RELATED POLICIES: HC-C04101

1. PURPOSE/ PRINCIPLE:

This document describes data policies including the measures that the Center takes to keep its customers'/patients' data secure and confidential, both on a process and product level. It also describes policies for sharing, storing and transferring next generation sequencing (NGS) data generated, analyzed and reported, and the data retention policy of all cytogenetic, clinical microarray and next generation sequencing data generated within the Center

2. SCOPE:

This document applies to all personnel accessing, processing or creating data for interpretive output in the Center.

3. RESPONSIBILITIES:

The Director of the Center is responsible for directing the Bioinformatics Director, the Laboratory Manager/Supervisor, and other medical directors. Collectively, they are responsible for different aspects of performing various genomic tests. Specifically, the Bioinformatics Director and/or Bioinformatics Supervisor are responsible for properly processing, analyzing, transferring and storing all data generated in a CAP-, CLIA-, and HIPPA-compliant environment. The Laboratory Manager/Supervisor is responsible for implementing and ensuring compliance of CAP, CLIA, HIPPA and all written policies within the Center. The medical directors have the responsibility of appropriate usage of the data strictly for clinical interpretation of these patient data in a CAP- and CLIA-certified diagnostic laboratory.

4. GENERAL:

4.1. Definitions

- 4.1.1. AWS: Amazon Web Services
- 4.1.2. BAA: HIPAA Business Associate Agreement
- 4.1.3. cd: unix command to change directory
- 4.1.4. CMA: Clinical Microarray
- 4.1.5. Checksum: Count of the number of bits in a transmission unit that is included with the unit so that the receiver can check to see whether the same number of bits arrived. If the counts match, it's assumed that the complete transmission was received.
- 4.1.6. Cloud/Cloud Computing: The practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.
- 4.1.7. HC: Hospital Center
- 4.1.8. HCUID: Hospital Center Universal Identifier
- 4.1.9. devel: Development
- 4.1.10. EC2: Elastic Cloud Computing
- 4.1.11. FASTQ: Raw NGS sequencing file
- 4.1.12. HIPAA: Health Insurance Portability and Accountability Act
- 4.1.13. HPCC: High Performance Computing Center

4.1.14. https: Secure Hypertext Transfer Protocol
4.1.15. IRB: International Review Board
4.1.16. Local/On-premises: Software is installed and runs on computers on the premises (in the building) rather than at a remote facility, such as a server farm or cloud (somewhere on the internet). This may also describe storage hardware.
4.1.17. ls: unix command to list directory contents
4.1.18. NGS: Next Generation Sequencing
4.1.19. Pipeline: Combination of different software packages, applications, scripts and databases. It includes the specific version of each component and associated configurations (e.g. command lines or other configuration items).
4.1.20. Raw data: Raw data is defined differently for different platforms. Raw data for NGS sequencing performed on Illumina platforms is considered to be fastq files. Raw data for NGS sequencing performed on Thermo Fisher Platforms is considered to be bam files since bams store flow space information which is stripped when bams are converted to fastqs.
4.1.21. rsync: Rsync is an algorithm typically used to synchronize files and directories between two different systems while minimizing network usage. Rsync will use SSH to connect as user to a remote-host.
4.1.22. SFTP: Secure file transfer protocol
4.1.23. SSH: Secure Shell, an industry standard method of connecting to remote systems (typically Unix) for command-line access.
4.1.24. SSL: Secure socket layer is a standard security technology for establishing an encrypted link between a web server and a browser.
4.1.25. TSM: IBM Spectrum Protect, or Tivoli Storage Manager (TSM) is a data protection platform that gives enterprises a single point of control and administration for backup and recovery.
4.1.26. UUID: A universally unique identifier (UUID) is a 128-bit number used to uniquely identify information in computer systems.
4.1.27. VCF: variant calling file
4.1.28. VPN: Virtual Private Network
4.1.29. Workstation: A local desktop

