DEMYSTIFYING BIOMEDICAL BIG DATA

A Course from GeorgetownX (BIOX-201)

COURSE SYLLABUS

TABLE OF CONTENTS

Demystifying Biomedical Big Data
Whether you are a student, basic scientist, researcher, clinician, or librarian, this course is designed to help you understand, analyze, and interpret biomedical big data.

COURSE DIRECTORS

Bassem Haddad, MD
Bassem R. Haddad, MD, is a Professor of Oncology and Obstetrics and Gynecology, and a molecular cytogeneticist at Georgetown University Medical Center and the Co-Director of the Medical School Molecular and Human Genetics course. Dr. Haddad has an active molecular cytogenetics research laboratory working in the area of translational research and cancer biomarker discovery. His research focuses on understanding the genetic aberrations and instability that occur in genetic diseases, particularly cancer. Dr. Haddad received his MD and Residency training in Obstetrics and Gynecology from the American University of Beirut, and completed his fellowship training in Molecular Cytogenetics at Boston University School of Medicine, Baylor College of Medicine, and the NIH. He is a longtime member of the Georgetown University Medical School faculty and has been involved in genetic teaching for many years. Dr. Haddad is very interested in innovative approaches in medical education and has previously developed and directed another MOOC course released on edX in June 2014, Genomic Medicine Gets Personal.

Yuriy Gusev, PhD
Dr. Yuriy Gusev is an Associate Professor of Bioinformatics and a Bioinformatics Lead at the Georgetown Innovation Center for Biomedical Informatics (ICBI). He has over 20 years of experience in academic and industry research in the fields of bioinformatics, systems biology, biomarker discovery, and computational modeling of biological systems. Dr, Gusev is leading efforts of a bioinformatics analytical team in the development of pipelines for multi-omics and Next Generation Sequencing data analysis and integration. His team has been involved in many enterprise scale projects including the Georgetown Database of Cancer (GDOC), the In Silico Research Center of Excellence, and Georgetown Center for Systems Biology of Cancer as well as external collaborations with NCI, FDA and Argonne National Labs. Dr. Gusev has many years of experience in teaching and training in bioinformatics and computational biology at several academic centers in US including Johns Hopkins School of Medicine and Whiting School of Engineering. He has also developed several new curricula for graduate and undergraduate programs at Georgetown. Dr. Gusev has received his training in applied mathematics and computational biology at St. Petersburg University, Russia and completed his postdoctoral training in cancer genetics at the Waksman Institute of Rutgers University.

Peter McGarvey, PhD
Dr. McGarvey is Deputy Director of the Innovation Center for Biomedical Informatics and Associate Professor in the Department of Biochemistry and Molecular and Cellular Biology. Dr. McGarvey has academic and commercial experience in molecular biology, biotechnology, bioinformatics and software development, and his research interests include genomic and proteomic analysis, biological databases, and data visualization. Currently, Dr. McGarvey helps manage the CPTAC Data Center, The Protein Information Resource, the UniProt Knowledgebase and is involved in various other funded projects such as the Clinical Genome Resource and the

GUMC CTSA. Dr. McGarvey has a PhD in Biological Sciences from the University of Michigan and MS in Technology Management from University of Maryland University College.

GUEST SPEAKERS

Krithika Bhuvaneshwar, MS
Ms. Krithika Bhuvaneshwar has over 8 years professional work experience and combines her interdisciplinary skills in bioinformatics and biostatistics for numerous projects within ICBI. Krithika has applied her bioinformatics expertise to develop novel algorithms and analysis pipelines for clinical, genomic and expression profiling data management, quality control and integrative downstream analysis. Her research responsibilities also include clinical data management; utilizing cloud computing for high throughput bioinformatic analysis of genome sequencing data and other omics data, statistical and bayesian analysis; systems biology and pathway analysis. Krithika has worked with collaborators to develop supporting evidence to confirm of their lab findings in publicly available studies. In this regard, she has performed computational analysis of gene expression and RNA-seq data from public datasets in G-DOC, The Cancer Genome Atlas (TCGA), NCBI dbGAP and other datasets from public repository Gene Expression Omnibus (GEO). She is also the project manager for the Georgetown Database of Cancer (G-DOC) project.

