14th Annual Bioinformatics Open Source Conference

BOSC 2013

Berlin, Germany
July 19-20, 2013

http://www.open-bio.org/wiki/BOSC_2013

Welcome to BOSC 2013! The Bioinformatics Open Source Conference, established in 2000, is held every year as a Special Interest Group (SIG) meeting in conjunction with the Intelligent Systems for Molecular Biology (ISMB) Conference.

BOSC is sponsored by the Open Bioinformatics Foundation (O|B|F), a non-profit group dedicated to promoting the practice and philosophy of Open Source software development within the biological research community.

This year's keynote speakers are Sean Eddy and Cameron Neylon. Sean Eddy, a group leader at the Howard Hughes Medical Institute's Janelia Farm, is the author of several well-known open source computational tools for sequence analysis including the HMMER and Infernal software suites, as well as a coauthor of the Pfam database of protein domains. Cameron Neylon is Advocacy Director for the Public Library of Science, a research biophysicist and well-known agitator for opening up the process of research. He speaks regularly on issues of Open Science including Open Access publication, Open Data, and Open Source.

Session topics this year include Software Interoperability, Visualization, Cloud and Genome-Scale Computing, and several other topics that previous BOSC attendees will recognize. New this year is a session on Open Science and Reproducible Research, which will end with a short talk entitled “Ten Simple Rules for the Open Development of Scientific Software” that will be followed by time for discussion by the audience. Our panel discussion on Day 2, Strategies for Funding and Maintaining Open Source Software, will include experts on various funding approaches ranging from grant-based to commercial with value-added services.

There are three scheduled poster sessions. We have space for several last-minute posters in addition to those listed in the program.

Thanks in part to generous support from Eagle Genomics, we were able to award Student Fellowships to the authors of the three best student abstracts. Congratulations to the student winners, all of whom received free admission to BOSC and $250 towards their travel expenses: Markus List, Joeri van der Velde, and Yuriy Vaskin.

BOSC is a community effort—we thank all those who made it possible, including the organizing committee, the program committee, the session chairs, and the ISMB SIG chair, Steven Leard. If you are interested in helping to organize BOSC 2014, please email bosc@open-bio.org.

2013 Organizing Committee:

Nomi Harris (Chair), Jan Aerts, Brad Chapman, Peter Cock, Christopher Fields, Jeremy Goecks, Hans-Rudolf Hotz, Hilmar Lapp

2013 Program Committee:

Heikki Lehväslaiho, Hans-Rudolf Hotz, Tiago Antao, Brad Chapman, Thomas Down, Peter Cock, Francesco Strozzi, Hilmar Lapp, Jeremy Goecks, Ben Temperton, Jan Aerts, Chris Fields, Shiran Pasternak, Kam Dahlquist, Kazuharu Arakawa, Scott Markel, Michael Reich, Timothy Booth, Sophia Cheng, Heiko Dietze, Hervé Ménager, Peter Robinson, Olivier Sallou, Raoul Bonnal, Imtiaz Khan, Ronald Taylor, Nomi Harris

BOSC 2013 Schedule

Day 1 (Friday, July 19, 2013)

Time	Title	Speaker or Session Chair
7:30-9:00	**Registration**	
9:00-9:15	**Introduction and Welcome**	Nomi Harris (Chair, BOSC 2013)
9:15-10:15	**Keynote: Network ready research--the role of open source and open thinking**	Cameron Neylon
10:15-10:45	***Coffee Break***	
10:45-12:30	**Session: Open Science**	Chair: Hilmar Lapp
10:45-11:00	Open Science Data Framework: A Cloud enabled system to store, access, and analyze scientific data	Anup Mahurkar
11:00-11:15	myExperiment Research Objects: Beyond Workflows and Packs	Stian Soiland-Reyes
11:15-11:30	Empowering Cancer Research Through Open Development	Juli Klemm
11:30-11:45	DNAdigest - a not-for-profit organisation to promote and enable open-access sharing of genomics data	Fiona Nielsen
11:45-11:50	Jug: Reproducible Research in Python	Luis Pedro Coelho
11:50-11:55	OpenLabFramework: A Next-Generation Open-Source Laboratory Information Management System for Efficient Sample Tracking	Markus List
12:00-12:30	Ten Simple Rules for the Open Development of Scientific Software [discussion]	Andreas Prlic
12:30-1:30	***Lunch***	
1:00-2:00	**Poster Session I**	
2:00-3:30	**Session: Visualization**	Chair: Jan Aerts
2:00-2:25	Refinery Platform - Integrating Visualization and Analysis of Large-Scale Biological Data	Nils Gehlenborg
2:25-2:40	MetaSee: An interactive visualization toolbox for metagenomic sample analysis and comparison	Kang Ning
2:40-2:55	DGE-Vis: Visualisation of RNA-seq data for Differential Gene Expression analysis	David Powell
2:55-3:10	Genomic Visualization Everywhere with Dalliance	Thomas Down
3:10-3:25	Robust quality control of Next Generation Sequencing alignment data	Konstantin Okonechnikov
3:25-3:30	Visualizing bacterial sequencing data with GenomeView	Thomas Abeel
3:30-4:00	***Coffee Break***	