4.2. Data Security and Confidentiality
4.2.1. In the absence of a HC IT Director, the Bioinformatics Director shall make all necessary decisions pertaining to data security and confidentiality at HC, but shall do so in accordance with the hospitals overall policies.
4.2.2. Access Rights
4.2.2.1. Non-Hospital integrated hardware: Access to computing equipment is granted by the Bioinformatics Director or Bioinformatics Supervisor. Usernames and passwords are generated by the Bioinformatics Director or his designee, e.g. Bioinformatics Supervisor or System Administrator.
4.2.2.2. Hospital integrated hardware: Username and passwords to equipment that is integrated into
the hospital LDAP configuration is controlled by hospital IT although access may be

granted by the IT Director or his designee.

4.2.2.3. User groups

4.2.2.3.1. root

4.2.2.3.1.1. The IT Director, Bioinformatics Director and their designees, e.g. Bioinformatics Supervisor and/or System Administrator, may have root access.

4.2.2.3.2. bioinfo

4.2.2.3.2.1. Users in this group shall have access rights and permission to access and modify non-clinical software, associated packages, configurations and databases according to standard operating procedures.

4.2.2.3.3. bioinfoclin

4.2.2.3.3.1. Users in this group shall have access rights and permission to access clinical data.

4.2.2.3.3.2. Users in this group shall have access rights and permission to access and modify clinical software, associated packages, configurations and databases according to standard operating procedures.

4.2.2.3.4. clinical

4.2.2.3.4.1. Users in this group shall have access rights and permission to access clinical data.

4.2.2.3.4.2. Users in this group shall have access rights and permission to access and modify clinical software, associated packages, configurations and databases according to standard operating procedures.

4.2.2.3.5. research

4.2.2.3.5.1. Users in this group cannot access clinical data or pipelines.

4.2.2.3.5.2. Users in this group shall have access rights and permissions to research data. If the samples are human subjects, the user may need IRB approval.

4.2.2.3.5.3. Users in this group have access to software deemed *research* access only.

4.2.2.3.6. hla

4.2.2.3.6.1. Users in this group can access clinical or research data or pipelines pertaining to the HLA lab only.

4.2.2.3.6.2. Users in this group shall have access rights and permissions to research data. If the samples are human subjects, the user may need IRB approval.

4.2.2.3.7. smbgroup

4.2.2.3.7.1. Users in this group include instruments that generate raw data, or require access to configuration files for operation.

4.2.2.4. Reserved user names

4.2.2.4.1. root: Reserved for System Administrator.

4.2.2.4.2. binfouser: Clinical pipelines may be run using binfouser.

4.2.2.4.3. clinicaluser: Reserved for users that require access to clinical data.

4.2.2.4.4. cmauser: Reserved for users that require access to clinical microarray data.

4.2.2.4.5. HCuser: Reserved for users that require access to clinical data. Clinical pipelines may be run using HCuser.

4.2.2.4.6. ruser: Reserved for System Administrator and Bioinformatics Supervisor for access to the Genetrix Production Workstation.

4.2.2.4.7. researchuser: Reserved for users that require access to research data only.

4.2.2.4.8. Deprecated user names

4.2.2.4.8.1. bmiuser: Reserved for System Administrator.

4.2.2.4.8.2. hlauser: Reserved for users that require access to human leukocyte antigen (HLA) data.

4.2.2.4.8.3. ngsuser: Reserved for clinical users, laboratory technicians or instrumentation that require access to data generated from NGS-based instruments.

4.2.2.4.8.4. sshuser: Reserved for general ssh user access.

4.2.2.4.9. Instrument-based user names are user accounts with access to specific instrument types. Access is reserved for clinical users, laboratory technicians or instrumentation that requires access to data generated from the specified instrument.

4.2.2.4.9.1. equipmentuser: Reserved user account with access to all equipment folders. Considered a super user for all equipment.

4.2.2.4.9.2. 3730user: 3730 DNA Analyzer

4.2.2.4.9.3. affyuser: Affymetrix array based instrument

4.2.2.4.9.4. hiseq4000user: HiSeq4000 sequencer

4.2.2.4.9.5. ionadmin: Applies to Thermo Fisher hardware only. Equivalent to root access.