Ben Busby, PhD
As the Genomic Outreach Coordinator at NCBI, Dr. Busby is responsible for genomics education, product management and normalization across large datasets. In these capacities, he frequently gives talks at major academic institutions, arranges online educational experiences and in-person hackathons, collects use cases and develop personas for users of NCBI genomics resources, manages user centered software design, and develop tools for normalization of RNAseq and other genomic data types.

As the Chair of the Bioinformatics Department at FAES, Dr. Busby has worked with over 300 students, both in group settings and individually. He guides individual students to establish projects working on computational aspects of their individual research interests, spanning a diverse spectrum of topics, with an emphasis on high-throughput data analysis.

Dr. Busby's research at NCBI was primarily focused on algorithm development for answering large-scale biological questions. In addition to his own biological research, he collaborated with several other researchers, and develop new projects based on novel algorithmic methods.

Ashlea Carver, JD
Ashlea Carver, JD, is the Deputy Conflicts Officer at Georgetown University Medical Center and Regulatory Associate at Georgetown University Medical Center and has experience with tobacco control and the regulation of food and drugs. She holds a juris doctor degree from the University of South Carolina School of Law.

Robert Clarke, PhD, DSc
Robert Clarke, Ph.D., D.Sc., is the Dean for Research at Georgetown University Medical Center and co-Director of the Breast Cancer Program at the Georgetown-Lombardi Comprehensive Cancer Center. Dr. Clarke's research interests include the cellular and molecular mechanisms driving endocrine resistance in breast cancer as well as drug resistance, drug/hormone

interactions, and the application of molecular profiling to predict breast cancer phenotypes and identify functionally relevant gene signaling networks. Dr. Clarke leads several multinational molecular medicine studies in breast cancer, in collaboration with colleagues at Lombardi Comprehensive Cancer Center, Virginia Tech and the University of Edinburgh. Dr. Clarke received a PhD and DSc in Biochemistry from The Queen's University of Belfast, UK.

Jerry S.H. Lee, PhD
Dr. Lee serves as a Health Sciences Director within the National Cancer Institute's (NCI) Office of the Director where his chief responsibility is to help direct the NCI's Center for Strategic Scientific Initiatives (CSSI). In this role, he provides leadership and input in planning, developing, and implementing rapid strategic scientific and technology initiatives that keep the Institute ahead of the scientific curve of potential new exciting areas and discoveries. This may involve direct development and application of advanced technologies, creation of new trans-disciplinary teams, and/or use of available federal funding mechanisms to forge novel partnerships that emphasize innovation and convergence of scientific disciplines.

Since its inception in 2003, the Center has supported more than 2,000 trans-disciplinary projects through programs such as the Innovation Molecular Analysis Technologies (IMAT), NCI Alliance for Nanotechnology in Cancer, The Cancer Genome Atlas (TCGA), Clinical Proteomic Tumor Analysis Consortium (CPTAC), Physical Sciences-Oncology Centers (PS-OC), Provocative Questions (PQ), and Cancer Target Discovery and Development (CTD2) network. These exploratory initiatives that focus on the integration of advanced technologies, trans-disciplinary approaches, infrastructures, and standards, to accelerate the creation of publicly available, broadly accessible, multi-dimensional data, knowledge, and tools to empower the entire cancer research continuum for patient benefit.

Prior to joining the NCI, his research involved elucidating mechanisms of age-related diseases and has co-authored over twenty papers, five book chapters, and one book. He continues research as an adjunct associate professor at Johns Hopkins University, where he also earned his bachelor`s degree in biomedical engineering and Ph.D. in chemical and biomolecular engineering. He also holds an appointment at the Washington D.C. Veterans Affairs Medical Center and collaborates with clinicians on next generation patient-centered outcomes research.

