Research Scientist: Bioinformatics

Brad Chapman

Job Description

As a member of the Harvard T.H. Chan School of Public Health Bioinformatics core, the Research Scientist will provide custom bioinformatics analysis and software development for researchers in the school and wider biological community, with a focus on next-generation sequence data management and analysis. This includes:

- Developing collaborations with researchers, providing bioinformatics support for ongoing investigations.
- Building reusable analysis tools resulting from collaborations, leading to widely useful software infrastructure.
- Designing custom reporting and visualization of computational results, to communicate these results through publication and conference presentation.
- Initiating new research directions important for future Harvard Chan School activities.
- Stewardship to the wider scientific community by providing software and training. Collaborative mentoring with other bioinformaticians in the core.

The ideal candidate will integrate within the Bioinformatics Core team and initiate self-directed research and development.

Requirements

- PhD in biological sciences and bioinformatics
- 10+ years of programming experience
- Demonstrated utilization of best-practice software methodologies, including revision control, testing frameworks and reproducible development practices
- Strong background in the data management, processing and analysis of sequencing experiments (Exome/Whole Genome Sequencing, RNA-Seq and other variants)
- Experience building and maintaining large scale analysis tools on distributed platforms
- History of providing reproducible analysis environments on cloud computing infrastructure like Amazon Web Services
- Ability to communicate results effectively, both formally and informally

Activities Statement

The prevalence of high throughput sequencing data in Public Health research requires validated, scalable infrastructure able to effectively process, prioritize and distribute results to biological collaborators. This infrastructure is a complex coordination of rapidly changing open source software tools from the scientific community. Most on-going projects don't contain budget or time for development of this infrastructure, but it is critical to answering underlying biological questions driving much current research.

My work enables great science at Harvard Chan School by developing this re-usable infrastructure for use by the Bioinformatics core and the wider scientific community. The key component of my job is understanding the needs of existing projects, and providing a high quality, easy to use implementation that helps improve our ability to collaborate with research scientists. For example, by having infrastructure that reliable calls variants on large scale whole genome sequencing projects we can devote core hours to downstream analysis and presentation of data instead of putting together ad hoc tools to do the processing for each new incoming project.

This infrastructure builds on expertise and available tools from the wider bioinformatics community. We aggressively re-use existing software and also contribute back to that software in the form of bug reports, fixes and validation. By being an active and respected member of the community, we are able to establish collaborative relationships outside of the immediate Harvard community. This drives the development of new areas of research and continued building of infrastructure.

My work involves three major areas of focus:

- Developing external collaborations to develop infrastructure and expertise that we can re-use on multiple core projects.
- Establishing scientific standards for assessing data and developing best-practice analysis approaches.
- Working with the scientific community to build tested, reliable software that is widely useful and usable.

Infrastructure and collaborations

High quality, scalable infrastructure is essential to working effectively in biological research. A major challenge is that this work is not well funded or rewarded in the context of traditional academic research which focuses on answering specific biological questions. My personal focus is on developing this critical reusable infrastructure by establishing larger collaborations which require the functionality. The development happens within the context of the project but is then available for subsequent work done by other members of the bioinformatics core.

I'd like to highlight two projects where an external collaboration helped the Harvard Chan school establish local expertise while simultaneously solving difficult scientific problems in the scope of the work. Both of these projects led to re-usable tools that are in daily use by myself and other members of the Bioinformatics core.

The first is a 1500 whole genome sequencing project of families with Alzheimer's disease, in collaboration with Rudy Tanzi's group at Massachusetts General Hospital. This project was a logistical challenge initiated after our initial trials analyzing 50 whole genome datasets for Peter Kraft. We effectively scaled our analysis to handle the data volume, developing relationships with

Dell, Intel and Harvard FAS Research Computing to help organize compute resources to handle the processing. We tuned variant callers for improved sensitivity and precision based on reference standards, improving our understanding of calling in complex genomic regions. Finally we integrated and developed downstream tools for analysis of these large datasets. We contributed to development of GEMINI, used for querying variations in the context of other annotations of interest.

