Downloading Your Data

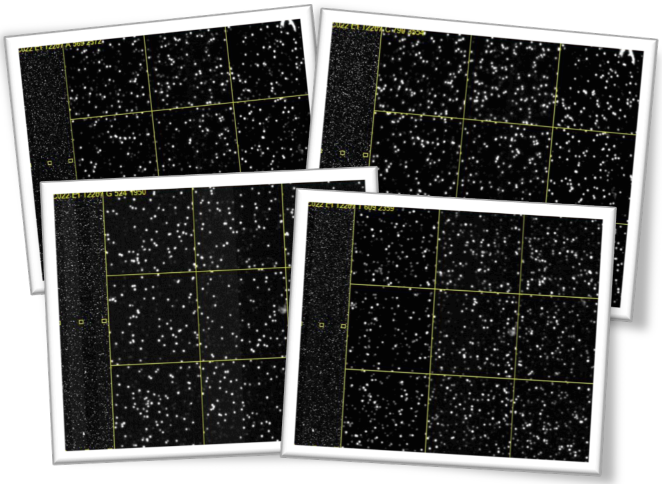
UB Next-Generation Sequencing and Expression Analysis Core

From Our HiSeq2000

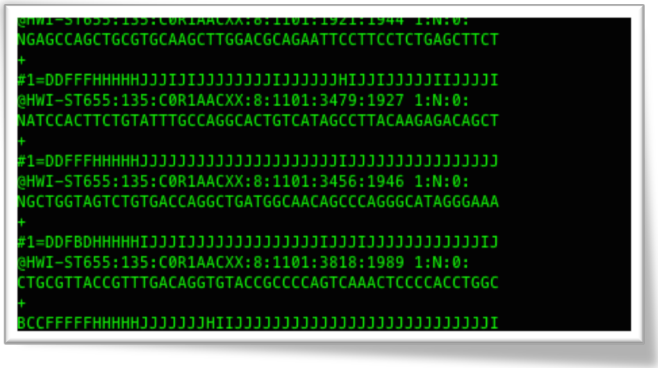
Your data undergoes quite a journey after the sequencing process has finished. It makes its way from the HiSeq2000 platform down to our main server clusters at the Center for Computational Research (CCR). In order to accomplish this large undertaking, we have a direct Ethernet line that connects these two machines to facilitate the fastest and most stable transfer possible.

Once arriving on our servers downstairs, your data is not quite ready for distribution. It arrives in illumina standard formats, which need to be converted to widely accepted industry standards like FASTQ and compressed FASTQ files. In order to do this, the Sequencing Core utilizes illumina CASAVA 1.8.2. The CASAVA utilities enable us to convert the data to FASTQ, while also preforming any necessary demultiplexing of barcodes for pooled libraries. This step varies in time based on the number of multiplexed pools on the flowcell, as well as available processing power.

Finally, your data will be made available to you via a secure FTP web server, where you can download your data to your machine using several different methods.



The machine takes images every cycle, resulting in an image for A,C,G and T bases. This will eventually be converted into your FASTQ files.



```
@HWI-ST655:135:C0R1AACXX:8:1101:1921:1944 1:N:0:
NGAGCCAGCTGCGTGCAAGCTTGGACGCAGAATTCCTTCCTCTGAGCTTCT
+
#1=DDFFFHHHHHJJJIJIJJJJJJJJJIJJJJJJHIJJIJJJJJJIIJJJJI
@HWI-ST655:135:C0R1AACXX:8:1101:3479:1927 1:N:0:
NATCCACTTCTGTATTTGCCAGGCACTGTCATAGCCTTACAAGAGACAGCT
+
#1=DDFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJIJJJJJJJJJJJJJJJJ
@HWI-ST655:135:C0R1AACXX:8:1101:3456:1946 1:N:0:
NGCTGGTAGTCTGTGACCAGGCTGATGGCAACAGCCCAGGGCATAGGGAAA
+
#1=DDFBDHHHHHIJJJIJJJJJJJJJJJJJJJIJJJIJJJJJJJJJJJJJIJ
@HWI-ST655:135:C0R1AACXX:8:1101:3818:1989 1:N:0:
CTGCGTTACCGTTTGACAGGTGTACCGCCCCAGTCAAACTCCCCACCTGGC
+
BCCFFFFFHHHHHJJJJJJHIIJJJJJJJJJJJJJJJJJJJJJJJJJJJJI
```

FASTQ format contains identification information, sequence data and quality scores.