Dr. Lee is a member of the Innovation Policy Forum of the National Academies Board on Science, Technology, and Economic Policy, the Foundation for the NIH's Biomarkers' Consortium Cancer Steering Committee, the Health and Environmental Sciences Institute's Board of Trustees, and the editorial board of Convergence Science Physical Oncology journal. In 2016, Dr. Lee was assigned to Office of the Vice President to serve as the Deputy Director for Cancer Research and Technology for the White House Cancer Moonshot Task Force. A few key efforts he helped coordinate include the Applied Proteogenomics OrganizationaL Learning and Outcomes Network (APOLLO), international collaborations to share molecular characterization datasets, the Blood Profiling Atlas in Cancer pilot, as well as co-chairing an interagency group focused on cancer data and technology policy issues.

Subha Madhavan, PhD
Dr. Madhavan is the Founding Director of the Innovation Center for Biomedical Informatics (ICBI) at the Georgetown University Medical Center and Associate Professor of Oncology. She is a world-class leader in data science, clinical informatics and health IT who is responsible for several

biomedical informatics efforts including the Georgetown Database of Cancer (G-DOC) a resource for both researchers and clinicians to realize the goals of personalized medicine, the NCI In Silico Research Center for Excellence, and co-directs Lombardi Cancer Center's Biostatistics and Bioinformatics shared resource. She leads the Biomedical Informatics program of the Georgetown-Howard Universities CTSA. Her group's research focuses on extracting knowledge from big data to enrich evidence for use and effectiveness of precision therapies. She has contributed to novel information sciences findings in research articles published in journals such as Nature, Bioinformatics, Molecular and Cell Biology (MCB), AJPM, Frontiers in Oncology, Bioinformatics, Cancer Informatics, and Molecular Cancer Research (MCR).

Dr. Madhavan has a Master of Science in Information Technology from University of Maryland and a Ph.D. in Molecular Biology and Biological Sciences from the Uniformed Services University for the Health Sciences through a highly ranked Indo-US Collaborative program. She conducted post-doctoral research at Johns Hopkins University.

John L. Marshall, MD
Dr. John Marshall is a Professor of Medicine and Oncology, the Chief of Hematology and Oncology, and the Director of the Otto J Ruesch Center for the Cure of Gastrointestinal Cancers, at the Lombardi Comprehensive Cancer Center, Georgetown University Medical Center. Dr. Marshall received his training at Duke University, the University of Louisville, and Georgetown University. Dr. Marshall is an internationally recognized expert in new drug development for GI cancer, with expertise in phase I, II, and III trial design, and has served as Principal Investigator for more than one hundred clinical trials. While he has an interest in many areas of cancer research, his primary focus has been on the development of vaccines to treat cancer. Dr. Marshall has become an outspoken advocate for GI cancer patients and the importance of clinical research participation. Most recently, he has established the Otto J Ruesch Center for the Cure of GI Cancers, an organization solely focused on improving the lives of GI cancer patients through innovative research, personalized medicine, and focused advocacy.

Sheila Cohen Zimmet, BSN, JD
Sheila Cohen Zimmet, BSN, JD, is the Senior Associate Vice President for Regulatory Affairs at Georgetown University Medical Center, an Adjunct Associate Professor in the Department of Pharmacology and Physiology, a Faculty Associate in the Georgetown University Center for Clinical Bioethics, and a member of the National Advisory Child Health and Human Development Council of the National Institutes of Health. She holds a BS in nursing degree from Georgetown University School of Nursing and a juris doctor degree from Georgetown University Law Center.

FACULTY SUPPORT TEAM

Course TAs
Stephanie Huang
Jonathan Lagdameo

Center for New Designs in Learning & Scholarship Development Team
Barrinton Baynes (Videography)
Xiaoke Ding (Animation)
Linda Huber (Project Coordination & Instructional Design)

Joe King (Videography)
Anna Kruse (Project Coordination & Instructional Design)
Yiran Sun (Graphic Design)
Yianna Vovides (Instructional Design)

You can contact the faculty support team directly at gux@georgetown.edu.

WHAT IS THE COURSE ABOUT?

