Mojavensis: Issues of Polymorphisms

Chris Shaffer
GEP 2009
Washington University

Mojavensis assembly

- Isolate genomic DNA
- Create 3 libraries
  - Plasmid library: ~3.5 kb inserts (1 million)
  - Fosmid library: ~37 kb inserts (250,000 clones)
  - Bac library: ~150 kb inserts (30,000 clones)
- Sequence both ends
- Take all data and attempt whole genome assembly

Mojavensis assembly

- After whole genome shotgun and assembly:
  - 6842 scaffolds; 319,616,621 total bases
  - 5033 gaps; estimated 13,608,479 bases
  - Average gap size: 2705 bases
- Room for improvement
- Better quality needed for Repeat analysis

Mojavensis assembly

- Differences from Mouse and virilis examples:
  - There have been no vector sequences seen so don't expect them
  - There is significant overlap of each fosmid so it is not vital that you finish the ends of your sequence to high quality
  - Digests algorithm will add two extra cut sites, one at each "vector"/"insert" junction. See the wiki on work around.

Mojavensis Assembly

- Since the libraries were made from whole genomic DNA there are two dot chromosomes (maternal and paternal) that give rise to sequence data
- Sequence differences between the two homologs can confound phred/phrap/consed (originally designed for assembly of clones (Bacs and fosmids)
- Siblings from mojavensis and other Drosophila species were crossed for 10 generations to remove polymorphisms
- Unfortunately we have found a higher than expected rate of polymorphisms (about one third of projects in spring 07 had polymorphisms)

Duplication or Polymorphism

- When you find two nearly identical sequences these could either be polymorphism or a duplication
- How can you tell which you have?
- Restriction mapping data!

Case 1: single polymorphism

- Small isolated single base polymorphisms will usually be found as high quality discrepancies
- Use your restriction digest data for the region containing the putative polymorphism to guide you
  - Sizes match: tag as a polymorphism
  - Sizes do not match: tag "tell phrap not to overlap discrepant reads"

Demo

- Computers
- Open X11; start terminal if needed
  - cd Desktop/465-C16/edit_dir
  - Start consed
  - Open contig 5
  - Open high quality discrepancy navigator
  - Notice all the discrepancies at 4522

Demo (cont')

- Is this a misassembly or a polymorphism?
- Check digest
  - Click "find main window"
  - Click "digests"
  - In the "Huge list of enzymes" select
    - EcoRI and SacI to add these to HindIII and EcoRV
  - Under what to digest select "entire single contig" and enter 5 in the box labeled "single entire contig:"
  - Click "OK"

Demo (cont')

- For error message that it cannot link vector ends to end of contig click "Dismiss"
- Select "Display digests" from Window menu
- Select zoom from Window menu
- The problem area is around base 4522
- Select "position" in sort by (lower left)
- Note base 4522 is within the 3982 base predicted fragment which matches an observed fragment if 3977

Demo (cont')

- Since it matches more likely a polymorphism
- The more digests that match the better
- Check again later when assembled into larger contig for other digests
- If at end still looks like polymorphism add polymorphism tag

Case 2: polymorphism clusters

- Consed will not assemble together
- Will look like a spanned gap in assembly view with sequence matches at each end
- Force join the two contigs
- Check the digest for this region
- Sizes match: tag all polymorphisms
- Does not match; undo join or quit and do not save, attempt to call oligo's to span gap
- Seek help if needed for oligo design

Case 3: Insert Polymorphism

- Very complicated
- Very unlikely in green and low yellow
- Want to confirm assembly with digests
- If needed pull out all discrepant reads into new contig to get digests to match
- Seek help!