Research Techniques Made Simple: Bioinformatics for Genome-Scale Biology

Amy C. Foulkes[1], David S. Watson[2], Christopher E.M. Griffiths[1], Richard B. Warren[1], Wolfgang Huber[3] and Michael R. Barnes[2]

High-throughput biology presents unique opportunities and challenges for dermatological research. Drawing on a small handful of exemplary studies, we review some of the major lessons of these new technologies. We caution against several common errors and introduce helpful statistical concepts that may be unfamiliar to researchers without experience in bioinformatics. We recommend specific software tools that can aid dermatologists at varying levels of computational literacy, including platforms with command line and graphical user interfaces. The future of dermatology lies in integrative research, in which clinicians, laboratory scientists, and data analysts come together to plan, execute, and publish their work in open forums that promote critical discussion and reproducibility. In this article, we offer guidelines that we hope will steer researchers toward best practices for this new and dynamic era of data intensive dermatology.

Journal of Investigative Dermatology (2017) **137,** e163–e168; doi:10.1016/j.jid.2017.07.095

CME Activity Dates: 21 August 2017
Expiration Date: 20 August 2018
Estimated Time to Complete: 1 hour

Planning Committee/Speaker Disclosure: Amy Foulkes is a consultant/advisor for AbbVie, Almirral, Eli Lilly, Leo Pharma, Novartis, Pfizer, Janssen and UCB. Christopher Griffiths is on the speakers' bureau and is a consultant/advisor for AbbVie, GSK, Janssen, Pfizer, Lilly, Novartis, Celgene, Leo Pharma, UCB, Sun Pharmaceuticals, and Almirral; in addition, Dr. Griffiths receives research grant support from AbbVie, GSK, Janssen, Pfizer, Lilly, Novartis, Sandoz, Celgene, and Leo Pharma. All other authors, planning committee members, CME committee members and staff involved with this activity as content validation reviewers have no financial relationships with commercial interests to disclose relative to the content of this CME activity.

Commercial Support Acknowledgment: This CME activity is supported by an educational grant from Lilly USA, LLC.

Description: This article, designed for dermatologists, residents, fellows, and related healthcare providers, seeks to reduce the growing divide between dermatology clinical practice and the basic science/current research methodologies on which many diagnostic and therapeutic advances are built.

Objectives: At the conclusion of this activity, learners should be better able to:

- Recognize the newest techniques in biomedical research.
- Describe how these techniques can be utilized and their limitations.
- Describe the potential impact of these techniques.

CME Accreditation and Credit Designation: This activity has been planned and implemented in accordance with the accreditation requirements and policies of the Accreditation Council for Continuing Medical Education through the joint providership of William Beaumont Hospital and the Society for Investigative Dermatology. William Beaumont Hospital is accredited by the ACCME to provide continuing medical education for physicians.
William Beaumont Hospital designates this enduring material for a maximum of 1.0 *AMA PRA Category 1 Credit(s)*™. Physicians should claim only the credit commensurate with the extent of their participation in the activity.

Method of Physician Participation in Learning Process: The content can be read from the Journal of Investigative Dermatology website: http://www.jidonline.org/current. Tests for CME credits may only be submitted online at https://beaumont.cloud-cme.com/RTMS-Sept17 – click 'CME on Demand' and locate the article to complete the test. Fax or other copies will not be accepted. To receive credits, learners must review the CME accreditation information; view the entire article, complete the post-test with a minimum performance level of 60%; and complete the online evaluation form in order to claim CME credit. The CME credit code for this activity is: 21310. For questions about CME credit email cme@beaumont.edu.

INTRODUCTION

Modern dermatology has been revolutionized by the many so-called 'omic' profiling platforms enabled by high-throughput sequencing (HTS, also referred to as next-generation sequencing). Plunging data generation costs have enabled dermatology researchers to generate genome scale data relating to genome sequence variation (Scott et al., 2013), epigenomes (Zhou et al., 2016), and transcriptomes

[1]*The Dermatology Centre, Salford Royal NHS Foundation Trust, The University of Manchester, Manchester Academic Health Science Centre, Manchester, UK;* [2]*William Harvey Research Institute, Centre for Translational Bioinformatics, Barts and The London School of Medicine and Dentistry, Charterhouse Square, London, UK; and* [3]*European Molecular Biology Laboratory, Heidelberg, Germany*

Correspondence: A.C. Foulkes, NIHR Academic Clinical Lecturer in Dermatology, The Dermatology Centre, Salford Royal NHS Foundation Trust, The University of Manchester, Manchester Academic Health Science Centre, M6 8HD. E-mail: Amy.foulkes@manchester.ac.uk

Abbreviations: HTS, high-throughput sequencing; RNA-seq, RNA sequencing

ADVANTAGES

- Bioinformatics methods allow efficient and powerful analysis of multi-omic data in a way that could not be achieved using simpler methods.
- Bioinformatics software are customizable to all ranges of computational ability; however, some informatics tasks are difficult and require experience.
- Involving bioinformatician colleagues from project conception should improve project design, maximizing the opportunity to detect relevant association.
- Sharing data, metadata, and code, and propagating the culture of bioinformaticians, will fuel best practices in dermatology research, promoting open research and reproducibility.