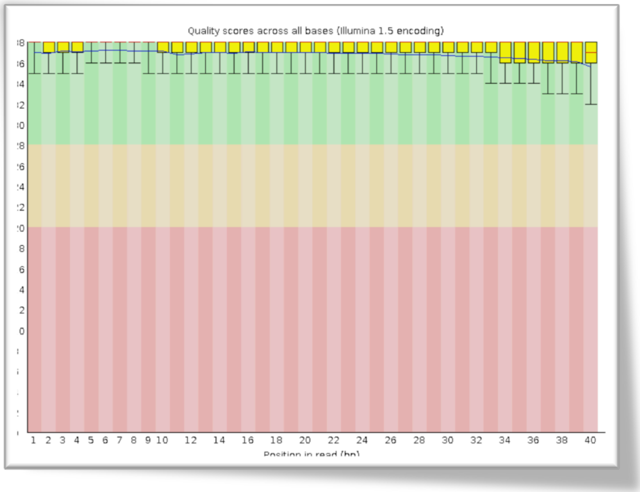
File Formats

As NGS sequencing technologies mature so too do the file formats that support it. It is difficult to keep up with the deluge of information, so this brief guide will only cover the standard file formats that the UB Next-Gen Core distributes routinely.

FASTQ

FASTQ format represents your sequencing data in its most raw form. It stores both the sequencing reads as well as their associated quality encodings. FASTQ file can be utilized in several different ways.

FASTQ files can be run through many different alignment algorithms, such as BOWTIE, BWA or MAQ, for analysis involving a published reference genome. In addition, for analysis that do not have published reference genomes, FASTQ can be used by many different de-novo assembly tools, such as Velvet, MIRA, and SoapDenovo. These programs utilize FASTQ to assemble your sequencing reads into contigs and scaffolds.



FASTQ formatted sequencing data is essential to many downstream analysis tools and pipelines. This picture is of the quality encodings, provided by the FastQC program

.GZ Compression

Due to the increasing volume of sequencing data being generated by more powerful sequencing platforms, the industry has adopted strategies for coping with data, and more specifically data transfer across networks. One tool in our arsenal is the use of the .GZ compression scheme. This format allows us to retain all of the original file, while drastically reducing its size and allowing for quick unpacking.

To decompress a .GZ file, Unix and Linux users can utilize the following command:

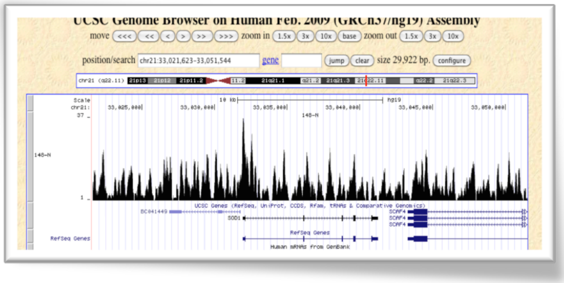
```
gunzip file.txt.gz
```

For those researchers who choose to operate on a PC, there are several free tools available that can decompress this file format, though due to the size of these files, once uncompressed they can be quite unmanageable. For more information please see this article on opening .GZ files on PC:

http://pcsupport.about.com/od/fileextensions/f/gzip-file.htm

BAM

The BAM file format has quickly become the industry standard for post-alignment sequencing information. It is a binary compressed sequence alignment map (SAM) whose most powerful characteristic is its ability to quickly utilize an index to jump from location to location across the genome. This allows a researcher to quickly look at a specific gene or feature without having to read through the entire length of the file.



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

position/search chr21:33,021,623-33,051,544 size 29,922 bp.

BAM Format can be uploaded for direct viewing of your sequencing data to the UCSC Genome Browser, as well as several other viewing tools. For more information please see our guide on Viewing Data in the UCSC Browser

BAM files are not human readable, in other words -- you cannot open these files up in a text editor and read through them.

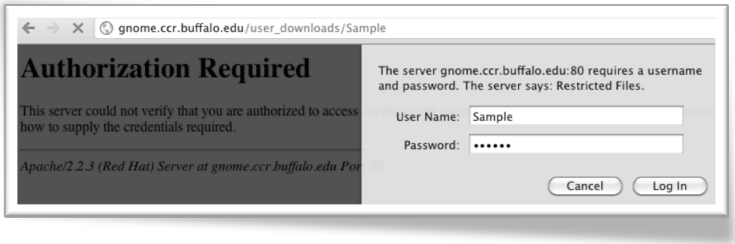
Nearly all major analysis tools have been updated to use this industry standard file type. If you have alignments done at the UB Next-Gen Core, this will be the resulting output file type.

