Title: Bioinformatics meets open data: An efficient and cost-effective framework to boost basic science in countries and universities with limited resources

Serghei Mangul[1#], Lana S. Martin[2#], Ben Langmead[3], Javier Sanchez Galan[4], Ian Toma[5], Fereydoun Hormozdiari[6], Pavel Pevzner[7], Eleazar Eskin[1]

[1] Department of Computer Science, University of California, Los Angeles
[2] Institute for Quantitative and Computational Biosciences, University of California, Los Angeles
[3] Department of Computer Science, Johns Hopkins University
[4] Grupo de Investigación en Biotecnología, Bioinformática y Biología de Sistemas, Universidad Tecnológica de Panama
[5] Department of Clinical Research and Leadership, George Washington School of Medicine and Health Sciences
[6]Department of Biochemistry and Molecular Medicine, School of Medicine, University of California, Davis
[7] Department of Computer Science, University of California, San Diego

Corresponding author: smangul@ucla.edu
# - the authors contributed equally

Abstract

Publishing cutting-edge research in today's STEM journals requires resources and administrative coordination that are not available to many academic institutions and national economies. Adequate training and support could help scientists in lower-income regions and resource-poor institutions produce career-enhancing STEM research. We present the rationale for leveraging existing resources to enable scientists in lower-resource institutions and countries to re-analyze published "-omics" data given training, support, and access to standard computing hardware and cloud-based resources.

Introduction

Bioinformatics algorithms are now crucial for processing high throughput "-omics" data and deriving meaningful interpretations in most biomedical and life science research domains. Bioinformatics-related training and research mostly take place in nations with higher income economies and resource-rich institutions that offer adequate training and administrative support. Indeed, some aspects of science (e.g., wet labs) require significant resources to establish and maintain. However, today's availability of low-cost computing allows any individual with sufficient training and cloud access to develop novel computational methods and perform in silico analyses—key aspects of many modern scientific endeavors.

The growing popularity of cloud computing and availability of online training materials provides an excellent opportunity for aspiring bioinformaticians in countries and institutions with limited resources. In comparison to well-endowed universities in higher income nations, academic institutions in limited resource regions typically receive less funding, operate with less administrative coordination, and produce fewer papers in high-ranking academic journals. Scaling bioinformatics capacity upward and outward in countries and academic institutions with limited resources could build highly trained STEM talent within economically disadvantaged communities *while* broadening and diversifying the global scientific community.

Downstream impacts of training and involving bioinformaticians from lower income nations and resource-limited institutions include (a) improving quality of research by stimulating innovation and expanding the number of problems addressable by the global scientific community, (b) diversifying the scientific community, (c) stimulating innovation by

fostering novel perspectives on old problems; (d) further engaging scientists and the public in cross-cultural collaboration, (e) potentially easing economic and social tensions related the unemployment and emigration through provision of local jobs in regions currently suffering from depressed STEM employment.

Science in limited resource countries

Geographic disparities among lower and higher income nations are apparent in the types of scientific activities performed, the quantity of scientific publications produced annually, and the ranking of journals in which papers are published. The World Bank classifies national economies as lower middle-income countries (LMIC) when the gross national income (GNI) per capita falls between US$996 and $3,895; and defines low-income countries (LIC) as those with a GNI under $995 (1). Out of 219 economies surveyed in 2018, the World Bank classified 36.99% as high income, 15.53% as upper-middle income, 21.46% as LMIC (47 countries), and 25.57% as LIC (34 countries). Scientists in high and upper-middle income countries depend on novel high throughput technologies and large cohort sizes—economically intense resources—to produce a large volume of scientific discoveries in today's "big data" world of science.

