Check-In Procedure for GBS Data from Hudson Alpha

Prerequisites

The following shell scripts and programs that are stored in /homes/mlucas/scripts are required:

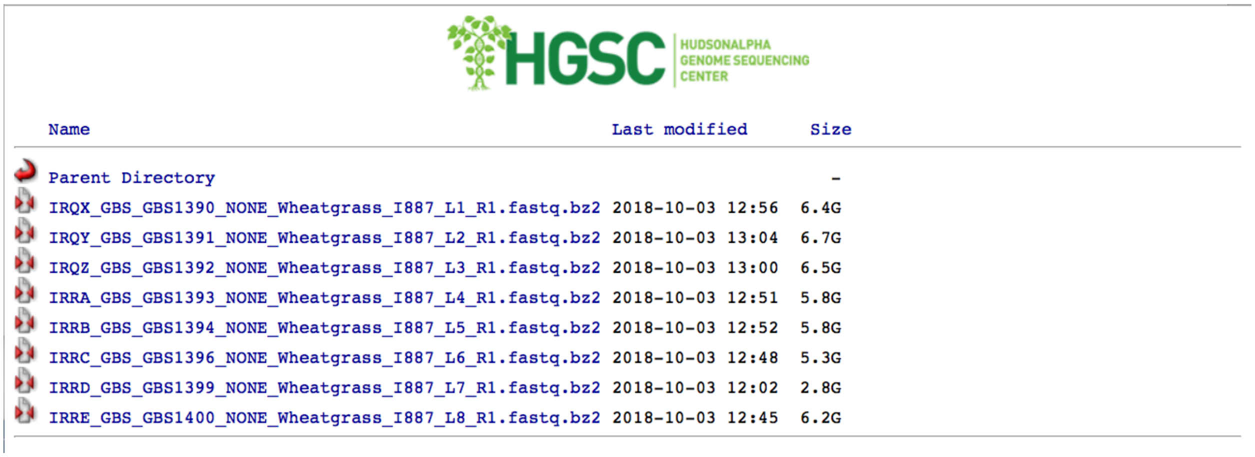
```
checkin_HA_gbs_data.sh

rename_gbs_file

compute_gbs_file_metadata

generate_barcode_distribution

generate_blank_dna_quantification_report
```

Download GBS Data from HA

1. Notification of availability of new data will be received via email from Hudson Alpha – (usually from Jane Grimwood.) The email will contain a URL to access the files along with a username and password.
2. Create a file called .wgetrc in your home directory on Beocat. The file will contain the username and password for the site that you will be downloading from and will allow you to download the files via a shell script containing wget commands without having to enter the Hudson Alpha username/password on the command line. The syntax is:

   ```
   user=<HA username>
   password=<HA password>
   ```
3. Save the file and set permissions to be only read/write for the user.

4. Login to Hudson Alpha website using the link provided in the email from Hudson Alpha. A list of files similar to one shown in Figure 1 below will be displayed.



Name	Last modified	Size
Parent Directory		-
IRQX_GBS_GBS1390_NONE_Wheatgrass_I887_L1_R1.fastq.bz2	2018-10-03 12:56	6.4G
IRQY_GBS_GBS1391_NONE_Wheatgrass_I887_L2_R1.fastq.bz2	2018-10-03 13:04	6.7G
IRQZ_GBS_GBS1392_NONE_Wheatgrass_I887_L3_R1.fastq.bz2	2018-10-03 13:00	6.5G
IRRA_GBS_GBS1393_NONE_Wheatgrass_I887_L4_R1.fastq.bz2	2018-10-03 12:51	5.8G
IRRB_GBS_GBS1394_NONE_Wheatgrass_I887_L5_R1.fastq.bz2	2018-10-03 12:52	5.8G
IRRC_GBS_GBS1396_NONE_Wheatgrass_I887_L6_R1.fastq.bz2	2018-10-03 12:48	5.3G
IRRD_GBS_GBS1399_NONE_Wheatgrass_I887_L7_R1.fastq.bz2	2018-10-03 12:02	2.8G
IRRE_GBS_GBS1400_NONE_Wheatgrass_I887_L8_R1.fastq.bz2	2018-10-03 12:45	6.2G

Figure 1 Example Hudson Alpha File Download Page

5. Create a shell script containing wget commands to download the required files. An example list of commands for the script example *get_HA_GBS1390-GBS1400.sh* is shown below:

```
#!/bin/bash
wget –qc http://www.hagsc.org/restricted_access/wheatgrass/GBS.2018.2/IRQX_GBS_GBS1390_NONE_Wheatgrass_I887_L1_R1.fastq.bz2
wget –qc http://www.hagsc.org/restricted_access/wheatgrass/GBS.2018.2/IRQY_GBS_GBS1391_NONE_Wheatgrass_I887_L2_R1.fastq.bz2
wget –qc http://www.hagsc.org/restricted_access/wheatgrass/GBS.2018.2/IRQZ_GBS_GBS1392_NONE_Wheatgrass_I887_L3_R1.fastq.bz2
wget –qc http://www.hagsc.org/restricted_access/wheatgrass/GBS.2018.2/IRRA_GBS_GBS1393_NONE_Wheatgrass_I887_L4_R1.fastq.bz2
wget –qc http://www.hagsc.org/restricted_access/wheatgrass/GBS.2018.2/IRRB_GBS_GBS1394_NONE_Wheatgrass_I887_L5_R1.fastq.bz2
wget –qc http://www.hagsc.org/restricted_access/wheatgrass/GBS.2018.2/IRRC_GBS_GBS1396_NONE_Wheatgrass_I887_L6_R1.fastq.bz2
wget –qc http://www.hagsc.org/restricted_access/wheatgrass/GBS.2018.2/IRRD_GBS_GBS1399_NONE_Wheatgrass_I887_L7_R1.fastq.bz2
wget –qc http://www.hagsc.org/restricted_access/wheatgrass/GBS.2018.2/IRRE_GBS_GBS1400_NONE_Wheatgrass_I887_L8_R1.fastq.bz2
exit
```

