Biostatistics and Bioinformatics Shared Resource

Data Type

Types and amount of scientific data expected to be generated in the project: ***Summarize the types and estimated amount of scientific data expected to be generated in the project.***

Describe data in general terms that address the type and amount/size of scientific data expected to be collected and used in the project (e.g., 256-channel EEG data and fMRI images from ~50 research participants). Descriptions may indicate the data modality (e.g., imaging, genomic, mobile, survey), level of aggregation (e.g., individual, aggregated, summarized), and/or the degree of data processing that has occurred (i.e., how raw or processed the data will be)

The BBSR performs data analysis for data generated across a wide array of genomics platforms including bulk RNA-seq, DNA-seq, SNP genotyping, DNA methylation as well as single cell RNA and ATAC-seq and spatial transcriptomics. The BBSR maintains 20Tb of data stored on Dartmouth's DartFS secure, redundant and networked storage system, managed by Dartmouth's Research computing team. DartFS includes automatic daily, weekly and monthly back ups of all stored data.

While raw data is stored by the Genomics and Molecular Biology Shared Resource (GMBSR), the BBSR stores processed data files generated during a typical data analysis. These include intermediate file types including processed sequence reads (FASTQ) alignment files (BAM), read counts (TXT), peak calls (BED), as well as final output file types such as aggregated gene expression matrices (TXT) and analysis results (in XLSX or CSV formats). For single cell projects, the vast majority of analysis is conducted in the R Statistical Programming Environment, therefore intermediate files generated in these analysis are stored as RDS format objects.

The expected amount of data to be generated is unknown and depends on the size of each individual project. Based on previous core usage, maintaining 20TB of storage on DartFS is sufficient to store data for ongoing projects before they are submitted to public repositories or transferred to investigator specific storage volumes.

Scientific data that will be preserved and shared, and the rationale for doing so: ***Describe which scientific data from the project will be preserved and shared and provide the rationale for this decision.***

From the date we return the results of an analysis to a core user, we will retain intermediate files for 1-month, and final output files will be kept for 1 year. Typical intermediate file types generated in bioinformatics analysis are often very large and expensive to store long term. Unlike the final output

files, most intermediate files are not required to be shared when submitting data to public repositories. After 1 year, it is the responsibility of individual users to develop a long-term storage plan for their data.

If the BBSR is still actively working on a project 1 month after the results were returned to the user, we will facilitate transfer of data files to investigator specific storage volumes on DartFS. BBSR staff can be given access to investigator specific storage volumes on DartFS, allowing us to continue working on the project until the related manuscript is submitted to a journal for publication, at which time the data will be uploaded to a public repository with assistance from BBSR staff. Submission of project data to a public repository usually also includes submitting the raw data, stored by the GMBSR. BBSR has access to the GMBSR storage volume of DartFS, allowing BBSR staff to assist with the inclusion of raw data when submitting data to a public repository.

Metadata, other relevant data, and associated documentation: Briefly list the metadata, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) that will be made accessible to facilitate interpretation of the scientific data.

Metadata for raw data files is maintained in the GMBSR's Laboratory Information Management System (LIMS), hosted on a Virtual Machine (VM) managed by Dartmouth's Research Computing group. For processed data files, project metadata is stored in a combination of TXT, XLSX, and CSV files. Computational code use to perform all analysis provides details of specific bioinformatic and statistical methods applied to the data, and therefore contains additional project metadata. Computational code is made available to core users upon request, and often serves as an educational tool for training students in the labs of core users in computational data analysis.

Related Tools, Software and/or Code

State whether specialized tools, software, and/or code are needed to access or manipulate shared scientific data, and if so, provide the name(s) of the needed tool(s) and software and specify how they can be accessed.

The BBSR uses a wide range of computational tools and software to perform data analysis. The vast majority of preprocessing for genomic data analysis is conducted using publicly available programs that are implemented through the UNIX based command line. Software is stored and managed through a combination of downloaded binaries, software image containers using Singularity, and the package manager Conda. For developing and maintaining standardized pipelines, the BBSR uses

Snakemake, a Python-based workflow management system, which executes complex analysis pipelines by optimizing resource usage, file generation, software management, and detailed analysis logs. The majority of downstream statistical data analysis is implemented using freely available R and Python software, both of which have large users communities that openly disseminate new data type specific analysis methodologies and utilities, many of which are utilized by the BBSR for specific analyses.