Scientists in LMIC and IC have historically made fewer contributions to modern science (2). Of the 159 countries included in the *Nature Index* database, 103 countries produced fewer than 100 scientific papers from 2015 to 2016. In comparison to higher income countries, LMIC

and IC operate with different historical economic developments. Implementing administrative teams and building wet labs in a developing country could produce high throughput "-omics" data sufficient for large-scale analysis, but projects of this scale would require coordination infrastructure and financial aid well beyond amounts already provided by domestic and international sources.

During the past decade, several limited resource countries have offered financial incentives with the aim of stimulating domestic scientific productivity. For example, the Indonesian federal government offers a cash prize award of 100 million Rupiah (US$7,400) to scientists who publish in high-impact journals (3). An estimated 90% of universities in China offer scientists awards up to 500,000 yuan (US$70,000) for a paper published in a high impact journal (4). Despite financial incentivization, papers published by Indonesia's scientists comprise 0.16% of papers published by U.S. scientists. Papers published by China's scientists remain about half of those published by U.S. scientists (2). Cash awards for publication may be insufficient compensation for graduate students who lack full-time employment and for a limited infrastructure of research support. Additionally, the administrative burden of securing long-term access to appropriate training and support services may represent a barrier to scientists operating with smaller budgets.

Improving the accessibility of bioinformatics training and research

We propose a new model of education, training, and support that would enable scientists living in lower-income countries to re-analyze publicly available omics data with bioinformatics methods—and, ultimately, increase domestic scientific research and publication productivity. Adequate computer infrastructure and high-speed internet connections are often available in the computer science and engineering departments of educational institutions in LMIC and IC. At present, these resources are rarely used for bioinformatics analysis as the computer scientists running high-performance clusters are not trained in bioinformatics. Alternatively, researchers in resource-poor institutions can use cloud resources to analyze large omics data sets, which eases the logistical burden of assembling a high-performance computing infrastructure.

Re-analysis of existing high throughput biomedical data can be cost-effective, produce novel scientific discoveries, provide important insights into complex biological systems, and potentially correct any statistical or computational issues identified in the original publication (2) (5). For example, Paez-Espino et al. (6) used bioinformatics methods to characterize the Earth's virome solely based on existing metagenomics data; Iyer et al. (7) used existing RNA sequencing data to catalog non-coding RNAs; and Gutzwiller et al. (8) used gene expression data to study the relationship between fruit fly and its endosymbiont *Wolbachia pipientis*.

Unlike in wet labs, where training can be slow due to the numerous safety issues, teaching and supporting scientists to perform bioinformatics analyses with "-omics" data is quick, safe and can be done one with a few lessons. Aspiring bioinformaticians in lower-income

countries can leverage existing resources available at local educational institutions with online educational materials to produce novel methods and perform re-analysis of published data. Platforms for virtual training of bioinformaticians already exist (see https://github.com/smangul1/online.bioinformatics/wiki). Scientists living in LMIC and LIC can source foundational education needs through such online workshops, resources, and review articles. For example, novice computational analysts may self-tutor using online UNIX workshops geared for first-time command-line users. Lay-friendly review articles on niche scientific topics can help undergraduates and established scholars alike gain a basic understanding of a concept or field without enrolling in time- and cost-prohibitive coursework. In addition, the increasing availability of processed (i.e., summarized) versions of genomics datasets provides for bioinformatics trainees new "onramps" to increase and broaden the relevance of data for more training levels and specialties. Summarized datasets are typically smaller and require less network bandwidth to handle.

We anticipate an even greater demand for analysis of bioinformatics data in coming years. We believe that establishment of a global bioinformatics training and support consortium, which unifies existing platforms and materials, will incentivize scientists in lower-income countries and institutions to participate in cutting-edge STEM research. We have developed an online resource guide comprised of educational materials, example bioinformatics code, example data sets, cloud-based resources, and an interface to access interesting and potentially important data sets (see https://github.com/smangul1/online.bioinformatics/wiki). We have used two major resources

globally, which have already been used in 14 LIC and 41 LMIC (see Supplementary Note 1). This platform can feasibly introduce scientists in academic institutions and regions with poor STEM resources to career-enhancing aspects of modern biotechnology that have previously been restricted to wealthier colleges, universities, and national economies.

