PAUL ST. AMAND

USDA BIOINFORMATICS SERVER TRAINING

PRELIMINARIES

- People with UNIX experience, please disperse throughout classroom to help others.
- Does anyone need an account on our server?
- Please ask questions as they arise.

PRELIMINARIES

- This presentation will be a complete walk-through and how-to on:
- 1. Services available on our USDA Bioinformatics Server
- 2. Basic UNIX commands
- 3. TASSEL3 GBS-UNEAK PIPELINE
- 4. TASSEL5 GBS-Reference PIPELINE
- 5. TASSEL MRASeq PIPELINE
- 6. Blast
- 7. Flapjack (time permitting)

PRELIMINARIES

- This presentation will be directly related to the tools on our USDA Bioinformatics server, therefore it is limited to only those working in our USDA labs.
- Should take ~3 hours.
- Scripts are for Ion Torrent Proton sequencing data only. They would need modification for Illumina data.

BIOINFORMATIC TOOLS ON THE USDA SERVER

- https vs http: Server certificates not all in place
- Some pages may need http: other may need https:
- It is OK to have your web browser trust certificates for the following as needed:
  - 129.130.90.3 (USDA Bioinformatics server)
  - 129.130.90.13 (Ion Proton sequencer)

BIOINFORMATIC TOOLS ON THE USDA SERVER

- Lab web site:
- (https://hwwgenotyping.ksu.edu)
- (https://hwwgenotyping.ksu.edu/protocols/)
- (https://hwwgenotyping.ksu.edu/protocols/GBS_protocols/)

BIOINFORMATIC TOOLS ON THE USDA SERVER

- Web services on the USDA server
  - R-Studio (standard ID & password)
    - (http://129.130.90.3:8787/auth-sign-in)
  - Galaxy (additional ID & password)
    - (http://129.130.90.3/galaxy)

BIOINFORMATIC TOOLS ON THE USDA SERVER

- bamtools
- bedtools
- bowtie
- bowtie2
- clustal
- dialign
- docker
- emboss
- fastqc
- fastx-toolkit
- flexbar
- galaxy
- git
- gnuplot
- hmmer
- java
- mummer
- muscle
- ncbi-blast+
- pandaseq
- perl
- phylip
- plink
- primer3
- python
- r
- rna-star
- samtools
- tassel
- tmux
- tophat
- trimmomatic
- trinity
- vcftools

BIOINFORMATIC TOOLS ON THE USDA SERVER

- USDA Bioinformatics server:
- Ubuntu 18.01 LTS
- 12 cores, Intel Xeon CPU X5680 @ 3.33GHz (TB 3.60GHz)
- 24 cores (with hyperthreading)
- Ram 192 GB
- RAID 40 TB, free 16.25 TB

BIOINFORMATIC TOOLS ON THE USDA SERVER

- Rules:
- 1. NO PASSWORD SHARING EVEN WITH OTHERS IN THE LAB.
- 2. No data is backed up on this server. YOU are responsible for backing up your own data.
- 3. Do not install ANY software (even in your own account). Ask Paul to install it for you.
- 4. You should change the password using the passwd command.

BIOINFORMATIC TOOLS ON THE USDA SERVER

- Locations:
- User home: /home/userID
- Share: /home/share
- Scripts: /home/share/tools
- Ref genomes: /home/share/tools/refs

BASIC UNIX

- Connecting to our server: ssh programs (secure-shell, on-campus or via VPN):
- Mac 'terminal'
- PC 'putty' (https://www.putty.org)
- ssh userID@129.130.90.3

BASIC UNIX

- Connecting to our server: FTP programs (file transfer programs, easy to file copying).
- ftp userID@129.130.90.3
  - Mac:
    - 'Fetch' (https://fetchsoftworks.com)
    - 'CyberDuck' (https://cyberduck.io)
  - PC :
    - 'Filezilla' (https://filezilla-project.org)
    - ‘CyberDuck' (https://cyberduck.io)

BASIC UNIX

- Command Description
- pwd print working directory (prints to screen, displays current full path of your location on the filesystem)
- ls list contents of current directory
- ls –lh list contents of current directory with extra details
- ls /home/pst/*.txt lists all files in /home/pst ending in .txt (* is a wildcard that matches anything)
- passwd change password of current user

BASIC UNIX

Command	Description
mkdir temp directory	makes a directory called temp within the current
cd	change directory to your home directory
cd ~	change directory to your home directory
cd ..	change directory UP one level
cd /home/pst/temp	change directory to specific directory, full path
cd temp	change directory to specific directory, relative path
cd -	change directory to previous directory

