty, and armed with a vast array of modeling and inferential techniques, will be leading this charge and through their deep impact, be recognized as the leaders that they should be.

Acknowledgments

This work has been supported by grants from the National Cancer Institute (R01- CA178744, P30-CA016672, R01-CA160736, R01CA194391) and the National Science Foundation (1550088, 1463233).

References

Bhadra A, Rao A, Baladandayuthapani V. Inferring network structure in non-normal and mixed discrete-continuous genomic data. Biometrics. 2017

Zhu Y, Qiu P, Ji Y. TCGA-assembler: open-source software for retrieving and processing TCGA data. Nat Methods. 2014; 11(6):599–600. [PubMed: 24874569]