FAQ for Seven Bridges data transfer

1. **What is happening?**
   We are moving to the cloud! Existing datasets will be moved to cloud storage and PIs will need to take ownership of their data and manage it in the cloud. Moving forward, sequencing and analysis data will only be retained in our local repository for six months. If you do not want your data in the cloud, you will need to make your own arrangements for long-term storage; otherwise, continue reading for more details on using the Seven Bridges Platform.

2. **Why is our data being moved?**
   Our GNomEx data repository is growing nearly exponentially over the past 13 years, particularly with improvements to Illumina sequencing technology. While the Bioinformatics Shared Resource has absorbed the costs of this storage, this has become unsustainable, particularly as our storage infrastructure ages. None of our peer institutions provide long-term data storage. After extensive evaluation and vetting by members of the faculty advisory committee for the High-Throughput Genomics and Bioinformatics Shared Resource and in close consultation with the Directors of those shared resources, as well as Research Informatics, we have decided to move data storage from local to cloud using the Seven Bridges platform for data processing and storage management.

3. **Why Seven Bridges?**
   Seven Bridges provides a number of benefits besides just cloud storage including: fast bioinformatics analysis enabled by access to Amazon Web Services compute resources; point-and-click interface to run many common bioinformatics analyses using common workflow language; point-and-click interface to archive data to Amazon Glacier storage to save on storage costs, ability to securely share data with collaborators around the world, access to public datasets from TCGA (cancer genomes), TARGET (cancer therapies), ENCODE (functional and epi-genetics), CCLE (cancer cell lines), SGDP (ethnic whole genomes), CPTAC (proteomics), TCIA (cancer imaging), and dbGaP (genetic variants, with appropriate authentication).

4. **What about the other GNomEx instances, including UGP, B2B, and the new Clinical GNomEx?**
   These are not affected. This only concerns the main research GNomEx instance.

5. **I have Sanger-sequenced DNA through the DNA Sequencing core (directed by Derek Warner). Will that be transferred?**
   No, this only concerns the sequencing from the High Throughput Genomics core (directed by Brian Dalley), as well as Analyses associated with such data. Microarray data from the HTG core prior to 2010 will also be included in the transfer.

6. **I am not part of the University of Utah. What about my data?**
   For external clients of the High Throughput Sequencing Core and external users of GNomEx, sequencing and analysis data will be maintained for 6 months. Users will need

to make their own arrangements to store their data elsewhere. Data will be deleted from local servers after 6 months.

7. **Is Seven Bridges free?**
   No. HCI is paying the licensing fees associated with Seven Bridges for any University of Utah GNomEx user. The storage and compute costs for individual labs or users will be invoiced through Seven Bridges.

8. **What happens to the data currently on GNomEx?**
   After all legacy data has been transferred to Seven Bridges, users will have 6 months to claim their data, after that time it will be deleted locally, with some exceptions (see below).

9. **How will our data be moved to Seven Bridges?**
   We will be transferring all data to Amazon cloud storage on your behalf. Transfers will be completed in May 2018. Initially, all data will be maintained by HCI for a period of six months (until November 2018). During this time labs should sign up for a Seven Bridges account and take ownership of their data and begin paying for their storage costs. Unclaimed data will be removed in November 2018.

10. **What if we don't want to pay for our data to be stored in Seven Bridges?**
    If you would like to keep your data, but do not want to use Seven Bridges, within 6 months you should transfer your data to another cloud data storage provider of your choosing (e.g. DNAnexus, Amazon S3/Glacier, Google Cloud, Microsoft Azure Cloud). Or you can download your data to a local compute facility (e.g. CHPC, departmental IT). Be aware that storing data on desktop or laptop computers is consider bad practice due to their propensity for failure, loss, etc. Regardless of your choice, your data will be deleted from our local servers after 6 months.

11. **Will you help me move my data to another storage facility instead of Seven Bridges?**
    No. Seven Bridges is allowing us to bulk transfer the GNomEx repository via Amazon Snowballs in an easy and efficient manner. With over 450 labs, we cannot individually manage transfers to other systems.

