Published in final edited form as:

Med Ref Serv Q. 2020 ; 39(1): 90–99. doi:10.1080/02763869.2020.1688623.

Developing a Suite of Online Learning Modules on the Components of Next-Generation Sequencing Projects

Robert A. Wright [Basic Science Informationist]

Johns Hopkins University School of Medicine, William H. Welch Medical Library, Suite 1-200,
2024 E. Monument Street Baltimore, MD 21205.

Abstract

This column describes a project funded by the *Big Data to Knowledge* initiative to develop online learning modules for researchers, clinicians, and informationists/librarians on aspects of next-generation sequencing projects. The modules describe a framework for next-generation sequencing projects, with particular emphasis placed on experimental design, ethical considerations, data storage, and data sharing. The author outlines how the modules were developed and summarizes their contents.

Keywords

Big data; bioinformatics; educational materials; next-generation sequencing; user training

INTRODUCTION

The creation, sharing, reanalysis, and recombination of biomedical big data offer unprecedented opportunities for new discoveries in biology and human health. Next-generation sequencing (NGS) data is a particular type of biomedical big data that reveals expression, structure, and change in the genome. Because NGS data address these fundamental properties, researchers can use them to understand the nature of biological processes, the determination of traits, and the causes of disease. For example, researchers recently reanalyzed a large NGS dataset, looked for small changes in the genomes of Alzheimer's disease (AD) patients, and identified a gene potentially linked to AD risk.¹ This study used data from a repository that includes human genomic sequences, the database of Genotypes and Phenotypes (dbGaP) <<https://www.ncbi.nlm.nih.gov/gap/>>, which is maintained by the National Center for Biotechnology Information (NCBI). The NGS data from over 11,000 AD patients and healthy controls are included in the dbGaP dataset that the researchers reanalyzed.²

Compared with earlier sequencing technologies, NGS is a much faster process, with improved overall accuracy and a much-reduced cost.^{3,4} Because of these improvements, researchers have applied NGS technologies to a wide variety of biological and clinical questions, leading to an accompanying explosion of NGS data. The size and growth of the

(rwright32@jhmi.edu).

Sequence Read Archive (SRA), another large repository of shared NGS data maintained by NCBI, reflects this explosion. The SRA contained about ten terabases of data (10^{13} bases) in 2010. It grew 1,000 times in seven years, reaching about 10 petabases (10^{16} bases) by 2017.⁵

Next-generation sequencing projects are complex. In part, this is because the datasets that they generate or reuse are very large and rich with meaning. These projects have additional complexities, however, that involve how researchers design effective experiments, inform participants of risks and benefits, use computational tools in sequence analysis, and store, manage, and document big data. In addition, NGS projects are composed of many steps that may require specific expertise, are deeply inter-related, and may have important consequences for other downstream steps.

To address this complexity, to assist with effective planning for NGS projects, and to support the sharing of data from these projects, the Welch Medical Library (Welch) and the Division of Health Science Informatics (DHSI)—both units within the Johns Hopkins School of Medicine—developed a suite of five online learning modules. These modules fulfill the requirements of a three-year R25 grant (2015–2018) from the National Institutes of Health *Big Data to Knowledge* (BD2K) initiative for developing open educational resources to facilitate the sharing of biomedical big data.⁶ A framework module provides an NGS project overview. Four other modules on experimental design, ethics, data storage, and data sharing cover these topics in detail. See the module descriptions below for URLs to individual module webpages.

The modules have content designed for researchers and clinicians who are new to NGS projects or who want to learn more about particular aspects of these projects. Module authors also direct content to the informationists and librarians who work with these researchers and clinicians. This content orients informationists and librarians to the core activities, information flows, key expertise, and planning requirements associated with NGS projects. It also suggests important roles that informationists and librarians can play to support these projects, including roles in literature searching, finding funding, data management, data sharing, and the provision of guides to resources on relevant topics. This column will describe in more detail the content of the five modules after first outlining the module development process.

THE MODULE DEVELOPMENT PROCESS

Work Organization

The following team produced the five learning modules:

- *Principal investigators* – The directors of Welch and DHSI served as the grant's principal investigators, with the authority and responsibility for leading and directing the project. They conceived the project's aims, scope, governance, and module development plan.
- *Co-investigator* – A researcher with expertise in human genetics and the application of NGS technologies served as co-investigator and module director.