Case Study: Central America

The Central American Region is a mix of LMIC and LIC countries, each bearing a rich history of natural products discoveries, and medical, biomedical, and public health research developments. Traditionally, scientists in LMIC and LIC have conducted research both domestically and in collaboration with scientists from higher-income countries. Inadequate funding structures, and a lack of administrative experience managing interdisciplinary training programs, have challenged the integration of life sciences and computational biosciences in Central America. Life scientists are typically trained in a purist scholastic tradition, following time-honored academic divisions, and computer scientists are typically trained for employment in private software development and telecommunication industries.

Building interdisciplinary training and support in science and technology departments could help Central American institutions move toward more innovative STEM practices such as advanced computing technology and fields such as bioinformatics and systems biology. Most major educational institutions in Central America already possess adequate computational infrastructure to perform analytical work and method development. Given specialized training and cloud-based computing resources, students and researchers in Central America could

develop the techniques and knowledge required to modernize domestic academic units—much as many smaller institutions in the United States have during the past decade. Ultimately, bridging this gap could advance established molecular biology national laboratories in Central America. As domestic institutions become powerhouses for data collection, management, re-analysis, and open data dissemination, national clinics and hospitals could establish and develop their own develop high-throughput genetic and genomic research units.

Moving forward

Cloud-based resources can enable scientists in developing countries to train, research, and make novel scientific discoveries using publicly available "-omics" data. We believe that these resources, together with installed capacities and infrastructures already available in countries with limited resources, can support expansion of self-sustaining, cutting-edge STEM communities around the world. This model could have several important positive impacts for domestic local economies and education systems and would most likely be supported by domestic governments as part of their national STEM training programs. In addition to the positive downstream impacts mentioned above, such training programs will allow policymakers to introduce novel scientific domains into existing educational curricula, could incentive federal, state, and local funding of STEM resources, and could increase the global scientific community's focus on solving problems related to neglected tropical diseases or local genetic anomalies.

References

1. **Development, Organisation for Economic Co-operation and.** Gross national income. [Online] https://data.oecd.org/natincome/gross-national-income.htm.
2. *Nature Index 2015 Global.* **May, Mike and Brody, Herb.** S1, s.l. : Nature, 2015, Vol. 522.
3. *The developing world needs more than numbers.* **Rochmyaningsih, Dyna.** 7639, s.l. : Nature, 2017, Vol. 542.
4. *Don't pay prizes for published science.* **Editorial.** 137, s.l. : Nature, 2017, Vol. 547.
5. *Data Sharing.* **Longo, Dan L. and Drazen, Jeffrey M.** s.l. : New England Journal of Medicine, 2016, Vol. 374.
6. *Uncovering Earth's virome.* **David Paez-Espino, Emiley A. Eloe-Fadrosh, Georgios A. Pavlopoulos, Alex D. Thomas, Marcel Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N. Ivanova & Nikos C. Kyrpides.** s.l. : Nature, 2016, Vol. 536.
7. *The landscape of long noncoding RNAs in the human transcriptome.* **Matthew K Iyer, Yashar S Niknafs, Rohit Malik, Udit Singhal, Anirban Sahu, Yasuyuki Hosono, Terrence R Barrette, John R Prensner, Joseph R Evans, Shuang Zhao, Anton Poliakov, Xuhong Cao, Saravana M Dhanasekaran, Yi-Mi Wu, Dan R Robinson, David G Beer, Fel.** s.l. : Nature Genetics, 2015, Vol. 47.
8. *Dynamics of Wolbachia pipientis Gene Expression Across the Drosophila melanogaster Life Cycle.* **Florence Gutzwiller, Danny W. Rice, Irene L. G. Newton, R. Scott Hawley, View ORCID ProfileLuis Teixeira and View ORCID ProfileCasey M. Bergman.** 12, s.l. : G3: Genes, Genomes, Genetics, 2015, Vol. 5.

