Statistics in Biosciences
https://doi.org/10.1007/s12561-024-09455-4

COMMENTARY

Leadership at the Intersection of Statistics & Genomics: A COPSS-NISS Leadership Webinar with Drs. Rafael Irizarry and Mingyao Li

Jingyi Jessica Li[1,2,3,4]

Received: 29 May 2024 / Revised: 26 August 2024 / Accepted: 27 August 2024

Abstract
In a COPSS-NISS webinar focused on leadership at the intersection of statistics and genomics, esteemed panelists Drs. Rafael Irizarry and Mingyao Li shared their leadership journeys and provided insights into this interdisciplinary field to inspire future leaders. They discussed the value of statistics in distinguishing signal from noise in the artificial intelligence (AI) era, the strengths of statisticians in ensuring rigor and robustness in genomics research, and the trade-offs between model expressiveness and interpretability. Additionally, they offered advice on how junior faculty can seek collaborations and increase their visibility, balance staying current with technological advancements, while developing methods carefully and thoroughly, and best practices for collaborating with domain experts. The recording of the webinar is available at https://www.youtube.com/watch?v=t6SsAoh95ig.

Keywords Leadership · Statistics · Genomics · Interdisciplinary research · Webinar

1 Introduction

On March 29, 2024, the Committee of Presidents of Statistical Societies (COPSS) and National Institute of Statistical Sciences (NISS) co-organized a webinar that featured two statisticians, Dr. Rafael Irizarry and Dr. Mingyao Li, who have

✉ Jingyi Jessica Li
jli@stat.ucla.edu

[1] Department of Statistics and Data Science, University of California, Los Angeles 90095-1554, USA

[2] Department of Biostatistics, University of California, Los Angeles 90095-1772, USA

[3] Department of Computational Medicine, University of California, Los Angeles 90095-1766, USA

[4] Department of Human Genetics, University of California, Los Angeles 90095-7088, USA

Published online: 12 September 2024

demonstrated exemplary leadership at the intersection of statistics and genomics, two fields that are increasingly intersecting due to rapid technological advancements in data generation. The focus was on discussing how to successfully bridge the interests of statistics and genomics, and how statisticians can make impactful contributions in both worlds.

2 Panelists' Leadership Journeys

Dr. Rafael Irizarry, Professor and Chair of Data Science at the Dana-Farber Cancer Institute and Harvard School of Public Health, is a renowned statistician known for his contributions to genomics and the development of open-source statistical methods through R Bioconductor. He has mentored numerous trainees, offered popular online courses in data analysis, and contributed to various professional committees. His accolades include the COPSS Presidents' Award and the Benjamin Franklin Award, with his work impacting fields from infectious diseases to COVID-19 vaccine effectiveness. With a Bachelor's in Mathematics from the University of Puerto Rico and a Ph.D. in Statistics from University of California, Berkeley, Dr. Irizarry began his academic journey in the Department of Biostatistics at Johns Hopkins University, where he first encountered the intersection of statistics and genomics. In 2013, he joined Harvard University and the Dana-Farber Cancer Institute, focusing on applying statistics and data science to cancer research. His leadership and expertise led to his appointment as Chair of the Department of Data Science in 2018. As chair, he has fostered an environment of excellence and innovation. He emphasizes attracting top talent and ensures the department's capacity to handle vast amounts of data generated by the cancer center, providing rigorous analysis that leads to reliable discoveries.