Time	Title	Speaker or Session Chair
4:00-5:30	**Session: Bioinformatics Open Source Project Updates**	Chair: Hans-Rudolf Hotz
4:00-4:15	BioRuby project updates - power of modularity in the community-based open source development model	Toshiaki Katayama
4:15-4:30	Biopython project update	Peter Cock
4:30-4:45	InterMine - Collaborative Data Mining	Alex Kalderimis
4:45-5:00	GenoCAD 2.2 Grammar Editor	Jean Peccoud
5:00-5:15	Improvements and new features in the 7th major release of the Bio-Linux distro	Timothy Booth
5:15-5:20	**Announcements**	Nomi Harris
5:20-6:30	**Poster Session II**	
5:20-6:30	**BOFs**	
7:00	Pay-your-own-way BOSC dinner, Hendrik's (www.hendriks-berlin.de), Straße des 17. Juni 13, 10623 Berlin. RSVP at bit.ly/BOSC2013-dinner	

Day 2 (Saturday, July 20, 2013)

Time	Title	Speaker or Session Chair
8:45-8:50	**Announcements**	Nomi Harris
8:50-9:00	Codefest 2013 Report	Brad Chapman (Codefest 2013 Organizer)
9:00-9:15	Open Bioinformatics Foundation: A Community For, By, and Of You	Hilmar Lapp (President, O|B|F)
9:15-10:15	**Keynote: Biological sequence analysis in the post-data era**	Sean Eddy
10:15-10:45	***Coffee Break***	
10:45-12:30	**Session: Software Interoperability**	Chair: Jeremy Goecks
10:45-11:10	BioBlend - Enabling Pipeline Dreams	Enis Afgan
11:10-11:35	Taverna Components: Semantically annotated and shareable units of functionality	Alan Williams
11:35-11:50	UGENE Workflow Designer – flexible control and extension of pipelines with scripts	Yuriy Vaskin
11:50-12:05	Oqtans: A Multifunctional Workbench for RNA-seq Data Analysis	Gunnar Rätsch
12:05-12:10	PhyloCommons: community storage, annotation and reuse of phylogenies	Hilmar Lapp
12:10-12:15	GEMBASSY: an EMBOSS associated package for genome analysis using G-language SOAP/REST web services	Kazuharu Arakawa

Time	Title	Speaker or Session Chair
12:15-12:30	Rubra - flexible distributed pipelines for bioinformatics	Clare Sloggett
12:30-1:30	***Lunch***	
12:30-1:30	**Poster Session III**	
1:30-3:30	**Session: Cloud and Genome-Scale Computing**	Chair: Peter Cock
1:30-1:45	Towards Enabling Big Data and Federated Computing in the Cloud	Enis Afgan
1:45-2:00	MyGene.info: Making Elastic and Extensible Gene-centric Web Services	Chunlei Wu
2:00-2:15	An update on the Seal Hadoop-based sequence processing toolbox	Luca Pireddu
2:15-2:30	Open Source Configuration of Bioinformatics Infrastructure	John Chilton
2:30-2:55	An Open Source Framework for Gene Prioritization	Hoan Nguyen
2:55-3:10	RAMPART (aRobustAutomaticMultipleAssembleRToolkit)	Daniel Mapleson
3:10-3:30	OmicsConnect: flexible multi-omics data capture and integration tools for high-throughput biology	Joeri van der Velde
3:30-4:00	***Coffee Break***	
4:00-4:40	**Session: Translational Genomics**	Chair: Nomi Harris
4:00-4:25	Community development of human variant calling and validation pipelines	Brad Chapman
4:25-4:40	Understanding Cancer Genomes Using Galaxy	Jeremy Goecks
4:40-5:30	**Panel: Strategies for Funding and Maintaining Open Source Software**	*Moderator*: Brad Chapman *Panelists*: Peter Cock, Sean Eddy, Carole Goble, Richard Holland, Scott Markel, Jean Peccoud
5:30-5:40	Presentation of Student Travel Awards	Nomi Harris
5:40-6:40	**BOFs**	

Any last-minute schedule updates will be posted at
http://www.open-bio.org/wiki/BOSC_2013_Schedule

Keynote Speakers

Sean Eddy

Sean Eddy is a group leader at the Howard Hughes Medical Institute's Janelia Farm. He is interested in deciphering the evolutionary history of life by comparison of genomic DNA sequences. His expertise is in the development of computational algorithms and software tools for biological sequence analysis. He is the author of several computational tools for sequence analysis including the HMMER and Infernal software suites, as well as a coauthor of the Pfam database of protein domains. He serves as an advisor to several foundations and US science agencies, including the National Institutes of Health and the National Academy of Sciences, often on matters of large-scale computation and data analysis in biology.

