Common Misassembly Protocols

Wilson Leung

Table of Contents

Compare total size of real and in-silico digests

1. Determine the total size of real restriction digests
    a. Open an `xterm` and navigate to the `edit_dir` of your project
    b. Run the script `calc_total_fragSizes.pl`
        - Use the Digests window to examine each digest if there are substantial differences in the total sizes of the four digests
        - Differences in real digest sizes can usually be attributed to:
            - Miscalled doublets
            - Partial digests
            - Multiple samples in a lane
2. Calculate the total size of the major contigs in the project
    a. Open Assembly View and identify the major contigs in the project
    b. Go to the Contig List in the Consed Main Window to determine the size of each of these major contigs
    c. Calculate the sum of the major contigs sizes plus the vector sequence (pcc01.fasta = 8,139 bp)
3. Compare total size of the real restriction digests with the size of the *in-silico* digests (using either the **Just Part of Clone** or **Entire Single Contig** options)

Download the NCBI Trace Archive read package

In the NCBI Trace Archive search results page:

1. Change the **Save result of search as** field to an unique archive name
2. Select the **.gz file** checkbox
3. Unselect the **FASTA** checkbox and select the **SCF** checkbox
4. Click on the **Save query's result** button to save the results
5. Save the reads package in your project directory

Examine mate pair and clone depth in Assembly View

1. Open Assembly View
2. Select **What to Show**, then select **Fwd/Rev Pairs**
3. Enable the following options:
   - show consistent fwd/rev pair depth
   - show each consistent fwd/rev pair within contigs
   - show gap-spanning fwd/rev pairs
   - show consistent fwd/rev pairs between diff scaffolds
   - show legs on squares for consistent fwd/rev pairs
4. Click on the **Apply** button

Incorporating NCBI Trace Archive reads

1. Open an `xterm` and navigate to the `edit_dir` of your project
2. Use the `add_trace_archive_reads.sh` script to rename and move the new reads into the appropriate directories:

   ```
   add_trace_archive_reads.sh \
     <path to the trace archive read package> \
     <name of the fof file with all the new reads>

   Example:
   add_trace_archive_reads.sh \
     ../unpaired_traces_2402D22.tar.gz \
     unpaired_traces_2402D22.fof
   ```

   This command will rename the reads in the `unpaired_traces_2402D22.tar.gz` archive (found in the parent directory) and move them to the appropriate directory. The script will then create an fof file (`unpaired_traces_2402D22.fof`) with the new read names and place it in the current directory

3. Use `phredPhrap` or `Add New Reads` to incorporate the new reads into your project

Pull out inconsistent mate pairs

1. Open Assembly View and run `crossmatch`
2. Highlight and then left click on the inconsistent mate pairs (red/purple lines beneath the contigs) in Assembly View
3. Click on the **Pull Out Reads** button in the Clicked Forward/Reverse Pairs window
4. In the Put Reads Into Their Own Contigs window, highlight the reads you would like to put into their own contigs (e.g. based on the repeat structure)
   - To select a group of reads:
     - Left click on the first read you want to pull out
     - Navigate to the last read you want to pull out (scroll down if necessary)
     - Hold down the shift key and left click on the last read you want to pull out
   - To select multiple reads individually:
     - Left click on the first read you want to pull out
     - Hold down the control key and left click to select additional reads
5. Click on the **Remove Highlighted Reads** button to pull out the selected reads
6. Save the project and **Close All Windows**

Reassemble regions with multiple high quality discrepancies

1. Select the Consed Main Window
2. Select **Navigate** -> **Multiple High Quality Discrepancies**
3. Use the Multiple High Quality Discrepancies window to navigate to each region with multiple discrepancies
   - Click on the **Tell Phrap No Overlap** button if the discrepancy is a real base difference and is likely to be caused by a misassembly
     - Check the quality of the trace data
     - Look for multiple reads with the same discrepancy
     - Compare the size of the *in-silico* and real restriction digest fragments to distinguish polymorphisms from misassemblies
   - Add a comment tag to discrepancies that could be attributed to either single nucleotide polymorphisms or base calling errors
   - Click on the **Next** button to navigate to the next discrepant region
4. Follow the "Run Miniassembly" protocol to reassemble contigs with the **Tell Phrap No Overlap** tags

Retrieve reads from the NCBI Trace Archive using a query string

1. Open a web browser and navigate to the NCBI Trace Archive web site at
2. Copy the query string from the `xterm` onto the clipboard, then paste the query into the **Enter a query string** textbox and click on the **Submit** button
3. Follow the "Download the NCBI Trace Archive read package" protocol to download the new reads
4. Follow the "Incorporating NCBI Trace Archive reads" protocol to add the new reads into your project