12. **Is our data safe in the cloud?**
    Yes, data is encrypted both during transfer and in storage in Seven Bridges. Seven Bridges utilizes the Amazon cloud for data storage. Without authorization, the data cannot be accessed or read. The Seven Bridges platform and Amazon maintain the highest standards of data security possible and are trusted by clinical testing labs, hospitals and pharma companies around the world. The data in the cloud is duplicated in multiple data centers in different locations, in case of catastrophic computer failure or natural disaster (redundancy), and the professional hardware is maintained by world-leading IT experts.

13. **Our data is under an (old) IRB protocol that prevents data from leaving the University.**

Many IRBs will allow us to store the data securely offsite. If yours does not, please contact us immediately. Note that the new Clinical GNomEx is not part of this move.

14. **Is GNomEx going away?**
    No, it is not. All High Throughput Sequencing requests and billing will continue to go through GNomEx, and the database entry of all existing and past sequencing requests will still be available. Further, a list of the files and their metadata for all past experiments and analyses will be available on GNomEx.

15. **What about all future high throughput sequencing data?**
    Future sequencing results will be delivered first to the GNomEx repository and then be uploaded automatically to your Seven Bridges account. Sequence files will remain locally for a period of six months, and then will be automatically deleted. Six months should be enough time to analyze, download, and archive files if you are not using Seven Bridges.

16. **What will be left on GNomEx?**
    A manifest file of the files that used to be there and were uploaded to Seven Bridges. For Analysis projects, bigWig and tabix-indexed VCF files will also remain as a courtesy for users who use GNomEx as a track distribution hub for IGV or UCSC browsers (the files will be uploaded too, just not deleted locally).

17. **If we upload new data files to a GNomEx Analysis project (e.g. A5678) in the future, when will the files be deleted?**
    If you deposit files into a GNomEx Analysis project, either manually, or by using the Pysano #a directive, or by a Bioinformatics Shared Resource Core member doing an analysis for you, the files are still subject to the six month time limit and will be removed after that time frame. If the PI has a Seven Bridges account, it will be first archived under the PI's account, before being removed.

18. **How do we claim files on Seven Bridges?**
    When the files are transferred to Seven Bridges, the Sequence Request project identifier numbers (1234R) and Analysis Project identifier numbers (A4567) for each lab will be maintained and organized into a Seven Bridges project for the lab. The Principal Investigator or designated member of the lab will need to sign up for a Seven Bridges account for the lab and provide a payment method (PO#) to Seven Bridges. Seven Bridges has been added to UShop as an authorized vendor (use the non-catalog request form to generate a standing PO). Once an account is established, the lab's GNomEx files will be claimed, owned, and maintained by the lab.

19. **How is Seven Bridges organized?**
    Each lab will be organized as a "division" within the Seven Bridges platform under an enterprise HCI account. Each lab division will have an administrator (initially the lab PI, although additional administrators may be assigned) who will be responsible for the billing, adding or removing users, and creating or removing projects. Each lab member who wants to use or access Seven Bridges will need an account. Projects may be created within the lab division, and users may work on one or more projects. Unlike GNomEx,

projects are not differentiated into raw data (Requests, or 1234R) and processed data (Analysis, or A4567). Data files and workflows may be shared between projects within the lab division, but not between lab divisions. However, users may be invited to other labs, with read and/or write privileges to individual projects within those.

20. **May an individual be part of more than one lab division?**
    Yes, individuals can be invited as a member to another lab division and granted read and/or write access to individual projects by the lab division administrator. Any files written to the project or workflows executed on that data will be charged to the lab who owns the project.

21. **How will Bioinformatics Shared Resource Core members access our data in Seven Bridges?**
    The Bioinformatics Shared Resource will automatically have a user account in every lab division and will be able to see and work on the data, just as with our current GNomEx setup. If we are requested by the lab to perform analysis on their data using the Seven Bridges platform, we will execute workflows as a member of that lab division, and compute costs will be incurred by the lab. If we perform analysis locally, we can upload the results in the project on Seven Bridges.

22. **When and how do I sign up for a Seven Bridges account?**
    We will be sending an invitation via email to all University of Utah PIs who currently have data in GNomEx to sign up with Seven Bridges. PIs must provide a PO # to Seven Bridges in order to take ownership of their data. Once the PI has an account, he or she may send additional invitations to members of their lab to sign up accounts. Please stay tuned for further updates.