He provided essential intellectual direction for the project as a whole and critical oversight of module content.

- *Project manager* – An associate director at Welch was the grant’s project manager, setting goals and monitoring progress towards project milestones.
- *Instructional coordinators* – The project had three instructional coordinators (ICs), two from the Johns Hopkins Sheridan Libraries (a data services librarian and a senior data management consultant) and one from Welch (a basic science informationist). They worked with the co-investigator to develop module content and coordinated with the instructional designers on the modules’ look and feel. See below for a more detailed description of the roles played by the ICs.
- *Evaluation panel* – The principal investigators recruited a five-member panel of information specialists who provide services related to NGS, bioinformatics, and data management in health sciences libraries. This panel provided valuable feedback on module content, organization, design, and usability.
- *Instructional designers* – Staff members from the Office of Online Education at the Johns Hopkins School of Medicine provided invaluable instructional design support for the modules. They produced sixty-five embedded videos, advised the ICs regarding design elements and narrative flow, and created the final products using the Articulate Rise 360 e-learning software <<https://articulate.com/360/rise>>.

The Roles of the Instructional Coordinators

The ICs were the members of the module development team most responsible for assembling the content of the modules. It is worth examining the roles they played while creating this content in order to capture a sense of the effort needed to meet the project’s goals. The roles highlighted in the following discussion and outlined in Figure 1 were common to the three ICs. However, many of the details will focus on the roles the author played as the IC of the framework module.

Scoping.—As has been mentioned, next-generation sequencing projects are inherently complex and comprised of many steps. Nested within this complexity are additional layers of complexity in the form of processes for applying for institutional review board approval, procedures for acquiring and tracking human samples, protocols and software for analyzing the products of sequencing, etc. Because of this layered complexity, each module is generally broad in its focus and not too specific about approaches, tools, or techniques. One exception to this level of focus encompasses sections of the data sharing module that cover in considerable detail the steps for submitting NGS data to dbGaP. Overall, however, the modules aim to introduce key topics and concepts while providing resources that users can access to explore the additional complexities.

With this approach in mind, the IC and the co-investigator determined what would be covered and at what depth early in the development process for each module. In the case of the framework module, the IC created a list of up to four core activities for each step in an

NGS project. The co-investigator provided feedback on this list, and the IC made the necessary changes. The framework's 22 steps and 49 core activities served as the module's content outline.

Researching.—The ICs filled knowledge gaps related to NGS by meeting with the co-investigator and by searching for answers in the literature. A primary literature source was the co-investigator's textbook on bioinformatics, which has sections on NGS technologies, data analysis, data storage, etc.⁷ The IC for the framework module found over 100 additional sources by searching PubMed and other databases, the Johns Hopkins Libraries' Catalog, and the web. Lists of these sources appear throughout the module and include review articles, research articles, book chapters, videos, databases, and websites.

Synthesizing and Summarizing.—After assembling supporting material through research, the ICs synthesized and summarized this material. For example, the IC used eight sources to write the content for the section of the framework module on the sample processing step, with core activities related to preparing DNA for sequencing. These sources were primarily research protocols providing systematic instructions for sample processing. The two non-protocol sources were a review article on recent advances in sample processing in clinical NGS and a standards document for clinical NGS laboratories. For the module, the IC used only the most relevant material from these eight sources and reduced that output to its essential elements. The results are the four short paragraphs in the module describing three key elements of sample processing: (1) generating a library of short pieces of DNA, (2) uniquely identifying DNA samples with barcodes, and (3) targeting certain parts of the genome for sequencing.

In addition to extracting material from sources relevant to core activities, the ICs also highlighted material relevant to the modules' three audiences: researchers, clinicians, and informationists/librarians. The focus of the framework module is to provide an NGS project overview. In this sense, it is inherently relevant to its three audiences as a tool for understanding a complex process with a bird's-eye view. The IC also highlighted parts of the framework of particular importance to certain audience members. For example, the IC points out the informationist's role in expert searching in the financial issues and question formulation steps of the framework.

