Big data in small places

To the Editor:

Recently, several articles have focused on the need for flexible, scalable approaches to bioinformatics provision in smaller research institutes and university departments[1–3]. As a small institute of around 80 researchers, the Sainsbury Laboratory (Norwich, UK) has been working for the past four years to adapt to the influx of big data sets from high-throughput approaches. In that time, we have successfully transitioned from a 'top-down' model of bioinformatics provision to a 'bottom up' model that incorporates several features discussed in recent articles[1–3]. As a result, we have sped up the analysis cycle and can now handle increasing workloads in a timely, productive manner with a modest core support team. Here we provide a description of how we achieved this upgrade in the hope that our experience will prove useful for other small institutions seeking to address the informatics challenges posed by large-scale biological research approaches.

Dealing with big data sets can be abstracted into three main tasks: we must be able to manage, understand and analyze. 'Managing' is to carry out the computer science–based transfer and storage of data. 'Understanding' implies a clear knowledge of the biological context and caveats of the data as well as the functioning and limitations of the methods. And 'analyzing' refers to the application of the various bioinformatics methods to specific biological questions and data. Our support model distributes labor between bioinformaticians and bench scientists to optimize the delivery of these three tasks.

To ensure that data handling runs smoothly, bioinformaticians can help bench researchers by removing the burden of worrying about the mechanics of dealing with the data that their experiments produce. We have found that several simple tools and tricks reduce the perceived barrier to access and improve data management. Mounting storage devices directly to desktop machines by means of a secure shell (SSH) file system (http://en.wikipedia.org/wiki/SSHFS) makes it possible for terabytes of data to be accessed as if they were on a USB stick. Acting as 'lab manager' to the produced data by providing local rules about data descriptions and providing tools to make sure files will validate against these rules helps maintain order over the produced data. Tools like Galaxy[4] (our favored workflow-engineering environment) lower the barrier to access by allowing a user to create and share complex analysis pipelines through a straightforward graphical user interface (GUI). In parallel, bioinformaticians can develop tools for immediate deployment in a familiar and flexible framework.

For the majority of research projects, bioinformatics can be considered a subdiscipline of molecular biology and biologists must learn bioinformatics methods along side basic wet-lab methods. Given proper training and demystification of what bioinformatics methods actually are, biologists are perfectly capable of working their own informatics. A critical advantage of this model is that bioinformatics and biological concepts are now being thought of by the same brain, which significantly accelerates project turnover and reduces the likelihood of missed insights and misunderstandings.

In our experience, many biologists initially approach bioinformatics methods as a set of black box tricks in which the basic rules of rigorous experimentation somehow don't apply. Perhaps it is the mathematical comfort zone provided by E-scores and *P*-values that gives a false sense of absolute accuracy to the results of bioinformatics analyses, but it is puzzling how careful bench biologists turn into naive experimentalists once they sit at the computer. Bioinformaticians can have the strongest effect on proper use of bioinformatics practices by helping in the design and execution of experiments and controls.

A particularly relevant example stems from our experience with the detection of single-nucleotide polymorphisms (SNPs) from next-generation DNA sequence data. We have completed several projects cataloging genetic variation in microbes and plants and using next-generation sequencing alignment and SNP calling algorithms. None of the SNP identification programs give perfect results so the amount of error must be quantified. However, this serious limitation is not initially obvious to a biologist whose main focus is the end goal of generating lists of SNPs. In-depth explanations of the methods may not help as they can mire the discussion in statistical or technical details not fully appreciated by the biologist. We approach this problem by encouraging the use of controls to demonstrate and estimate the

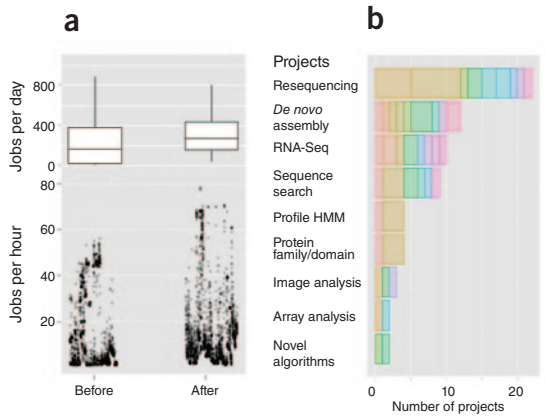

Figure 1 Bioinformatics jobs done and methods used at the Sainsbury Laboratory. (**a**) Our model increases productivity: we see a general rise in the bioinformatics jobs being run per hour and per day (*t*-test, $P = 0.027$), 'before' and 'after' we implemented our new systems for data management, our Galaxy instance and carried out training. (**b**) We have many biologists using many bioinformatics approaches; each color represents a user, each bar the cumulative number of projects for a method. Most biologists work on two to four informatics projects, applying two to three methods. HMM, hidden Markov model.

error rates, for example, by computationally introducing SNPs into a reference sequence and showing the extent to which recall of these SNPs is accurate (see Supplementary Fig. S2 and methods in ref. 5). It is our experience as informaticians, that such an exercise usually has a profound impact on our biology colleagues because they appreciate the value of controlled experimentation and informed criticism of data. It unequivocally demonstrates that bioinformatics methods have error. It frees experimentalists to see the approach as just another way of estimating something and to approach bioinformatics as a set of methods that can be dissected with the familiar knife of experimentation.

