

Open access publishing

A researcher's perspective

Karl Broman

Biostatistics & Medical Informatics
University of Wisconsin–Madison

biostat.wisc.edu/~kbroman
github.com/kbroman

These are slides for a talk I will give on 24 Oct 2013, at a symposium on open access publishing, organized by the Ebling Library, UW–Madison.

I'm a statistician. My research focuses on genetics, and most of my papers are in genetics journals.

So in commenting on open access, I'm focusing on scientific publications, and perhaps more narrowly, on the biological sciences.

Access in action

Interesting reference

- [8] Kang C. and Speller R. The effect of region of interest selection on dual energy x-ray absorptiometry emasurements of the calcaneus in 55 post-menopausal women. *The british Journal of Radiology*, 72:864–871, 1999.
- [9] The 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. *Nature*, 467:1061–1073, 2010.
- [10] John C.V., Mark D.A., Eugene W.M., Peter W.L., Richard J.M., Granger G.S., and Hamilton O.S. The sequence of the human genome. *Science*, 291:1304–1351, 2001.
- [11] Schwartz D.C. and Waterman M.S. [New generations: sequencing machines and their computational challenges](#). *Journal of Computer Science and Technology*, 25(1):3–9, 2010.
- [12] Church D.M., Goodstadt L., Hillier L.W., Zody M.C., Goldstein S., She X., Bult C.J., Agarwala R., Cherry J.L., DiCuccio M., Hlavina W., Kapustin Y., Meric P., Maglott D., Birtle Z., Marques A.C., Graves T., Zhou S., Teague B., Potamousis K., Churas C., Place M., Herschleb J., Runnheim R., Forrest D., Amos-Landgraf J., Schwartz D.C., Cheng Z., Lindblad-Toh K., Eichler E.E., and Ponting C.P. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biology*, 7.5:e1000112, 2009.
- [13] Tor D.W., Matthew C.K., Steven C.L., and John J. Increased sensitivity in neuroimaging analyses using robust regression. *Neuroimage*, 26:99–113, 2005.

2

I'll begin with an illustration of what I mean by access.

The other day I was reading a manuscript and saw an article of interest.

Access in action

Google Scholar

The screenshot shows a Google Scholar search result for the article "New Generations: Sequencing Machines and Their Computational Challenges" by David C. Schwartz and Michael S. Waterman. The result is from the "Journal of Computer Science and Technology" (J CST) in January 2010, Volume 25, Issue 1, pp 3-9. The page includes a "Buy now" button for \$39.95 / €34.95 / £29.95*, a "Get Access" button, and a "Look Inside" link. The right sidebar offers sharing options (Facebook, Twitter, LinkedIn) and other actions like "Export citations" and "About This Journal".

3

If I paste the article title into Google Scholar, I immediately find the paper and can go directly to the journal.

But I was sitting at home on my couch.

And they charge \$40 for a 7 page paper!

What's the deal with the prices?

| | |
|---|---------|
| Broman K, Speed T, Tigges M (1996) Estimation of antigen-responsive T cell frequencies in PBMC from human subjects. J Immunol Meth 198:119–132 | \$39.95 |
| Broman KW, Weber JL (1999) Method for constructing confidently ordered linkage maps. Genet Epidemiol 16:337–343 | \$35.00 |
| Broman KW, Feingold E (2004) SNPs made routine. Nat Methods 1:104–105 | \$18.00 |
| Broman KW (2005) Mapping expression in randomized rodent genomes. Nat Genet 37:209–210 | \$18.00 |

4

I went back to some of my early papers, and found these outrageous prices.

