Teaching ‘big data’ analysis to young immunologists

Joachim L Schultze

Everyone and everything seems to go ‘big data’ these days. The task ahead will be to train young immunologists to formulate intelligent hypotheses using big data resources.

Imagine two immunologists, Bill and Steve, who meet in 2030. Steve asks, “Bill, how is your science going?” Bill answers, “We now have access to 500 petabytes of storage, a 100,000-node GPU computer cluster with 500 terabyte RAM, and the newest data-integration engine that allows us to instantaneously access all publically available data collections worldwide. We finally generate and prioritize our hypotheses based on big data.” There are numerous developments already ongoing that make this scenario very real. The big data revolution incorporates the ‘three Vs’: volume of data, velocity of processing the data, and variability of data sources; therefore, preparation is required to take advantage of the technological advances and tools available to efficiently interrogate large data sets to maximal effect. Indeed, every immunologist will have access to high-quality and publically available big data regarding primary immune cells from numerous species. Though they will not need to become computer scientists per se, young immunologists will need to know how to make use of the wealth of data that are currently being generated and that are anticipated in the coming decades, and thus training will be required not only in molecular and systems immunology but also in big data science. This will require changes in graduate and undergraduate education in order to achieve the goals of training students to embrace the big data era of science and teaching them how to use vast data resources in hypothesis generation. Here,

I will discuss some of the developments in big data science, its impact on immunology and how we need to adapt our education programs to cope with the expected changes.

Big data in the life sciences

When it comes to big data in the life sciences, ‘omics’ technologies are currently the major data drivers^{1–3}. The increase in next-generation sequencing data; sequencing capacity; and different sequencing approaches to assess DNA sequence, structure or methylation, DNA-protein interactions, or different RNA species and RNA-protein interactions is staggering^{4–6}. The quality of information is further improved by integration of different omics technologies. For example, the role of transcription factor binding to DNA in gene expression is much better described when integrating transcriptome data from the same sample⁷. Other technologies, including proteomics, lipidomics, microbiomics and metabolomics^{8,9} and high-resolution microscopy¹⁰ are adding further to the big data avalanche.

One needs to keep in mind that the generation of ever-larger data sets is not restricted to pharmaceutical companies or large-genome centers. Technological advances now allow any academic life scientist to compile terabytes of data. Efficient storage and retrieval of data is certainly one bottleneck, but even bigger bottlenecks include asking the right questions with all these data at hand, quickly exploring the data in an intelligent way, visualizing the data in an intuitive manner and drawing logical conclusions from the models derived from big data analysis. The availability of big data will change the way we ask scientific questions (Fig. 1). Furthermore, the generation of big data will have to be based on sound biological

observations, for example, the change of a cellular function in response to an environmental stimulus. Only if the design of a big data experiment reflects the biological observation can the big data be meaningful for generating computational models that prioritize hypotheses to explain the initial observations. In such a scenario, the use of big data is a risk-minimizing strategy to quickly and directly derive the most likely hypothesis explaining the underlying biology. Because the next generation of immunologists should be able to fulfill such tasks, we need to train them in both wet lab skills and computational skills.

Big data in immunology

There are several areas in immunological research that will greatly benefit from a sophisticated big data analysis, including genome, epigenome, transcriptome, metabolome, lipidome, proteome, cytochrome (CyTOF) and even microbiome data (Fig. 2). An example is the analysis of specific B or T cell antigen receptor repertoires^{11,12}. Recent technological advances in cytometry, single-cell manipulation technologies, mass spectrometry, and high-throughput sequencing of the B and T cell receptor (BCR and TCR) repertoires will make it possible to analyze B and T cell responses by following changes in clonal and population dynamics and function, thereby providing a fuller picture of the immune response to a given stimulus or therapeutic intervention. Big data derived from human BCR and TCR repertoires at the level of single receptors, especially when paired with technologies that can identify the antigenic epitopes recognized by those receptors, will bring clinical diagnosis, antibody drug discovery and vaccine development to a new level and will give insight into the ability of processed

Joachim L. Schultze is at the Department of Genomics and Immunoregulation, Life and Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany.
e-mail: j.schultze@uni-bonn.de