Sean's talk is entitled *Biological sequence analysis in the post-data era*.

> Biological systems are almost unfathomably complex, yet their complexity is reproducibly specified by a small digital genome. We understand many basics of development and evolution but we lack a truly satisfying quantitative understanding of how biological complexity is specified and how it evolves. One important line of attack on the problem is to reconstruct the history of molecular evolution by comparative genome sequence analysis. Biological sequence comparison has a long intellectual history, but only recently, with the advent of inexpensive large scale DNA sequencing, have we gained comprehensive access to genome sequences from essentially all species. Though welcome, this influx of genome sequence data is exposing structural flaws in computational biology research tools. Because the research community values innovative science over infrastructure in any short-term decision, academic researchers have difficulty investing sufficient effort in robust software and datasets that may enable even more innovative science over the long term. Meanwhile, professional commercialization of the software and data infrastructure also continues to prove difficult, in part because open source code and data availability is a fundamental principle of scientific publication of reproducible, reusable results. I'll discuss what I see as some of the key tensions, challenges, and opportunities in these regards, in part in the context of our work at Janelia Farm on the HMMER and Infernal codebases, and our nascent work on the genomic specification of neural circuits in Drosophila.

Cameron Neylon

Cameron Neylon is Advocacy Director for the Public Library of Science, a research biophysicist and a well-known agitator for opening up the process of research. He speaks regularly on issues of Open Science including Open Access publication, Open Data, and Open Source as well as the wider technical and social issues of applying the opportunities the internet brings to the practice of science. He was named as a SPARC Innovator in July 2010 for work on the Panton Principles and is a recipient of the Blue Obelisk for contributions to open data. He writes regularly at his blog, Science in the Open.

Cameron will speak about *Network ready research: The role of open source and open thinking*.

> The highest principle of network architecture design is interoperability. If Metcalfe's Law tells us that a network's value can scale as some exponent of the number of connections then our job in building networks is to ensure that those connections are as numerous, as operational, and as easy to create as possible. Where we make it easy for anyone to wire in new connections we maximise the ability of others to contribute to the value of our shared networks.
>
> Bioinformatics has, from time to time, been derided as "slidedecks full of hairballs", yet those hairballs, and their ubiquity are emblematic of the fact that at its heart bioinformatics is a science of networks. Networks of physical interactions, of genetic control, of degree of similarity, or of ecological interactions amongst many others. Bioinformatics is also amongst the most networked of research communities and amongst the most open in the

sharing of research papers, of research data, tools, and even research in process in online conversations and writing.

Lifting our gaze from the networks we work on to the networks we occupy is a challenge. Our human networks are messy and contingent and our machine networks clogged with things we can't use, even if we could access them. What principles can we apply so as to build our research into networks that make the most of the network infrastructure we have around us. Where are the pitfalls, and where are the opportunities? What will it take to configure our work so as to enable "network ready research"?

O|B|F Membership

Professionals, scientists, students, and others active in the Open Source Software arena in the life sciences are invited to join the Open Bioinformatics Foundation (the O|B|F). The O|B|F grew out of the volunteer projects BioPerl, BioJava and Biopython and was formally incorporated in 2001 in order to handle modest requirements of hardware ownership, domain name management and funding for conferences and workshops. In 2005, we enacted bylaws for the first time, and along with it created a formal membership.

In 2012, we decided to give up our own incorporation to associate ourselves with Software In The Public Interest, Inc., a fiscal sponsorship organization that we felt aligned well with our own values and culture. The bylaws underwent a series of changes, in part to better reflect our current practices, and in part to pave the way for joining SPI. The changes were approved on Sep 11, 2012, our membership overwhelmingly approved of associating with SPI, and as of October 12, 2012, O|B|F is an SPI-associated project.

This program includes a form you can fill out to join the O|B|F. (Yes, we realize that a paper form is kind of retro, but at the moment, this is the way we meet the requirements for documenting our membership.) Also, f you are interested in meeting and talking to some of the O|B|F Directors and members, please join us at the BOSC dinner (see below).

Talk and Poster Abstracts

Talk abstracts are included in this program in the order in which they will be presented at the conference. Some, but not all, of the talks will also be presented as posters. There are also a few spaces available for last-minute posters. If you would like to present one, please email your abstract (which must meet the BOSC criteria of available source and recognized open source license) to bosc@open-bio.org.