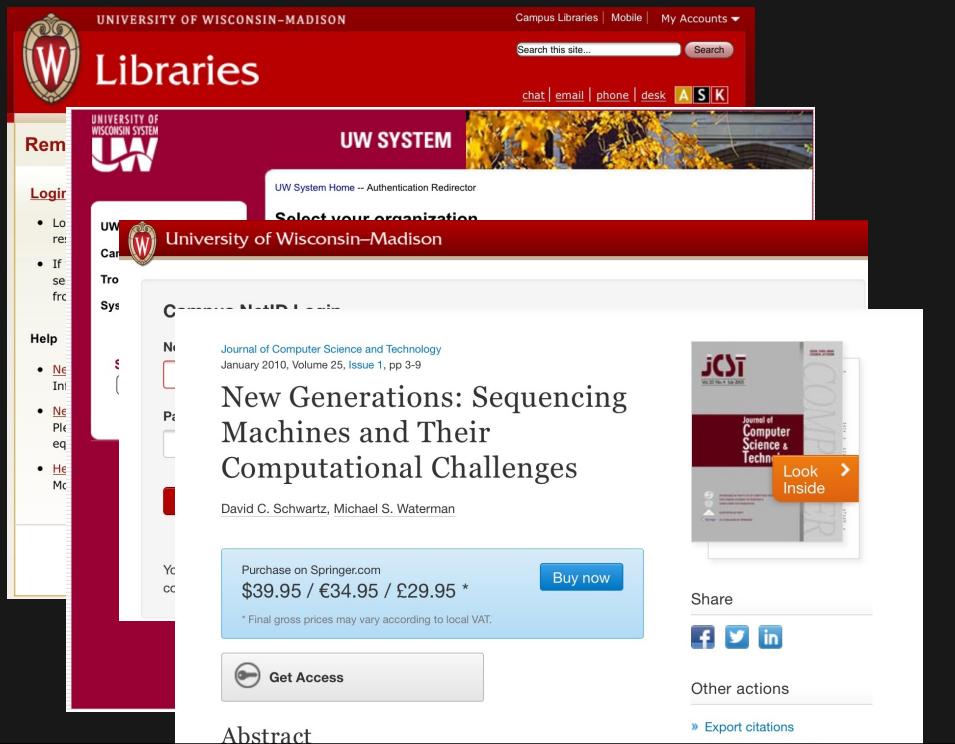
\$18 for a 2-page paper?

I understand that the publishing industry has a long history of screwy pricing, but you'd have to be either desperate or stupid to pay this.

And for that 1999 Genetic Epidemiology article, published by Wiley, you have to register in order to find out that it's \$35 for just 24 hours of access.

Access in action

journal.com.ezproxy.library.wisc.edu/blah



The screenshot shows a web browser displaying a login page for the University of Wisconsin-Madison Libraries. The URL in the address bar is `journal.com.ezproxy.library.wisc.edu/blah`. The page has a red header with the UW logo and "Libraries". A central modal window shows an article from the "Journal of Computer Science and Technology" titled "New Generations: Sequencing Machines and Their Computational Challenges" by David C. Schwartz and Michael S. Waterman. The modal includes a "Buy now" button and a "Get Access" button. To the left of the modal, there's a sidebar with links like "Remainder", "Log in", "Help", and "Contact Us". On the right, there's a sidebar with library links like "Campus Libraries", "Mobile", "My Accounts", "Search this site...", "chat", "email", "phone", "desk", and "ASK". The bottom right corner of the slide has the number "5".

One useful trick that I've learned (for folks at UW–Madison): If you paste `ezproxy.library.wisc.edu` into the URL for an article, then after entering your password, you can sometimes get access to the article.

But it didn't work in this case.

Access in action

Library catalog

The screenshot shows a library catalog interface for the University of Wisconsin-Madison. The main title "Libraries" is displayed prominently. On the left, there's a sidebar with various links like "Catalog", "Search", "Journals", "Books", and "Explore". The main content area is titled "Find It Publication Information" for the "Journal of Computer Science and Technology". Key details shown include:

- Publication title:** Journal of Computer Science and Technology
- Coverage (any format):** Jan 1997 (Vol. 12, no. 1) - present (delayed 1 year)
- Show format availability:** Full text available (1000-9000)
- ISSN:** 1063-423X
- Language:** English
- Subjects:** Computers--Computer Architecture ; Computers

Below this, there's a search bar for "Search within this publication" and a link to "Advanced Search". At the bottom, there's a section for "Browse specific issues" with a list of years from 2010 to 2013.