Dr. Mingyao Li, Professor of Biostatistics at the University of Pennsylvania, is acclaimed for her versatile, impactful research spanning multiple intersections of statistics, genetics, transcriptomics, and pathology. She holds a Bachelor's and Master's degree in Mathematics from Nankai University, China, and a second Master's and Ph.D. in Biostatistics from the University of Michigan. Joining University of Pennsylvania in 2006, she became a full professor in 2017. Her research focuses on cellular heterogeneity in human tissues through gene expression diversity and single-cell transcriptomics. Recognized as an International Statistical Institute Member, Fellow of the American Statistical Association, and Fellow of the American Association for the Advancement of Science, Dr. Li's leadership journey is marked by interdisciplinary collaboration. After her early work in genome-wide association studies (GWAS), she recognized challenges like crowded research landscapes and dominance of resource-heavy consortia. Hence, in 2010, guided by colleagues like Dr. Rafael Irizarry, she shifted to RNA-seq and epigenetics. Her research later expanded to single-cell RNA-seq and spatial transcriptomics, reflecting her adaptability to technological advancements. Driven by her dedication to impactful research, she delved into new fields once an existing field became crowded, maintaining her innovative edge. Collaborating with diverse experts, she emphasized continual learning, identifying scientific problems first, and engaging deeply in

disease studies. Her recent affiliation with the Department of Pathology and Laboratory Medicine highlights her commitment to bridging basic science and translational research.

3 Key Messages from the Q&A Session

The webinar included a Q&A session, where the panelists answered questions from the moderator Dr. Jingyi Jessica Li and the audience. Below are some highlights.

3.1 Artificial Intelligence (AI) Challenges and Opportunities for Statistics

Q: What are your opinions regarding whether the current AI challenges statistical models in the genomics field? Compared with computer science training, what are the unique opportunities for people with statistics training in the genomics field?

Rafael Irizarry: ***"AI is not a competitor but a tool for statisticians. The 'epsilon' persists."***

Dr. Rafael Irizarry stated that AI is not conceptually new. A long-standing challenge is computing expectations. For example, in prediction problems, we use covariates to predict the expected outcome. Over the past several decades, with the development of computers, prediction models have become more flexible, evolving from traditional regression models fitted by manual calculation to machine learning algorithms like random forests, which provide greater flexibility. With advancements like Graphics Processing Units (GPUs) and smart algorithms for optimization and parallelization, we can now create more flexible models in a finite amount of time. AI is just another tool in our toolbox to address the problems we care about, but it is not a competitor. Our statistical abilities remain essential. While we can estimate expectations using increasingly complex models, the inherent noise (the "epsilon") persists. Balancing signals and noise is a fundamental aspect of statistics and will not disappear. Therefore, it is not a choice between AI and statistics; rather, AI is a valuable tool that we statisticians need to learn and integrate into our research.

Mingyao Li: ***"AI is not a threat but complements statistical models. Statisticians bring rigor and robustness."***

Dr. Mingyao Li added that AI is not a threat to statistics. The key is to identify a problem and determine which tool is more suitable for addressing it. We should not be constrained by our statistical training to only use statistical models. Instead, we should be open to leveraging AI when it is suitable for the task at hand. Statisticians bring a unique strength to the table: a deep understanding of the rigor and robustness required in methods. This rigorous thinking is essential in ensuring the validity and reliability of the results, whether using traditional statistical models or AI-driven methods. By combining the strengths of statistics and AI, we can enhance our ability to solve complex problems.

3.2 Model Choice between Expressiveness and Interpretability

Rafael Irizarry: ***"Understanding data is crucial."***
Mingyao Li: ***"Intuition and domain knowledge guide model choice."***

About choosing between complex, less interpretable models and simpler, more interpretable ones, Dr. Rafael Irizarry and Dr. Mingyao Li shared their perspectives. Dr. Irizarry emphasized the importance of understanding the data and not limiting oneself to only interpretable models, especially when prediction accuracy is crucial. However, he noted that prediction might not often matter in basic science. Meanwhile, Dr. Li mentioned she never struggled with this choice, relying on her intuition and domain knowledge to guide her modeling decisions. For example, in cell-type deconvolution tasks, she used her biological knowledge to inform model choice and interpret results. To summarize both panelists' perspectives, balancing model interpretability and complexity, along with practical considerations, intuition, and expertise, is essential for effective data analysis in genomics.

3.3 Roles and Strengths of Statisticians in Genomics

Q: What are the roles or strengths of statisticians in the face of numerous computational methods in the genomics field, where data ownership and biological expertise are not exclusive to statisticians?