Figure 1 The circle of systems immunology. The big data-driven circle of systems immunology as a strategy to use big data to generate hypotheses in a data-driven fashion to use prioritized database hypotheses for further experimental validation by classical approaches using loss- and gain-of-function experiments, *in vivo* modeling of disease, murine genetic models and other functional and observational approaches such as live cell imaging or flow cytometry.

autoantigen peptides to bind these receptors. Another example is the combination of transcriptome data with extended bioinformatics to dissect activation of immune cells on a multidimensional scale¹³, along a time scale¹⁴ or both. Transcriptomes are an excellent starting point to determine transcriptional regulatory networks during immune cell activation. Such networks can be enriched for specific classes of genes (for example, transcription factors) that can be examined in subsequent experiments following a prioritization based on the hierarchy defined by unbiased computational modeling of transcriptome data. Another source of big data is the recently introduced single-cell RNA sequencing technologies that will revolutionize the way we will define immune cell subsets in the near future^{15,16}.

Future training programs for immunology

To design future training programs in immunological big data science, we need a clear vision and understanding about the role immunology should have in the future. There might be differences between centers or universities and from country to country, but big data will play a role in all settings. The advent of omics-based big data science will allow the assessment of human immune parameters in unprecedented detail and in an integrated fashion using data from different high-throughput technologies. Moreover, it is becoming increasingly clear that species-specific genetic, epigenetic and microbiome-mediated mechanisms are important modulators of immune-related mechanisms and diseases and are best studied in humans using omics-based technologies^{17–19}. As stated by others, this will create new interest in human immunology in the decades to come^{20,21}, and therefore the use

of omics technologies in human immunology needs to be reflected in the curricula of our training programs. Moreover, our ability to measure genomic differences between individuals and species and to map microbiomes from different organs (all requiring big data generation and analysis) will lead to mandatory reporting of such omics data in animal models addressing human diseases. Particularly for those immune-related diseases for which genetic, epigenetic or microbiome information can be obtained with a reasonable effort in humans, it can be foreseen that increasing requirements for reporting full genome and microbiome information in the respective animal models will shift research toward human immunology.

With such a setting in mind, the following scenario can be proposed. An integration of bioinformatics, genomics, big data science and systems biology into our undergraduate and graduate programs in immunology would be most favorable (Box 1). Currently, nobody would ask for a specialized master program in genetic mouse modeling. Yet an understanding of at least the basics of genetic engineering in the mouse is a prerequisite for cutting-edge immunological research, and it is an integral part of study programs in immunology. Big data science and systems approaches now need to become equally integral components of our immunological study programs. For most institutions, this will require close collaborations in teaching with computer science, informatics, bioinformatics and mathematics departments.

Another issue is timing. When should we start to train young immunologists? Given

that high school training will be difficult to achieve, the next starting point would be undergraduate programs at colleges and universities. Programs that specialize in molecular life sciences should at least integrate immunology and bioinformatics. Learning some basics in big data science and programming would allow the next generation of scientists to use the wealth of data ahead much better than we can do it today (Box 2). These young scientists, who will already be digital natives concerning the use of the internet, should become natives of the interface between immunology and big data science and computational biology. As with any type of dual citizenship, speaking both languages perfectly will be the key to success. Further specialization can follow after the bachelor's degree. Depending on the education system, this should be either integrated into master's studies (as in Europe) or directly into graduate programs during the PhD phase (as in the US). One could envision a master's program on molecular and systems immunology covering classical aspects of immunology (innate immunology, adaptive immunology, infection immunology and tumor immunology) but also integrating clinical immunology, genomics and other omics technologies, bioinformatics, big data science and systems immunology.