6

So I go back to the library catalog, search for the journal, get to the journal site again, find the paper, and...

Access in action

Finally.

Schwartz DC, Waterman MS. New generations: Sequencing machines and their computational challenges. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 25(1): 3-9 Jan. 2010

New Generations: Sequencing Machines and Their Computational Challenges

David C. Schwartz¹ and Michael S. Waterman^{2,3}

¹Laboratory for Molecular and Computational Genomics, Department of Chemistry and Laboratory of Genetics
University of Wisconsin-Madison, WI 53706, U.S.A.

²Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089-2910, U.S.A.

³Department of Automation, Tsinghua University, Beijing 100084, China

E-mail: dcschwartz@faststaff.wisc.edu; msw@usc.edu

Received September 5, 2009; revised November 24, 2009.

Abstract New generation sequencing systems are changing how molecular biology is practiced. The widely promoted \$1000 genome will be a reality with attendant changes for healthcare, including personalized medicine. More broadly the genomes of many new organisms with large samplings from populations will be commonplace. What is less appreciated is the explosive demands on computation, both for CPU cycles and storage as well as the need for new computational methods. In this article we will survey some of these developments and demands.

Keywords genome sequencing, new generation sequencing, read mapping, optical mapping, sequence assembly, Eulerian graphs

1 Introduction

It may be somewhat futile to attempt to track perfectly an explosion. But here we hope to give some hints about the technological and computational challenges that will surely be addressed along the path to the commoditization of sequence information. As the cost of sequence information drops, its utility will grow exponentially across areas of science, medicine, and safety of our food supply. And, of course, more unfathomable applications who would have predicted 50 years ago that lasers would find broad application as "pointers"? Accordingly, we expect that the experimental and computational challenges will become progressively intertwined in ways that may foster development of completely new directions for science. In this regard, challenges that are now unthinkable. In this regard, we present here a brief overview of the current state of DNA sequencing, and our best guesses for how technology and computation may interact for creating this future.

2 Current Technology

Although commercial next generation platforms differ from each other in how sequence is actually obtained, they share the common advantage of not

requiring bacterial clone libraries. In many ways, the obviation of clone library construction and handling is a major reason why genomic sequencing costs have plummeted while genome throughput has been increasing. Templates for large scale DNA sequencing are made from a library spread across massive culture plates and individual clones are isolated by "picking robots" for downstream sequencing reactions. Such operations, for large genomes such as human, require factory floor settings bristling with robots and technicians performing tasks that are now largely automated. Next generation platforms construct "dumb" libraries directly from individual genomic DNA molecules, which are amplified by emulsion or bridge PCR (polymerase chain reaction). Entire genome libraries consist of small vesicles, or surfaces laden with amplicons, but there is one company^[1] whose libraries comprise unamplified genomic templates that are bound to surfaces.

2.1 Next-Generation Sequencing

Today, an investigator can choose between four commercially available systems, each offering a panoply of technical strengths and weaknesses that need to be considered against overall cost and application: 1) Illumina's Genome Analyzer, 2) Life Technologies' SOLiD

I finally have a PDF of the paper.

Access in action

There's also PubMed

The screenshot shows a web page from the NIH Public Access website. At the top, there's a banner for "NIH Public Access Author Manuscript" with the text "Accepted for publication in a peer reviewed journal". Below the banner, the URL "Journal List > NIHPA Author Manuscripts > PMC3222932" is visible. The main content area displays a manuscript titled "New Generations: Sequencing Machines and Their Computational Challenges" by David C. Schwartz and Michael S. Waterman. The abstract discusses the impact of new sequencing technologies on healthcare. The page includes sections for "Keywords", "1 Introduction", and "Abstract". On the left side, there's a sidebar with links like "Using PubMed", "PubMed Q", "Full Text A", "PubMed F", "PubMed Tu", and "New and Notable". The right side has a vertical sidebar labeled "NIHPA Manuscripts". The overall layout is clean and professional.