Acknowledgments

J. E. S-G is supported by the Sistema Nacional de Investigadores (SNI) from La Secretaría Nacional de Ciencia, Tecnología e Innovación (SENACYT), which is administered by the federal government of Panama. S.M. and L.M. receive support from the Institute for Quantitative and Computational Biosciences (QCBio) at UCLA.

Competing interests

The authors declare no competing interests.

Supplementary Note 1

Several top-ranked educational institutes in the United States already disseminate their educational materials to other countries. For example, the MicroMaster Program at UCSanDiegoX trains researchers in the biomedical, medical, and life sciences to use bioinformatics techniques to analyze the data generated by their laboratories. In addition, the program trains computer scientists how to approach real life biology problems—even if the individual lacks a background in biology.

The materials developed for both approaches have been used by professors in at least 30 countries (Supplementary Table 1 and Supplementary Figure 1). The majority of countries (53%) who have implemented the UCSanDiegoX bioinformatics materials are classified as "high-income." However, 23% of countries adopting the program are upper-middle income, and 23% of countries are lower-middle-income. If countries with few resources for conducting domestic scientific research—including Bangladesh, Indonesia, Pakistan, Tunisia, and West Bank—have adopted online learning for bioinformatics training, then we suggest other lower-middle income countries may benefit as well.

In addition, the online course "Algorithms for DNA Sequencing," developed by a professor at Johns Hopkins University and offered worldwide by Coursera, trains researchers on selecting appropriate algorithms to use when analyzing biomedical, medical, or life sciences data. The materials developed for this course have been used in 147 countries (Supplementary Table 2 and Supplementary Figure 2). The majority of countries who have used the course are classified as high income (33%) and upper-middle income (27%). However, a substantial portion of course participants are located in lower-middle income (23%) and low-income (9.52%)

countries—including historically resource-poor nations in the African continent. Eight percent of countries included in roster of course users do not provide GNI per capita data (e.g., Cuba, Puerto Rico, Somalia, Macao, Syria, Taiwan).

Supplementary Tables and Figures

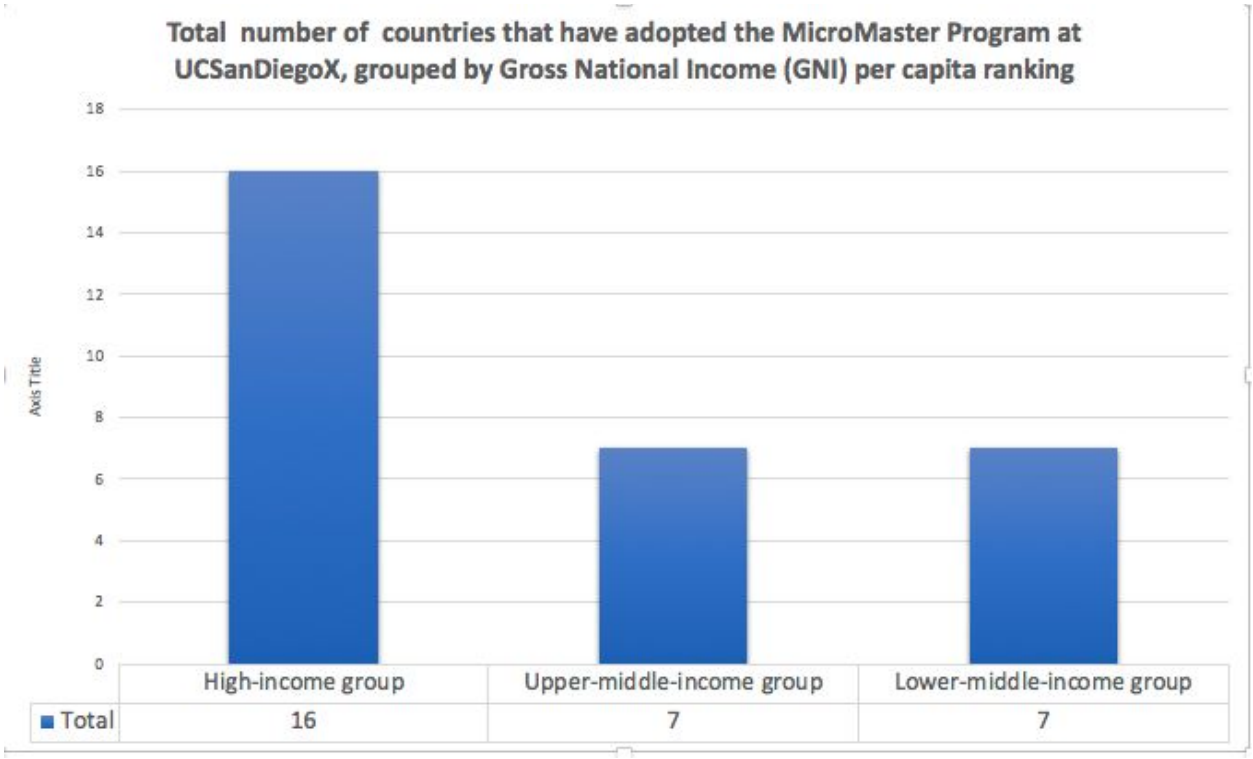
Supplementary Table 1. Gross national income (GNI) per capita ranking of countries that have adopted the MicroMaster Program at UCSanDiegoX.