BASIC UNIX

- Command — Description
- tab — tab-completion, automatically fills in partially typed commands/ file names
- up/down arrows — previous or next commands

- man any-command — gives you help on any-command, usually
- touch myfile — creates a new, empty file called myfile OR updates the timestamp on the file if it already exists, without modifying its contents
- nano new.txt — opens nano editing a file called new.txt. see ribbon at bottom for help. ^x means CTRL-x.
- cp myfile myfile2 — copies myfile to myfile2. if myfile2 exists, this will overwrite it!***

BASIC UNIX

- Command Description
- cat new.txt displays the contents of new.txt
- more new.txt displays the contents of new.txt screen by screen.
  spacebar to page down, q to quit
- head new.txt displays first 10 lines of new.txt
- head -n 100 new.txt displays first 100 lines of new.txt
- tail new.txt displays last 10 lines of new.txt
- tail -n 100 new.txt displays last 100 lines of new.txt

BASIC UNIX

- Command — Description
- mv myfile newname — renames file to newname. If a file called newname exists, this will overwrite it!
- mv myfile newlocdir — moves myfile into the destination directory newlocdir
- rmdir temp — removes directory called temp. temp must be empty
- rm –rf temp — this will delete directory temp along with all its content without asking you for confirmation!***
- rm myfile — removes file called myfile
- anycommand > myfile — redirects the output of anycommand writing it to a file called myfile
- anycommand >> myfile appends the output of anycommand to a file called myfile

BASIC UNIX

- Command — Description
- CTRL-c — kills whatever process you're currently doing
- who — shows all other logged-in users (not R-studio, not Galaxy)
- top — displays all the processes running on the machine, and shows available resources, q to quit, use briefly, then quit
- htop — displays all the processes running on the machine, and shows available resources, more graphic, q to quit, use briefly, then quit
- chmod ugo=rwx sample.sh — sets read, write, execute permissions for user, group, and other
- chmod u=rwx sample.sh — sets read, write, execute permissions for user

BASIC UNIX

- Command Description
- **curl** fromURL toCurrentLoc copy files via web-URL, useful to get Proton fastq files onto server. http://129.130.90.13/report/1047/
- **curl** --user userID:userPW -O http://129.130.90.13/path/file.fastq
- **| tee** -a logfile.txt send output of any command to screen AND to a file
- **nohup** No-Hangup, allows jobs to continue on server even after you disconnect.
- **exit** disconnect from server. ends running programs! always exit and quit your ssh program when done.

BIOINFORMATICS FILE TYPES

- fasta: (.fasta .fa .fas .seq)
- Stores name and sequence for 1 or multiple sequences. Used by many programs.
- Two rows per sequence
- ">" Name
- Sequence

```
>SEQUENCE_1
ACGCGTTCGAGATCGGCGCT
>SEQUENCE_2
CCCGTCGTCTTGAGAGGAGACTCTGCGCAGG
```

BIOINFORMATICS FILE TYPES

- fastq: (.fastq .fq)
- Sequencer output. Stores a formatted sequence and its quality data. 4 lines/sequence read.
- Line 1 begins with a '@' character and is followed by a sequence identifier.
- Line 2 is the raw sequence bases.
- Line 3 begins with a '+' character and is optionally (rare) followed by the same sequence identifier again.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

BIOINFORMATICS FILE TYPES

- hap: (.hap)
- Genotyping data, TASSEL stores calls in this format.
- 1 header row
- Rows are markers
- Genotypes are in columns

BIOINFORMATICS FILE TYPES

- Line endings and file types
- unix = \n (newline)
- windows = \r\n (carriage return, newline)
- mac = \r (carriage return)
- Mac 'BBEdit' https://www.barebones.com/products/bbedit/
- PC 'Notepad++' https://notepad-plus-plus.org

WC AND GREP

- wc file — count characters or words or lines in a file
- wc -l file.fastq — count lines in fastq (divide by 4 for sequences)
- grep — Globally search a Regular Expression and Print
- grep "pattern" file — searches for the pattern in files, and displays lines matching the pattern
- grep "CTAAGGTAACGAT" file — find specific barcodes in reads
- grep -c “>" file — count sequences in a fasta
- grep ">" file — show sequence names in a fasta

QUESTIONS OR BREAK?