8

If I'd used PubMed rather than Google Scholar, I could have gotten to the published paper in just a few clicks, because the manuscript is in PubMed Central.

PubMed Central is only for federally-funded research, has a one year embargo, and (as here) might not include the published version of the paper.

PubMed Central is a good thing, but one generally can't wait a year, it's unfortunate that the published versions aren't always included, and from an author's point of view it can be a real hassle.

Another example

The screenshot shows a search result for a paper titled "Clustering Tooth Surface Biologically Informative" published in J DENT RES January 2013. The result is from the "Journal of Dental Research" section of PubMed Central. The page includes the journal's logo, the abstract, and several links for access options like "IADR Member Sign In" and "Purchase Short-Term Access".

RESEARCH REPORTS
Clinical
Institution: University of Wisconsin-Madison

J.D. Shaffer^{1,2}, X. Wang^{2,4,5}, R.J. Weyant², D.W. McNeill³

Impact Factor: 3.826 | Ranking: 2/82 in Dentistry, Oral Surgery & Medicine | Eigenfactor® Score: 0.02325 | 5-Year Impact Factor: 4.286

This item requires a subscription to Journal of Dental Research.

ABSTRACT
Dental caries affect uniformly across all tooth surfaces. Therefore, we used cluster analysis to group surface types that differ in posterior non-pit and fissure caries. We replicated a number of carious teeth (i.e., under genetic or environmental influences) through hierarchical clustering to represent the number of affected teeth. In the second Survey, the second

To view this item, select one of the options

► IADR Member Sign In
IADR Members, please sign in at IADR to access this item.

► Sign In
Already an individual subscriber? If so, please sign in to SAGE Journals with your User Name and Password.

User Name Sign In
Password

Remember my user name & password.
Forgot your user name or password?

► Purchase Short-Term Access
Pay per Article - You may purchase this article for your purchase, which is yours to keep, within 24 hours.
Regain Access - You can regain access to a non-expired article if your access period has not yet expired.

DOI: 10.1177/002203451246960 [Publ Date: 2013-01-01]

Received
Accepted

A supplement is available online only.

► OpenAthens Users
Sign in via OpenAthens : If your organization uses OpenAthens, you can log in using your OpenAthens username and password. Contact your library for more details.

► Login via Your Institution
Login via your institution : You may be able to gain access using your login credentials for your institution. Contact your library if you do not have a username and password.

► International and American Associations for Dental Research
Become a Member - Join the International and American Associations for Dental Research.
Click here to subscribe to the print and/or online journal.
Click here to recommend to your library.

RESEARCH REPORTS
Institution: University of Wisconsin-Madison

JDR JOURNAL OF DENTAL RESEARCH®

Home OnlineFirst All Issues Subscribe RSS Email Alerts

Impact Factor: 3.826 | Ranking: 2/82 in Dentistry, Oral Surgery & Medicine | Eigenfactor® Score: 0.02325 | 5-Year Impact Factor: 4.286

< Return to Search Results | Edit My Last Search | Search this journal

This item requires a subscription to Journal of Dental Research.

Current Issue
J Dent Res 92(1): October 2013, 92 (10)

► Alert me to new issues of Journal of Dental Research

Submit a Manuscript
Free Sample Copy
Email Alerts
RSS feed
More about this journal
Other IADR Journals
Advances in Dental Research
Critical Reviews in Oral Biology & Medicine
International and American Associations for Dental Research
Interested in article-level

32

9

As another example, I was interested in a paper from the Journal of Dental Research.

It's less than a year old, so it's not available in PubMed Central.

I ordered a copy by inter-library loan, but it didn't include the supplemental methods, and those are behind a paywall at the journal!