Mingyao Li: ***"Statisticians can leverage public data and think ahead of collaborators."***

Dr. Mingyao Li emphasized a key advantage in the genomics field: the wide availability of publicly accessible data. She highlighted the importance of statisticians thinking ahead of their collaborators to lead in addressing new research questions computationally. By leveraging their expertise in data analysis and computational methods, statisticians can guide and direct research efforts, ensuring that impactful questions can be feasibly addressed by data analysis.

Rafael Irizarry: ***"Statisticians should focus on methodologies in a mutually respected collaboration."***

Dr. Rafael Irizarry shared his perspective on the importance of having good collaborators. He acknowledged that predicting the next big research questions is challenging, but being in a leading institution like Dana-Farber Cancer Institute offers significant advantages. He benefited from having researchers come to him with questions and providing early access to data. Dr. Irizarry emphasized the value of letting biologists focus on the biological aspects while he concentrates on the methodological side. He pointed out that methods arising from collaborations should aim to be widely applicable, whether they are specific algorithms for new technologies or general methods that can be applied across various genomics technologies. This collaborative approach allows statisticians to make significant contributions to the field by focusing on their strengths in methodology and data analysis.

3.4 Career Advice for Department and Journal Choices

Q: Students and postdocs might have a choice between a statistics/biostatistics department and a medical school department for their first faculty job. Any advice on the pros and cons of the two choices? Advice on choosing journals to publish in?

Rafael Irizarry: ***"Candidates should consider what they truly want in their careers."***

Dr. Rafael Irizarry advised faculty job candidates to consider what they truly want in their careers. If they are interested in joining a statistics/biostatistics department, they should be aware that generalists—those who develop general-purpose methods with theory—often have better chances of being hired. There are exceptions, such as if a research area is particularly fashionable and hot, like single-cell or spatial data analysis, which might prompt a dean to push for a hire in a statistics/biostatistics department. His general advice is to develop methods that are general and abstract enough to be recognizable by statisticians, but not so abstract that they become unnatural. He acknowledged that this strategy might not always be best for science, but it has been effective in helping his trainees land faculty jobs. Regarding journal choices, Dr. Irizarry suggested writing down the statistical thoughts and aiming to publish in statistical journals. However, he noted that some science journals, like *Genome Biology*, also accept methodological work.

Mingyao Li: ***"Any choice has pros and cons, so following your passion is crucial."***

Dr. Mingyao Li shared her personal experience from a job interview in 2005. One institution told her that her area was too specialized, but the University of Pennsylvania biostatistics department was specifically looking for someone in statistical genetics, and she was hired because of that fit. She emphasized that there are always pros and cons to the direction one chooses, and ultimately, it is important to follow your passion. By doing so, you are more likely to find a position that aligns with your interests and strengths, whether it is in a statistics/biostatistics department or a medical school department.

3.5 Next Revolutionary Technologies

Mingyao Li:***"Spatial omics and imaging."***

Dr. Mingyao Li believed that one of the next revolutionary technologies is spatial omics and imaging. This field integrates spatial information with various omics data—genomics, transcriptomics, proteomics, and metabolomics—to provide a comprehensive understanding of spatial organization and function within biological systems. These technologies enable researchers to map molecules within tissues and cells, offering insights into the complexities of cellular environments and tissue

architecture. This advancement could lead to significant breakthroughs in understanding diseases, developmental biology, and tissue engineering.

