Transparent Informatics: A Foundation for Precision Medicine

A.W.Zaranek[1,2,3], M.P.Ball[1], T.Clegg[1,3], J.Sheffi[3], W.Vandewege[1,3], A.Berrey[2,3], G.M.Church[1]

[1]Personal Genome Project, Harvard Medical School, Boston, MA; [2]Arvados Foundation, Cambridge, MA; [3]Clinical Future, Cambridge, MA

Introduction

Precision medicine calls for the deep integration of genomic data into the normal course of clinical care. [National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease, 2011]. The scale, complexity, and applications of these data will require a fundamental change in how the biomedical informatics community approaches computing.

Answering many of the most interesting research and clinical questions will require complex, large scale, and interconnected systems spanning multiple organizations.

We propose that "transparent informatics" resources built with open source software and public domain data should be the foundation for the next generation of computing infrastructure in both biomedical research and clinical practice (Figure 1).

Further, we encourage widespread adoption of Arvados and the transparent informatics resource already available from the Harvard Personal Genome Project (PGP).

What's a Transparent Informatics Resource?

As increasingly large and complex datasets are used to deliver routine clinical care and do biomedical research, there is a need for well-integrated systems of software and data that can be used by everyone.

We call such a system a "transparent informatics resource" if it provides:

- **Software with downloadable source code** that is made publicly available under the AGPLv3 / GPLv2 (or any later version) or a compatible license such as Apache 2.0.
- **Integrated Human Datasets** (i.e. sequencing, health, and other biomedical data) in the public domain via the Creative Commons CC0 copyright waiver or an equivalent legal construct; other information (e.g. wikis, scientific articles) available under CC-BY-SA or a compatible license such as CC-BY or CC0.

Results

As of June 2013, the Harvard Personal Genome Project has publicly shared phenotypic, environmental, genomic, and annotation data from a total of approximately 2,700 research participants. All the methods for managing, generating, and interpreting these data are available as open-source software. The resulting resource can become a foundation for a wide range of biomedical research and precision medicine applications. (Figure 1 inset)

The PGP has built several applications [Ball, Thakuria, Zaranek et al. 2012] on top of a common software platform to operate the study. This platform, called Arvados, specifically addresses the biomedical data management, analysis, provenance, governance and sharing requirements of precision medicine applications. (see Figure 2).

First developed in 2007 [Zaranek et al., 2008], Arvados provides capabilities for data management & storage, pipeline development & execution, real-time querying of genomic data, maintenance of data provenance, pipeline reproducibility, and security (see Figure 3 for details). It is similar to Hadoop [Shvachko 2010]. Arvados is available under the AGPLv3 license, with its Software Development Kits licensed under Apache 2.

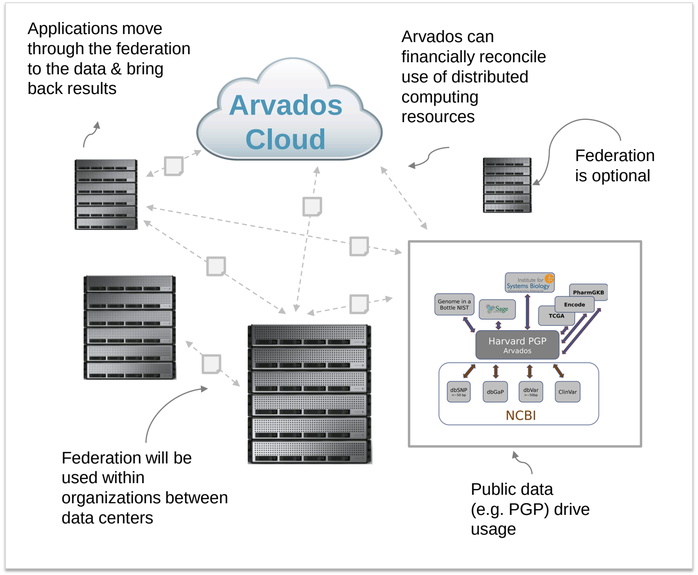

Figure 1: *Federated transparent informatics resources can provide a foundation for exascale computing infrastructure in both biomedical research and clinical practice. The Arvados project and the Harvard Personal Genome Project data could become important building blocks for such a federation.*

