Specific Aim 1.2. Assemble BACs. PI Schnable and Co-PI Aluru have extensive experience in the assembly of maize sequences, beginning with the MAGI project during which we assembled methylation and high-Cot filtered maize sequences [19]; Schnable's assemblies of RNA-Seq and genomic reads during the current NSF PGRP project; and our work on joint project NSF CCF 1162472 "Parallel Algorithms and Software for High-Throughput Sequence Assembly". We recently conducted a "bake-off" to identify which of eight available assembly tools best assemble maize BACs. Beginning with 5 Mo17 BACs that had been fully sequenced using Sanger technology, we used ART software [22] to simulate Illumina reads, which were then assembled using a variety of assembly tools. Considering a number of metrics of assembly quality Abyss [52] and Ray [8] emerged as the clear winners for this application.

We propose the following pipeline for Oh43 assembly. Our strategy comprises of utilizing the following data sets in a staged fashion, one after another in that order: RE digested BAC reads, S/P BAC reads, WGS reads, and mate-pair reads. In order to assemble a specific BAC, we will use the RE reads from the corresponding BAC in the first step because the unique barcodes used by BAC-Seq technique provide high confidence that these reads actually came from the BAC. In the second step, we will use the S/P BAC reads because we know that these reads came from a collection of a small number of BACs (few hundreds in number). We will then use the >50X coverage WGS reads to further extend the BAC contigs or combine contigs that are very closely co-located on the BAC, when it can be done so unambiguously. Finally, we will integrate the mate-paired reads into the assembly process to scaffold contigs, and help resolve the problems imposed by the abundant repeats.

Error Correction: Prior to the assembly process, we will correct read errors as it simplifies the assembly process and result in high quality assemblies. Recently, our group conducted a survey of error correction methods for next generation sequencing [Yang-2013]. We will leverage the evaluation toolkit developed for the survey to assess the performance of multiple error correction techniques [Liu-2013, Kelley-2010, Yang-2010, and Li-2010] on Oh43 read data. Majority of these techniques are specifically geared towards addressing substitution errors which are the predominant kind of errors on the MiSeq instrument. The evaluation included Reptile [Yang-2010] software developed by our group, which uses contextual information to perform accurate error correction. Owing to the algorithmic techniques Reptile is built upon, it handles large data sets efficiently. In case of repeat rich genomes such as Maize, it is difficult to identify sequencing errors because the observed *k*mer frequencies are not homogenous. Our group developed a rigorous statistical approach to solve this problem (REDEEM [Yang-2011]), which shows good results but can only be applied up to a few million reads due to its compute-intensive modeling. Due to BAC-wise separation of reads, we can benefit from this technique on Oh43 read data.

Assembly of RE digested BAC reads: The first step in assembling individual BACs is to use error-corrected RE reads. Initially, we will check for suffix-prefix overlaps between the two reads of each paired-end (PE) read. We expect that RE digested fragments that are shorter than 500 bp can be recovered as single reads because the paired-end read samples 250 bp from each end. After performing this for each set of RE reads separately, we attempt to assemble the BAC taking advantage of the overlaps between fragments from the two different restriction enzymes. The sequences of shorter fragments (< 500 bp) are fully known, while longer fragments are available as paired-end reads. Given that the fragments/reads are sufficiently long, we will use the SGA assembler [Simpson-2012] which is based on the string graph framework, an enhancement over the overlap-layout-consensus framework popular during the Sanger era. During an independent assessment conducted as a part of the Assemblathon-2 competition [Bradnam-2013], SGA assembler was shown to perform better among those that were evaluated.