With the continuous generation of massive amounts of biomedical data on a daily basis, whether from research laboratories or clinical labs, we need to improve our ability to understand and analyze the data in order to take full advantage of its power in scientific discoveries and patient care. For non-bioinformaticians, "handling" big data remains a daunting task. The purpose of this course is to facilitate the understanding, analysis, and interpretation of biomedical big data to those in the biomedical field with limited or no significant experience in bioinformatics. Our goal is to "demystify" the process through a series of lectures and online hands-on training sessions and demos on how to use publically available on-line resources and tools for genomic, transcriptomic, and proteomic data analysis, as well as other analytic tools and online resources. This course is funded by a research grant from the National Institutes of Health (NIH)-Big Data to Knowledge (BD2K) Initiative.

WHAT DOES THE COURSE INCLUDE?

Each week begins with a **Weekly Guide** and ends with a **Conclusion** sub-section. In between these sub-sections, you'll find both the **lectures and activities** related to the section's topic. For most topics, you'll find a sequence that includes one or more lecture videos, followed by a demonstration of a relevant open-access web tool, and an exercise that challenges you to try out what you've just learned using the web tool.

Course Section Outline

1. Introduction
In the first week, you will meet the course Directors, Drs. Haddad, Gusev, and McGarvey, who will introduce the course in detail. You will also hear from three Georgetown University senior faculty members.

2. Translational Research and Big Data
In the second week, we will offer an overview and introduction to translational research and the role of big data and translational bioinformatics in enabling translational research and personalized medicine.

3. DNA and Big Data
In the third week, we will cover several topics related to Big Data generated from studies of DNA.

4. RNA and Big Data
In the fourth week, we will cover the major types of Big data that are generated in studies of

transcriptomics i.e. RNA-based studies such as genome wide gene expression and microRNA expression profiling.

5. Proteins and Big Data Part I

In this week, we will cover three of the major types of big data that are generated in studies of proteins: protein sequence, protein interaction and mass spectometry proteomics data. We will review the sources of the data, databases and resources that store the data and explore challenges in managing and interpreting proteomics data.

6. Proteins and Big Data Part II

In this week, we will cover another important topic for all biomedical big data: data sharing, metadata, data formats, ontologies and related terminologies. We will end the week with a hands on demonstration covering aspects from week 5 and 6 on functional interpretation of large-scale omics data outlining an approach you can take with a variety of gene/protein data types to functionally analyze and characterize the results. We will use some online tools to illustrate how this can be done.

7. Systems Biology and Big Data

In the seventh week, we will cover an introduction to systems biology and the role of big data in the analysis of biological systems. An overview and definitions of systems biology is presented in Part 1 of the lecture, as well as the major types of methodologies of systems biology. Part 2 of the lecture provides a tour of major public online resources related to systems biology.

8. Course Conclusion

In the final week, you will hear from two scientists from the National Institutes of Health (NIH) who have been intimately involved with some of the most exciting initiatives and projects that heavily involve biomedical big data.

9. Final Exam

WHAT WILL I LEARN IN THE COURSE?

The course is designed to support you in achieving the following **learning goals**. By the end of the course, you will have:
- Come to understand how biomedical data are being generated and processed
- Learned about various biomedical big data resources (e.g. TCGA, G-DOC, UNIPROT, etc.)
- Explored and analyze genomic, transcriptomic, and proteomic data using various online analysis tools
- Made sense of big data using systems biology resources and tools
- Begun to appreciate the value of big data in biomedical research and clinical practice (e.g. enabling precision medicine)

WHAT SHOULD WE EXPECT FROM EACH OTHER?

What You Can Expect from the Course Team

Teaching Assistants will regularly participate in discussion boards to provide content clarification, guidance, and support.

You can also email us with important content-related questions at gux@georgetown.edu.

What You Can Expect from edX

In the event of a technical problem, you should click the "Help" tab located on the left border of the screen. This "Help" tab opens an instruction box that directs you to student Frequently Asked Questions (FAQs) for general edX questions. You can also:
- Report a problem
- Make a suggestion
- Ask a question

You may also contact technical@edx.org directly to report bugs.

What We Expect from You

You should expect to spend around three to six (3-6) hours per week to review optional resources, watch lectures, work through practice questions and other activities, including the discussion threads.