6. Create a shell script containing commands to convert the .bz2 format files to .gz format. An example list of commands for the example script *convert_HA_GBS1390-GBS1400.sh* is shown below:

```
#!/bin/bash
bzcat IRQX_GBS_GBS1390_NONE_Wheatgrass_I887_L1_R1.fastq.bz2  | gzip –c – >
IRQX_GBS_GBS1390_NONE_Wheatgrass_I887_L1_R1.fastq.gz

bzcat IRQY_GBS_GBS1391_NONE_Wheatgrass_I887_L2_R1.fastq.bz2  | gzip –c – >
IRQY_GBS_GBS1391_NONE_Wheatgrass_I887_L2_R1.fastq.gz

bzcat IRQZ_GBS_GBS1392_NONE_Wheatgrass_I887_L3_R1.fastq.bz2  | gzip –c – >
IRQZ_GBS_GBS1392_NONE_Wheatgrass_I887_L3_R1.fastq.gz

bzcat IRRA_GBS_GBS1393_NONE_Wheatgrass_I887_L4_R1.fastq.bz2  | gzip –c – >
IRRA_GBS_GBS1393_NONE_Wheatgrass_I887_L4_R1.fastq.gz

bzcat IRRB_GBS_GBS1394_NONE_Wheatgrass_I887_L5_R1.fastq.bz2  | gzip –c – >
IRRB_GBS_GBS1394_NONE_Wheatgrass_I887_L5_R1.fastq.gz

bzcat IRRC_GBS_GBS1396_NONE_Wheatgrass_I887_L6_R1.fastq.bz2  | gzip –c – >
IRRC_GBS_GBS1396_NONE_Wheatgrass_I887_L6_R1.fastq.gz

bzcat IRRD_GBS_GBS1399_NONE_Wheatgrass_I887_L7_R1.fastq.bz2  | gzip –c – >
IRRD_GBS_GBS1399_NONE_Wheatgrass_I887_L7_R1.fastq.gz

bzcat IRRE_GBS_GBS1400_NONE_Wheatgrass_I887_L8_R1.fastq.bz2  | gzip –c – >
IRRE_GBS_GBS1400_NONE_Wheatgrass_I887_L8_R1.fastq.gz
exit
```

Execute the Scripts to Check-in the GBS Data

7. Logon to Beocat

8. Create a folder in /bulk/mlucas/incoming to store the new HA files, e.g. **HA_IWG_20180901**

9. Store the scripts to download the files and to reformat the files from bz2 to gz in this folder.

10. Execute the script to download the files
    Example:

    ./get_HA_GBS1390-GBS1400.sh

11. Execute the script to reformat the files from bz2 to gz
    Example:

    ./convert_HA_GBS1390-GBS1400.sh

12. cd to /homes/mlucas/scripts and locate the script **checkin_HA_data.sh**

13. Execute the script on the command line with the Illumina project name
    Example:

    ./checkin_HA_gbs_data.sh HA_GBS1390-GBS1400

    The script will perform the following steps:

    a. Update the gbs database table for the associated gbs_id with the flowcell and lane values.
    b. Rename each of the 8 GBS files to a name conforming to the standard GBS file naming standard.

       Example:

       ```
       GBS1390x18SALCycle8P42P43_CCHVGANXX_s_1_fastq.txt.gz
       ```

    c. Compute the MD5 checksum and line count for each GBS file in the set and update the gbs table ms5sum and num_lines columns for the gbs_id associated with each file.

    d. Generate read-barcode distribution report

       This report will allow the user to check % valid reads and % reads found in any blank well in the GBS file.

       The report will have the following naming format:

       GBSnnnn_sample_summary.txt

       Example:

       ```
       GBS1390_sample_summary.txt
       ```

    e. Generate DNA quantification report

       This report will allow the user to check the DNA quantification values for blank wells in the GBS library

       The report will have the following naming format:

       GBSnnnn_blank_dna_quant_report.csv

       Example:

       ```
       GBS1390_blank_dna_quant_report.csv
       ```

Review QC Reports and Cleanup

14. Review the GBSnnnn_sample_summary.txt report and verify that the following thresholds have not been exceeded:

    % Valid reads > 90%
    % Reads in any BLANK well < 0.01%

    If either threshold is violated, investigate potential causes:
    1. Incorrect blank well in DNA plate record
    2. Poor sequencing run quality

15. Review the GBSnnnn_blank_dna_quant_report.csv to make sure that DNA quantification values in the blank wells are within tolerance.

    If the values reported are all NULL, this means that the dnaQuant table has not been updated yet for this GBS plate.

16. Change the group on the GBS file to ksu-plantpath-jpoland and remove write permissions from the file.

    ```
    chgrp ksu-plantpath-jpoland GBS1390x18SALCycle8P42P43_CCHVGANXX_s_1_fastq.txt.gz
    
    chmod a-w GBS1390x18SALCycle8P42P43_CCHVGANXX_s_1_fastq.txt.gz
    ```

17. Move the GBS file to /bulk/jpoland/sequence directory on Beocat.

    ```
    mv GBS1390x18SALCycle8P42P43_CCHVGANXX_s_1_fastq.txt.gz /bulk/jpoland/sequence/.
    ```