During the PhD phase, an additional level of training in big data science, systems immunology or computational biology should be offered to young immunologists. There are at least four levels of expertise to be reached. A minimum level of expertise would be to know about existing big data in the public domain,

Box 1 Suggestions for future undergraduate immunology programs

Minimum requirements

Initiate interdisciplinary, interfaculty or interdepartmental study programs—at least the following departments or disciplines should be involved:

- Immunology
- Computational sciences
- Molecular medicine or biology
- Genetics or genomics

Integrate lectures, courses and seminars in bioinformatics and genomics into the study program:

- These lectures, courses and seminars should be mandatory (not elective)
- Team up with the computational science; make sure the courses are geared towards computational approaches applicable to immunological research
- Practical courses should include experience in real data analysis
- Include lab rotations in computational science labs as part of the study program, together with a lab rotation report

Further suggestions

Offer additional courses in big data sciences and systems immunology:

- These lectures could be elective
- Team up with systems biology big data science departments; make sure the courses are geared towards approaches applicable to immunological research
- Offer access to learning a programming language (elective)

Figure 2 Some of the areas that generate big data in the life sciences. Innovative database integration engines will allow the analysis and integration of different sources of big data in the life sciences in the future.

science. This would be particularly suitable for students who were not trained in computational sciences before their PhD phase. A similar track could be designed for students in the computational sciences—if they want to enter the immunological arena, they need to learn the basic concepts of our field. Accordingly, they need to be trained in the basics of immunology and in immunological techniques and model systems, hands-on laboratory experience included. This would allow two paths of entry into a career in immunological big data science.

What about postdoctoral scientists? Science at the interface of different disciplines such as immunology, genomics, informatics, big data science, computational biology and bioinformatics will require lifelong continuous education. Even at the established faculty level, we are learning new approaches in these disciplines on a daily basis. For those of us who did not grow up with big data and who were not exposed to computational science during their studies, learning by doing will be rather important. Online education platforms such as Class Central (<https://www.class-central.com>) already offer numerous online courses on big data science, bioinformatics and programming. Taking time for learning disciplines not being taught at university will become mainstream. Nevertheless, the better the educational foundation established early on, the better one's scientific future in this most exciting area of research will be (**Box 2**).

the technologies they were generated with and the principles of big data analysis. A next step would be an introduction to previously published examples of good practice of big data science analysis in immunology. A critical further step will be hands-on training on applications of big data analysis using previously published good practice examples. Only if those steps are mastered will it be possible to reach the highest level of big data analysis allowing one to fully harness the potential of big data, to prioritize the most important questions, to integrate intuition with biological relevance and to guide the best experimental design for a big data experiment. This highest level will require teaching of a deep understanding of immunology while

allowing for sufficient time to practice big data analysis. If possible, PhD students could be embedded for a certain time into collaborating labs that entirely focus on big data science, for example, genomics labs. During this time, students would directly interact with computational scientists on a daily basis to learn and practice the necessary computational skills. The better the students are trained beforehand during their bachelor's and master's programs, the easier they will find such embedment. At the same time, this model fosters a new way to interact, by sharing knowledge, expertise and data in a very collaborative fashion between different groups. Another option for PhD students could be a structured program in big data

Box 2 Major considerations for scientists interested in big data immunology

- Participate in undergraduate or graduate programs that provide classes, seminars and courses in both classical fields of immunology and in bioinformatics, computational biology, genomics and systems immunology
- Learn the basics of the big data-driven circle of systems immunology. Understand how big data can be used to compute models that help to prioritize hypotheses in a data-driven fashion (see **Fig. 1**)
- Learn another language, namely a programming language—R, Perl and Python are good choices
- For your PhD thesis, look for projects that are at the interface of immunology and computational biology; the best scenario would be to join institutes or labs that offer both wet and dry lab experiences. If you can generate your own high-throughput data, you can drive your own hypothesis that you derived from your own calculations with your own data. If you then are able to validate them with your own wet lab experiments, you made a full circle in systems immunology
- Attend summer schools on computational biology: examples are the Lipari School on Bioinformatics and Computational Biology, the Summer Program in Biostatistics & Computational Biology at Harvard, the Summer School for Big Data in Biology at the University of Texas at Austin and the Dresden Summer School in Systems Biology. There are more programs available, and you can easily find them on the internet
- During your postdoctoral time, spend extra hours on computational skills
- As a young research group leader, engage with experts in both fields, immunology and computational sciences
- As an established group leader, continue to educate yourself in these novel technologies and in data-driven approaches that can quickly prioritize your own hypotheses. Initiate exchange programs with computational scientists who are willing to host your co-workers for a defined period of time to learn more about big data sciences