USDA TASSEL 'UNIVERSAL NETWORK ENABLED ANALYSIS KIT' (UNEAK) PIPELINE PROTOCOL

- UNEAK is a non-reference Genotyping by Sequencing (GBS) SNP discovery pipeline
- UNEAK is NOT supported in the latest versions of TASSEL.
- https://bytebucket.org/tasseladmin/tassel-5-source/wiki/docs/TasselPipelineUNEAK.pdf

BASIC UNEAK METHODOLOGY:

- 1. Find good reads: expected barcode, RE cut site, no N's. Trims off barcodes, truncates sequences that have a second cut site, or read into the common adapter. Trims reads to 64 bases (including the RE site). Pads truncated reads with poly-A.
- 2. Identify tag pairs for SNP calling via pairwise alignment and the network filter. All tags are sorted, duplicates removed, and all tag pairs with 1 bp mismatch are considered as candidate SNPs.
- 3. Finds the tag distribution in all of the samples. Assigns genotypes to each sample and converts each into a HapMap record.

UNEAK IMPLICATIONS:

- 1. No reference map is needed.
- 2. Only markers polymorphic among current samples are found.
- 3. Mismatches more than 1 bp are ignored.
- 4. Markers in unique sequences (rare, alien, not in reference) can be found.

BASIC UNEAK STEPS:

- 1. Get fastq files.
- 2. Make key file.
- 3. Add poly-A to each read.
- 4. Rename fastq files.
- 5. Run UNEAK jar file.
- 6. Merge and sort hapmap files.
- 7. Get quick output stats.
- 8. Copy files from server and get marker stats in Excel.

DETAILED UNEAK STEPS: 1. GET FASTQ FILES

- Amy will email user with link(s) to fastq files.
- Example: http://129.130.90.13/report/1040/
- user:password, DO NOT GIVE USERID OR PASSWORD TO OTHERS.
- Check notes to see what BC sets were used.
- Login to USDA bioinformatics server, make a working dir, cd to working dir
- Copy URL and use curl to get fastq file(s). curl --user user:password -O URLHERE
- Cancel the curl command for this class.
- /home/share/tools/USDA_Bioinformatics_Server_Training/Small_1*.fastq

DETAILED UNEAK STEPS: 2. MAKE KEY FILE

- Links barcode used to the sample & well.
- Lane is the fastq reads file (library) NUMBER for a set of samples.
- All reads with the same lane & barcode are merged.
- Reads with different lane numbers, but the same barcode are NOT merged.
- Must be a tab-delimited UNIX file type.
- Must have the first 7 required columns with these headers.
- You can have additional columns after the well column. Barcode set number and barcode number are useful for bookkeeping.

DETAILED UNEAK STEPS: 2. MAKE KEY FILE

- Duplicate sample NAMES will have their data merged. Useful for parents.
- You should have more than 1 sample well for each parent or you will have low marker numbers.
- Blank sample wells are not necessary, but help with error checking.
- Flowcell name must NOT have underscore "_", space " ", slash "/", or backslash "\", and must be the same on all lines of key file.
- Barcodes must be correct for each well.
- Copy KEY file to working directory.

DETAILED UNEAK STEPS: 3. ADD POLY-A TO EACH READ

- Ion Proton sequencing reads are of variable length. Illumina reads are all 100 bp.
- TASSEL will not work with reads shorter than 79 bases (BC+RE+Genomic).
- Adding poly-A insures all reads are at least 79 bases long.
- Poly-A addition improves marker finding by keeping slightly shorter reads.
- TASSEL ignores poly-A sections on the 3' end of read.

DETAILED UNEAK STEPS: 3. ADD POLY-A TO EACH READ

- Run the AddPolyA.sh script on EACH of your fastq files: "/home/share/tools/AddPolyA.sh YourFileName.fastq"
- AddPolyA.sh script takes ~10-40 min/fastq.
- AddPolyA.sh script does nothing to the original file, it makes a new file with "AAAA-" prefix on filename.
- AddPolyA.sh script does NOT edit sequence quality string, so do NOT use read quality filters on "AAAA-" files!

DETAILED UNEAK STEPS: 4. RENAME FASTQ FILES

- TASSEL will automatically use ALL ".fastq" files in the working directory.
- Must move or rename fastq files that you do not want in analysis.
- Renaming ".fastq" to ".raw" works.
- TASSEL expects fastq file names in this format: FLOWCELL_LANE_FILENUM.fastq
- FLOWCELL portion of file name MUST match that in KEY file exactly.
- LANE number portion of file name MUST match that in KEY file exactly.
- FLOWCELL name must NOT have underscore "_", space " ", slash "/", or backslash "\".
- FILENUM portion of file name is the incremental library sequencing run (rep).

