Variable selection and parameter estimation for spatially varying coefficient models

Wesley Brooks

Department of Statistics
University of Wisconsin–Madison

somesquares.org

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 focus 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.

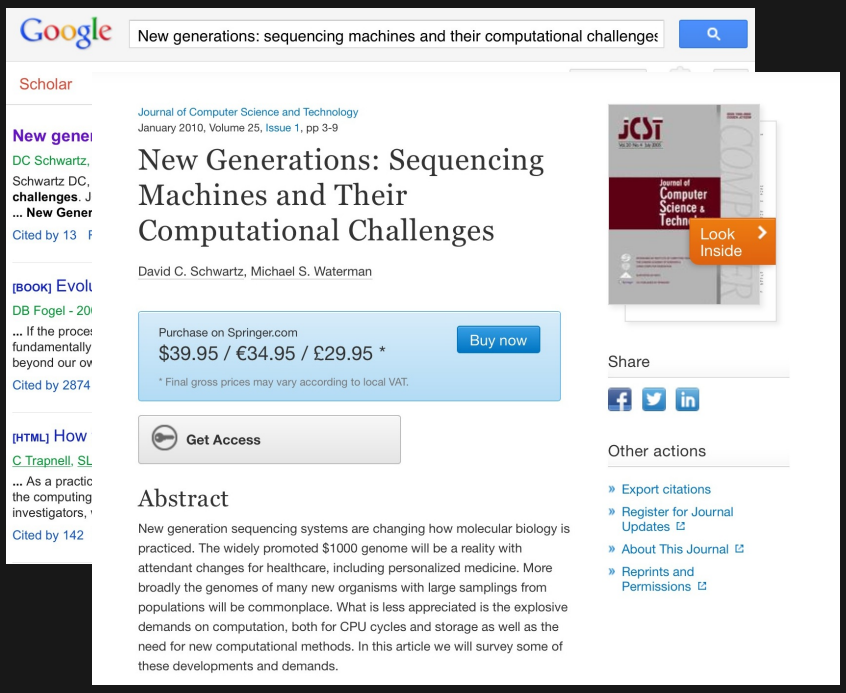
[13] Tor D.W., Matthew C.K., Steven C.L., and John J. Increased sensitivity in neuroimaging analyses using robust regression. *Neuroimage*, 26:99–113, 2005.

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



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

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

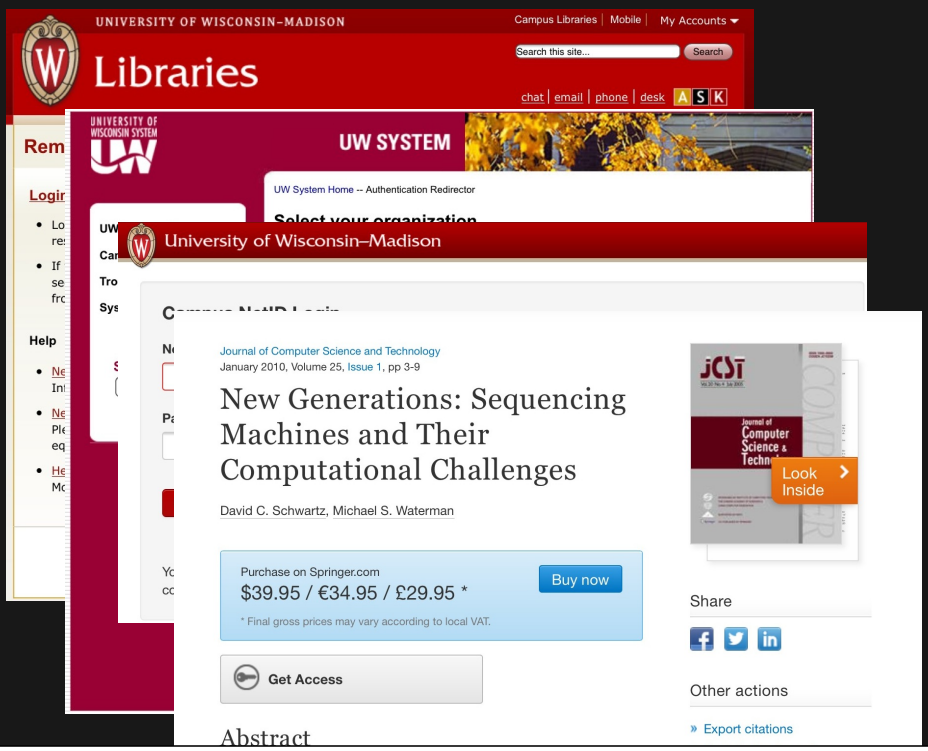
$18 for a 2-page paper?

