Sequence Improvement Protocol for GEP Hybrid Assembly Projects

Developed by Christopher Shaffer, with input from GEP members Don Paetkau, Michael Rubin, and Laura Reed

Abstract

This document is a general description of the GEP protocol for improving the hybrid assemblies produced by the Human Genome Sequencing Center at Baylor College of Medicine (HGSC-BCM) as part of the modENCODE project. The background and procedural details discussed here are reiterated in a walkthrough [GEP Hybrid Assembly Walkthrough]. The walkthrough gives detailed examples and screen shots and is designed for use as a non-graded "in class" training activity. In addition, a 1-page copy of the protocol below is also available [Finishing GEP Hybrid assemblies: Recommended protocol]. The protocol summary is designed as a reference for finishers as they work on their individual projects.

Introduction

The modENCODE project funded the sequencing of 8 new Drosophila species (https://www.hgsc.bcm.edu/arthropods/drosophila-modencode-project). Unlike earlier genomes, all sequencing data was generated using second generation sequencing systems (i.e. 454 and Illumina). The initial genomic assembly (version 1) was done using only reads from Roche's 454 system derived from both unpaired and paired end libraries. These version 1 assemblies have a large number of consensus errors found almost exclusively within long mononucleotide runs (MNRs). This is a known weakness with 454 sequencing technology.

In an attempt to improve the consensus sequences (particularly in regions with long MNRs), HGSC-BCM subsequently generated a large amount of Illumina genomic short read data for each species. However, this additional Illumina data was not combined with the 454 data to create a new de novo assembly. Instead, the Illumina reads were mapped back to the version 1 assembly. A modified version of the GTAK genome analysis toolkit (see: http://www.broadinstitute.org/gatk/) was then used to computationally identify regions with large numbers of Illumina reads that disagree with the consensus. The consensus sequences at these discrepant regions were then revised using a custom script to produce the version 2 assembly.

While many MNR errors were found and corrected in the version 2 assembly, preliminary analysis of the version 2 assembly of *Drosophila biarmipes* shows that many errors remain. The goal of this finishing project will be to undertake a careful analysis of selected regions

of this assembly (i.e. Muller F element and the base of the Muller D element) to find and correct as many of these errors as possible.

> **Note:** This document provides a general overview of the basic steps in analyzing 454/Illumina hybrid assemblies produced by the modENCODE project. Typical hybrid assemblies would be based on *de novo* assembly of both 454 and Illumina data. **The techniques described in this document might not apply to hybrid assemblies produced by a different approach (e.g. hybrid assembly produced by Newbler or wgs-assembler).**

Basic protocol

This project has three major objectives: the primary goal is to find and correct errors within MNR's. The secondary goal is to assess and close gaps with the option of attempting to correct low consensus quality (LCQ) regions. Finally, if time allows, finishers could search for and tag regions with putative polymorphisms. However this should only be done if all other goals have been successfully completed. Note that resolving misassemblies is **not** a major goal of this project because of the lack of evidence that would allow searching for a solution.

Primary goal: Base errors in regions of mononucleotide runs

The primary goal of this project will be to find and correct any errors within MNR's. For training purposes, the most basic protocol for completing this goal will be discussed below.

More advanced techniques (discussed on the private GEP wiki) can increase efficiency of the finisher in completing the GEP projects. However, we recommend that finishers start by becoming familiar with the basic protocol before incorporating the more advanced computational techniques. Once the finishers are familiar with the overall workflow, many improvements may become obvious to those with more advanced computer skills. We strongly encourage finishers who find more efficient procedures to share them on the wiki or bulletin board for other GEP members.