DETAILED UNEAK STEPS: 5. RUN UNEAK JAR FILE

- Must issue java command from within working directory.
- Command format: java -jar JARFILE -k KEYFILE -s WORKINGDIR -p PROJECTNAME -PARAMETERS
- Jar file is always here: /home/share/tools/UNEAKpipeline3.jar
- Working directory (note, no "/" at end): /home/userID/DIRNAME
- Key file is in your working dir: /home/userID/DIRNAME/SAMPLE-KEY.txt
- Project name will be an automatically created sub-folder in the working directory (don't mkdir yourself).

DETAILED UNEAK STEPS: 5. RUN UNEAK JAR FILE

- Parameters: -B 0 -D 0 -c 100.0 -e PstI-MspI -F 0.0001 -H 1.0 -M .8 -n 1 -r 1

-B,--isBiparental <arg> Biparental mapping population? 1 or 0

-D,--isDHpopulation <arg> Double Haploid population? 1 or 0

-c,--chiSquare <arg> Chisquare value required for SNP calling, such as 0.01, (set to 100 to turn off)

-e,--enzyme <arg> Enzyme used to make the GBS library: PstI-MspI

-F,--MAF <arg> Minor allele frequency cutoff, 0-0.5, markers below this freq are removed

-H,--maxHeterozygous <arg> Max heterozygosity allowed for a tag, such as 0.1

-M,--maxMissing <arg> The maximum missing values allowed for each snp, 0-1, percent.

-n,--numDifferent <arg> Max number of base differences allowed when calling SNPs, such as 1

-r,--minReads <arg> Minimum number of reads required for a tag, such as 1

- Recommended parameters: -B 0 -D 0 -c 100.0 -e PstI-MspI -F 0.0001 -H 1.0 -M .8 -n 1 -r 1

DETAILED UNEAK STEPS: 6. MERGE AND SORT HAPMAP FILES

- TASSEL UNEAK will create 3 hapmap files with duplicates.
- "sort" command will merge 3 .hap files, sort on sequence and position, remove duplicates, & create a final.hap
- Must issue sort command from within hapmap directory.
- Command format: sort -u -k1r,1 -k6n,6 *.hap > final.hap

DETAILED UNEAK STEPS: 7. GET QUICK OUTPUT STATS

- Need a quick way to see & check results, use GBSTagStats.sh.
- Must issue command from within hapmap directory.
- Command format, onscreen only: /home/share/tools/GBSTagStats.sh final.hap
- Output includes all samples (blanks) and all markers.
- Command format, to file only: /home/share/tools/GBSTagStats.sh final.hap > GBSTagStats.txt
- Command format, to screen and file: /home/share/tools/GBSTagStats.sh final.hap | tee -a GBSTagStats.txt

DETAILED UNEAK STEPS: 8. COPY FILES FROM SERVER AND GET MARKER STATS IN EXCEL

- No files on the server are backed up. COPY AND SAVE YOUR DATA ELSEWHERE!
- Files to backup: *.fastq, *.hap, *.tbt.byte.log, *key.txt, GBSTagStats.txt.
- Excel can not read unix file types, change before copying from server.
- Unix to Win (any text file): unix2dos final.hap
- Unix to Mac (any text file): unix2mac final.hap
- Excel templates to filter markers: 2-parent pops or association mapping pops
- Excel template to impute missing parent data for 2-parent pops.

DETAILED UNEAK STEPS: 8. COPY FILES FROM SERVER AND GET MARKER STATS IN EXCEL

- Open 'final.hap' in Excel, sort genotype cols left-to-right, move 2-parents to beginning, check and remove blank, paste into template. Copy formulas down.
- Template: '(2 Parent Crosses) using UNEAK PIPELINE DATA, GBS tag stats and conversion.xlsx'
- Template: '(AM Crosses) using UNEAK PIPELINE DATA, GBS tag stats.xlsx'
- NOTE: the UNEAK hapmap 'assembly' col is the SNP position, 0-based, for sequence in 'rs' col, need to add 1 for standard counting.
- NOTE: the excel 'Position' col is the SNP position, 1-based, no correction needed on the 'ActualTagSeqInGenomeIncludingPstI' col.
- NOTE: the 'ActualTagSeqInGenomeIncludingPstI' col adds the cut "C" into the PstI site.