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 24-access.

Access in action

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

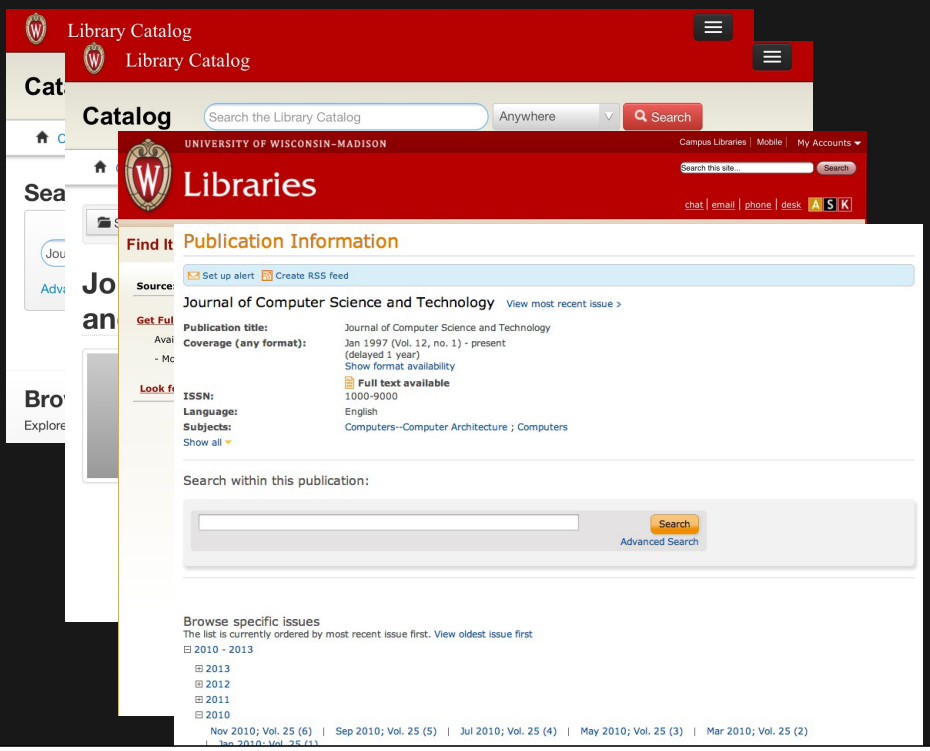


One useful trick that I've learned (for folks at UW–Madison): If you paste `ezproxy.library.wisc.edu` into the URL for an article, then after entering your password, you can sometimes get access to the article.

But it didn't work in this case.

Access in action

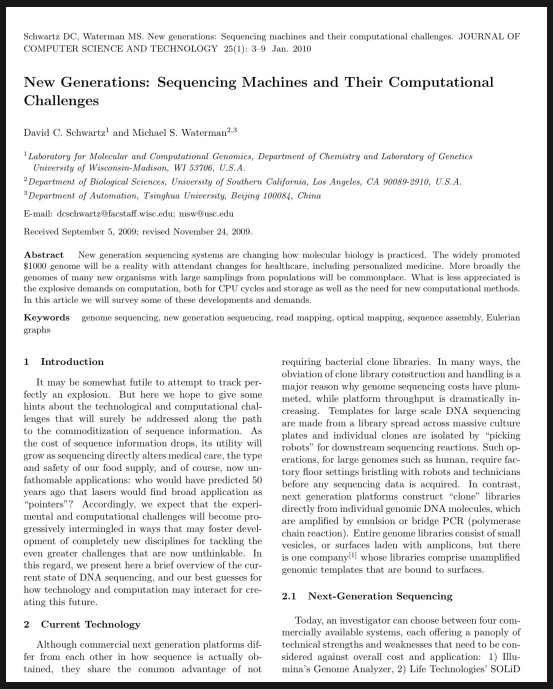
Library catalog



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. Schwartz[1] and Michael S. Waterman[2,3]