Authors should put up their posters in their assigned poster spot before the first poster session (which starts at 12:30 on the first day). After that time, any unused poster slots will be made available for last-minute posters. The ISMB staff specify that posters should not exceed the following dimensions: 0.95 m wide x 1.30 m high.

Optional BOSC Dinner

We invite you to join BOSC organizers and attendees at a pay-your-own-way dinner the first evening of BOSC (Friday, July 19, at 7pm) at Hendrik's (www.hendriks-berlin.de), Straße des 17. Juni 131, 10623 Berlin. Take the S-Bahn S5 or S7 line from the Westkreuz station to Tiergarten and walk back 100m (across Straße des 17. Juni). The restaurant is under the railway arches.
If you want to join us for dinner, RSVP at http://bit.ly/BOSC2013-dinner before Friday at noon. The restaurant has space for 30 BOSC guests; only those who RSVP will be admitted.

Talks and Posters

Title	Author	Poster #
Open Science Data Framework: A Cloud enabled system to store, access, and analyze scientific data	Anup Mahurkar	
myExperiment Research Objects: Beyond Workflows and Packs	Stian Soiland-Reyes	1
Empowering Cancer Research Through Open Development	Juli Klemm	
DNAdigest - a not-for-profit organisation to promote and enable open-access sharing of genomics data	Fiona Nielsen	2
Jug: Reproducible Research in Python	Luis Pedro Coelho	3
OpenLabFramework: A Next-Generation Open-Source Laboratory Information Management System for Efficient Sample Tracking	Markus List	4
Ten Simple Rules for the Open Development of Scientific Software	Andreas Prlic	
Refinery Platform - Integrating Visualization and Analysis of Large-Scale Biological Data	Nils Gehlenborg	5
MetaSee: An interactive visualization toolbox for metagenomic sample analysis and comparison	Kang Ning	6
DGE-Vis: Visualisation of RNA-seq data for Differential Gene Expression analysis	David Powell	
Genomic Visualization Everywhere with Dalliance	Thomas Down	
Robust quality control of Next Generation Sequencing alignment data	Konstantin Okonechnikov	7
Visualizing bacterial sequencing data with GenomeView	Thomas Abeel	8
BioRuby project updates - power of modularity in the community-based open source development model	Toshiaki Katayama	
Biopython project update	Peter Cock	
InterMine - Collaborative Data Mining	Alex Kalderimis	
GenoCAD 2.2 Grammar Editor	Jean Peccoud	9
Improvements and new features in the 7th major release of the Bio-Linux distro	Timothy Booth	10
BioBlend - Enabling Pipeline Dreams	Enis Afgan	
Taverna Components: Semantically annotated and shareable units of functionality	Alan Williams	11
UGENE Workflow Designer – flexible control and extension of pipelines with scripts	Yuriy Vaskin	
Oqtans: A Multifunctional Workbench for RNA-seq Data Analysis	Gunnar Rätsch	12
PhyloCommons: community storage, annotation and reuse of phylogenies	Hilmar Lapp	
GEMBASSY: an EMBOSS associated package for genome analysis using G-language SOAP/REST web services	Kazuharu Arakawa	
Rubra - flexible distributed pipelines for bioinformatics	Clare Sloggett	13
Towards Enabling Big Data and Federated Computing in the Cloud	Enis Afgan	
MyGene.info: Making Elastic and Extensible Gene-centric Web Services	Chunlei Wu	14

O|B|F - Open Bioinformatics Foundation

Membership Application

I wish to apply for membership in the Open Bioinformatics Foundation (O|B|F).

First and Last Name: _____

Street Address: _____

City, State, Zip Code: _____

Country of Residence: _____

Email Address: _____

All fields are mandatory. The O|B|F will treat all personal information as strictly confidential and will not share personal information with anyone except members of the O|B|F Board of Directors, or entities or persons appointed by the Board to administer membership communication. This may be subject to change; please see below.

I am an attendee of BOSC 201___: ☐ Yes ☐ No

If you answered No, please state why you meet the membership eligibility requirement of being interested in the objectives of the O|B|F:

(Use back of page if you need more space)

I understand that membership rights and duties are laid down in the O|B|F Bylaws which may be downloaded from the O|B|F homepage at http://www.open-bio.org/. I understand that if the O|B|F's privacy statement changes I will be notified at my email address (as known to O|B|F), and if I do not express disagreement with the proposed change(s) by terminating my membership within 10 days of receipt of the notification, I consent to the change(s).

_____

Signature

Talk and Poster Abstracts

In the pages that follow, talk abstracts appear in the order in which the talks will be presented. Some authors will also present their work as posters. Those abstracts have a poster number at the bottom of the page. Poster-only abstracts appear after the talk abstracts.