QUESTIONS OR BREAK?

USDA TASSEL-5 GBSV2 REFERENCE PIPELINE PROTOCOL

- TASSEL 5 is a reference-based Genotyping by Sequencing (GBS) SNP discovery pipeline

- http://www.maizegenetics.net/tassel
- https://bitbucket.org/tasseladmin/tassel-5-source/wiki/UserManual
- https://bitbucket.org/tasseladmin/tassel-5-source/wiki/Tassel5GBSv2Pipeline

BASIC TASSEL-5 METHODOLOGY:

- 1. Find good reads: expected barcode, RE cut site, no N's. Trims off barcodes, truncates sequences that have a second cut site, or read into the common adapter. Trims reads to 64 bases (including the RE site). Stores potential tags in a database.
- 2. Align the tags to the reference genome using BWA, store the genome positions with tags in the database.
- 3. Finds the tag distribution in all of the samples, assigns genotypes, and stores them in the database.
- 4. Outputs from the database a HapMap file and other files with statistics on genotypes and markers.

TASSEL-5 IMPLICATIONS:

- 1. A reference map is required.
- 2. All marker types can be found (monomorphic, polymorphic, short-indels).
- 3. No markers can be found for sequences that are NOT part of the reference genome (alien introgressions).

BASIC TASSEL-5 REFERENCE STEPS:

- 1. Get fastq files.
- 2. Make key and lines files.
- 3. Add poly-A to each read.
- 4. Rename fastq files.
- 5. Edit TASSEL script.
- 6. Run TASSEL script.
- 7. Get quick output stats and several other outputs.
- 8. Copy files from server and get marker stats in Excel.

DETAILED TASSEL-5 REFERENCE STEPS: 1. GET FASTQ FILES

- Amy will email user with link(s) to fastq files.
- Example: http://129.130.90.13/report/1040/
- user:password, DO NOT GIVE USERID OR PASSWORD TO OTHERS.
- Check notes to see what BC sets were used.
- Login to USDA bioinformatics server, make a working dir, cd to working dir
- Copy URL and use curl to get fastq file(s). curl --user user:password -O URLHERE
- Cancel the curl command for this class.
- /home/share/tools/USDA_Bioinformatics_Server_Training/Small_1*.fastq

DETAILED TASSEL-5 REFERENCE STEPS: 2. MAKE KEY AND LINES FILES

- Key file links barcode used to the sample & well.
- Lane is the fastq reads file (library) NUMBER for a set of samples.
- All reads with the same lane & barcode are merged.
- Reads with different lane numbers, but the same barcode are NOT merged.
- Must be a tab-delimited UNIX file type.
- Must have the first 7 required columns with these headers.
- You can have additional columns after the well column. Barcode set number and barcode number are useful for bookkeeping.
- Duplicate sample NAMES will have their data merged. Useful for parents.

DETAILED TASSEL-5 REFERENCE STEPS: 2. MAKE KEY AND LINES FILES

- You should have more than 1 sample well for each parent or you will have low marker numbers.
- Blank sample wells are not necessary, but help with error checking.
- Flowcell name must NOT have underscore "_", space " ", slash "/", or backslash "\", and must be the same on all lines of key file.
- Barcodes must be correct for each well .
- Lines file lists samples to analyze.
- No duplicate names, can be subset or all lines, remove 'Blank' for accurate stats.
- Must be a tab-delimited UNIX file type with only 1 column.
- Copy KEY and LINES files to working directory.

DETAILED TASSEL-5 REFERENCE STEPS: 3. ADD POLY-A TO EACH READ

- Ion Proton sequencing reads are of variable length. Illumina reads are all 100 bp.
- TASSEL will not work with reads shorter than 79 bases (BC+RE+Genomic).
- Adding poly-A insures all reads are at least 79 bases long.
- Poly-A addition improves marker finding by keeping slightly shorter reads.
- TASSEL ignores poly-A sections on the 3' end of read.

DETAILED TASSEL-5 REFERENCE STEPS: 3. ADD POLY-A TO EACH READ

- Run the AddPolyA.sh script on EACH of your fastq files: "/home/share/tools/AddPolyA.sh YourFileName.fastq"
- AddPolyA.sh script takes ~10-40 min/fastq.
- AddPolyA.sh script does nothing to the original file, it makes a new file with "AAAA-" prefix on filename.
- AddPolyA.sh script does NOT edit sequence quality string, so do NOT use read quality filters on "AAAA-" files!