Country*	**GNI per capita rank****
Australia	High-income group
Belgium	High-income group
Canada	High-income group
Czech Republic	High-income group
Finland	High-income group
Germany	High-income group
Iceland	High-income group
Israel	High-income group
Netherlands	High-income group
Norway	High-income group
Saudi Arabia	High-income group
Slovakia	High-income group
South Korea	High-income group
Spain	High-income group

United Kingdom	High-income group
United States	High-income group
Albania	Upper-middle-income group
China	Upper-middle-income group
Iran	Upper-middle-income group
Russia	Upper-middle-income group
Serbia	Upper-middle-income group
Thailand	Upper-middle-income group
Turkey	Upper-middle-income group
Bangladesh	Lower-middle-income group
Egypt	Lower-middle-income group
India	Lower-middle-income group
Indonesia	Lower-middle-income group
Pakistan	Lower-middle-income group
Tunisia	Lower-middle-income group
West Bank	Lower-middle-income group

** Data obtained on November 8, 2018, via personal communication with Pavel Pevzner, developer of courses offered by UCSanDiegoX, https://www.edx.org/bio/pavel-pevzner.

** Data obtained on November 13, 2018, from https://data.oecd.org/natincome/gross-national-income.htm.



Supplementary Figure 1. Total number of countries* that have adopted the MicroMaster Program at UCSanDiegoX, grouped by Gross National Income (GNI) per capita ranking**.

** Data obtained on November 8, 2018, via personal communication with Pavel Pevzner, developer of courses offered by UCSanDiegoX, https://www.edx.org/bio/pavel-pevzner.

** Data obtained on November 13, 2018, from https://data.oecd.org/natincome/gross-national-income.htm.

Supplementary Table 2. Gross national income (GNI) per capita ranking of countries that have adopted "Algorithms for DNA Sequencing" via Coursera.

