Gerstein Lab Protected Data Protocol – 1st Draft

LL

Genome Tech

2012-02-07

Current Outline

1) Location and access of protected data
2) Documentation of protected data
3) Protected data on Louise
4) Encryption of protected data
5) Deletion of protected data

Current Outline

1) Location and access of protected data
2) Documentation of protected data
3) Protected data on Louise
4) Encryption of protected data
5) Deletion of protected data

1) Location and Access of Protected Data

- All protected data that the Gerstein lab has access to will be primarily stored on a file server that will not be accessible over the Internet or through the Yale VPN.
- It will only be accessible through MBGNet.
- The physical location of the server will be in a locked room or closet for which only MF will have access.

1) Location and Access of Protected Data

- At the top level of the directory structure, data will be divided by source. Hence, all data from a particular source will be stored under one top-level branch of the directory tree.
- At deeper levels, data will be divided into directories that naturally separate the data, such as experiment type, disease type, etc.

1) Location and Access of Protected Data

- Access to these directories will be controlled by Unix user groups.
- Individuals will log in with their Yale NetIDs and passwords.
- NetIDs will be assigned to user groups that correspond to datasets from a particular source.
- User membership in particular groups determine their access to the various protected data sets.

Current Outline

1) Location and access of protected data
2) Documentation of protected data
3) Protected data on Louise
4) Encryption of protected data
5) Deletion of protected data

2) Documentation of Protected Data

- A spreadsheet will be used to keep track of information related to each protected dataset, hereafter referred to as the **metadata spreadsheet**.
- It will exist as a Google Docs spreadsheet to facilitate multi-user editing and collaboration. This spreadsheet will keep track of the following information.

2) Documentation of Protected Data

- **Source:** Source of protected datasets
- **Name:** A descriptive name for the dataset, which should give some indication of what kind of data the dataset contains
- **Description:** Additional information on what kind of data is contained in the given dataset.
- **Authorization Documents:** Links to relevant documents concerning the Gerstein lab's approved access of the data (e.g. ethics approval letter)
- **Authorized Uses:** Links to files explaining the parameters of the Gerstein lab's use of the data
- **Authorized Individuals:** Individuals that are allowed to access the data
- **Authorization Protocol:** Link to a portion of the Authorization Protocol Gdoc that explains how authorized access was obtained, for future reference when obtaining protected data from this source.
- **Expiration date:** Date that the Gerstein lab's access privileges will expire, hence the date by which renewal must be approved, or the data must be deleted.
- ***[possibly more?]***

2) Documentation of Protected Data

- This spreadsheet must reflect the granting/denying of access to users to protected datasets. Additionally, MF must update the user groups on the protected data file server accordingly.
- Additionally, an **Authorization Protocol** Gdoc (wiki page?) will be established that documents the procedures followed to obtain each dataset. This information is intended to facilitate the future acquisition of protected datasets. This document should include information on dealing with Yale's IRB, and the relevant institutional Signing Officials (SOs) involved in getting access to protected data.

2) Documentation of Protected Data

- Current metadata spreadsheet:
  - https://docs.google.com/spreadsheet/ccc?key=0AmSq_qEpEM6jdEhGTWRZWUxVanVFSkhuMXhVaUV3aUE&hl=en_US - gid=0
- Current Authorization Protocol Gdoc:
  - https://docs.google.com/document/d/1nCNWss4j-LhAZli1aj60qF8V-wnWts7epQk_tSbcJwc/edit?hl=en_US

Current Outline

1) Location and access of protected data
2) Documentation of protected data
3) Protected data on Louise
4) Encryption of protected data
5) Deletion of protected data

3) Protected Data on Louise

- The “gerstein2” group on Louise exists to allow restriction of certain directories to only those individuals authorized access to protected data.
- A special directory will be established for the specific purpose of performing high-performance scientific computations on protected data.
- Membership of this group will be controlled by the administrators of Louise, in coordination with MF.
- Members of this group are also listed in a Google spreadsheet called “Gerstein2 group on louise” for informational purposes.
- Need to add information on where exactly this directory is

Current Outline

1) Location and access of protected data
2) Documentation of protected data
3) Protected data on Louise
4) Encryption of protected data
5) Deletion of protected data

4) Encryption of Protected Data

- If protected data is to be copied to any location that is not among the secure locations listed above, it must be encrypted with a strong OpenSSL encryption scheme (des3 is an incredibly strong encryption scheme).

Current Outline

1) Location and access of protected data
2) Documentation of protected data
3) Protected data on Louise
4) Encryption of protected data
5) Deletion of protected data

5) Deletion of Protected Data

- Standard file removal utilities only unlink the file from the visible filesystem, which leaves the data susceptible to recovery with the right tools.
- Hence, protected data deletion must involve "file shredding": a utility that overwrites the data with random 0s and 1s before unlinking the file.
- The srm (secure remove file) Unix command (http://sourceforge.net/projects/srm/) is designed to function just like the standard "rm" command, except it removes by file shredding. This utility must be used to delete all protected data.

Summary of Protected Cancer Data

- We downloaded a subset of TCGA's somatic mutation data corresponding to cancer samples indicated to us by Rick Wilson.
  - 21 matched tumor-normal breast cancer (BRCA) samples
  - 31 matched tumor-normal Uterine Corpus Endometrioid Cancer (EMC) samples
  - 15 matched tumor-normal glioblastoma multiforme (GBM) samples
  - 5 matched tumor-normal ovarian serous cystadenocarcinoma (OV) samples.
- The data for each of these cancers indicates
  - The genomic coordinates of each mutation
  - Reference and tumor alleles
  - Provides various mutation classifications (missense or nonsense, SNP or indel, etc.) and,
  - Experiment metadata
- An analysis of the distribution of the mutations throughout the genome indicate that
  - GBM, OV and AML datasets appear to be whole genome (i.e. the mutations have no bias toward gene-coding regions)
  - BRCA and EMC datasets have ~90% of their mutations overlapping a gene-coding region, suggesting that they are exome datasets, or have been filtered for exon-localized mutations

- Furthermore, we have obtained somatic mutation data on a single malignant melanoma cancer patient used in a Nature paper by Pleasance *et al.* (PMID: 20016485).
- Data on some 33,344 substitutions and 982 indels were obtained by the authors.
- Data attributes include:
  - The genomic coordinates of each mutation
  - The reference and mutant alleles
  - The coding effect of each mutation
- Distribution analysis of these mutations indicates that there is no bias toward gene-coding regions.

- The whole genome datasets will be useful for classifying mutations mapping to various functional genomic elements (TFBSes, genes, pseudogenes, etc.). The exon-focused datasets, however, will be better suited to studying the specific disruption effects to certain genes.