DETAILED TASSEL-5 REFERENCE STEPS: 4. RENAME FASTQ FILES

- TASSEL will automatically use ALL ".fastq" files in the working directory.
- Must move or rename fastq files that you do not want in analysis.
- Renaming ".fastq" to ".raw" works.
- TASSEL expects fastq file names in this format: FLOWCELL_LANE_FILENUM.fastq
- FLOWCELL portion of file name MUST match that in KEY file exactly.
- LANE number portion of file name MUST match that in KEY file exactly.
- FLOWCELL name must NOT have underscore "_", space " ", slash "/", or backslash "\".
- FILENUM portion of file name is the incremental library sequencing run (rep).

DETAILED TASSEL-5 REFERENCE STEPS: 5. EDIT TASSEL SCRIPT

- Get a copy of the Tassel5GBSv2_pipeline_Paulv3.sh from our web server. https://hwwgenotyping.ksu.edu/protocols/GBS_protocols/
- You MUST change several definitions in the script before running it (notes in script):

WD=working directory

Study=project name

DKF=key file name

PKF=production key file name

MRC=minimum read count

MCL=minimum locus coverage

MAF=minimum minor allele frequency

TF=LINES (taxa) file name.

MQS=0 (do not change) minimum quality score

DETAILED TASSEL-5 REFERENCE STEPS: 5. EDIT TASSEL SCRIPT

RG=reference genome

WHEAT: RG=/home/share/tools/refs/161010_Chinese_Spring_v1.0_pseudomolecules.fa

RYE: RG=/home/share/tools/refs/Secale_cereale_Lo7_v2_ordered.fa

BARLEY: RG=/home/share/tools/refs/Hordeum_vulgare.Hv_IBSC_PGSB_v2.dna.all.fa

PEARL MILLET: RG=/home/share/tools/refs/pearl_millet_v1.1.merged.genome.fa

HESSIAN FLY: RG=/home/share/tools/refs/hf_v1.0.merged.genome.fa

- Once edited, save the file as a UNIX file type, and copy the script back to your working directory
- Make certain that the script is executable: "chmod +x Tassel5GBSv2_pipeline_Paulv3.sh"

DETAILED TASSEL-5 REFERENCE STEPS: 6. RUN TASSEL SCRIPT

- Make sure your working directory has: *.fastq, *key.txt, *lines.txt, and script.
- Run the script: “nohup ./Tassel5GBSv2_pipeline_Paulv3.sh | tee -a Tassel5GBSv2_pipeline_Paulv3-log.txt"
- Check output for parameters and errors. You can cancel the script with Control-C. The script should take from 20 min to 8 hours.
- If you cancel the script, fix errors then delete newly created sub-directories and log before re-running.

DETAILED TASSEL-5 REFERENCE STEPS: 7. OUTPUT STATS AND OTHER OUTPUTS

- Once script is done, cd into hapmap directory.
- Check results: "/home/share/tools/GBSTagStats.sh YourFileNameHere.hmp.txt | tee -a GBSTagStats.txt"
  - Output includes all samples (blanks) and all markers.
- Get marker counts per genome and per chromosome: "/home/share/tools/GetMarkerCountsPerChrom.sh"
- Plot marker distribution per chromosome: "/home/share/tools/PlotMarkerDistributionFromHapmapVarY.sh YourFileNameHere.hmp.txt 1.0"
  - You can specify the bin size to any cm size (1.0cm and 0.25cm are useful).
  - If you prefer a fixed Y-axis use: PlotMarkerDistributionFromHapmap.sh

DETAILED TASSEL-5 REFERENCE STEPS: 7. OUTPUT STATS AND OTHER OUTPUTS

- Get the reference sequence for all markers: "/home/share/tools/GetRefSeqForALLGBSMarkers.sh YourHapMapFileName.hmp.txt"
  - Creates "markerSeqs.fa" with 400 bases for each SNP from the reference. Base 200 is the SNP position (insertions are between 199 and 200?).
    - NOTE: this is NOT the marker sequence or the reads found in the GBS, but is from the reference file.
- If you plan to use FlapJack with this data, create FlapJack genotype and map files: "/home/share/tools/HapmapToFJ.sh YourFileNameHere.hmp.txt"