Rafael Irizarry: ***"Spatial transcriptomics, cell-free DNA, molecular imaging, long-read sequencing, methylation, and immunology."***

Dr. Rafael Irizarry identified several emerging technologies poised to revolutionize biology and medicine. He highlighted spatial transcriptomics, which analyzes the spatial distribution of RNA molecules within tissues, with continuous improvements in spatial resolution and gene coverage. This technology has the potential to enhance our understanding of tumor microenvironments and complex tissues. Dr. Irizarry also emphasized early cancer detection through cell-free circulating DNA (cfDNA) analysis, which involves distinguishing between very low levels of cancer-derived DNA and background noise—a challenge calling for statisticians. He noted advancements in molecular imaging for non-invasive visualization of biological processes, enhancing diagnostics and disease monitoring. Long-read sequencing technologies also hold promise to reveal structural variations and other genome features that shorter reads might miss. Additionally, methylation analysis is becoming crucial for understanding epigenetic changes linked to gene regulation and disease development, such as cancer. Dr. Irizarry also highlighted immunology, particularly therapies that harness the immune system to target cancer cells. These technologies, brought to his attention by collaborators, represent the forefront of biomedical research.

3.6 Best Practices for Collaborating with Domain Experts

Mingyao Li: ***"Learning domain-specific knowledge is essential."***

Dr. Mingyao Li emphasized the importance of learning domain-specific knowledge for productive collaboration with experts. She shared her experience from 2006 at the University of Pennsylvania when she worked on a proposal about heart failure. Initially, it was challenging to grasp the subject's intricacies, but her efforts paid off. Now, she can effectively communicate with cardiologists and other biomedical scientists, enabling her to contribute significantly to interdisciplinary projects.

Rafael Irizarry: ***"Choose collaborators who respect statistical expertise, identify meaningful research questions, and trust your own expertise in data analysis."***

Dr. Rafael Irizarry outlined several best practices for collaborating with domain experts:

1. Choosing Collaborators: Select collaborators who respect and understand the importance of statistical thinking. While it is acceptable for domain experts not to know statistics, they should acknowledge this gap and be open to learning. Mutual respect is essential for effective collaboration, especially as the field of statistics has gained significant respect over the past 10–15 years.
2. Identifying Research Questions: Collaborators may sometimes lack clear research questions, complicating the collaboration. Generating data without a specific pur-

pose or merely trying to mimic existing papers rarely leads to fruitful outcomes. Educating collaborators on formulating meaningful research questions can be difficult, especially for early-stage biologists, and can be unfair to both parties due to the significant time and energy required.

3. Trusting Your Expertise: Trust your own expertise in data analysis. Dr. Irizarry advised against being swayed by computational talks that might not make sense. He shared his experience of realizing that some results in such talks were invalid. He advised that if something within your area of expertise seems questionable, trust your intuition.

3.7 Advice for Junior Faculty to Find Collaborators

Rafael Irizarry: ***"Seek departmental support and avoid chasing well-established PIs."***

Dr. Rafael Irizarry advised that if junior researchers are struggling to find collaborators, it might indicate that their department chair is not providing adequate support. He suggested that chasing after famous principal investigators (PIs), such as leaders of NIH consortia, is often unproductive. These well-established PIs typically already have computational experts in their labs and may be reluctant to collaborate, sometimes due to their controlling nature. Dr. Irizarry noted that while actively seeking out collaborators can sometimes be successful, it is not always the best approach.

Mingyao Li: ***"Increase visibility by giving talks."***

Dr. Mingyao Li recommended that junior researchers give as many talks as possible. Presenting their work at conferences, seminars, and workshops increases their visibility within the research community and can lead to potential collaborations. By sharing their expertise and findings, junior researchers can attract interest from peers and established scientists who may be interested in collaborative projects.

3.8 Balancing Technology Advancements with Method Development

Q: What are your thoughts on the challenge of following technology in method development, as Dr. Mingyao Li stated, considering we often don't have enough time to thoroughly evaluate the robustness of each method?

Mingyao Li: ***"It is a challenge."***

Dr. Mingyao Li acknowledged that balancing the need to stay current with technological advancements in method development and having sufficient time to thoroughly evaluate the robustness of each method is a significant challenge. She admitted this is a difficult question she struggled with herself and did not have a definitive answer.