- Scaling by evaluation of required computational methods: `http://bcb.io/2013/10/21/updated-comparison`
- Scaling with the use of better compute and filesystems: `http://bcb.io/2013/05/22/scaling`
- Scaling by use of cloud infrastructure: `http://bcb.io/2014/12/19/awsbench/`
- GEMINI analysis tool paper: `http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003153`

The second is an ongoing collaboration with the AstraZeneca Oncology Research and Development. We developed a framework for validation of cancer variant callers using synthetic datasets from the ICGC-TCGA DREAM challenge. The led to open sourcing of an AstraZeneca developed variant caller, VarDict, which provides improved detection of indels compared with existing somatic callers. Additionally, we integrated multiple structural variant and CNV callers improving our ability to detect larger variants in difficult regions. In the continued collaboration, we plan to expand work on detection and prioritization of difficult variants. All of this work contributes to the collaboration, as well as expanding our internal ability to detection variation for a wider range of projects.

- Cancer caller and structural variant validation: `http://bcb.io/2015/03/05/cancerval/`

Scientific standards

A central theme of the infrastructure and development projects is the need to have high resolution validation of methods and implementations. High throughput analysis work is rapidly changing and requires continuous integration of updated methods, while simultaneously expanding to incorporate new types of input data and assays. As part of providing service in collaborations, the core needs to be able to assess sensitivity and precision of methods to understand both the outputs of analysis processes as well as cases where we cannot reliably detect biological signal.

As part of this validation work, the core has becoming increasingly involved with global community developing reference standards. Growing out of our initial work on the Archon Genomic XPrize, we contributed to efforts at the National Institutes of Standards and Technology to produce reference materials for germline variant calling. This resulted in a community developed set of standards and a subsequent paper in Nature Biotechnology. We've continued our involvement with this community through the transition to the Global Alliance for Genomics and Health (GA4GH) benchmarking group. This expands reference materials to cover increasingly diverse patient populations and new sequencing technologies like PacBio and 10X Genomics.

The key component of this work is establishing reproducible, measurable metrics to assess how well high throughput sequencing approaches identify variants. Our involvement in these initiatives establishes the Harvard Chan School as a key contributor to developing fully transparent standards

and open source software for the ongoing transition to personalized, precision medicine. This is critical for patient care, establishing Harvard as a center of expertise for clinical variant assessment.

- Genome in a Bottle paper in Nature: `http://www.nature.com/nbt/journal/v32/n3/full/nbt.2835.html`
- Validation of variant calling methods: `http://bcb.io/2014/10/07/joint-calling/`
- GA4GH benchmarking: `http://www.ga4gh.org/#/benchmarking-team`
- Archon Genomics XPrize: `http://genomics.xprize.org/about/overview`

Community development

Building analysis tools involves connecting a wide variety of specialized open source scientific software. This extensive reuse enables the rapid modifications and continuous change inherent in research software development. In addition to using these resources, an important part of my work is contributing back the communities that produce and maintain them. As part of our work in the core we make our analysis toolkit, called bcbio – short for Blue Collar Bioinformatics, a blog where I discuss on-going development work – available to the community. Our active development and support of this tool includes a ready to run version on cloud infrastructure, as well as regular releases that run in local HPC environments, including maintained versions on our local Odyssey and Orchestra clusters.

Beyond development, I'm also involved with community building as part of the Open-Bio foundations. I've been an organizer of the Bioinformatics Open Source Conference (BOSC) for the past six years, which brings together developers interested in improving re-usability of biological software and increasing diversity in the bioinformatics community. An important aspect of this has been my organization of Codefest, a free two day coding session prior to the BOSC conference. This mentoring and development workshop brings new members into the community through individual work with existing programmers, as well as encouraging development of new cross-project collaborations. A recent example is the development of the Common Workflow Language. This initiative, derived from Codefest 2014 meetings, connects multiple commercial and academic providers of tools for building computational workflows in an effort to improve standardization and reproducibility of analyses.

This organizational and technology building provide a strong statement from Harvard Chan School about our concern for making our research tools and results widely accessible. My work in the bioinformatics core thus produces both great science within our research groups, and also enables this same great science in the larger scientific community.

- bcbio – our analysis tool: `https://github.com/chapmanb/bcbio-nextgen`
- bcbio availability on Amazon Web Services: `http://bcb.io/2014/12/19/awsbench/`
- Bioinformatics Open Source Conference: `http://www.open-bio.org/wiki/BOSC_2015`
- Codefest: `http://www.open-bio.org/wiki/Codefest_2015`
- Common Workflow Language: `https://github.com/common-workflow-language/common-workflow-langu`