BAC Assembly: In the next step of the assembly process, we will make use of the S/P BAC reads. We will use the de Bruijn graph framework to build an assembly graph on the S/P BAC reads. Next, we will annotate the contigs generated from assembling RE reads onto the de Bruijn assembly graph. The paired-end constraints available from the fragment library used for generating S/P BAC reads will be used in conjunction with the overlaps between the S/P BAC reads and the contigs resulting from the RE reads assembly to extend the latter. This step will result in an improved coverage of the BAC. The know-how essential to completing the steps described previously is embedded into the design of the modern de novo assemblers such as ABySS and ALLPATHS-LG [Simpson-2009 and Gnerre-2011]. Therefore, we will leverage the existing body of knowledge and tools available to develop software for accomplishing a speedy development of the solution for this task. We will then try to extend or bridge the contigs from WGS reads, whenever this can be done without ambiguity. As a final step in the assembly process, we will use the mate-pairs to orient and order the contigs resulting from the previous step as well as to close the gaps between them appropriately. We will use the recommendations made in the recent comprehensive evaluation of scaffolding tools for accomplishing the last objective [Hunt-2014].

Specific Aim 1.3. Assemble multi-BAC contigs. Following the procedure described in Specific Aim 1.2, we will have separately assembled each BAC. Next, we will identify overlaps among assembled BACs to assemble multi-BAC contigs, each of which will then be assigned to a chromosome and ordered and oriented. In this effort we will be assisted by both mate-pair libraries and the existing genetic map based on 200 B73xOh43 RILs from Ed Buckler's NAM project and the RNA-Seq based genotyping data we will generate from 1,000 DHs (Specific Aim 4). The many thousands of sequence tags that comprise these genetic maps will enable us to assign BAC assemblies and contigs of BAC to chromosomes and order and orient these contigs of BACs within chromosomes.

References:

[Yang-2013] Xiao Yang, Sriram P. Chockalingam, and Srinivas Aluru. "A survey of error-correction methods for next-generation sequencing." Briefings in bioinformatics 14, no. 1 (2013): 56-66

[Liu-2013] Yongchao Liu, Jan Schröder, and Bertil Schmidt. "Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data." Bioinformatics 29, no. 3 (2013): 308-315

[Kelley-2010] David R. Kelley, Michael C. Schatz, and Steven L. Salzberg. "Quake: quality-aware detection and correction of sequencing errors." Genome Biol 11, no. 11 (2010): R116

[Yang-2010] Xiao Yang, Karin S. Dorman, and Srinivas Aluru. "Reptile: representative tiling for short read error correction." Bioinformatics 26, no. 20 (2010): 2526-2533

[Li-2010] Ruiqiang Li, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li et al. "De novo assembly of human genomes with massively parallel short read sequencing." Genome research 20, no. 2 (2010): 265-272

[Yang-2011] Xiao Yang, Srinivas Aluru, and Karin S. Dorman. "Repeat-aware modeling and correction of short read errors." BMC bioinformatics 12, no. Suppl 1 (2011): S52

[Simpson-2012] Jared T. Simpson, and Richard Durbin. "Efficient de novo assembly of large genomes using compressed data structures." Genome research 22, no. 3 (2012): 549-556

[Bradnam-2013] Keith R. Bradnam, Joseph N. Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert et al. "Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species." GigaScience 2, no. 1 (2013): 1-31

[Simpson-2009] Jared T. Simpson, Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven JM Jones, and İnanç Birol. "ABySS: a parallel assembler for short read sequence data." Genome research 19, no. 6 (2009): 1117-1123

[Gnerre-2011] Sante Gnerre, Iain MacCallum, Dariusz Przybylski, Filipe J. Ribeiro, Joshua N. Burton, Bruce J. Walker, Ted Sharpe et al. "High-quality draft assemblies of mammalian genomes from massively parallel sequence data." Proceedings of the National Academy of Sciences 108, no. 4 (2011): 1513-1518

[Hunt-2014] Martin Hunt, Chris Newbold, Matthew Berriman, and Thomas D. Otto. "A comprehensive evaluation of assembly scaffolding tools." Genome biology 15, no. 3 (2014): R42