1. Introduction

A. What does archiving mean in the digital age?

Archiving digital data is the act of permanently or semi-permanently moving stale data from a highly utilized device to a long term alternate storage device.

B. What is stale data?

Stale data is digital data that has not been used within a specified time frame. The length of the time frame can vary greatly from weeks to years and is dependent on data volume and available storage capacity.

C. Why use alternate storage?

The point of alternate storage verses increasing server space is two-fold:

1) **Money!** High performance storage attached to high performance servers is purchased at a premium and contains many features to ensure data is always instantly available. Archive media is cheaper than high performance storage and is designed to hold onto data, not provide access to data instantly on demand.
2) **Conservation!** Data archived to offline media reduces energy costs and reduces demand for premium server space, climate control systems, and backup servers. It also allows for more efficient utilization of high performance storage for active projects.

D. What's the catch?

The catch is that not all data will always and forever be immediately accessible. However, in return, the data will be retrievable from archive and can be replaced on the server should the need arise.

E. Why set our own policy?

Quite frankly, we have too much data. Some of the data does not even belong to the Genetics Department but is using resources funded by the Genetics Department. We have been forced into a position where we need to set some ground rules for how long we can hold onto data and try to implement it in a reasonable way.

F. Scope?

This policy only applies to departmental multi-user computer systems (like microscope computers and servers, as identified in Section 3 below). Personal computers and computers intended for individual use by specific users or in individual labs or offices are not covered by this policy.

G. How was this policy developed?

The computer systems administrator (Ulrich Neuss) developed an initial draft policy. Input from the most computationally involved faculty was solicited and changes were made. The reviewed policy was then submitted for review by the infrastructure committee and presented to the entire faculty. Inputs from each review are reflected in this data policy.

2. De-identification of Data

All data brought to the department, used with department equipment, must be de-identified. We are not compliant and cannot allow PII of any sort on our systems. You accept the responsibility of ensuring that all information and samples are de-identified before use.

3. Archive Policy (Data Collection Computers & Transient Data Systems)

A. Data Collection Computers

- Data Collection Systems will be defined as any computer and/or device that is directly involved in the generation/collection of raw data. For example this includes all computers directly attached to microscopes, the Nano-drop and the HiScan.
- Data should be removed from Data Collection systems immediately.
- Data Collection systems are NOT backed up by the department.
- Data Collection systems should not be thought of as storage devices.

B. High Volume Transient Data Systems

- **The GA Pipeline**
  - The system will not store raw data files nor are the raw data files backed up (*.tif)
  - Bcl or raw data files required to rerun CASAVA will be retained for 30 days and then deleted from the GApipeline
  - Results files (sequence, export, config, log, demultiplexing files) will be stored for 6 months after the run, users will receive a warning 3 months before the 6 month expiration deadline.
  - Users wishing to retain data beyond these cutoffs must provide their own storage and arrange for data transfer before the respective deadlines.

- **The Slide Scanner**
  - Data generated on the slide scanner is to be immediately removed and transferred to another storage location.
  - The Slide Scanner will NOT be backed up.
  - Data moved to the image analysis server will be stored for 3 months after the run. Users will receive a warning 1 month before the 3 month expiration deadline.
  - Users wishing to retain data beyond these cutoffs must provide their own storage and arrange for data transfer before the 3 month deadline.

4. Analysis Servers (Evolution/Genomecruncher/Helix/Mendel)

A. Archive Policy

- Data older than 1.5 years should be archived to external media.
- Projects involving large downloads should be archived 30 days after project completion.
- Analysis Servers ARE backed up!
- Faculty still affiliated with the department will still follow the 1.5 year activity rule.
- Terminated Faculty will also have their data archived after 1.5 year.
- Terminated Techs and Students will follow the 1.5 year activity rule, as current lab members my still be accessing the terminated person's data.

B. Archive Media/Data retention

Data that has been archived will be retained in archive for 5 years. Users wishing to retain data in archive for longer must make their own arrangements to replace aging storage media.

C. Archive Policy Exceptions

Users having purchased their own storage for the use of their lab may elect to have their directly purchased storage exempted from the department archive policy or define a different archive policy.

D. Backup Policy

- Full backups happen once per month on all servers.
- Incremental backups happen daily after the full backup is completed on each server.
- Backups are retained for 1 month allowing for roll back of about 1 month.
- Only user data and configuration settings are backed up. No bare metal restore.
- Servers should have operating system drives imaged every 6 months during a patch cycle.
- Users having purchased their own storage must make arrangements to purchase sufficient additional storage to allow for backup of their own storage. Users are encouraged to add storage to the current department backup system; otherwise they must accept all consequences of potential data loss.

I acknowledge that I have read and agree to this policy.

☐ I am also a faculty member and I agree to these terms on behalf of my entire lab.

Name: _____ Lab: _____
(Please print)

Signature: _____ Date: _____