Sequencing File Mover: A Tool for the Management of Sequencing Data

Kevin A. Duerr, Michelle Z Mawhinney, Kurt N. Hetrick, Sean M.L. Griffith, Alice M. Sanchez, Alysen B. Robinson,
Brad Tibbils, Brian D. Craig, Janet L. Goldstein, Lee Watkins, Jr., Kimberly F. Doheny
Center for Inherited Disease Research (CIDR), Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD

Introduction:

CIDR:

The Center for Inherited Disease Research provides high quality sequencing and genotyping services and statistical genetics consultation to investigators working to discover genes that contribute to disease.

Sequencing File Mover:

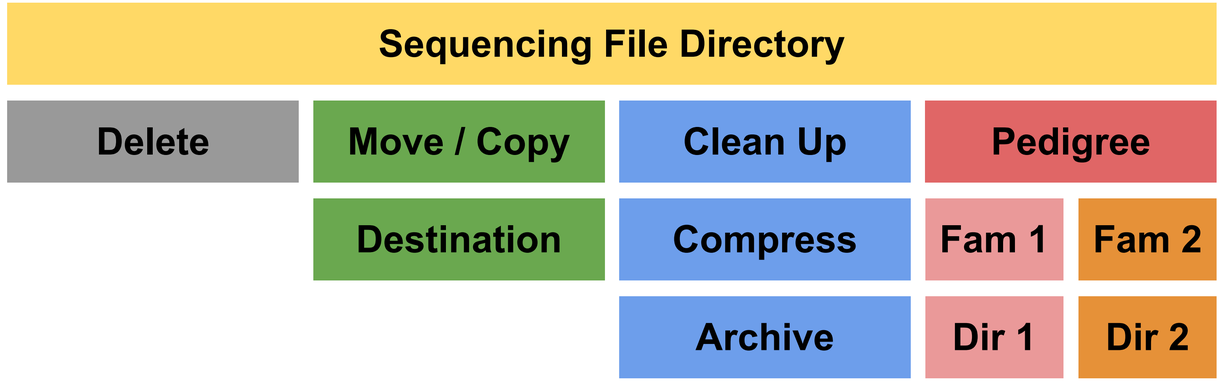
- Java desktop application created to aid lab management in filtering, moving, copying, and deleting large quantities of files of diverse types.
- Features a custom set of file manipulation modes that have been designed around the needs of sequencing lab managers.
- Each mode previews queued changes to give the user a clear representation of their submission.
- Performance parameters can easily be altered.

Frequent changes to lab workflows create a clutter of data, and when the data is massive, organization and proper storage can become difficult. The focus of this poster is to describe the steps to create a sequencing data management tool and the benefits of a specialized lab application.

Development:

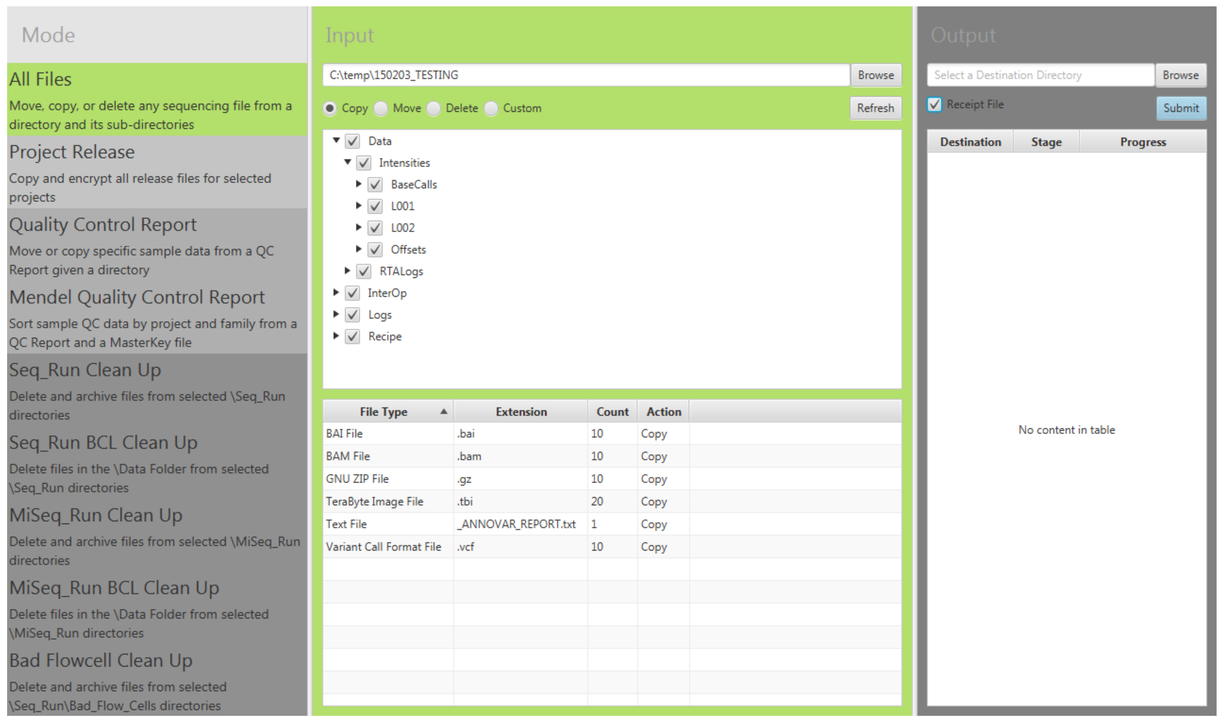
- Originally operating solely to copy or move a specific set of files from a selected directory, the application evolved to accept additional metadata, to be used during file management. For example, using pedigree information, the application can sort files and create a uniform directory structure for each family.
- Selecting a directory creates a tree view of all subdirectories, with the option to check or uncheck whichever directories the user would like to include in their submission.
- Multiple options were added to clean up vestigial sequencing data files and directories. Each of these options have been customized to delete residual files and transfer desired files to a compressed archive based on the type of data generated within a specific time period.
- Once parameters are set, the application gives counts of file type or directory size, specific to the mode selected, giving the user a clear view of what will happen upon submission. In cases where pedigree information is used, counts are made for family size to be referenced to file counts.
- Upon submission, the application will first check if the selected destination already contains any of the chosen files and will display a warning if needed.
- After submission, to gauge wait time, the application displays a progress bar, estimating the time remaining.
- To keep track of all file modifications, the application will create a receipt file that is updated for each file that is manipulated while running.
- If the user wishes to cancel the submission, the user has the option to revert any changes that have been made. The application will then either replace moved files or delete copied files.

File Pipeline:



Results:

- Previously, the amount of time taken to move or remove terabytes worth of data manually on a monthly basis was prohibitive and error-laden.
- With the help of this application, lab management can submit their mass file manipulation request in a few minutes, resulting in significant time savings and fewer errors.
- Data organization is now streamlined for consistency, creating a more efficient data environment.



Conclusion:

Correctly sorting through raw lab data is an essential and time consuming part of the lab management routine. The Sequencing File Mover successfully automates tedious lab data management and creates a more organized and efficient lab environment.

Future Enhancements:

- Incorporation of MD5 data integrity checks, currently executed separately.
- Streamlining our data encryption process by integrating our current security options.

JOHNS HOPKINS MEDICINE

CIDR Center for Inherited Disease Research, Johns Hopkins University

www.cidr.jhmi.edu