BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 9 2012, pages 1272–1273
doi:10.1093/bioinformatics/bts128

Genome analysis

Advance Access publication March 13, 2012

SEQanswers: an open access community for collaboratively decoding genomes

Jing-Woei Li[1], Robert Schmieder[2], R. Matthew Ward[3], Joann Delenick[4], Eric C. Olivares[5,*] and David Mittelman[3,6,*]

[1]School of Life Sciences, The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, [2]Computational Science Research Center and Department of Computer Science, San Diego University, San Diego, CA 92182, [3]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA 24061, [4]Department of Biology, Graduate School of Arts & Sciences, Yale University, New Haven, CT 06520, [5]SEQanswers.com, Union City, CA 94587 and [6]Department of Biological Sciences, Virginia Tech, Blacksburg, VA 24061, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: The affordability of high-throughput sequencing has created an unprecedented surge in the use of genomic data in basic, translational and clinical research. The rapid evolution of sequencing technology, coupled with its broad adoption across biology and medicine, necessitates fast, collaborative interdisciplinary discussion. SEQanswers provides a real-time knowledge-sharing resource to address this need, covering experimental and computational aspects of sequencing and sequence analysis. Developers of popular analysis tools are among the >4000 active members, and ~40 peer-reviewed publications have referenced SEQanswers.

Availability: The SEQanswers community is freely accessible at http://SEQanswers.com/

Contact: david.mittelman@vt.edu; ecolivares@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 24, 2012; revised on February 24, 2012; accepted on March 02, 2012

1 INTRODUCTION

The Human Genome Project represents one of the greatest concerted achievements of the life sciences. This massive global effort jump-started the genomics era and enabled more ambitious and collaborative projects such as the Cancer Genome Atlas (Cancer Genome Atlas Research Network, 2008), 1000 Genome Project (1000 Genomes Project Consortium, 2010) and Human Microbiome Project (The NIH HMP Working Group *et al*., 2009). These large population-scale studies, powered by high-throughput sequencing (HTS) technologies, have generated massive amounts of genomic data with the potential to revolutionize genetics and medicine.

The translation of these data to actionable medicine, however, is complicated by the challenges of extracting meaningful information from HTS data (Mardis, 2010). The challenge is not purely computational, as bioinformatics is bound by the experimental methods employed to produce genomic data (Alkan *et al*., 2011). A successful experiment minimizes false positives and depends on the optimization of an entire pipeline, from sample preparation to computational analysis.

As HTS begins to transform nearly all aspects of biological and medical science, more labs will incorporate the production and analysis of genomic data into their studies. However, these experimental and computational methods are evolving at an incredible pace and it is increasingly challenging for smaller research groups outside of major genome centers to stay current. Real-time, interdisciplinary collaboration helps large genome centers optimize analysis pipelines and methods, and allows smaller groups to exploit them, even if they did not have resources to facilitate the initial development.

2 THE SEQANSWERS COMMUNITY

SEQanswers was launched in 2007 as an open forum to enable scientists across disciplines to collaboratively advance genomics and, particularly, HTS technologies. To date, there are >4000 active users visiting the online community each month. There is a rapidly growing number of discussion threads (currently >10 000) that span topics from sequencing platforms, experimental design, data analysis and biological interpretation (Fig. 1A). The SEQanswers community is truly global (Supplementary Fig. S1) and includes members from major genome centers and individual groups, as well as key developers of popular data analysis tools and methods. The community currently hosts >300 new questions, and 1800 new responses per month. This incredibly high rate of participation has led to rapid responses to questions, shortening initial response time from a week in early 2008 to less than a day in 2011 (Fig. 1B). Collaborative and transparent discussion on SEQanswers has triggered the development of new experimental techniques, data analysis methods and pipelines, as well as collaborative assessment of analysis standards (Supplementary Table S1). This innovation is captured in part by >30 peer-reviewed publications that cite SEQanswers so far (http://seqanswers.com/wiki/Papers_Referencing_SEQanswers).

SEQanswers is not the only online resource for knowledge sharing and collaboration: major sequencing technology companies have platform-centric user communities, but these are often restricted to customers and exclude the greater scientific community. In contrast, BioStar (Parnell *et al*., 2011), an open, community-driven

*To whom correspondence should be addressed.

bioinformatics resource, currently hosts >2300 threads, which far exceeds the sum of discussions found on communities operated by sequencing companies (Supplementary Table S2). BioStar's principle feature is to enable researchers to ask questions and obtain brief answers, ranked by community vote, to bioinformatics-related problems. BioStar's success can be attributed to its well-defined scope and focus on a simple question and answer format for bioinformatics. However, this format precludes other forms of collaboration, discussion and debate.