4.2.2.4.9.6. iscanuser: iScan array scanner

4.2.2.4.9.7. misequser: MiSeq sequencer

4.2.2.4.9.8. nextseq500user: NextSeq500 sequencer

4.2.2.4.9.9. pgmuser: personal genomics machine (PGM)

4.2.2.4.9.10. protonuser: Proton sequencer

4.2.2.4.9.11. sbsuser: Applies to Illumina hardware only. Equivalent to admin access.

4.2.2.4.9.12. s5user: S5/S5XL sequencer

4.2.2.5. Logging: Username and access rights shall be logged in the *FORM.HC-40101.1-Computing Equipment Access Log.*

4.2.3. Maintenance

All maintenance by vendors or personnel, including but not limited to operating system updates, replacement of hard drives, network cards, etc. will be logged in the *FORM.HC-C04101.2-Computing Equipment Maintenance Log*.

4.2.4. Security

4.2.4.1. All clinical data is transferred within Hospital's network except in certain cases when using vendor supplied cloud computing services such as Amazon EC2 (i.e. Cartagenia BENCHLab NGS, transferring data per patient or physician request, and archiving of deidentified data in Amazon S3 or Glacier storage services).

4.2.4.2. Access to Hospital's network is only available while onsite, unless using VPN access.

4.2.4.3. IDF/HC Server Room

4.2.4.3.1. Physical access is restricted to:

4.2.4.3.1.1. Director of Bioinformatics and his designees.

4.2.4.3.1.2. IT personnel

4.2.4.3.1.3. Pre-approved vendors for maintenance or other services.

4.2.4.3.2. Access is restricted by a locked (keyed) entry.

4.2.4.3.3. Access to and from the room is logged with name, time of entry and departure and reason for entry.

4.2.4.4. University of Southern California (USC) HPCC

4.2.4.4.1. Physical access is restricted to the Director of Bioinformatics and his designees.

4.2.4.4.2. Network access is restricted to:

4.2.4.4.2.1. Director of Bioinformatics

4.2.4.4.2.2. Director of Bioinformatics's designee

4.2.4.4.3. There are at least three biometric security measures when attempting to get physical access to the hardware.

4.2.4.4.4. Hospital hardware and clinical data are segregated from research data and non-Hospital hardware, and is HIPAA compliant.

4.2.4.4.5. There are no dark internet connections.

4.2.4.5. All personnel shall follow Hospital and Hospital IT policies with respect to data access, confidentiality and security. Any questions or discrepancies in policies shall be brought to attention of the HC Bioinformatics Director.

4.2.5. HIPPA Compliance

All users with access to clinical data must undergo HIPAA training according to the standards and guidelines set by the institution.

4.3. Data Transfers

4.3.1. All clinically reported cytogenetic, microarray and NGS data are generated in-house.

4.3.2. No data pertaining to clinical NGS tests are transferred to external reference laboratories or other service providers except for clinical reports that may be part of the patient record.

4.3.3. Data may be requested by the ordering physician or patient/guardian in the form of fastq or Filtered VCF files for NGS data.

4.3.3.1. Data files that are sent are stripped of HCUID which is replaced by a UUID.

4.3.3.2. The sample name, which includes HCUID is built into the file name and header of the files.

4.3.3.3. In situations where data is requested for duo, trio, quad or other family exome, or similar NGS test, data are provided for the proband and related samples.

4.3.3.4. All data files are accompanied with a checksum to ensure that the full file is transmitted and not corrupted.

4.3.4. Data transmission

4.3.4.1. Data transferred outside of the Hospital network are encrypted during transfer. Example protocols include SFTP, HTTPS and rsync via SSH.

4.3.4.2. AWS

4.3.4.2.1. Data may be transferred to AWS using one of two methods, by internet or via an encrypted storage device through their AWS Snowball Appliance service.

