**Summary**

Description of files included in BAC\_Seq folder

* BAC\_Seq\_Discussion\_020714.pdf: Slides used by Iowa State group during the first joint meeting for the project. The slides include additional annotations
* BAC\_Seq\_Proposal.pdf: Georgia Tech portion of the writeup for the proposal submitted to request funding
* Information\_About\_NGS\_Toolkits: A listing of tools which can be used to perform operations on FASTA/FASTQ files
* Answers\_By\_ISU\_For\_GT: Answers provided by Iowa State (ISU) group for questions asked by Georgia Tech group
* Short\_Reads\_Trimming-20140430.pptx: A description of how ‘trim\_fastq.pl’ script was used for trimming purpose
* BAC\_Reference\_Assemblies.txt: Reference assemblies for 12 BACs (for verification purpose)
* Pre\_Pipeline\_For\_RE\_Reads.pptx: Preprocessing pipeline for RE reads
* Illumina\_Adapter.pptx: Adapter used by MiSeq for fixing and amplifying sequences
* 0509discussion.pptx: Slide deck used by Iowa State group to provide an update about the project
* BACseq\_cope\_sub2.sh: Shell script outling the updated RE assembly pipeline

Problem: Our collaborators at Iowa State, who work in plant genomics, have a requirement to sequence the genome a Maize (Corn) species. Performing Sanger sequencing would be time-consuming and expensive. On the other hand, it is not feasible to do a de novo genome assembly of short-reads generated using WGS method (and NGS instruments) alone as plant genomes are known to have a lot of repeat structures. The solution strategy which was proposed to address these limitations is as follows. First, the genome is fragmented into BACs, each of length 120-150 kb. The BACs are generated such that there are overlaps among them. Once the individual BACs are assembled, these overlaps together with auxiliary information can be used to assemble the genome from the BACs. Each BAC is digested with multiple restriction enzymes (REs) and the pool of digests corresponding to a BAC and RE is tagged with unique identifier. A set of barcoded pools of digests are combined and sequenced using Illumina MiSeq. As the reads contain unique identifiers, each read coming out of the NGS instrument can be assigned back to a BAC. It is therefore possible to assemble a BAC using reads associated with it. Additional WGS short reads and mate-pair reads are also generated, which can be used to improve the coverage of the individual BAC assemblies. Please refer to “BAC\_Seq\_Proposal.pdf” included in this folder for further details.

Description of input files

* As a first step, we obtained FASTQ files containing PE reads, which correspond to RE digests
* We obtained the data for 12 BACs and 3 different RE digests for each BAC. The REs are BanII, NlaIII, and NspI
* The reads coming out of the Illumina MiSeq instrument were sorted based on index (or barcode) into individual files. The script used for this task is not available yet
* Further, the reads were trimmed based on quality score. The script used for this task is called ‘trim\_fastq.pl’. A description of how this script was used for trimming purpose is available in ‘Short\_Reads\_Trimming-20140430.pptx’
* There are 4 or 5 files available for each BAC+RE digest. The reads in all files have been subjected to trimming. ‘\*.trimmed-paired-1.fq’ and ‘\*.trimmed-paired-2.fq’ files contain the two reads corresponding to a PE read respectively. If after the trimming process, one of the reads had to be discarded completely, then the remaining read is available in ‘\*.trimmed-singletons-1.fq’ or ‘\*.trimmed-singletons-2.fq’. Whenever 5 files are available for a BAC+RE digest, the 5th file is a union of the four files previously described and therefore redundant with respect to its information content
* Reference assemblies for 12 BACs (for verification purpose) are available in ‘BAC\_Reference\_Assemblies.txt’ file
* Illumina adapter sequence: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCGTATCATTAAAAAAAAAACAAAACTATAATTTGATTT
* Refer to ‘Illumina\_Adapter.pptx’ file and “RE Part” slide from ‘BAC\_Seq\_Discussion\_020714.pdf’ for further details
* Iowa State group used ‘bowtie2’ for mapping and ‘IGV’ for visualization

Solution ideas and Analysis

* Error detection/correction strategy
* Strategy for assembly of RE reads
* Analysis: Relationship between assembly quality and the number of (and the combination of) REs used
* Analysis: When attempting to overlap PE reads, how many overlaps are false? We need to be cognizant of repeats within reads which could lead to false overlaps
* Analysis: What is the coverage for various BAC+RE combinations?

**Meeting on Feb 14:**

Questions from Rahul

* Is there some information about the bounds (upper/lower) on the distance between two cuts in a BAC for the sequencer to be able to read the paired reads.
* Along with the paired reads, is there information available about distance between them, in terms of distribution, observable bounds.
* Do the restrictive enzymes create sticky end reads or blunt end reads. In case of sticky end reads, how do we deal with chimeras. Specifically, at what point in the BAC sequencing is the barcoding done, since that might affect the chimeras (For eg. If the corresponding adaptor is added right after the restrictive enzyme, then expected number of chimeras would be less as opposed to the case where the adaptors are added at a later stage).
* In case of a partial digest, how can we differentiate between a read as a result of partial digest vs a read as a result of chimera (since both will have the cut pattern within the read).
* If there is an error in reading the barcode, that read might get assigned to a wrong BAC. How frequently can that occur? We would like to have some idea of the impact of this factor.

Questions/Notes from Kishore

* Minimum tiling path of BACs
* We need to be cognizant of repeats within the reads which could result in false overlaps
* Look into the pipeline that is being currently used
* How are BACs obtained from genome?
* How many times is the combination of BAC and enzyme sequenced? What is the coverage for a BAC + enzyme?
* What is the relationship between the quality of assembly and the # of enzymes used?
* What is the cost associated with reducing the BAC size?
* Is there an optimum BAC size?
* When it comes to combining information from multiple BACs: Do we combine them two at a time or all at the same time?
* How is the barcoding actually done? Why not shotgun/barcode whole BAC?
* What are the various limitations of the process?
* Start of a WGS read could falsely cause it be identified with a BAC (start sequence matches the barcode)
* What is the length and composition of the barcode?
* What is the probability that a barcode appears in the original genome?