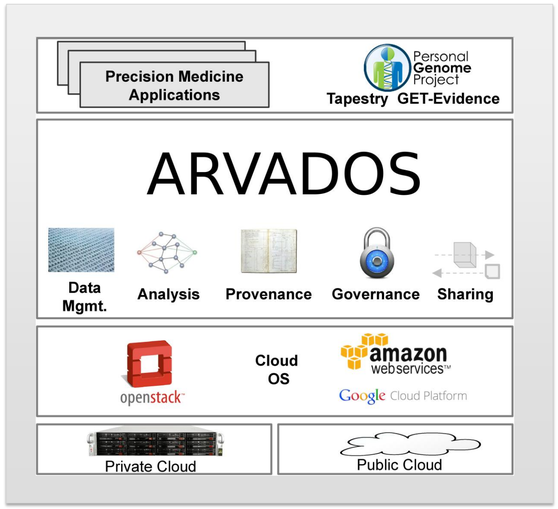

Figure 2: *Arvados is a community-driven free software project that extends popular cloud operating systems such as OpenStack or Amazon Web Services. It is designed to run on-premise or hosted. Arvados creates an abstraction between the informatician and the biomedical data management, analysis, provenance, governance and sharing requirements of precision medicine applications. It also enables easy portability across cloud operating systems.*

Tapestry and GET-Evidence are Arvados applications that run the Harvard Personal Genome Project; these applications and data evolve with private innovations in the industry and can be used to develop and validate precision medicine applications that are non-public.

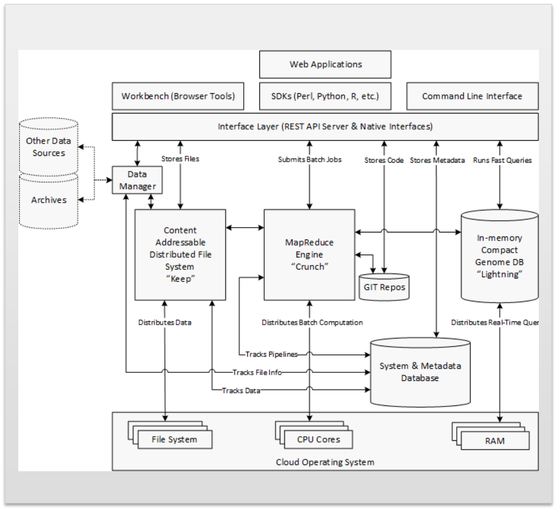

Figure 3*: Technical overview of Arvados. The system exposes a standard set of services to biomedical applications and users that accelerate time-to-results, improve accuracy, and reduce costs while permitting continued use of existing tools.*

Core services include "Keep," a content-addressable distributed filesystem, "Crunch," a MapReduce style distributed computation system, and "Lightning," an in-memory database for querying genomic data.

A common software platform, when combined with public data used as a lingua franca, makes it feasible for organizations to share highly private data in a controlled manner.

Conclusions

As DNA sequencing becomes more and more affordable, we are confronted by new barriers to widespread adoption of genomics in medicine besides sequencing cost.

Use of transparent informatics resources as a core building block will unlock the potential of precision medicine by ensuring results are repeatable, eliminating barriers to sharing and collaboration as well as creating a safe environment for hardening applications for use with highly private data.

We believe the Arvados project could become an important enabling infrastructure platform that addresses the needs of the genomic research and clinical communities.

Further, by adopting transparent informatics resources as a foundation, the industry can address the computing challenges posed by genomic and other similarly large biomedical data sets.

Literature cited

National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. National Academies Press (US), 2011.

M. P. Ball, J.V. Thakuria, A.W. Zaranek et al. A public resource facilitating clinical use of genomes. Proc Natl Acad Sci U S A, 109(30):11920–11927, Jul 2012. doi: 10.1073/pnas.1201904109.

K. Shvachko, et al. The hadoop distributed file system. In Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on, pages 1–10, 2010. doi: 10.1109/MSST.2010.5496972.

A. W. Zaranek, T. Clegg, W. Vandewege, and G. M. Church. Free factories: Unified infrastructure for data intensive web services. Proc USENIX Annu Tech Conf, 2008:391–404, May 2008.

Acknowledgements

We are grateful to members of the Personal Genome Project, PersonalGenomes.org, and to the entire team at Clinical Future, Inc. for making this research possible.