Analysis will be based on the presence of High quality (>=40) discrepancies (designated HQD) to find regions of interest. This list is typically very long because of the very high depth of read coverage. While we certainly expect consensus errors to be on the HQD list, most of the items on the list are false positives that will not be associated with consensus errors. Because our goal is to improve the quality of the consensus, any HQD's that do not lead to a change in the consensus are considered noise in the system and can be safely ignored. The primary job of the finisher is to distinguish between regions where the consensus needs to be corrected and regions where the consensus is correct. Finishers should use the following guidelines to discriminate between these two situations. Detailed examples are given in the associated walkthrough:

Almost all false positive HQD's (i.e. HQD's that are on the list but do not actually lead to a change in consensus) can be attributed to either incorrect mapping of reads (~80-90%) or polymorphisms in the strain of flies used to generate the sequencing libraries (at most ~10-20%). To avoid most false positives we will only be interested in sites with **at least three HQD's** (i.e. the same consensus position appears at least three times in the HQD list).

Given the error profile of 454 sequencing, a finisher should cross reference any region with at least 3 HQD's with the presence of a MNR of at least 5 bases. These locations with a MNR and a nearby position with at least 3 HWD's should be inspected carefully. Remember that because of the way Consed aligns and numbers base positions that the location of the 3 HQD's does not need to exactly coincide with the coordinates of the MNR. Any location with 3 HQD's that is within 5 bases of a MNR should be examined. Based upon that examination, the consensus sequence should be corrected if indicated by the Illumina reads. Finishers should ignore any region with 3 or more HQD if there is no nearby MNR.. Using this technique, finishers should be able to find and correct most of the base calling errors within MNR's.

It is also possible to have errors in regions where the coverage is so low that there do not have 3 HQD's. To address this issue check finishers will need to carefully double check **all** MNR's found within regions with less than 40 x coverage. These regions can be identified by using the "Main Window -> Navigate -> Search for High (low) Depth of Coverage".

Optionally finishers can examine regions not associated with MNR's to help find putative polymorphisms, however this is not a specific goal of the GEP research and can be considered an optional tertiary goal of the sequence improvement project and the data will not be used in the final analysis.

Secondary goal: gaps and low consensus quality

Gaps: The computational pipeline used to create version 2 assemblies resulted in only a handful of closed gaps. Initial inspection of the version 2 assembly of *D. biarmipes* suggests that some gaps can be resolved manually by a finisher using the Illumina data already present in the projects. However, most gaps will require additional Sanger sequencing data generated from PCR amplicons of the region covering the gap. Finishers are requested to design one pair of PCR primers flanking each gap in the project. Actual generation of PCR/Sanger data for these regions is an optional activity.

Low Consensus Quality (LCQ): Given the high coverage of the genome with both 454 and Illumina data, there will only be a small number of regions with low consensus quality. Initial inspection of 10 projects revealed that all the LCQ regions reported by Consed were either associated with a gap or are at the ends of the project. This suggests that problems of this type will be rare. The LQC areas found at the ends of the project are not examples of problems of this type.

Similar to the Sanger sequence improvement projects, we partition the large genomic scaffolds into 100kb chunks (with 5-10kb overlap) to create project packages. Reads that extend beyond the ends of the project might not be included in the project. Hence the low read coverage and low consensus quality at the ends of the project is an artifact of the partitioning scheme and can be safely ignored. Likewise, the regions surrounding gaps will, by definition, have low quality and low read coverage. Closing the gap will resolve the LQC bases near gaps.

For detecting LCQ regions, finishers should set the "Threshold for Low Consensus Quality (highest low)" field (under "General Preferences") to 30. Any region with consensus scores at or below this quality value should be examined manually and "data needed" tags should be added if deemed necessary. Since each project overlaps adjacent projects by **at least** 5kb, you do not need to address **any problems** in the first or last 2.5 kb because they will be resolved by the finishers working on the adjacent project (if present). The use of the low consensus quality navigator is covered in "Using Consed Graphically" and the "Drosophila Finishing Problem Set". Problem areas of this type are expected to be so rare that they will not be discussed further.

