Senior Research Scientist: Bioinformatics

Brad Chapman

Job Description

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

- Developing research collaborations, providing bioinformatics support for ongoing investigations.
- Building reusable analysis tools resulting from collaborations, leading to widely useful software infrastructure.
- Integration of developed analyses within open source and commercial frameworks, allowing wide usage across a diverse set of research needs.
- Designing custom reporting and visualization of computational results, to communicate these results through publication and conference presentation.
- 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
- 15+ 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, Google Compute Engine and Microsoft Azure
- 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. Researchers want to take advantage of high throughput sequencing for existing cohorts like the Nurses' Health Study and need ready to run tools and expertise to effectively process, quality control and analyze their data.

The necessary infrastructure is a complex coordination of rapidly changing open source software tools from the scientific community. Improvement and support for this infrastructure requires regular maintenance and development work. This is in often in contrast to project budgets and timelines which focus on the underlying biological questions.

My work enables science at Harvard Chan School by developing 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. Additional, we support our analysis tools in a wide variety of computational environments, ranging from local HPC to cloud infrastructure to commercial analysis platforms. This maximizes re-use and interoperability, making the software development process as efficient as possible. By being an active and respected member of the community, we are able to establish collaborative relationships outside of the immediate Harvard community, driving the development of new areas of research and continued building of infrastructure.

My work involves three major areas of focus:

- Developing tooling for software interoperability.
- Establishing scientific standards for validation.
- Building external collaborations to develop infrastructure and expertise that we can re-use on multiple core projects.

Interoperability

The largest barrier to efficient software development is integrating a wide variety of specialized open source scientific software. As part of our work in the core we make bcbio – our analysis toolkit named after Blue Collar Bioinformatics, a blog where I discuss on-going development work – available to the community. Extensive reuse of open source tools allows us continually improve, maintain and support bcbio.

While we've effectively been able to integrate tools within bcbio, a second challenge has been to include bcbio inside other platforms. Users increasingly want to run the same analyses on local HPC environments; on cloud environments like Amazon Web Services, Google Compute Engine and Microsoft Azure; and on commercial providers like DNAnexus, SevenBridges and Arvados.

To avoid the support overhead of maintaining bcbio in multiple environment, I've become actively involved with the Common Workflow Language (CWL) community. This is a community effort initiated at the Open Bioinformatics Codefest, a free two day coding session prior to the Bioinformatics Open Source Conference (BOSC). I've been an organizer of Codefest and BOSC since 2010 and 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.

Emerging from this collaborative environment, the Common Workflow Language became a widely used community standard supported by both open source workflow tools, commercial analysis providers and the Global Alliance for Genomics and Health (GA4GH). Practically, it provides a way to run bcbio within multiple computational environments. To ensure this, we contributed bcbio workflows to the GA4GH-DREAM workflow execution challenge, which ensures our community built tools run reliably for multiple users. This provides flexibility for research analysis at Harvard, allowing us to choose between local and remote compute solutions based on cost and turnaround requirements.

- bcbio, our analysis framework – `https://bcb.io`
- Common Workflow Language: `http://www.commonwl.org/`
- Codefest: `http://www.open-bio.org/wiki/Codefest_2017`
- Bioinformatics Open Source Conference: `http://www.open-bio.org/wiki/BOSC_2017`
- GA4GH-DREAM Workflow Execution Challenge `https://www.synapse.org/#!Synapse:syn8507133/wiki/415976`

Validation

As we've continued to expand the reach and usage of bcbio, it becomes increasingly critical to ensure we have high confidence methods for validating workflows. 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, my work in the core has becoming increasingly involved with the global community developing reference standards. We continued our work with the National Institute of Standards and Technology's Genome in a Bottle and GA4GH benchmarking teams on expanding reference standard for human genome sequencing. This includes providing ethnically diverse test datasets which ensure methods work on the type of patient populations we investigate at Harvard. We've also provided test datasets and validations for moving to the more accurate Human Genome Build 38, and helped to build validation sets for difficult biological problems like structural variant calling. In the context of bcbio, this has included expanding our structural variant calling capabilities as well as integrating efficient gVCF based germline variant calling.

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.

- Validation of human genome build 38: `http://bcb.io/2015/09/17/hg38-validation/`
- NIST Genome in a Bottle: `http://genomeinabottle.org/`
- GA4GH benchmarking: `https://genomicsandhealth.org/working-groups/benchmarking`

Infrastructure and collaborations

High quality, scalable and validated 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 emphasizes answering specific biological questions. My personal goal is to continue to develop 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 to both analyses in the core as well as the larger research community.

I'd like to highlight three projects where external collaborations 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 an ongoing collaboration with AstraZeneca Oncology Research and Development. This is a multi-year project focused around improving calling for low frequency somatic variants and handling difficult data types. Recent work allowed use of Unique Molecular Indexes (UMIs), which helps effectively sequence to high depths for identifying low frequency variants in somatic tissue or circulating tumor DNA. Coupled with filters for identifying artifacts caused by oxidative damage or FFPE deamination, this allows us to work on hard to analyze samples. At Harvard Chan School, this work supports exome FFPE analysis of Nurses' Health Study samples with Peter Kraft.

- Validation and improvement of cancer calling: `http://bcb.io/2016/04/04/vardict-filtering/`
- DNA damage filters: `https://github.com/bcbio/bcbio.github.io/blob/master/_posts/2017-01-31-damage-filters.md`

The second collaboration, with the University of Melbourne Center for Cancer Research, develops automated pipelines for cancer analysis. This extends bcbio's somatic calling functionality to clinical patient samples, requiring additional automation, validation and accreditation. By allocating effort to improving single sample turnaround and provenance, we build the reproducibility workflow components essential for both research work and clinical applications.

The third is a set of work with commercial analysis providers: Curoverse and Veritas Genetics, Seven Bridges Genomics and DNAnexus. These projects integrate CWL pipelines generated by bcbio into their workflow environments. This demonstrates the usefulness of bcbio, and Harvard's community contributions, to a wide audience of researchers.

Community based collaborative development provides 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