Downloading From FTP via Web Browser

This section contains a step-by-step guide to downloading your data via our FTP server. When your sequencing data is finished through our pipelines described in this document, you will receive a notification email stating that your HiSeq2000 data is now available for download. This email will contain a link to your project specific FTP site, as well as login credentials. Please keep these credentials safe as they are your key

Step 1

Navigate your favorite internet browser to the FTP download page provided to you in the notification email. Please provide your login credentials when prompted.



gnome.ccr.buffalo.edu/user_downloads/Sample

Authorization Required

This server could not verify that you are authorized to access how to supply the credentials required.

Apache/2.2.3 (Red Hat) Server at gnome.ccr.buffalo.edu Por

The server gnome.ccr.buffalo.edu:80 requires a username and password. The server says: Restricted Files.

User Name: Sample

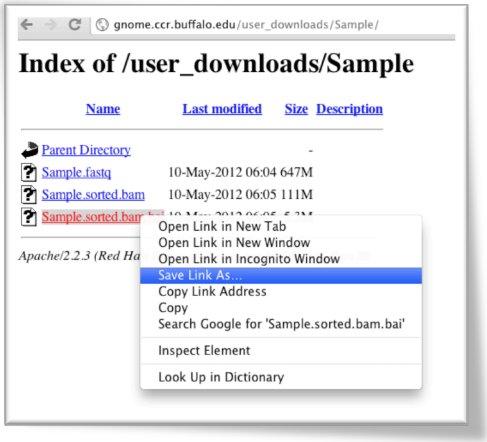
Password: ••••••

Please provide your login credentials exactly as described in your notification email.

Step 2

To download, right click on the file that you are interested in, and click the "Save Link As" option. This will initiate the download.

It is important to note some caveats here. First, Certain internet browsers, particularly older versions of internet explorer will not allow you to download files over a certain size due to software limitations. If this occurs installing an updated web browser will fix this issue. Secondly, you cannot currently download entire directories through the web browser, only individual files. We are actively working to address these limitations.



gnome.ccr.buffalo.edu/user_downloads/Sample/

Index of /user_downloads/Sample

Name	Last modified	Size	Description
Parent Directory		-	
Sample.fastq	10-May-2012 06:04	647M	
Sample.sorted.bam	10-May-2012 06:05	111M	
Sample.sorted.bam.bai	10-May-2012 06:05	5.3M	

Apache/2.2.3 (Red Ha

"Save Link As" will allow you to save files to your computer.

Downloading From FTP via “Wget”

By far the most simple and quickest way to download your data is through the “Wget” command. Wget is a utility through the GNU Project and supports downloading from HTTP, HTTPS and FTP protocols. The only drawback of this tool is that it is a unix/linux utility, thus not available for PC users.

The most powerful tool is Wget’s ability to recursively download. In other words, it can navigate through directories downloading every available file, allowing for one command to pull all of your data all at once.

Using Wget

Using Wget to access your data is actually very simple. Often the hardest part will be to install the utility on your machine if it is not already available. Installing Wget is beyond the scope of this guide.

To download your data from the UB Next-Gen Core FTP site, please issue the following command, substituting “Sample” for your Username and Password:

```
# Download From the UB Next-Gen Core
wget -r --user=Sample --password='Sample' http://gnome.ccr.buffalo.edu/user_downloads/Sample/
```

Downloading Data via External Hard Drive

Though the core does not generally allow for data to be shipped on External Drives, we do make exceptions for large projects where it is not feasible for the data to be downloaded quickly. In this situation we ask the researcher to provide us with a working external drive that the data can be loaded on. We will try to be flexible in the types of drives, and the platform requirements for what you provide to us, however the core is not liable for any issues that may result with the external drives that are provided.

Data Storage Policy

The core provides storage for your data set for thirty days after the notification of completion email is sent. We have limited capacity to restore data removed from our storage system, and if it is possible for your data set, there is a restoration fee that will be assessed based on the amount of data required. We try to give users a notification that their data is slated to be removed when their time is up, however it is the researchers responsibility to download and store their data in a safe and reliable way. If you have any questions on how to do this, please inquire for more information.

More Information

For more information about downloading your data, please contact jbard@buffalo.edu. Additionally, for more information on the file formats talked about in this document, a great resource is the UCSC Genome Browser Data File Format FAQ, which covers the majority of the commonly used file formats.

http://genome.ucsc.edu/FAQ/FAQformat.html