Rafael Irizarry: ***"If not aiming for high-impact journals, prioritize rigor over speed."***

Dr. Rafael Irizarry suggested that it is acceptable to proceed slowly and prioritize doing things correctly, especially if one is not aiming to publish in high-impact journals like those from Nature. He cited the example of the limma method for gene expression microarray data analysis. This method, developed by Gordon Smyth in 2004 [1], was not published in a high-impact journal. However, it has received many citations and is probably the most widely used tool in statistical genomics. Despite being published nearly a decade after microarray technology was developed in 1995, it has achieved significant impact due to its robustness and reliability. Dr. Irizarry emphasized that learning to balance speed and rigor is crucial and comes with experience.

3.9 Paper and Book Recommendations

Mingyao Li: ***"The 'Point of Significance' column by Naomi Altman in Nature Methods."***

Dr. Mingyao Li recommended the "Points of Significance" column in *Nature Methods*, authored by Naomi Altman, a professor of statistics at Pennsylvania State University. Dr. Li highlighted this column as an excellent resource for researchers seeking to improve their understanding and application of statistical methods in biological research. Notable contributions include topics such as "Statistics versus Machine Learning" [2], "Machine Learning: Supervised Methods" [3], and "Bayesian Statistics" [4]. These articles provide practical suggestions and educational insights into best practices in statistical analysis and reporting, making them invaluable for enhancing statistical literacy among biologists.

Rafael Irizarry: ***"The book 'Practical Statistics for Medical Research' by Douglas G. Altman."***

Hearing the name "Altman," Dr. Rafael Irizarry recommended the book "Practical Statistics for Medical Research" [5] by Douglas G. Altman. However, he and Dr. Mingyao Li soon realized they were referring to different Altmans. Dr. Irizarry highlighted Douglas Altman's book as an essential resource for anyone involved in medical research. The book offers comprehensive guidance on applying statistical methods to medical research, covering a wide range of topics. It is designed to be accessible to both statisticians and non-statisticians working in the medical and health sciences, making it an invaluable reference for enhancing statistical understanding in this field.

4 Summary

The webinar, co-organized by COPSS and NISS, focused on leadership at the intersection of statistics and genomics. Featuring Dr. Rafael Irizarry from the Dana-Farber Cancer Institute and Harvard School of Public Health, and Dr. Mingyao Li from

the University of Pennsylvania, the event highlighted their leadership journeys and insights into this interdisciplinary field.

During the Q&A session, both panelists addressed the integration of AI in genomics, balancing model complexity and interpretability, and the roles of statisticians in the face of computational advancements. They also provided advice on finding collaborators and choosing career paths for junior researchers.

This event provided valuable insights and practical advice for aspiring leaders and researchers in the fields of statistics and genomics, underscoring the importance of interdisciplinary collaboration and continuous learning.

Acknowledgements The author would like to thank everyone who contributed to the success of this webinar: Dr. Lorin Crawford from Microsoft Research and Brown University for co-organizing the webinar; Dr. Amita Manatunga, Chair of COPSS and Professor at Emory University, for suggesting the webinar; Dr. Natalie Dean from Emory University for sharing experience with webinar organization; Megan Glenn from NISS and Dr. Mary Ryan from the University of Wisconsin-Madison for their efforts in advertising and promoting the webinar; Jim Rosenberger from NISS for the support.

Declarations

Disclosures The author declares no competing interests.

References

1. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3:Article 3. https://doi.org/10.2202/1544-6115.1027
2. Bzdok D, Altman N, Krzywinski M (2018) Statistics versus machine learning. Nat Methods 15(4):233–234. https://doi.org/10.1038/nmeth.4642
3. Bzdok D, Krzywinski M, Altman N (2018) Machine learning: supervised methods. Nat Methods 15(1):5–6. https://doi.org/10.1038/nmeth.4551
4. López Puga J, Krzywinski M, Altman N (2015) Points of significance: Bayesian statistics. Nat Methods 12(5):377–378. https://doi.org/10.1038/nmeth.3368
5. Altman DG (1990) Practical statistics for medical research. Chapman and Hall/CRC, Boca Raton