Twitter is useful (for venting and more)



RESEARCH REPORTS

Clinical

J.R. Shaffer^{1,2*}, E. Feingold^{1,3}, X. Wang^{2,4,5},
D.E. Weeks^{1,4}, R.J. Weyant^{2,6}, R. Crout^{2,7},
D.W. McNeil⁸, and M.L. Marazita^{1,2,4,5,9}

¹Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA; ²Center for Oral Health Research in Appalachia, University of Pittsburgh, Pittsburgh, PA 15261, USA, and West Virginia University, Morgantown, WV 26506, USA; ³Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA; ⁴Center for Craniofacial and Dental Genetics, School of Dental Medicine, University of Pittsburgh, Pittsburgh, PA 15219, USA; ⁵Department of Oral Biology, School of Dental Medicine, University of Pittsburgh, Pittsburgh, PA 15261, USA; ⁶Department of Dental Public Health and Information Management, University of Pittsburgh, School of Dental Medicine, Pittsburgh, PA 15261, USA; ⁷Department of Periodontics, West Virginia University School of Dentistry, Morgantown, WV 26506, USA; ⁸Dental Practice and Rural Health, West Virginia University School of Dentistry, Morgantown, WV 26506, USA; and ⁹Clinical and Translational Institute, and Department of Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA; *corresponding author, jm51@pitt.edu

J Dent Res DOI: 10.1177/0022034512463241

Clustering Tooth Surfaces into Biologically Informative Caries Outcomes

Fig. 1. Similarly, to assess any effects of the inclusion of biological relatives in our sample, we repeated hierarchical clustering in the maximal subset of unrelated individuals (see Appendix Fig. 2), which were nearly identical to the total COHRA and NHANES 1999–2000 samples.

In both the total COHRA sample and in the NHANES 1999–2000 sample, the fifth cluster was subdivided into maxillary and mandibular components (indicating possibly 6 instead of 5 clusters). However, these subdivisions were not consistently observed via cross-validation of the COHRA sample. That is, in some randomly chosen halves of the COHRA sample, the sixth cluster did not distinguish maxillary and mandibular components of the fifth cluster. Therefore, we have conservatively shown results for the 5 stable clusters, as well as the maxillary and mandibular subdivisions of the fifth cluster.

Overwhelmingly, the cluster results were stable within the COHRA sample and inconsistent between the COHRA sample and NHANES 1999–2000 sample. That being stated, there were subtle differences observed among cluster results. For example, left-right asymmetry was observed, albeit rarely, in the random halves (e.g., Appendix Fig. 1A, tooth #20 and tooth #29). Likewise, some tooth surfaces physically positioned on the border between two adjacent clusters shifted membership (e.g., Appendix Figs. 1A and 2, tooth #21 and tooth #28). In many cases, the result of such shifts in cluster membership echoed the subtle differences between COHRA and NHANES 1999–2000. We speculate that the all-or-nothing nature of our clustering approach represents an oversimplification of the relationships among tooth surfaces with respect to dental caries risk factors

1

10

I was reduced to venting on twitter.

But then I got the appendix I wanted by email (twice!), within an hour of my tweet. (Thanks, MM and KW!)

It's all about money

(Costs in scientific publishing)

- ▶ Research
- ▶ Writing
- ▶ Peer review, editorial oversight
- ▶ Journal administration
- ▶ Copy editing, typesetting
- ▶ Distribution
- ▶ Profit

11

Open access is all about money.

Most of the costs behind a research paper are paid by grants or institutional funds. For most journals, peer review and editorial oversight are unpaid.

There are real costs associated with journals, but in the end they are all paid from the same sources (grants and institutional funds).

Do we really want to give away the product of our research and then buy it back repeatedly, at great profit to the publishers?

And shouldn't the literature be available generally and not just to those with access to well-funded research libraries?

It's not about

- ▶ Peer review
- ▶ Predatory publishing
- ▶ Impact factors
- ▶ Evaluating researchers
(for grants & promotions)

Well, it sort of is...

12

The Open Access discussion often gets tied up with discussion about peer review, predatory publishing, and journal impact factors.

But to me, it is a completely separate issue, whether we want stringent peer review before publication or instead leave the evaluation entirely to post-publication review.

On the other hand, the current culture is to evaluate researchers based on the perceived quality of the journals in which they've published. This makes it difficult to change to open access.

If everyone's still going to send their best work to Science, Nature, & Cell, then that work will continue to be locked up behind pay walls.