4.3.4.2.2. Connections to AWS via web browser are done via https.

4.3.4.2.3. Connections to AWS may also be done via SSH protocol.

4.3.4.2.4. AWS maintains HIPAA compliant servers/storage in the cloud via their various storage and process services, i.e. EC2, S3 and Glacier. A valid BAA exists with AWS.

4.3.4.2.5. The primary archive for NGS data is AWS Glacier.

4.3.4.3. Cartagenia BENCHLab NGS

4.3.4.3.1. Connections to BENCHLab NGS are done via https.

4.3.4.3.2. BENCHLab NGS database is located on a secured, HIPAA compliant server on Amazon EC2. A valid BAA exists with Agilent Technologies.

4.3.4.4. University of Southern California (USC) High Performance Computing Center (HPCC)

4.3.4.4.1. Server backups including software backups are done via TSM to USC HPCC. Scheduling and frequency of backups are listed below in the TSM section.

4.3.4.4.2. Storage of Hospital data is on a hardware and data segregated HIPPA compliant location.

4.3.4.4.3. Access to the primary storage is restricted to the System Administrator, Bioinformatics Director and his designee.

4.3.4.4.4. Access to the primary storage is limited as specified in the Data Recovery section below.

4.3.4.5. Data Access

4.3.4.5.1. Access to clinical and research data is specified by *HC-C04101 Computing Equipment*. The combination of user access rights/permissions of folders, data and equipment, and software continuity ensure that clinical patient data are secure and accessible by only those trained individuals who meet HIPPA compliance.

4.3.4.5.2. Transfer of clinical data generated within the Center to personal external drives, transmission via personal email, or storage on cloud sharing services such as but not limited to Dropbox or Google Drive is strictly prohibited per hospital policy.

4.3.4.5.3. All transfer and storage of data should meet Hospital guidelines and meet HIPAA or IRB compliance where applicable.

4.3.4.6. De-identification

4.3.4.6.1. Samples are de-identified and assigned a sample identifier as specified by *HC-C04102 File Naming Convention*.

4.3.4.6.2. Data that are shared for research purposes or shared with vendors for the purpose of debugging software or analysis issues are assigned a PLMUID before transfer.

4.3.4.6.3. Data that are shared by request from the patient, guardian or physician are assigned a UUID.

4.3.4.6.4. There is no identifying patient information in any raw NGS data or downstream generated data analysis. Anonymous identifier, HCUID is used instead, which may be embedded in the header of VCF files and other analysis files, or used to name raw sequence and other data files.

4.3.4.7. Encryption

4.3.4.7.1. Remote access to servers via command line and remote command execution via scripts occurs via ssh. Ssh is a cryptographic network protocol for secure data communication.

4.3.4.8. Communication

4.3.4.8.1. All communication regarding patient samples are to abide by Hospital policy and conform with HIPAA compliance.

4.3.5. Any electronic communication is to occur via Hospital email only. Personal email accounts are not acceptable.

4.4. Data Retention

4.4.1. Storage Policy

All raw and processed data generated on vendor supplied instruments/equipment are copied/moved to local storage for temporary storage and/or downstream analysis and processing. For Illumina NGS data, raw data are considered fastqs and logs. For Thermo Fisher NGS data, raw data are considered bams and logs. There are files that come before fastqs (bcl) for Illumina sequencers, and bams (dat) for Thermo Fisher Scientific sequencers, but they are prohibitively large to store long term. It should also be noted that bams exist as the raw data format for Thermo Fisher Scientific sequencers because they contain flowspace information. This data is stripped if the bam is converted to fastq. All raw data moved to local storage is automatically archived offsite to Primary Archive in AWS Glacier. Raw data is stored indefinitely. VCFs and other downstream processed data such as coverage and alignment files are not required to be stored in archive. VCF files may be stored for other purposes such as for populating a variant database for use in clinical interpretation, or test development.

4.4.2. Data Retention Period

4.4.2.1. Temporary: Defined as temporary files generated by a software application that may be deleted immediately after the associated wet lab process/test as specified by the SOPs for that specific test is complete. These files are not required to regenerate raw data for analysis of a patient sample.