In each course section we have included activities to support you in reaching the specified learning objectives for that section. These ungraded activities are available to you as **practice questions** and not counted toward your grade, nor do they have due dates. They include:
- Multiple choice questions;
- Multiple answer (checkbox) questions; and
- Text response questions.

Your **Final Exam** will comprise 100% of your grade and will be made available between April 11 and May 2, 2017. The final exam will look very much like the demo exercise activities that you complete as you work through the course; with the exam, however, the step-by-step instructions will not be available, so we encourage you to make yourself familiar with use of the free online web tools that you practice throughout.

Your final exam must be completed between April 11 and May 2 (by 23:59 UTC). This is your chance for a course certificate, if you signed up for the Verified Track.

All activities included in the course are designed to help you gauge your learning as a result of your interaction with the course content, specifically the video lectures and the web tool demonstration for each section. Instructions on how to complete the activities are included within each course section.

Timeline

As part of this instructor-paced course, we ask you to follow along each week as we release new material. We do have one due date for this course, **May 2, 2017,** when your final exam must be

completed should you be interested in receiving a score for this course (and a certificate if you signed up on the Verified Track). The course will remain open as a live course until May 9, 2017, for those who wish to further explore the material. This iteration of the course will then be archived, and Demystifying Biomedical Big Data: A User's Guide will be reintroduced as a self-paced course. As a self-paced course, the course will be seen as a live resource; it will be minimally staffed, so we especially encourage learning interactions amongst students.

Netiquette Guidelines

Please be respectful
To promote the best educational experience possible, we ask each student to respect the opinions and thoughts of other students and be courteous in the way that you choose to express yourself. Informed debate should never give way to insult, rudeness, or anything that might detract from the learning process. BIOX-201-01x students should be respectful and considerate of all opinions.

In order for us to have meaningful discussions, we must learn to understand what others are saying and be open-minded about others' opinions. If you want to persuade someone to see things differently, it is much more effective to do so in a polite, non-threatening way rather than to do so antagonistically. Everyone has insights to offer based on his/her experiences, and we can all learn from each other. Civility is essential: our teaching assistants can and will remove students from the class who detract from the learning process with insulting comments on the course-wide discussion boards.

Look before you write
Prior to posting a question or comment on the discussion board, the GeorgetownX course team asks that you look to see if any of your classmates have the same question. Upvote questions that are similar to your own or that are also of interest to you, instead of starting a new thread. This will greatly help our TAs best monitor the discussions and bring important questions to the course team's attention.

Use the discussion board for course-related posts only
Although we encourage students to get to know each other, please use the discussion board for course content conversations only.

Properly and promptly notify us of technical issues
Although we do not predict technical issues, they can and may happen. To make sure these receive prompt attention, please use the "Help" tab to troubleshoot, and if you still need assistance, you can contact us at gux@georgetown.edu.

Academic Integrity

Observe the honor policy
Although collaboration and conversation will certainly contribute to your learning in the course, we ask students to refrain from collaborating with or consulting one another on any graded material for the course. Violations of the honor policy undermine the purpose of education and the academic integrity of the course. We expect that all work submitted will be a reflection of one's own original work and thoughts.

GeorgetownX faculty and stuff expect all members of the community to strive for excellence in scholarship and character.

APPENDIX A: DETAILED COURSE OUTLINE

1. Introduction (released February 14, 2017)

In the first week, you will meet the course Directors, Drs. Haddad, Gusev, and McGarvey, who will introduce the course in detail. You will also hear from three Georgetown University senior faculty members. Dr. Robert Clarke will speak about "Bioinformatics at Georgetown University Medical Center." Dr. Subha Madhavan will speak about "Biomedical Informatics: Enabling Research and Health Care." Dr. John Marshall will speak about "Biomedical Big Data: Enabling Personalized Medicine."

Learning Objectives
- Describe several ways that biomedical big data is enabling biomedical research and precision
- Highlight key big data repositories, web portals and tools and other online resources

2. Translational Research and Big Data (released February 21, 2017)

In the second week, we will offer an overview and introduction to translational research and the role of big data and translational bioinformatics in enabling translational research and personalized medicine.