Country*	GNI per capita rank**
Argentina	High-income group
Australia	High-income group
Austria	High-income group
Bahrain	High-income group
Belgium	High-income group
Canada	High-income group
Chile	High-income group
Croatia	High-income group
Cyprus	High-income group
Czech Republic	High-income group
Denmark	High-income group
Estonia	High-income group
Finland	High-income group
France	High-income group
Germany	High-income group
Greece	High-income group
Hong Kong	High-income group
Hungary	High-income group
Iceland	High-income group
Ireland	High-income group

Israel	High-income group
Italy	High-income group
Japan	High-income group
Kuwait	High-income group
Latvia	High-income group
Lithuania	High-income group
Luxembourg	High-income group
Malta	High-income group
Netherlands	High-income group
New Zealand	High-income group
Norway	High-income group
Oman	High-income group
Panama	High-income group
Poland	High-income group
Portugal	High-income group
Qatar	High-income group
Saudi Arabia	High-income group
Singapore	High-income group
Slovakia	High-income group
Slovenia	High-income group
South Korea	High-income group
Spain	High-income group
Sweden	High-income group

Switzerland	High-income group
Trinidad and Tobago	High-income group
United Arab Emirates	High-income group
United Kingdom	High-income group
United States	High-income group
Uruguay	High-income group
Albania	Upper-middle-income group
Algeria	Upper-middle-income group
Armenia	Upper-middle-income group
Azerbaijan	Upper-middle-income group
Belarus	Upper-middle-income group
Bosnia and Herzegovina	Upper-middle-income group
Botswana	Upper-middle-income group
Brazil	Upper-middle-income group
Bulgaria	Upper-middle-income group
China	Upper-middle-income group
Colombia	Upper-middle-income group
Costa Rica	Upper-middle-income group
Ecuador	Upper-middle-income group
Fiji	Upper-middle-income group
Guatemala	Upper-middle-income group
Guyana	Upper-middle-income group
Iran	Upper-middle-income group

Iraq	Upper-middle-income group
Jamaica	Upper-middle-income group
Jordan	Upper-middle-income group
Kazakhstan	Upper-middle-income group
Lebanon	Upper-middle-income group
Libya	Upper-middle-income group
Macedonia	Upper-middle-income group
Malaysia	Upper-middle-income group
Maldives	Upper-middle-income group
Mauritius	Upper-middle-income group
Mexico	Upper-middle-income group
Montenegro	Upper-middle-income group
Namibia	Upper-middle-income group
Paraguay	Upper-middle-income group
Peru	Upper-middle-income group
Romania	Upper-middle-income group
Russia	Upper-middle-income group
Saint Lucia	Upper-middle-income group
Serbia	Upper-middle-income group
South Africa	Upper-middle-income group
Thailand	Upper-middle-income group
Turkey	Upper-middle-income group
Bangladesh	Lower-middle-income group