We have seen that the attitude of 'bioinformatics as assay' propagates rapidly within a research group. Once a concept is adopted in laboratory meetings and research discussions and the issues are explained by biologists to other biologists, we reach a virtuous cycle that is self-reinforcing in a laboratory. Our ultimate aim is that the biologists use bioinformatics in a mature and critically aware way.

A key to achieving this sea change is to sustain a productive working dialog between the two parties by giving the biologist the vocabulary needed to work in the field and discuss issues as a peer of the bioinformatician. At the Sainsbury Laboratory, we focus on training and in getting our bioinformaticians to discuss their tricks and toys in a relaxed yet formal fashion. We began by implementing a wide range of courses aimed at the novice but covering enough ground to introduce all the vital aspects of each topic. Specifically, we teach introductory courses on broad topics like *de novo* assembly, RNA-Seq and so forth. Advanced training in things like command-line use, scripting languages and statistics are always useful for a smaller number of biologists—research is unpredictable and often existing tools with a GUI will lag behind the cutting edge in a way that researchers don't want to. The ability to run a brand new tool on the command-line and parse its output with a small custom script is an excellent advantage for researchers who need the cutting edge right away.

After formal training sessions, follow-up is vital. Help and resources should be available on-demand and the trainer needs to operate an open-door policy for questions. Answers to questions and discussions on request will help to prevent the learner's enthusiasm from stalling early on.

Wider laboratory culture changes can be sustained and extended through a range of familiar exercises and resources. Journal club meetings specifically designed to tackle discrete bioinformatics topics help enormously to reinforce awareness of what is being done in the field and what the details of execution are. Laboratory meetings in which the biologist presents their bioinformatics work to an interested audience provide a vital opportunity to develop critical appraisal of informatics methods.

Our approach has borne fruit. We have found that when biologists are able to handle part or all of the bioinformatics load on their projects, our productivity increases (**Fig. 1**). The turnover of jobs run on our computer cluster increased substantially after opening it up to trained biologists. The number of concurrent bioinformatics projects we are now handling is high, too. In total, 25% of our researchers (20 individuals) are now actively involved in running their own bioinformatics projects, way above the 2.5% (two researchers) the old model permitted. The number of biologists, not the size of the core support team, limits the number of projects that we can handle. Extra analysis capacity can now be brought in at the project level when hiring new biologists and is not throttled by the size of the core support team; in addition, the expertise can scale as the number of projects requiring bioinformatics methods being carried out increases.

Computing power can be a limitation to bioinformatics, but this problem is not as acute in smaller institutions as it is in larger sequencing centers. A massive infrastructure investment is not necessary and it is possible to provide expandable computing infrastructure. At the Sainsbury Laboratory, when the small core team was responsible for the majority of work with our hardware, there was often a lot of spare processing capacity and analyses did not run flat-out. With job-scheduling software, however, modest computing clusters can be made to support the activities of many researchers by distributing resources evenly through time. It is possible to acquire for moderate costs a few high-powered servers and a storage device that can be built into a cluster easily. Well-designed clusters can be expanded by adding new servers and extra disks to storage appliances whenever projects require it.

It may seem that the scheme we have developed benefits experimental biologists at the expense of bioinformaticians. Put another way, our scheme moves bioinformaticians' work away from data analysis to training and systems administration—a role that may not suit those with a keen interest in research. In fact, we have found that the main advantage of implementing the bottom-up model is that it pays back in saved time for the bioinformaticians, thereby creating a new set of opportunities. Bioinformaticians working within this model will free up time to follow their own projects, such as research into new methods. With a newly qualified, captive beta-testing audience and more time to exploit the growing data, it becomes possible to simultaneously push forward the institution's research projects as well as the bioinformaticians' reputation.

The bottom-up model of bioinformatics provision is flexible and scalable. As well as paying off in the current environment, such flexibility also helps to pave the way for future changes. Coming generations of biology students, taught in a data-rich environment, will be more knowledgeable about the techniques for handling big data and will be primed to begin such projects without needing much central assistance. As biology becomes a more data-rich science, it will attract more students and researchers trained in computer science.

Institutes and laboratories must provide an environment that supports a wide range of skill levels in biology and computer science to take full advantage of their main asset—people. Our approach grants this flexibility and means that we are not reliant on any single entity in our organization to make progress. We have quicker paths to insight because the same brains that think about the biology are processing bioinformatics concepts. Once a critical and informed approach to bioinformatics takes hold, it rapidly spreads from member to member in a virtuous cycle and service provision becomes creating environments for insights, an ecosystem of colleagues with deep experience of methods in a sustaining self-reproducing community of peers.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Daniel MacLean & Sophien Kamoun

The Sainsbury Laboratory, Norwich Research Park, Norwich, UK.
e-mail: dan.maclean@tsl.ac.uk

1. Lewitter, F. & Rebhan, M. *PLoS Comput Biol.* **5**, e1000368 (2009).
2. Multiple authors. Special issue: Education in Bioinformatics. *Brief. Bioinformatics* **11**(6) (2010).
3. Kallioniemi, O., Wessels, L. & Valencia, A. *Bioinformatics* **27**, 1345 (2011)
4. Goecks, J., Nekrutenko, A. & Taylor, J. *Genome Biol.* **11**, R86 (2011).
5. Raffaele, S. *et al. Science* **330**, 1540–1543 (2010).