4.4.2.2. Short-term: 30 days post data processing.

4.4.2.3. Intermediate: one (1) year post data processing.

4.4.2.4. Long-term: If a patient was younger than 18 years of age when last treated, until the patient reaches age 21, or for seven years from the date of last treatment, whichever is longer.

4.4.2.5. Indefinite: Undefined but not shorter than the period defined by *long-term*.

4.4.3. Data to be backed up for Indefinite Storage

4.4.3.1. Data stored long-term or indefinitely shall be defined on a test-by-test basis. It is to include at a minimum the file or set of files required to regenerate all downstream files needed for analysis and interpretation. In cases where the files are prohibitively large or prone to data corruption, accommodations shall be made to ensure data integrity. This includes calculating checksum using MD5, CRC32 or another appropriate algorithm as decided by the Bioinformatics Director, and storage redundancy including but not limited to RAID configuration (hardware or software) also as defined by the Bioinformatics Director.

Checksums shall be calculated and confirmed anytime data is transferred between sites (i.e., local versus cloud, and vice versa).

4.4.3.2. Clinical Exome Sequencing test and related tests (focus exomes and derived tests) – fastqs and logs stored indefinitely

4.4.3.3. Onco Sequencing – bams and logs stored indefinitely

4.4.4. Data to be kept locally for Intermediate Storage

4.4.4.1. Processed files are stored locally for at least one (1) year post processing. Processed files include any file output from an analysis pipeline. This can include raw data, coverage files, logs, and quality control data. Data that is deleted from local storage can be regenerated from raw data stored in archive upon request to the Bioinformatics Director, Bioinformatics Supervisor or their designee.

4.4.5. Data to be kept locally for Short-Term Storage.

4.4.5.1. Data generated for the purposes of the Genome Core/Research unless specified by the lab manager/supervisor.

4.4.5.2. Data that are restored to the Restore folder may be kept short term.

4.4.6. Software and Annotations

Software and annotations required to regenerate the final analysis used for final interpretation are retained and versioned. All logs indicating which software version, parameters and commands, and annotation versions are kept in order to regenerate the final analysis. Software and annotations used in clinical data production are kept indefinitely, or at a minimum until all patients whose data were generated by a specific pipeline reaches age 21, or for seven years from the date of last treatment, whichever is longer. Some exceptions may apply including Torrent Suite software, Ion Reporter Software and other software that are dependent on a specific hardware requirement, and whose maintenance is determined by a vendor. For instance, in some cases, software that are updated on a specific hardware cannot be rolled back to prior versions.

4.4.7. TSM Server

4.4.7.1. Use the backup-archive client to store backup versions of your files on the Tivoli Storage Manager server. You can restore these backup versions if the original files are lost or damaged.

4.4.7.2. TSM server manages the IBM LTO4 multi tape drive robotic library.

4.4.7.3. All client backup and restore procedures also apply to the web client, except that the web client does not support a Preferences editor. The web client also does not offer a Setup wizard, which is available in the backup-archive client GUI, on Windows clients.

4.4.7.4. Tivoli Storage Manager provides backup and archive services for all files on the following file systems: File Allocation Table (FAT), FAT 32, NTFS, and ReFS.

4.4.8. TSM Schedule

4.4.8.1. Administrator have set up schedules to automatically back up files.

4.4.8.2. On the Unix side, these two files (dsm.opt & dsm.sys) require to specified which file system is needed to be backup and them TSM backup master server information.