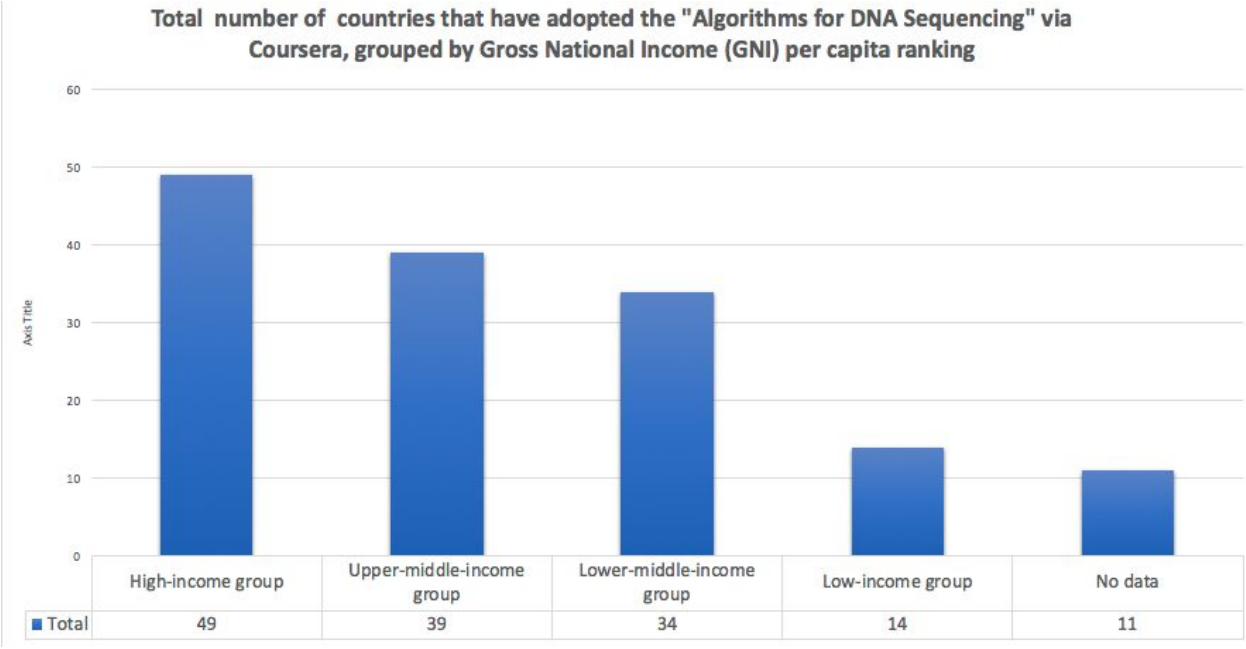
Bhutan	Lower-middle-income group
Bolivia	Lower-middle-income group
Cambodia	Lower-middle-income group
Cameroon	Lower-middle-income group
Congo	Lower-middle-income group
Cote d'Ivoire	Lower-middle-income group
Egypt	Lower-middle-income group
El Salvador	Lower-middle-income group
Georgia	Lower-middle-income group
Ghana	Lower-middle-income group
Honduras	Lower-middle-income group
India	Lower-middle-income group
Indonesia	Lower-middle-income group
Kenya	Lower-middle-income group
Kyrgyzstan	Lower-middle-income group
Laos	Lower-middle-income group
Lesotho	Lower-middle-income group
Moldova	Lower-middle-income group
Mongolia	Lower-middle-income group
Morocco	Lower-middle-income group
Myanmar	Lower-middle-income group
Nicaragua	Lower-middle-income group
Nigeria	Lower-middle-income group

Pakistan	Lower-middle-income group
Palestine	Lower-middle-income group
Philippines	Lower-middle-income group
Sao Tome and Principe	Lower-middle-income group
Sri Lanka	Lower-middle-income group
Sudan	Lower-middle-income group
Tunisia	Lower-middle-income group
Ukraine	Lower-middle-income group
Vietnam	Lower-middle-income group
Zambia	Lower-middle-income group
Afghanistan	Low-income group
Benin	Low-income group
Burkina Faso	Low-income group
Burundi	Low-income group
Ethiopia	Low-income group
Guinea	Low-income group
Madagascar	Low-income group
Mali	Low-income group
Nepal	Low-income group
Niger	Low-income group
Rwanda	Low-income group
Senegal	Low-income group
Tanzania	Low-income group

Uganda	Low-income group
Aruba	No data
Cuba	No data
Faroe Islands	No data
French Guiana	No data
Guernsey	No data
Macao	No data
Puerto Rico	No data
Somalia	No data
Syria	No data
Taiwan	No data
Venezuela	No data

** Data obtained on November 14, 2018, via personal communication with Ben Langmead, developer of courses offered by Coursera, https://www.coursera.org/instructor/benlangmead.

** Data obtained on November 14, 2018, from https://data.oecd.org/natincome/gross-national-income.htm.



Supplementary Figure 2. Total number of countries* that have adopted "Algorithms for DNA Sequencing" via Coursera, grouped by Gross National Income (GNI) per capita ranking**.

** Data obtained on November 14, 2018, via personal communication with Ben Langmead, developer of courses offered by Coursera, https://www.coursera.org/instructor/benlangmead.

** Data obtained on November 14, 2018, from https://data.oecd.org/natincome/gross-national-income.htm.