DRAFT Research Computing Resources for Basic Science – Cheat Sheet

Duke University School of Medicine June 3, 2024 -- *Send questions and feedback (including corrections and broken links) to Steve Rozen sr110@duke.edu*

For desktop/laptop and network support in the School of Medicine https://bit.ly/Duke-OASIS

Researchers in the School of Medicine can use resources from two IT systems

OASIS/DHTS (School of Medicine Office of Academic Solutions and Information Systems) https://medschool.duke.edu/research/research-support/research-support-offices/oasis/oasis-services/oasis-research-technology

Suitable for all data security classifications.

High performance computing cluster at the School of Medicine Research Computing Service

Information at: https://bit.ly/SoM-HPC

50 GB home directory, 10 TB working storage (60-day retention)

Red Hat Linux 9, Slurm scheduler (https://slurm.schedmd.com), InfiniBand networking within the cluster, Lustre file system

27 compute nodes (64 hyperthreaded cores, 1 TB RAM, $0.07 / CPU-hour)

1 GPU node (4 X NVIDIA L40S GPUs + 64 CPU cores and 1 TB RAM, $3.50 / GPU-hour)

To request access, technical help, consulting, or to discuss adding your own node to the HPC cluster: https://bit.ly/som-hpc-request (Duke internal web site)

OASIS consulting services https://medschool.duke.edu/research/research-support/research-support-offices/oasis/oasis-services/oasis-rts/research-3

DHTS Isilon data storage Please contact OASIS consulting services

Long term, inexpensive Amazon Web Services storage Please contact OASIS consulting services

OIT (Duke Office of Information Technology) https://oit.duke.edu

Not suitable for data classified as “sensitive” or “protected health information” (PHI).

Duke Compute Cluster (DCC) https://bit.ly/oit-dcc

High performance compute cluster with 1,300 nodes running Alma Linux and the Slurm scheduler (https://slurm.schedmd.com).

Two Slum partitions are free to use. In addition, labs can add their own nodes (provided they are compatible with the cluster) and always get priority access on their node. Other labs can use the node but are evicted immediately if the owner wants to use the node.

DCC provides most bioinformatics software packages using the “Environment Modules” system (https://modules.sourceforge.net), and other standard packages can usually be added to the module system on request.

PI can arrange with OIT to add their own nodes to DCC. Computing can also be reserved at $85 / core-year. Additional pricing at https://oit.duke.edu/help/articles/kb0030661.

DCC offers the “OPEN OnDemand” web-browser interface (https://openondemand.org) . This provides a Linux desktop plus RStudio, JupyterLab notebooks, and VS Code.

Data storage 1 TB free storage per lab, additional storage available for rent. Archival storage is $62 / TB-year without backup and $144 / TB-year with backup. Additional pricing for higher speed storage is at https://oit.duke.edu/help/articles/kb0030661.

For appointments regarding all OIT services, including requesting new accounts on DCC, see https://outlook.office365.com/owa/calendar/ResearchComputing@ProdDuke.onmicrosoft.com/bookings/.

Website hosting https://oit.duke.edu/service/website-hosting.

Other services less often used for basic science are linked to at https://secureit.duke.edu/data_services (internal Duke website). These include **PACE** (Protected Analytics Computing Environment, https://medschool.duke.edu/pace), **REDCap, Protected Network for Research** (OR&I, Office of Research and Innovation)

DRAFT for discussion; completely unofficial proposal on omic data security categories

May 31, 2024

Steve Rozen sr110@duke.edu

The proposed basic principle is that human genetic data (sequence, SNPs, STRs) with a Duke owner that poses a reasonable risk of re-identification must usually* be classified as "sensitive". It the data are not owned by Duke, the category is up to the owner. For example, the Illumina Platinum (human) Genome sequences are "public". In other situations, a non-Duke owner may require a minimum security category in a data-use agreement. In any situation, the owner of the data or the data steward at Duke can opt for a more stringent category. For example, the sequence of a bacterium engineered to produce a particular protein might be "sensitive" because of IP considerations, or IRB approval restrictions may require even non-re-identifiable omic data to be considered "sensitive". Ultimately it is the responsibility of the data owner and the Duke data steward to determine the data category.

Determining omic data security category

Type of data	Security Category
Nonhuman data of any kind with Duke or non-Duke owner	Determined by owner
The remaining data types refer to human data	
Sequence from publicly available human cell line, such as cell line from ATCC	
Processed transcriptomic data (e.g. single cell count matrices, expression microarray) without sequence data	
Other processed data without sequence data (e.g. ATAC-seq, ChIP seq, etc.)	
Proteomic or metabolomic data	
DNA methylation data without sequence data (e.g. from a methylation array)	
Somatic mutation data	
Sequence mapping to < 1 megabase of the genome	
Sequence mapping to > 1 megabase of the genome	Sensitive if Duke owner, otherwise determined by owner*
Germline SNP (Single nucleotide polymorphism) or STR (short tandem repeat, microsatellite) data	

"Sequence" refers to DNA or RNA sequence including short reads and mapped short reads.
"Somatic mutation data" refers to mutations that occurred after conception, such a mutations that arise during cancer development.
For data security categories, see: https://security.duke.edu/policies-procedures-and-standards/data-security/data-classification-standard/

- Presumably study-participant consent could allow a less stringent security category