DETAILED TASSEL-5 REFERENCE STEPS: 8. COPY FILES AND GET MARKER STATS IN EXCEL

- No files on the server are backed up. COPY AND SAVE YOUR DATA ELSEWHERE!
- Files to backup: *.fastq, *.hap, *_ReadsPerSample.log, *key.txt, *lines.txt, TAGlist.txt, summary*.txt, *log.txt, GBSTagStats.txt, *.pdf, markerSeqs.fa, MarkersPerChrom.txt, *.FJGenotypes.txt, *.FJMap.txt, *.vcf
- Excel can not read unix file types, change before copying from server.
  - Unix to Win (any text file): unix2dos final.hap
  - Unix to Mac (any text file): unix2mac final.hap

DETAILED TASSEL-5 REFERENCE STEPS: 8. COPY FILES AND GET MARKER STATS IN EXCEL

- Excel templates to filter markers: 2-parent pops or association mapping pops
- Excel template to impute missing parent data for 2-parent pops.
- Open 'final.hap' in Excel, sort genotype cols left-to-right, move 2-parents to beginning, check and remove blank, paste into template. Copy formulas down.
- Template: '2 Parent Cross using REF PIPELINE DATA, GBS tag stats and conversion.xlsx'
- Template: '(AM Crosses), Ref Pipeline.xlsx'
- NOTE: the marker position is given as chromosome and reference base.

QUESTIONS OR BREAK?

MRASEQ PRE-PROCESSING READS PROTOCOL

- MRASeq is similar to Genotyping by Sequencing (GBS) in purpose and applications.
- Multiplex Restriction Amplicon Sequencing (MRASeq), a new next generation sequencing based marker platform for wheat breeding (not yet published)
- Amy Bernardo, Paul St. Amand, Ha Quang Le, and Guihua Bai

MRASEQ IMPLICATIONS:

- Replace patented GBS methodology.
- Easier, more economic, 2-step PCR, no restriction or ligation.
- Finds fewer markers (10-50%) compared to GBS.
- Has greater read depth (1-5X) per marker.
- Markers are more uniformly distributed across each chromosome.
- Markers are non-methylation sensitive.
- Analysis is identical to GBS for TASSEL UNEAK or reference pipelines after reads pre-processing.

BASIC MRASEQ STEPS:

- 1. Pre-process reads files.
- 2. Use reads files in either TASSEL UNEAK or reference pipelines.

MRASEQ BACKGROUND:

- TASSEL expects each read to have: Barcode-PstI-Genomic
- MRASeq adds an M13 tail and either a specific or degenerate section between the barcode and PstI sites.
- We must remove that section from all sequences prior to using TASSEL.
- MRASeq also changes the MspI side of a read, but since TASSEL will discard the MspI site and all sequence after MspI, we do not need process that side of the read.
- Currently, the quality string in the fastq is NOT adjusted.
- MRASeq analysis is identical to GBS once the reads have been pre-processed
- Currently, we must not use read quality filter settings in TASSEL (usually are not used anyway).

MRASEQ PRIMERS:

- M13 tail: GATGTAAAACGACGGCCAGTG
- Specific or Degenerate section, 6-10 bases: BRYGWS
- PstI site: CTGCAG
- M13 + 6 degen + PstI: GATGTAAAACGACGGCCAGTG-BRYGWS-CTGCAG

MRASEQ PRE-PROCESSING READS FOR TASSEL:

- One search and replace command will convert all reads in 1 file.
- sed (stream editor) is efficient for this purpose.
- Find string: GATGTAA.*CTGCAG
- Replace string: CTGCAG
- Create a new file so original is unchanged.

MRASEQ PRE-PROCESSING READS FOR TASSEL:

- Command format: sed 's/GATGTAA.*CTGCAG/CTGCAG/' Original.fastq > NoDegen.fastq
- Currently, we use barcode sets 5-18 for MRASeq.
- Barcode sets 5-18 are under NDA from Ion Torrent. We can NOT legally give them out to anyone.
- Get sequences for barcode sets 5-18 from Paul and do NOT share them ever.
- Proceed to TASSEL UNEAK or Ref analysis methods.

QUESTIONS OR BREAK?

USDA BIOINFORMATICS SERVER BLAST

- Current Blast: Blast 2.6.0+
- Our current databases that can be used with blast:
- Wheat: /home/share/tools/refs/ 161010_Chinese_Spring_v1.0_pseudomolecules.fa
- Rye: /home/share/tools/refs/Secale_cereale_Lo7_v2_ordered.fa
- Barley: /home/share/tools/refs/ Hordeum_vulgare.Hv_IBSC_PGSB_v2.dna.all.fa
- Hessian Fly: /home/share/tools/refs/hf_v1.0.merged.genome.fa
- Pearl Millet: /home/share/tools/refs/pearl_millet_v1.1.merged.genome.fa

BASIC COMMAND:

- blastn -query QUERY -db DATABASE
- QUERY is usually a file, but can be a sequence.