LIMITATIONS

- Some statistical analysis methods require an understanding of underlying assumptions—erroneous assumptions can lead to false results.
- The use of some analytical pipelines requires access to high-performance computing facilities: this may be achieved by access to omic core facilities that provide researchers with compressed datasets that are amenable to computer-based analysis.

(Li et al., 2014; Swindell et al., 2016), and these developments have increased the dermatology-relevant data openly available in repositories (Table 1).

Bioinformatics refers to the tools used to collect, classify, and analyze such datasets, collectively enabling the field of *computational biology*. Bioinformatics techniques have been developed to make sense of the output of omic platforms, including HTS, microarrays, liquid chromatography-mass spectrometry, and others (Kimball et al., 2012).

Table 1. High-throughput sequencing repositories

Repository	Website	Curator
Europe		
European Nucleotide Archive (ENA)	http://www.ebi.ac.uk/ena	European Bioinformatics Institute
ArrayExpress	http://www.ebi.ac.uk/arrayexpress	European Bioinformatics Institute
European Genome-phenome Archive (EGA)	https://www.ebi.ac.uk/ega/home	European Bioinformatics Institute
United States		
dbGAP	https://www.ncbi.nlm.nih.gov/gap	The National Center for Biotechnology Information
Gene Expression Omnibus (GEO)	https://www.ncbi.nlm.nih.gov/geo	The National Center for Biotechnology Information
Short Read Archive (SRA)	https://www.ncbi.nlm.nih.gov/sra	The National Center for Biotechnology Information

Physicians are key instigators of research data collection requiring computational biology. Structured and validated analysis pipelines for most omic data have been implemented for researchers at various levels of complexity. Software has been designed for all ranges of computational ability, from simple "point and click" graphic user interfaces to highly customizable command line interfaces, with the latter approach offering superior flexibility and analytical complexity. Although programming may seem like a daunting challenge for those without backgrounds in math, computer science, or statistics, with practice, computational methods for exploratory and inferential analytics can become a familiar part of the research toolkit. Of course, there is no substitute for expertise, and we advise all research teams working with omic data to consult a bioinformatician early and often. Here we highlight several points of special relevance to the dermatologist and dermatology researcher, based on the first-hand experience of a junior clinician.

CONSIDERATIONS BEFORE DATA COLLECTION

Experimental Design

Researchers in dermatology use a wide variety of HTS techniques, many of which have been discussed previously in the Research Techniques Made Simple series. These include transcriptome analysis with RNA sequencing (RNA-seq) (Antonini et al., 2017; Whitley et al., 2016), immunosequencing (Matos et al., 2017), genome-wide epigenetics (Capell and Berger, 2013), proteomics, metabolomics, metagenomics, and assessment of the microbiome (Jo et al., 2016). Additionally, the Molecular Revolution in Cutaneous Biology series provided an overview of HTS techniques (Anbunathan and Bowcock, 2017; Botchkareva, 2017; Johnston et al., 2017; Kong and Segre, 2017; Sarig et al., 2017), as did Grada and Weinbrecht (2013) in an earlier Research Techniques Made Simple publication. However, researchers often do not reach out to data analysts until a study is practically complete. At that point, they may look for a mathematically inclined colleague to fill in the blanks of a statistical model and provide a friendly *P*-value suitable for publication. This order of events is all wrong. As Ronald Fisher famously put it back in 1938, "To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of" (Fisher, 1938).

The data analysis strategy, including the choice of statistical approaches, should be integral to planning any research study. Hypothesis testing, regression, and other statistical methods rely on rigorous collection and quality of the data, and any lapses here usually cannot be fixed retrospectively. How many samples are required to adequately power your experiment? If samples cannot be processed all at once, does it matter how they are grouped into separate batches? If the data do not corroborate your hypothesis, can a modified research question generate interesting results? Failure to consider these questions before data collection may doom a study before it even begins. Statistical expertise is required to answer these questions, which is why we urge researchers to team up with a data analyst who can help guide them through these tricky issues. This will typically either be a statistician, with a background in math and statistics, or a