4.4.8.3. Incremental backups schedule to set to Sunday to Friday night runs.

4.4.8.4. Full backups schedule to set to run every Saturday afternoon.

4.5. Data Recovery

4.5.1. Access

4.5.1.1. Data recovery may be performed by the Bioinformatics Director and his designee, e.g. Bioinformatics Supervisor.

4.5.1.2. Access to primary archive servers is specified in *FORM.HC-C04101.1 Computing Equipment Access Log*.

4.5.2. Data recovery instances may include but are not limited to:

4.5.2.1. Data has been removed from local storage based on data retention policies.

4.5.2.2. Data has been destroyed due to hardware failure on local storage.

4.5.2.3. Data has been corrupted during processing/downstream analysis.

4.5.2.4. Human error has led to the premature deletion or corruption of the local copy of data.

4.5.3. Data that has been previously archived need not be archived again once recovered to local storage.

5. PRECAUTIONS:

5.1. Users must not share their login credentials to any system and must log off any system when not in use to prevent unauthorized access to protected data. All staff shall undergo training and yearly review of policies to ensure proper compliance with the data security and confidentiality policies.

5.2. Since some diseases may be very rare, it may be possible to identify a patient's identity from raw sequencing data in the hands of a skilled analyst who may link the data with publicly available resources such as census data, Google, medical websites/journals or news outlets reporting health information

5.3. Once the local copy of patient data is deleted from Online/Local Storage, the only copy remaining is on the archive server.

5.4. If data has already been removed from local storage, the remaining copy may be located in the primary archive. Corruption or deletion of data in the primary archive may lead to irreversible loss of data. Although data may be regenerated from stored samples (e.g. blood or extracted DNA), depending on the disease state, there is no guarantee, that a repeat test, will yield the same results. Therefore access, to data on the primary archive is strictly prohibited to authorized personnel except in special instances illustrated above.

5.5. While the probability that a UUID will be duplicated is not zero, it is close enough to zero to be negligible.

6. PROCEDURES:

6.1. Data Transfers

6.1.1. Before data can be shared with a patient or requesting physician, the *Release of Sequence Data Consent Form* must be completed.

6.1.2. Identify the samples and file types to be retrieved.

6.1.2.1. Identify if data to be retrieved is for singleton, duo, trio, quad or other family case.

6.1.2.2. Identify test for which data is to be retrieved.

6.1.2.3. Use patient identifying information such as name and date of birth to map to patient HCUID if needed. The HC Sample Portal may be used for this purpose as specified in *HC-C04202 - HC Sample Portal Patient Data Loading and Processing Training.*

6.1.2.4. Identify if data has already been archived.

6.1.3. If the data have not been archived, navigate to the clinical storage location. Accessing this location via the command line directly on the storage server may appear differently than when the same folder location is SAMBA mounted on a pc with Windows or Mac.

6.1.3.1. Linux path: 10.241.8.11://gpfs/fs1/HC-raid/clinical/Patient-Analysis

6.1.3.2. Windows path via SAMBA: clinical:\\Patient-Analysis

6.1.4. If the data have already been archived and deleted from local/online storage, follow the *Data Recovery* procedure below in the next section and then continue with the procedure for *Data Transfers*.

6.1.5. Create a case folder in the following folder clinical:\\Data_Transfer\ using the next available sequential Data Transfer ID (e.g. HCcase1) in the *Data Transfer Log FORM.HC-C04107.2.*

6.1.5.1. e.g. clinical:\\Data_Transfer\HCcase# where '#' is the next available sequential case number.

6.1.5.2. Copy the requested data files to the HCcase folder where they will be renamed, and all HCUIDs replaced with a newly generated UUID.

6.1.6. Generate a UUID using the online site: https://www.uuidgenerator.net, e.g. 1ad6b77f-e9be-4b92-8d34-3240ab2de41c

6.1.7. Ensure that the UUID has not been used. This can be done by comparing UUIDs to all other IDs in the *Data Transfer Log FORM.HC-C04107.2*.

6.1.8. If the UUID has not been used, for all files replace the full sample name, e.g. HC0000001-B-D-20180101 with the UUID in the file name and header/content of the files. Repeat the above steps for all samples pertaining to a test case (i.e. all three samples for a trio case).