Optional goal

It is unknown at this time the frequency of single nucleotide polymorphisms (SNPs) in these modENCODE Drosophila species. It is possible that finishers will find sites where the evidence suggests that the high rate of HQD's is due to a sequence difference between the maternal and paternal chromosomes. While screening through the list of HQD's you may find regions with approximately 50% of the reads that show one allele and the remaining 50% that shows a different allele. While regions of this type probably occur most often due to mis-mapping of reads that actually belong elsewhere in the genome (e.g. transposons), some polymorphic sites are expected. These polymorphic sites can be of interest to people working on population structure and sequence diversity.

As an optional add-on, please add a "polymorphism" tag to sites when you find reasonable evidence of a putative polymorphism. For the GEP project, we will define a region as a putative polymorphism if it has between 40 to 60% of the reads showing one of two alleles. Because of the strategy used to construct the assembly, regions with less than 40% of the reads showing a different allele could be caused by incorrectly placed reads and should not be tagged. In addition, you should also ignore any discrepant region (regardless of the percentages) that overlaps with a (blue) repeat tag. The blue (repeat tags) indicates sequence similarity to a transposon and the discrepancies can likely be attributed to reads that have been misplaced.

The following page is a short hand summary of the recommended protocol. It can be printed and used while you finish to help guide you through the finishing protocol

Recommended protocol

For easy use, this protocol is also available as a single sheet document on the GEP website. Most of the goals of the project (steps 1 through 4 below) can be attempted in any order, the following order is advisable if you will be doing your own PCR/Sanger reactions. The order is designed to allow as much time as possible for generating the PCR/Sanger data.

1. Use search for string to search for "nnnn"; this will generate a navigator listing the locations of any gaps found in the project.
    a. If gaps are present, assess if an overlap exists; if an overlap is present, use tear/join technique (see demo in walkthrough) to remove gap.
    b. If no overlap can be found; design PCR primers flanking the gap. Optionally, order primer synthesis and attempt PCR/Sanger in lab and incorporate any new data into the project.
2. Confirm that the consensus quality threshold is set to 30: Main window -> Options -> General preferences. Now, in the aligned reads window, use Navigate ->Low consensus quality menu to list all LCQ regions. Any LCQ region in the first 2.5 kb, the last 2.5 kb, or associated with a gap does not need to be inspected.
    a. Inspect all appropriate LCQ regions, for regions with a consensus quality below 25 (or below 30 if the region is single-stranded); these regions should be manually tagged with a "Data needed". Optionally, design and order primers and attempt PCR/Sanger in lab and incorporate any new data into the project (be sure to ask your mentor if you are doing the optional primer design or wet-bench work on the project).
3. Use the Main Window -> Navigate -> "Search for Highly Discrepant Positions" to generate the list of potential problem areas with at least 3 discrepancies with Q scores ≥ 30.
    a. Go to every region on the list with at least 3 HQD's that is within 5 bases of a MNR that is at least 5 bases in length. Carefully examine the region to confirm or correct the length of the MNR in the consensus. If the region is not associated with a MNR, it can be ignored.
    b. (Optional goal) Examine any region where the proportion of the HQD's is between 40-60%. If the region is not marked with a repeat tag assess the likelihood of polymorphism vs. mis-mapping, and if appropriate, add a polymorphism tag to the location.
4. Use "Main Window -> Navigate -> Search for High (low) Depth of Coverage" to find all regions of score 10 with a 40 fold or less coverage. Carefully double check **all** MNR runs of length 5 or more found within these regions of low coverage. Minimum standard is 2 Illumina reads with all bases Q 20 in the MNR in order to overrule the consensus. Illumina reads with more than one HQD at any location not associated with the MNR should **NOT** be considered as evidence to support the consensus. This is to avoid introducing consensus errors from mis-mapped reads.
5. Be sure to complete and include in your submission the finishing report form. This is especially important if you improved the consensus but were not able to fully finish the project.