THE MODULES

The ICs and the co-investigator created five modules with unique content but common content elements. Thematic sections parse the modules' contents, and these sections contain discrete lessons on well-defined topics. The modules all have narrative text that summarizes, provides context, or describes steps in a process. They also include short videos with the co-investigator, which amplify the main messages of a module or lesson and may include specific real world examples drawn from his experience. The module creators used diagrams, bulleted lists, or tables to illustrate and organize complex content. Bibliographies and resource lists at the ends of modules or lessons are also common elements. What follows are summaries of the unique content in each module.

The Framework Module

The framework module (“A Framework for Next-Generation Sequencing Projects”) <<http://bd2k.welch.jhmi.edu/framework>>, developed by the co-investigator and the basic science informationist IC, has the broadest scope of the five modules. It provides an overview of the critical aspects of an NGS project, describing a framework of twenty-two individual steps grouped in five main stages: formulation, approval, preparation, experimentation, and publication (see Figure 2). These stages form the thematic sections of the module. The lessons of the formulation, approval, and preparation stages describe important steps to accomplish before performing any sequencing. These steps include assembling a research team, obtaining informed consent in clinical and research contexts, and deciding on the sequence-generating technologies and data analysis software that researchers will use. The lessons of the experimentation stage describe steps taken during the generation and interpretation of genomic sequences. These steps include acquiring data from DNA sequencers and securely storing it, analyzing the data using established workflows, and using statistical approaches to interpret the data for biologically or clinically significant associations. Finally, the lessons of the publication stage include steps taken by the researcher or clinician at the end of an NGS project. These steps include sharing data, returning results to patients, and publishing findings.

The Experimental Design Module

The experimental design module (“Experimental Design for Next-Generation Sequencing”) <<http://bd2k.welch.jhmi.edu/experimental-design>>, developed by the co-investigator and the basic science informationist IC, covers key factors to consider when designing experiments for NGS projects. Its three thematic sections relate to decisions about whom to sequence, which sequencing technique to use, and how to minimize error and bias. The first section’s lesson outlines the importance of family histories and pedigrees when deciding whose genomes to sequence for useful comparisons. The second section’s lessons describe genomic sequencing techniques, the pros and cons of these techniques, and the research and clinical scenarios in which they may be applied. Also included in this section is an introduction to other types of NGS approaches that address specific kinds of biological phenomena, such as differential gene expression and the regulation of genes through DNA-protein interactions. The lessons in the final section of the module highlight some of the known sources of error and bias in NGS, discuss the consequences of these for experimental results, and point users to approaches that address them.

The Ethics Module

The ethics module (“Ethics and Next-Generation Sequencing”) <<http://bd2k.welch.jhmi.edu/ethics>>, developed by the co-investigator and the data services librarian IC, addresses ethical issues faced by researchers and clinicians when they collect and use NGS data to understand disease mechanisms and treat patients. It has an introduction that describes two themes echoed throughout the module, the core principles of bioethics and key differences between research-based and clinical sequencing. In a section for researchers, the lessons outline the frequently broad scope of research-based sequencing and the rules and regulations governing human subjects and genomics research. The lessons also describe best practices for

obtaining consent and explaining risks and benefits, which include agreements with participants about how findings will be shared with them, and if and how data will be shared with the research community. A parallel section for clinicians addresses similar issues while pointing out important differences from the research context. These differences include the frequently narrow scope of clinical sequencing and specific guidelines for sharing medically actionable findings with patients. The final section of the module addresses special ethical situations, including working with vulnerable populations, and offers guidance for informationists and librarians on how they can support researchers and clinicians in meeting the ethical demands of working with NGS.

The Data Storage Module

The data storage module (“Data Storage for Next-Generation Sequencing”) <<http://bd2k.welch.jhmi.edu/data-storage>>, developed by the co-investigator and the basic science informationist IC, introduces data storage challenges and solutions for NGS projects. An introductory section lays out the scope of the challenge by describing the quantity of NGS data generated from a typical sequencing experiment, the barriers to transporting large data files, and the costs of data storage infrastructure. The lessons of the next section summarize data management best practices, including logical file naming conventions, data backup and archiving, and data security and privacy. The final section of the module covers data storage solutions, including individual lab setups, institution-level computing infrastructure, cloud computing, and public repositories of NGS data.