6.1.9. Create a sample folder for each sample in a case in the HCcase folder. e.g. clinical:\\Data_Transfer\HCcase#\UUID

6.1.10. Move the files to the appropriately named sample folder.

6.1.11. Copy data to AWS for data transfer

6.1.11.1. Using the same HCcase#, create a user account in AWS S3.

6.1.11.2. For each sample, create a folder using the UUID as the name. The parent folder should be named HCcase#. e.g. HCcase1\1ad6b77f-e9be-4b92-8d34-3240ab2de41c\

6.1.11.3. Using a secure connection (SSL, SSH or SFTP), copy the data files to the appropriate UUID folder.

6.1.11.4. Before releasing login information and instructions to download the data, verify HCUID sample name has in fact been stripped from the file names, headers and content. **In some operating systems, renaming a gzip file (current VCF file format), will not actually rename the embedded file name.**

6.1.12. Complete the *Data Transfer Log FORM.HC-C04107.2.*

6.2. Data Recovery

6.2.1. Notify the Bioinformatics Director or his designee, in the event that data must be recovered from the primary archive.

6.2.2. Complete FORM.HC-C04107.1. Indicate the:

6.2.2.1. HCUID

6.2.2.2. Sample ID

6.2.2.3. Project ID

6.2.2.4. Protocol ID

6.2.2.5. Date of sequencing run.

6.2.2.6. Reason for data recovery.

6.2.2.7. Files/folders to be recovered.

6.2.3. SSH into the Primary Archive machine using your username and password credentials. Refer to *HC-C04201- Servers Log On Process*.

6.2.4. Navigate to the archive location for the particular sample and project, e.g. *archive://clinical/Patient-Analysis/HCUID/TestID/sampleID*, by typing the command: cd *clinical/Patient-Analysis/HCUID/TestID/sampleID*

6.2.5. Type *ls -lah* in the command line to verify the folder contents.

6.2.6. To restore data:

6.2.6.1. In the event of an accidental deletion or corruption of data, in the command line type the command*: rsync -rtzv PROJECTID_PROTOCOLID_SAMPLEID_MMDDYY ngsuser@10.241.8.11:*//gpfs/fs1/HC-raid/clinical/Patient-Analysis/HCUID/TestID/sampleID where *PROJECTID_PROTOCOLID_SAMPLEID_MMDDYY* is the data to be copied and *10.241.8.11://gpfs/fs1/HC-raid/clinical/Patient-Analysis/HCUID/TestID/sampleID* is the destination of the new copy.

6.2.6.2. In the event of a data transfer request, in the command line type the command*: rsync -rtzv PROJECTID_PROTOCOLID_SAMPLEID_MMDDYY ngsuser@10.241.8.11://gpfs/fs1/HC-raid/clinical/**Data_Transfer**/Patient-Analysis/HCUID/TestID/sampleID*
where *PROJECTID_PROTOCOLID_SAMPLEID_MMDDYY* is the data to be copied and *10.241.8.11://gpfs/fs1/HC-raid/clinical/Data_Transfer/Patient-Analysis/HCUID/TestID/sampleID* is the destination of the new copy.

6.2.6.3. In the event of a data restoration request (for the purpose of reinterpretation), in the command line type the command*: rsync -rtzv PROJECTID_PROTOCOLID_SAMPLEID_MMDDYY ngsuser@10.241.8.11://gpfs/fs1/HC-raid/clinical/**Restore**/Patient-Analysis/HCUID/TestID/sampleID*
where *PROJECTID_PROTOCOLID_SAMPLEID_MMDDYY* is the data to be copied and *10.241.8.11://gpfs/fs1/HC-raid/clinical/Restore/Patient-Analysis/HCUID/TestID/sampleID* is the destination of the new copy.

6.2.7. In the event required intermediate files are not available, downstream and intermediate analysis files may be regenerated from raw data, as per original SOPs and software analysis pipelines.

7. REFERENCES:

7.1. Data Retention Workflow v6.20180809.pdf
7.2. Amazon Web Services (BAA) Business Associates Agreement
7.3. Cartagenia (BAA) Business Associates Agreement
7.4. Release of Sequence Data Consent Form

8. RECORDS:

Record Number	Title	Pages
FORM.HC-C04107.1	Data Recovery Log	1
FORM.HC-C04107.2	Data Transfer Log	1