[1] *Laboratory for Molecular and Computational Genomics, Department of Chemistry and Laboratory of Genetics University of Wisconsin-Madison, WI 53706, U.S.A.*
[2] *Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089-2910, U.S.A.*
[3] *Department of Automation, Tsinghua University, Beijing 100084, China*

E-mail: dcschwartz@facstaff.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 as sequencing directly alters medical care, the type and safety of our food supply, and of course, now 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 intermingled in ways that may foster development of completely new disciplines for tackling the even greater 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 genome sequencing costs have plummeted, while platform throughput is dramatically 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 before any sequencing data is acquired. In contrast, next generation platforms construct "clone" 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.

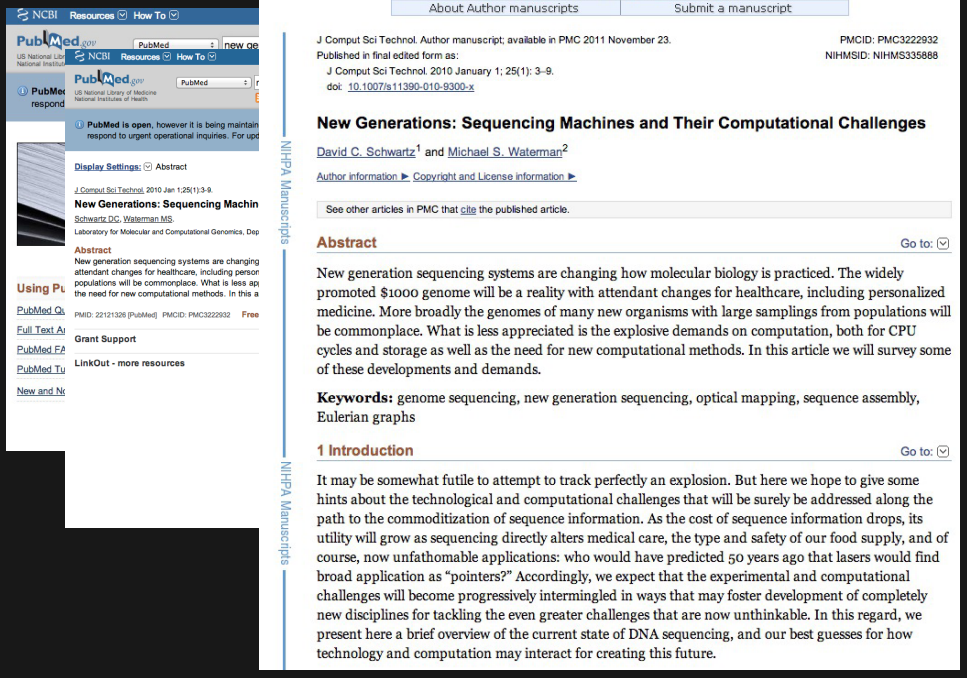
2.1 Next-Generation Sequencing

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



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.

It's all about money

(Costs in scientific publishing)

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

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…

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
  - charge submissions?
- Endowments
- Direct grants to journals

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. Genetics 190:403–412	$2,548
Broman KW (2012) Haplotype probabilities in advanced intercross populations. G3 2:199–202	$1,650
Broman KW, Kim S, Sen Ś, Ané C, Payseur BA (2012) Mapping quantitative trait loci onto a phylogenetic tree. Genetics 192:267–279	$2,891

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:	142448
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**

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

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

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.