23. **Will there be training sessions?**
    Our license agreement with Seven Bridges includes two onsite hands-on training for lab members and administrators. We anticipate the first training session will be held sometime in late May 2018. Please stay tuned. PI's, please choose a representative from your group to attend a training session and send their email address to Tim Parnell (Timothy.Parnell@hci.utah.edu), Associate Director of the Bioinformatics Shared Resource. Faculty are welcome to attend a training session as well.

24. **GNomEx is so easy to use! Will Seven Bridges be harder?**
    Believe it or not, Seven Bridges will actually be easier to use than GNomEx. The web interface is quite intuitive and simple to use to filter, search, and manipulate files. Files will be easily uploaded, downloaded, or archived for long term storage, using either a point-and-click web interface, or in bulk with command line tools.

25. **Can I perform analysis through Seven Bridges?**
    Yes! You will be able to run standard bioinformatics analyses (alignments, variant calling, gene expression comparisons, etc.) by clicking on 'workflows' on the Seven Bridges website… no programming skills necessary. Seven Bridges uses the Amazon compute cloud to run analyses and returns the results to your Seven Bridges account.

Seven Bridges has a number of publicly available workflows immediately available. The Bioinformatics Core will be uploading their pipelines as workflows so that you can run their specialized pipelines also. The compute instances in the Amazon compute cloud are adjusted such that only the necessary compute resources are requested for each job (or program) in a workflow, which minimizes the cost of the analysis. You can also create your own custom workflows by stringing together different programs using a visual interface. Each bioinformatics program is loaded via a Docker instance, keeping resources to a minimum. If you would rather write code than point-and-click, multiple analysis may be strung together into a custom workflow using the Common Workflow Language on the command-line, an open source consortium to standardize bioinformatics workflows.

26. **How much will it cost to store data and run analyses on Seven Bridges?**
    Once a PI takes ownership of the data in the Amazon cloud they will be responsible for paying for the long-term storage of their data and any bioinformatics costs incurred from analyses performed in the Seven Bridges platform. HCI's purchase of the Seven Bridges license enables users to take advantage of significant cost savings for cloud-based storage and computing. For example, if you pay $1,100 for one lane of paired end 125bp sequencing through the High-Throughput Genomics core facility, you generate approximately 50 GB of data. If you choose to have the data automatically uploaded to the Amazon cloud through the Seven Bridges platform it will cost approximately $7.00 to store the 50GB in Amazon S3 for 6 months while you perform bioinformatics analysis. If you use the Seven Bridges app for exome analysis it will cost approximately $2.00 per sample. When you have completed your analyses, you can use the Seven Bridges platform to archive your data in Amazon Glacier storage and it will cost approximately $3.00 per year to store the 50GB of data. This example serves to illustrate that the cost of data storage and compute is marginal compared to the cost of generating the data, and this system enables you to manage and analyze your data efficiently.

27. **Do I have to switch to using Seven Bridges for compute, or can I still use compute resources here locally?**
    Yes, you may continue to use our resources here if you so desire. We will continue to maintain interactive linux servers at HCI and powerful compute nodes at CHPC.

28. **How can I access data files on Seven Bridges to use in other compute resources?**
    You can generate secure anonymous, signed URLs for any file in a project, which can be used to either download or access through http at another compute resource, for example a genome browser or a local compute node at Utah, depending on context or size. Command line tools for bulk downloading files are also available from Seven Bridges.

29. **Storage costs on Amazon S3 are expensive! Is there a cheaper alternative?**
    Yes, you may (and are encouraged!) to archive older analysis and data files to Amazon Glacier, which is a slow-access storage system. This can be done easily through the point-and-click Seven Bridges website. It costs nothing to archive files into Glacier and has a fractional cost for storage compared to S3. However, retrieving files from Glacier

for new analyses does incur a per-file cost, and works via a queued system, i.e. it may take up to a day or more for the files to be retrieved.

30. **Will the legacy GNomEx data be stored in Glacier?**
    Yes, the legacy data being migrated from GNomEx will be stored in Glacier. To avoid a per file cost retrieval, small analysis files (Excel, PDF, image, text, etc files) will be zipped into a zip archive prior to depositing into GNomEx. A manifest file of what is stored in the zip archive will also be generated so that you know what is in there.