Perspective: a new world of immunology

Once we have trained all these young immunologists in big data analysis, what else will change in our future immunological research? We have to adapt quickly to collaborative research models that other research fields, such as particle physics and the human genomics and epigenetics consortia, have managed to establish. The more data we generate, the more mainstream data sharing before publication will become. Crowdsourcing might also become a natural habit. Sarah Fortune at the Harvard School of Public Health reported at POPTECH on cataloging bacterial cells (http://poptech.org/popcasts/fortune_and_biewald_crowdsourcing_tb_cell_annotation/), a task that is still best performed by humans. Using the internet to convince thousands of people to join in fulfilling this task accelerated the project by several years. Another example is 'Play to Cure: Genes in Space' (<http://scienceblog.cancerresearchuk.org/2014/02/04/download-our-revolutionary-mobile-game-to-help-speed-up-cancer-research/>). In a nutshell, the goal of this computer game for the player is to find the best route to pick up the most 'Element Alpha'; however, by doing so, players are actually plotting a course through genuine DNA microarray data, thereby helping

cancer scientists spot patterns in gigabytes of genetic information from thousands of tumors. Such collaborative models have the great potential to democratize basic scientific discovery. But for this to happen, we have to get out of our labs and talk to computer, web and social media specialists, to engineers and even to the public. Maybe if we are not ready ourselves, we might team up with the students trained in big data science to harness these great options outside our own scientific comfort zone. Since we will never be able to collect endless data in the future, we also need to be very responsible about our financial resources. The more data we collect, the more differences we will find between human disease and the model systems we use to study them. This will be particularly true for immune-mediated pathology. Big data science will definitely have an impact on our future priorities. Positively speaking, refocusing our efforts on human immunology will be a result of those technologies generating big data in the life sciences.

ACKNOWLEDGMENTS

I would like to thank to S. Barry and W. Kolanus in discussing big data science in immunology. I also would like to thank my co-workers M. Beyer and T. Ulas and my master's students P. Günther and K. Baßler for carefully reading the manuscript. J.L.S. is a member of the Excellence Cluster ImmunoSensation

at the University of Bonn and the Marie Curie Initial Training Network on Tumor Infiltrating Myeloid Cell Compartment (People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013 under Research Executive Agency grant agreement no. 317445). His work is mainly funded by the comprehensive research centers SFB704 and SFB645.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

1. Gerstein, M. *Nature* **489**, 208 (2012).
2. Marx, V. *Nature* **498**, 255–260 (2013).
3. Marx, V. *Nat. Methods* **10**, 293–297 (2013).
4. van Dijk, E.L., Auger, H., Jaszczyszyn, Y. & Thermes, C. *Trends Genet.* **30**, 418–426 (2014).
5. Anonymous. *Nat. Methods* **11**, 1 (2013).
6. Anonymous. *Nat. Methods* **5**, 1 (2008).
7. De Nardo, D. *et al. Nat. Immunol.* **15**, 152–160 (2014).
8. Orchard, S. *et al. Proteomics* **13**, 2931–2937 (2013).
9. Anonymous. *Nat. Chem. Biol.* **10**, 605 (2014).
10. Lichtman, J.W., Pfister, H. & Shavit, N. *Nat. Neurosci.* **17**, 1448–1454 (2014).
11. Newell, E.W. & Davis, M.M. *Nat. Biotechnol.* **32**, 149–157 (2014).
12. Georgiou, G. *et al. Nat. Biotechnol.* **32**, 158–168 (2014).
13. Xue, J. *et al. Immunity* **40**, 274–288 (2014).
14. Yosef, N. *et al. Nature* **496**, 461–468 (2013).
15. Jaitin, D.A., Keren-Shaul, H., Elefant, N. & Amit, I. *Semin. Immunol.* **27**, 67–71 (2015).
16. Jaitin, D.A. *et al. Science* **343**, 776–779 (2014).
17. Fairfax, B.P. *et al. Science* **343**, 1246949 (2014).
18. Belkaid, Y. & Hand, T.W. *Cell* **157**, 121–141 (2014).
19. Liang, L. *et al. Nature* **520**, 670–674 (2015).
20. Davis, M.M. *Immunity* **29**, 835–838 (2008).
21. Su, L.F. *et al. Cold Spring Harb. Symp. Quant. Biol.* **78**, 203–213 (2013).