BLAST ONE SHORT SEQUENCE:

- blastn -query <(echo -e “>QueryName\nACGTCTTGCCGACCACCGGCGGCCTGTTCGCCGGTGAGTTCCTGCAGCGGCCCGAAGAGGCAc") -db /home/share/tools/refs/161010_Chinese_Spring_v1.0_pseudomolecules.fa

BLAST OUTPUT FORMATS:

- blastn -query <(echo -e ">QueryName\nACGTCTTGCCGACCACCGGCGGCCTGTTCGCCGGTGAGTTCCTGCAGCGGCCCGAAGAGGCAc") -db /home/share/tools/refs/161010_Chinese_Spring_v1.0_pseudomolecules.fa
- -outfmt 0
- -outfmt 1
- -outfmt 7
- -outfmt 6

BLAST USING A FILE OF QUERIES:

- blastn -query InputFileName.fasta -db /home/share/tools/refs/161010_Chinese_Spring_v1.0_pseudomolecules.fa -outfmt 7

BLAST USING A FILE OF QUERIES:

- Example InputFileName.fasta (must be UNIX file type in fasta format)

>RhtB1-4BS-AL1
CCCATGGCCATCTCSAGCTG
>RhtB1-4BS-AL2
CCCATGGCCATCTCSAGCTA
>RhtB1-4BS-Rev
TCGGGTACAAGGTGCGGGCG
>RhtD1-4DS-AL1
CATGGCCATCTCGAGCTRCTC
>RhtD1-4DS-AL2
CATGGCCATCTCGAGCTRCTA
>RhtD1-4DS-Rev
CGGGTACAAGGTGCGCGCC

BLAST TASK TYPE:

- blastn Traditional BLASTN requiring an exact match of 11
- blastn-short BLASTN program optimized for sequences shorter than 50 bases
- megablast Traditional megablast used to find very similar (e.g., intraspecies or closely related species) sequences
- dc-megablast Discontiguous megablast used to find more distant (e.g., interspecies) sequences

BLAST TASK TYPE:

- blastn -query InputFileName.txt -db /home/share/tools/refs/161010_Chinese_Spring_v1.0_pseudomolecules.fa -outfmt 1 -task blastn
- blastn -query InputFileName.txt -db /home/share/tools/refs/161010_Chinese_Spring_v1.0_pseudomolecules.fa -outfmt 1 -task blastn-short

BLAST E-VALUES:

- The Expect value (E) is the number of hits one can "expect" to see by chance. An E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance.

- blastn -query InputFileName.txt -db /home/share/tools/refs/161010_Chinese_Spring_v1.0_pseudomolecules.fa -outfmt 7 -task blastn
- blastn -query InputFileName.txt -db /home/share/tools/refs/161010_Chinese_Spring_v1.0_pseudomolecules.fa -outfmt 7 -task blastn -evalue 1.0
- blastn -query InputFileName.txt -db /home/share/tools/refs/161010_Chinese_Spring_v1.0_pseudomolecules.fa -outfmt 7 -task blastn -evalue 5.0

BLAST OPTIONS:

- Type "blastn -help" to get a list of blastn options.

QUESTIONS OR BREAK?

FLAPJACK

- Flapjack: visualization tool for graphical genotyping, haplotype visualization on large data sets, allowing rapid navigation and comparisons between lines, markers, and chromosomes.

- Download: https://ics.hutton.ac.uk/flapjack/
- Manual: http://flapjack.hutton.ac.uk/en/latest/
- MABC: http://flapjack.hutton.ac.uk/en/latest/mabc.html
- Other tools: https://www.hutton.ac.uk/research/groups/information-and-computational-sciences/tools

FLAPJACK MABC TUTORIAL

- Example genotype data: Example.FJGenotype.txt
- Example map: Example.FJMap.txt
- Example QTL map: Example.FJQTL.txt

- Example genotype data: YGFHB1+5A-Overland.FJGenotypes.txt
- Example map: YGFHB1+5A.FJMap.txt
- Example QTL map: YGFHB-5A-small.FJQTL.txt