SEQanswers differs from BioStar both in format and scope. In an almost complementary capacity, SEQanswers eschews the Q&A format in favor of a more traditional forum format to facilitate collective discussion of technologies, methods and standards of practice. The traditional forum format emphasizes the chronology and evolution of collective thought, rather than focusing on identifying a single, best answer. The scope of SEQanswers differs from BioStar's exclusive bioinformatics focus, including all aspects of genomics, experimental and computational. Finally, in recognition of the sometimes lengthy and tediously detailed threads that can emerge from sequential discussion, we have developed a manually curated database, SEQwiki (Li *et al*., 2012), that consists of frequently asked questions, analysis methods, tutorials and sequencing service providers.

3 CONCLUSION

The massive amounts of data and rapid pace of genome technology development necessitates innovations in scientific communication. The current standard for scientific communication between disparate research groups focuses on peer-reviewed research published in traditional scientific journals. These journals have evolved for the Internet age, especially with the new emphasis on open access and fast publishing from both new, exclusively open access journals to traditional journals that have created new outlets for open access publication. While scientific journals will continue to have important roles as curators of research and referees for the peer-review process, there is an opportunity for open, internet-based platforms to supplement traditional journals by enabling the rapid exchange of results, techniques and data, the latter two being crucial for advancing research,[1] but notoriously difficult to access. SEQanswers was designed to address this need for genomics. The community has since developed into a thriving community that offers a wealth of information, including discussions that have facilitated the construction of analysis pipelines and consensus on standards in the genomics community.

ACKNOWLEDGEMENT

The authors would like to thank the members of the SEQanswers community for helpful suggestions on the manuscript.

Funding: This work was supported by an award through the NVIDIA Foundation's 'Compute the Cure' program to D.M.

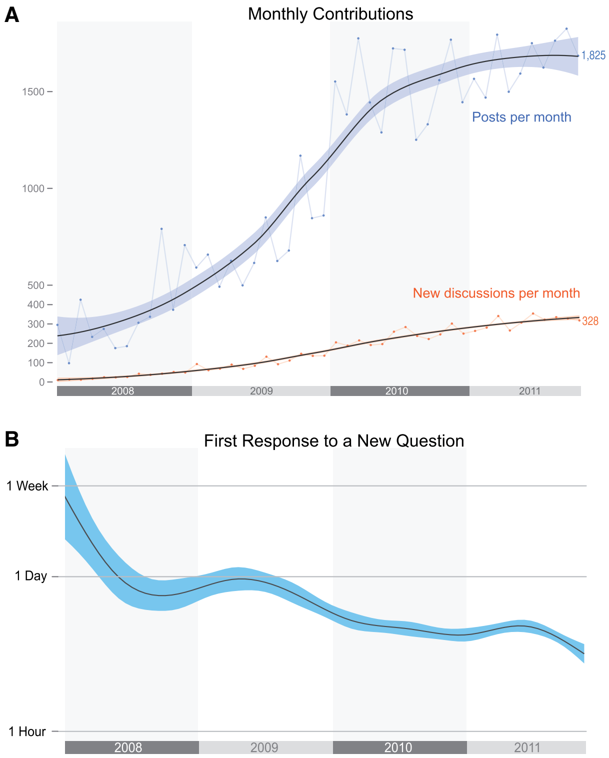

Fig. 1. SEQanswers is an active and fast growing community. (**A**) Monthly contributions to SEQanswers measured by the number of new posts (blue points/line) and discussions (orange points/line). Discussion counts include threads with at least two posts and exclude those with no answers. Also excluded are automated publication announcements. (**B**) The average response time to a new forum thread.

Conflict of Interest: none declared.

REFERENCES

1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Alkan,C. *et al*. (2011) Limitations of next-generation genome sequence assembly. *Nat. Methods*, **8**, 61–65.

Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Elsevier. (2010) *Access vs. Importance – A Global Study Assessing the Importance of and Ease of Access to Professional and Academic Information – Phase I Results*, Publishing Research Consortium.

Li,J.W. *et al*. (2012) The SEQanswers wiki: a wiki database of tools for high-throughput sequencing analysis. *Nucleic Acids Res*, **40**, D1313–D1317.

Mardis,E.R. (2010) The $1,000 genome, the $100,000 analysis? *Genome Med*, **2**, 84.

Parnell,L.D. *et al*. (2011) BioStar: an Online Question & Answer Resource for the Bioinformatics Community. *PLoS. Comput. Biol.*, **7**, e1002216.

The NIH HMP Working Group *et al*. (2009) The NIH Human Microbiome Project. *Genome Res*., **19**, 2317–2323.

[1] In a global survey, 62% of respondents claimed data accessibility is very important among all information to their research (Elsevier, 2010).