The Data Sharing Module

The data sharing module (“Data Sharing for Next-Generation Sequencing”) <<http://bd2k.welch.jhmi.edu/data-sharing>>, developed by the co-investigator and the senior data management consultant IC, has the most detail of the five modules. It introduces the context and reasons for sharing NGS data and describes the process of preparing data for submission to dbGaP. Its sections describe the environment for NGS data sharing, planning steps for sharing, approaches to data management, and detailed guidance for sharing data in dbGaP. Overview sections describe the importance of shared NGS data, the value of public data repositories, and data sharing requirements (see Figure 3). The next two sections point out ways that users can prepare for data sharing at every stage of an NGS project. Lessons in these sections describe specific activities, including developing a data sharing plan, obtaining informed consent from participants for data sharing, and acquiring data management expertise on the research team. A related section on data management describes additional tasks involving file metadata, file organization, data security, and repository specifications for data. The final section focuses on the details of submitting data to dbGaP, including an overview of submission forms, requirements for de-identifying data, required file types, and the dbGaP submission portal.

THE VALUE OF THE MODULES

The modules summarized in this column are tools for informing researchers and clinicians about NGS projects. The modules do this in two related ways. First, they provide a framework of important steps in a complex process. This framework can serve as a critical

planning guide, informing users of what they need to be aware of before starting an NGS project. Areas for planning, for instance, relate to research team expertise, ethical safeguards, technology, and infrastructure. Second, the modules provide more details about underlying issues and processes for particular key steps in the framework. They provide guidance, for instance, on obtaining informed consent, designing experiments, and sharing data in a repository.

Informationists and librarians can also find the modules valuable and use them in a number of ways. They can point their users to these modules for background information about NGS projects and include the modules in resource lists. They can use the modules as inspiration for their own online guides on NGS-related subjects. These guides might provide more detail on ethics best practices, sequencing technologies, or the landscape of data repositories. Finally, informationists and librarians can use the modules as a professional development tool that provides a solid foundation in the components of NGS projects. Armed with this grounding in the workflows and key considerations of NGS projects, they may become comfortable supporting question formulation, identifying units on campus with relevant expertise, or helping with data sharing.

User engagement will help determine whether the project modules contain useful content and effectively convey it. The author will track usage data, including the number and location of users and the modules used. User surveys planned by the author will help measure to what extent the modules support the work of their intended audiences: researchers, clinicians, and the informationists/librarians who work with them.

REFERENCES

1. Fleck D, Phu L, Verschueren E et al. "PTCD1 Is Required for Mitochondrial Oxidative-Phosphorylation: Possible Genetic Association with Alzheimer's Disease." *Journal of Neuroscience* 39, no. 24 (2019): 4636–4656. doi:10.1523/jneurosci.0116-19.2019 [PubMed: 30948477]
2. Gibbs RA, Lander ES, Wilson RK, Schellenberg GD, and Seshadri S. "Alzheimer's Disease Sequencing Project (ADSP) - dbGaP Study Accession: phs000572.v7.p4." Accessed October 1, 2019 https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000572.v7.p4.
3. Perry E "EBI: Next Generation Sequencing Practical Course - Improvements on the Previous Technology." Accessed October 1, 2019 <https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/improvements>.
4. Wetterstrand KA "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)." Accessed October 1, 2019 <https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>.
5. National Center for Biotechnology Information. "SRA Database Growth." Accessed October 1, 2019 <https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/>.
6. Seymour A, and Lehmann HP. "Training & Tools for Informationists to Facilitate Sharing of Next Generation Sequencing Data" Accessed October 9, 2019 https://projectreporter.nih.gov/project_info_description.cfm?aid=9036779&icde=0.
7. Pevsner J *Bioinformatics and Functional Genomics*. 3rd ed. Hoboken, NJ: John Wiley and Sons, Inc., 2015.

FIGURE 1.
The roles of the instructional coordinators

FIGURE 2.
The framework module

FIGURE 3.
The data sharing module