Retrieve reads from the NCBI Trace Archive using blastn

1. Open the Aligned Reads window and navigate to the region you want to search for additional data
2. Export the last 50 high quality bases before the gap
    a. Select **File** and then select **Export consensus sequence (with options)…**
    b. Select the **part** option under **Write Whole or Part of Consensus?**
    c. Enter the start and end positions for the region to extract, then click **OK**
    d. Specify the sequence file name in the **Save consensus bases to file** field
    e. Click **OK** to save the file
3. Navigate to the NCBI Trace Archive BLAST page at
    a. Click on **Browse** and select the file with the extracted sequence using the **Or, upload file** field in the **Enter Query Sequence** section
    b. Select ***Drosophila ananassae* – WGS** under the **Database** field
    c. Change the BLAST program to **Somewhat similar sequences (blastn)** under the **Program Selection** field
    d. Under the **Algorithm Parameters** section, change the **Expect threshold** to **1e-10** and uncheck the **Low complexity regions** filter
    e. Click on the **BLAST** button
4. Filtering BLAST results
    a. Click on the **Formatting options** link at the top of the BLAST results page
    b. Change the **Percent Identity Min** field to **100**
    c. Click on **Reformat** to filter the results
5. Examine each alignment and decide if you would like to retrieve the read
    a. Select the checkbox next to the description of the BLAST hit if you would like to retrieve the read
    b. Alternatively, Click on the "**All**" link above the BLAST **description table** in the **Descriptions** section to select all the reads in the BLASTN results
6. Click on the "**Trace**" link to send the list of read names to the query interface at the NCBI Trace Archive
7. Follow the "Download the NCBI Trace Archive read package" protocol to download the new reads
8. Use the `get_mate_pair_ti.sh` script to construct the NCBI Trace Archive query to retrieve the mate pairs of the reads we have just retrieved:

    ```
    get_mate_pair_ti.sh <path to trace archive read package>

    Example:
    get_mate_pair_ti.sh traces_Contig7_begin.tar.gz
    ```

    Examine the reads in the `traces_Contig7_begin.tar.gz` archive and construct the NCBI Trace Archive query for its mate pairs

9. Follow the "Retrieve reads from the NCBI Trace Archive using a query string" protocol to retrieve the mate pair reads from the NCBI Trace Archive
10. Follow the "Incorporating NCBI Trace Archive reads" protocol to add the new reads from both read packages into your project

Run crossmatch

1. Save the Consed assembly if there are any unsaved changes
2. Open Assembly View
3. Select **What to Show**, then select **Sequence Matches**
4. Click on the **run crossmatch** button
5. Depending on how repetitive the clone is, filter the sequence matches
   For example, to show only direct repeats between contigs:
   a. Check the box "ok to show sequence matches between contigs"
   b. Uncheck the following checkboxes:
      - ok to show sequence matches within contigs
      - ok to show inverted sequence matches
   c. Click on the **Apply** button at the bottom left corner

Run Miniassembly

1. Click on the **Miniassembly** button in the Consed Main Window
2. Verify that the contigs you want to assemble are listed under the **Contigs to Reassemble** section in the Reassemble Some Contigs window
3. If necessary, use these steps to add contigs to the **Contigs to Reassemble** list:
   - Under the **All Contigs** section, select the contig(s) you want to add
     - To select a group of contigs
       - Left click on the first contig you want to add
       - Navigate to the last contig you want to add
       - Hold down the shift key and left click on the last contig you want to add
     - To select multiple contigs individually:
       - Left click on the first contig you want to add
       - Hold down the control key and left click on the subsequent contigs you want to add
4. To remove contigs from the list of contigs to reassemble, left click on the contig you want to remove and click on the **Clear Highlighted** button
5. Click on the **Reassemble** button to run Miniassembly
6. Save the project and **Close All Windows**

Tag regions with multiple high quality discrepancies

1. Select **Navigate**, then select **Multiple High Quality Discrepancies** in the Consed Main Window
2. If there are discrepancies in the Multiple High Quality Discrepancies window:
   a. Left click on first discrepancy and click **Next**
   b. Examine the traces with high quality discrepancies:
      - If the discrepancy is real
        - Use the restriction digest to determine if the discrepancy can be attributed to polymorphisms (e.g. discrepancy is likely a polymorphism if the *in-silico* and real fragment sizes are consistent).

        - If so, add a polymorphism tag to the consensus
        - If not, click on the **Tell Phrap No Overlap** button in the Multiple High Quality Discrepancies window
    - If the discrepancy is spurious (e.g. low quality data)
        - Edit the discrepant base positions (if supported by the trace)
        - Add a comment tag on the consensus that explains why the discrepancy is not real
  c. Navigate to and examine the next discrepancy in the Multiple High Quality Discrepancies window
  d. Use the "Run Miniassembly" protocol to reassemble contigs with the **markedHighQuality** tags

Retrieving missing mate pairs

1. Create a list of reads to check for missing mate pairs
   a. In an `xterm`, use the `extract_read_names_from_ace.pl` script to extract read names from a list of regions:

```
extract_read_names_from_ace.pl \
  -r <comma-separated list of contigs and regions> \
  -i <path to the ace file> \
  -o <output file with all the read names>

Example:
extract_read_names_from_ace.pl \
  -r Contig6:32000-34656,Contig4,Contig5 \
  -i 2402D22.fasta.screen.ace.1 \
  -o check_mate_pairs.fof
```

   Put reads in Contig6 from 32000-34656 and all the reads in contigs 4 and 5 in the assembly that corresponds to the `2402D22.fasta.screen.ace.1` ACE file into the output file `check_mate_pairs.fof`

   b. To construct a list of read names for all the reads in your project, type the following command in the `xterm` (in `edit_dir` of your project)

```
ls ../chromat_dir > check_mate_pairs.fof
```

2. Using a file with a list of read names (e.g. `check_mate_pairs.fof`), generate the NCBI Trace Archive query for clones with missing mate pairs by issuing the following command on an `xterm`:

```
build_unpaired_cloneid_query.pl –i check_mate_pairs.fof
```

3. Select and copy the output (beginning from "CLONE_ID") into the clipboard
4. Follow the "Retrieve reads from the NCBI Trace Archive using a query string" protocol to download additional reads from the NCBI Trace Archive