

RESOURCE SHARING PLANS

Research tools and resources will be made available in full accordance with the NIH Grants Policy Statement, the Principles and Guidelines for Recipients of NIH Research Grants and Contracts, and the NIH Genomic Data Sharing Policy. Each of the Principal Investigators has significant experience participating in large-scale genome sequencing projects, and has an established record of sharing unique resources generated with NIH support with the academic research community, through publications and submission to data repositories without any license or research restrictions.

Genomic Data Sharing Plan

All genomic data (large-scale data as defined in the NIH Genomic Data Sharing Policy) resulting from the proposed project, as well as relevant phenotype-associated data, will be made available in a timely manner in full accordance with the NIH Final Genomic Data Sharing (GDS) Policy as well as any distributions plans established by the TOPMed Program. In particular, the catalog of structural variations identified from the individuals sequenced as part of the NHLBI-supported TOPMed Program by project investigators will be made publicly available within the timelines required by the GDS Policy and Supplemental Information. Project investigators that download unrestricted-access data from NIH-designated data repositories will acknowledge the specific data sets or applicable accession numbers(s) and the NIH-designated data repository. It is anticipated that genomic data made available from the TOPMed Program will be de-identified, and therefore, the derived data will meet the standards set forth in the HHS Regulations for the Protection of Human Subjects. In addition, this project will be registered in the Database of Genotypes and Phenotypes (dbGaP) <http://www.ncbi.nih.gov/gap> and the human genomic data will be submitted to the relevant NIH-designated public data archives advised by the steering committee of the TOPMed program.

Software Sharing Plan

This proposal describes three major software pipelines for SV analysis: an extended fusorSV framework for discovery, SVIM for functional annotation and impact assessment, and SV2Pheno for determining genotype-phenotype associations. As this proposal encompasses substantial work in several different areas, we plan on semi-regular code releases over the duration of the funding period (see timeline below). The release process will follow conventional practice: source code and documentation (including description, installation, use cases, etc.) will be packaged and available from a public distribution site, e.g. GitHub. Software images for use on cloud computing resources will also be disseminated to the broader community as appropriate. Each pipeline will be accompanied by journal publication describing the new enabled capabilities along with any new biological findings.

1. Software Licensing Plan. We will release the pipeline software developed as part of this grant under the Apache License, Version 2.0 (Apache-2.0). This license is a permissive free, open-source license ensuring the software is freely available to everyone, including biomedical researchers and educators in the non-profit sector, such as institutions of education, research institutions, and government laboratories. This license also allows everyone, including external researchers, collaborators, and students to modify the source code and to share these modifications with the broader community. This attribute ensures that code maintenance and development can continue under the auspices of others in the community in the very unlikely event that we are unwilling or unable to do so. Derivative works are not required also to be distributed under Apache-2.0, allowing dissemination and commercialization of enhanced or customized versions of the software and the incorporation of the software or pieces of it into other software packages.

2. Software Release Strategy. We will release official versions of the pipelines with major improvements on a semi-regular basis and minor updates as needed to ensure properly functioning software. All releases will be numbered and made available on GitHub. Improvements in the external software tools on which our pipelines depend will be reviewed for compatibility with our software, and we will issue numbered software releases as appropriate. We will also work to ensure new releases maintain compatibility across supported local data center and cloud platforms. Given typical development cycles, we estimate major releases perhaps a few times a year, minor releases on the cycle of months, and ad hoc patches on the cycle of weeks. Finally, we will publish companion scientific articles and application notes when appropriate. Announcements of updates will be made through the regular communication channels described in the proposal, e.g. Biostars community message board. We believe this overall strategy will be highly responsive to the needs of the user community.