Learning Objectives
- Summarize general approaches to analysis of big data in translational research
- List typical questions addressed by translational bioinformatics
- Conduct basic operations using several major online resources of big data for translational research, such as TCGA and G-DOC

3. DNA and Big Data (released February 28, 2017)

In the third week, we will cover several topics related to big data generated from studies of DNA.

Learning Objectives
- Describe the evolution of cytogenetics
- Explain the value that big data analysis brings to modern cytogenetics
- Explain how copy number data are explored and analyzed in cancer research
- Begin to describe genome sequencing, technologies, genomics terminology, and methodologies for analysis of DNA and RNA sequencing data
- Through hands-on experience, explore and analyze sequencing data on your own

4. RNA and Big Data (released March 7, 2017)

In the fourth week, we will cover the major types of big data that are generated in studies of transcriptomics i.e. RNA-based studies such as genome wide gene expression and microRNA expression profiling.

Learning Objectives
- Explain how biomedical big data are being generated from the RNA-based measurements in the field of biomedical research called transcriptomics
- Describe various types of transcriptomics data such as genes expression and microRNA expression
- Outline one or more applications of gene expression and microRNA expression data analysis in cancer research

5. Proteins and Big Data Part I (released March 14, 2017)

In this week, we will cover three of the major types of big data that are generated in studies of proteins: protein sequence, protein interaction and mass spectometry proteomics data. We will review the sources of the data, databases and resources that store the data and explore challenges in managing and interpreting proteomics data.

Learning Objectives
- Describe the size, growth and importance of protein sequence data
- List several sources of annotated protein sequence data
- Explain several big data management challenges related to protein sequence data
- Describe the relationship of protein sequencing to other omics data types
- Explain what protein interaction data is and some methods that generate the data
- Define types of protein interactions
- List interaction data sources and resources
- Describe the relationship of protein interaction data to other omics data types
- Describe the fundamentals of mass spec-based proteomics
- List sources of mass spec data

6. Proteins and Big Data Part II (released March 21, 2017)

In this week, we will cover another important topic for all biomedical big data: data sharing, metadata, data formats, ontologies and related terminologies. We will end the week with a hands on demonstration covering aspects from week 5 and 6 on functional interpretation of large-scale omics data outlining an approach you can take with a variety of gene/protein data types to functionally analyze and characterize the results. We will use some online tools to illustrate how this can be done.

Learning Objectives
- Argue for the importance of data sharing
- Describe what metadata is and why it is important
- Define data formats and describe why they are important
- Define biomedical ontologies and vocabularies and identify how they are used
- Explain how functional analysis of gene/protein data types is done

7. Systems Biology and Big Data (released March 28, 2017)

In the seventh week, we will cover an introduction to systems biology and the role of big data in the analysis of biological systems. An overview and definitions of systems biology is presented in Part 1 of the lecture, as well as the major types of methodologies of systems biology. Part 2 of the lecture provides a tour of major public online resources related to systems biology

Learning Objectives
  • Describe what a systems biology approach means
  • Explain how a systems biology approach helps to analyze big data
  • Describe how a systems biology approach enables discoveries in translational biomedical research
  • Apply systems biology tools to analyze molecular profiling data

8. Conclusion (released April 4, 2017)

In the final week, you will hear about regulatory issues related to biomedical big data from Sheila Zimmet, JD, and Ashley Carver, JD, office of Regulatory Affairs at Georgetown University Medical Center. You will also hear from two scientists from the National Institutes of Health (NIH) who have been intimately involved with some of the most exciting initiatives and projects that heavily involve biomedical big data. Dr. Ben Busby will speak about "Enabling Everyone to Share and Use Public Datasets." Dr. Jerry Lee will speak about "Cancer Moonshot."

Learning Objectives
  • Highlight publicly available resources related to biomedical big data (datasets, analysis tools, etc.)
  • Describe several exciting initiatives that heavily involve biomedical big data

9. Final Exam (released April 11, 2017)

The final exam will count for the entirety of your grade and will be open from April 11 to May 2.