Paying for it

- ▶ Traditional approach
 - subscriptions
 - page charges
- ▶ Open access
 - bigger page charges
 - submission charges?
- ▶ Endowments
- ▶ Direct grants to journals

13

The usual way in which publishing costs are paid are through a combination of subscriptions (both institutional and individual) and direct charges to the author.

In the new open access model, the page charges are increased in order to eliminate the subscription fees. One might have a fee for all submitted manuscripts and not just those accepted for publication.

I've not seen much discussion of other alternatives, but I would prefer to see endowments established, particularly for society journals. Alternatively, journals might be funded directly through grants.

\$7000 page charges

| | |
|--|---------|
| Broman KW (2012) Genotype probabilities at intermediate generations in the construction of recombinant inbred lines. <i>Genetics</i> 190:403–412 | \$2,548 |
| Broman KW (2012) Haplotype probabilities in advanced inter-cross populations. <i>G3</i> 2:199–202 | \$1,650 |
| Broman KW, Kim S, Sen Š, Ané C, Payseur BA (2012) Mapping quantitative trait loci onto a phylogenetic tree. <i>Genetics</i> 192:267–279 | \$2,891 |

14

To illustrate the costs, here are the pages charges for three of my papers from 2012, for a combined \$7000.

Invoice

GENETICS

□ Review Invoice

| Article Information | |
|-----------------------------------|--|
| Publisher: | Genetics Society Of America |
| Title: | Genetics |
| Issue: | Volume 192, Number 1 |
| Manuscript Title: | Mapping Quantitative Trait Loci onto a Phylogenetic Tree |
| Manuscript Number: | I42448 |
| Article Type: | Regular Research Papers |
| Corr. Author Name (e-mail addr.): | Karl W Broman (kbroman@biostat.wisc.edu) |
| Membership Status: | Member |

Review Estimated Publication Charges

| Items | Unit Price | Quantity | Amount |
|---|------------|----------|-------------------|
| Page Charges | \$70.00 | 13 | \$910.00 |
| Figure Charges | \$40.00 | 6 | \$240.00 |
| Supplemental Files (six pages or greater) | \$500.00 | 1 | \$500.00 |
| Open Access Option | \$1,200.00 | 1 | \$1,200.00 |
| Author Alterations | \$2.55 | 16 | \$40.80 |
| Subtotal: | | | \$2,890.80 |

15

Here's the invoice for the most expensive of those three papers.

The charges would have been "just" \$1700, but I paid an additional \$1200 to have it freely available (otherwise it would have been behind a pay wall for one year).

Choices for young investigators

- ▶ Pay for open access
- ▶ Support young open access journals

OR

- ▶ Let subscribers pay & do more experiments
- ▶ Continue to go after Science, Nature, & Cell

16

The page charges, and the continued reliance on impact factors, lead to difficult choices, particularly for young investigators.

Should I pay for open access, or should I let the subscribers pay and use the savings to do more experiments?

Should I support open access journals, or should I continue to go after Science, Nature, & Cell?

The best scientists may confidently maintain their pure publication record.

But more mediocre scientists, who may be just scraping by, probably don't feel they have that luxury. A Nature paper can "make you."

What can we do?

- ▶ Send our best work to open access journals
- ▶ Support junior faculty to keep their papers open
- ▶ Pay attention to the quality of the work
(not the impact factor of the journal)
- ▶ Raise endowments for trusted journals
- ▶ Reform copyright law

17

We need to send our best work to open access journals.

We need to find ways to support our junior colleagues, so that they may do so as well.

We need to evaluate people based on their work and not by the name of the journal in which it appeared. We all may say, “Science and Nature are often crap and there are lots of fantastic papers that appear elsewhere.” But somehow when we see Nature or Cell on someone’s CV, we still have an immediate, positive reaction.

I would like to see endowed journals, open forever.

The quickest way to free the product of federally funded research would be to reform copyright law. If the product of our research were forced open by law, the publishing industry would figure out how to pay for it in short order.

But given the state of politics in the US, I’m not too optimistic about that.