Computational code used to perform analyses will be kept indefinitely by the BBSR due to its small overall size and importance for reproducing an analysis for raw data files and maintaining provenance. Analysis code is made available to core users upon request, and often serves as an educational tool for training students in the labs of core users in computational data analysis. If new computational methodology is developed, the BBSR will facilitate sharing of the code through the core's GitHub page, or that if the individual investigator.

Standards

State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources, and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project. If applicable, indicate that no consensus standards exist

Analysis performed by the BBSR will make use of common bioinformatics file formats for which existing standards and composition is widely available and agreed upon. These include FASTQ, FASTA, BAM, BED, BIGWIG, VCF, & MEX.

Downstream statistical data analysis in R & Python will generate results in a diverse array of formats which do not necessarily have defined community data standards. These files will generally be in TXT and CSV formats. Whenever sufficient metadata will be provided with the files in order to make sure uses can understand and interpret all fields and file content.

Formal data standards for sharing spatial transcriptomics data have not been decided upon yet in the research community, however we plan to make available all raw data required to generate feature barcode matrices that can be used for downstream analysis, as well as the feature barcode matrices themselves and images in TIFF format.

Data Preservation, Access, and Associated Timelines

Repository where scientific data and metadata will be archived: Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived.

From the date we return the results of an analysis to a core user, the BBSR will retain intermediate file types (described above) for 1-month, and final output files for 1 year. Core users are advised of the BBSR storage policy when data & results are returned to them at the end of an analysis, and informed it is their responsibility to develop a long-term storage plan for their data. Users are reminded that files will be deleted 2 weeks prior to the 1 year deadline. For all data and results the BBSR assists core users with transfer of files to their own storage volumes.

For data included in published research, data will be submitted to public genomics repositories and the accession numbers for these submissions will be included in the associated publication. While data analyzed by the BBSR are expected to be de-identified, any data that may contain potentially sensitive health information (such as genotypes from SNP arrays and DNA-seq) data will be submitted to the database of Genotypes and Phenotypes (dbGaP). For genomics data that do not contain potentially sensitive health information data will be submitted to the Gene Expression Omnibus (GEO). This includes data from analysis of RNA-seq, DNA methylation, and single-cell RNA-/ATAC-seq.

How scientific data will be findable and identifiable: Describe how the scientific data will be findable and identifiable, i.e., via a persistent unique identifier or other standard indexing tools.

For data submitted to public genomics repositories (GEO & dbGaP), the accession numbers for these submissions will be included in the associated publication.

When and how long the scientific data will be made available: Describe when the scientific data will be made available to other users (i.e., no later than time of an associated publication or end of the performance period, whichever comes first) and for how long data will be available.

Relevant scientific data (based on the file formats discussed above) generated from this project will be made available at the time of publication. Data will be made available to reviewers upon request at the time of submission of each manuscript to a peer reviewed manuscript.

Access, Distribution, or Reuse Considerations

Factors affecting subsequent access, distribution, or reuse of scientific data: NIH expects

that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing.

For the vast majority of genomics data that will be shared, there are no anticipated factors or limitations that will affect the access, distribution or reuse of the scientific data generated by the proposal. For data that contains potentially sensitive health information, which will be shared through dbGaP, users will be required to apply for a receive permissions before they are able to access and reuse the data, and abide by dbGaPs usage conditions.

Whether access to scientific data will be controlled: State whether access to the scientific data will be controlled (i.e., made available by a data repository only after approval).

For data that contains potentially sensitive health information, which will be shared through dbGaP, users will be required to apply for a receive permissions before they are able to access and reuse the data, and abide by dbGaPs usage conditions.

Protections for privacy, rights, and confidentiality of human research participants:
If generating scientific data derived from humans, describe how the privacy, rights, and confidentiality of human research participants will be protected (e.g., through de-identification, Certificates of Confidentiality, and other protective measures).

Data from human subjects containing potentially sensitive health information will be de-identified prior to sharing with the BBSR. It is the responsibility of the sponsoring PI to ensure proper de-identification of human subjects material and the BBSR refuses to accept data for projects that do not meet this criteria. Once such data are submitted to a public repository, the data will only be accessed by approved individuals who have applied for specific access rights through dbGaP.

Oversight of Data Management and Sharing

Describe how compliance with this Plan will be monitored and managed, frequency of oversight, and by whom at your institution (e.g., titles, roles).

Owen Wilkins, Senior Research Scientist, PhD will oversee and/or manage data sharing with core users and submission of data to public genomics databases.
