AMOS Assembly Validation and Visualization

Michael Schatz

Center for Bioinformatics and Computational Biology
University of Maryland

April 7, 2006
Outline

- AMOS Introduction
  - Getting Data into AMOS

- AMOS Validation Pipeline
  - Mate-Based Validation
    - C/E Statistic
  - Read Alignment Validation
  - Read Depth Validation

- AMOS Assembly Investigator
  - Contigs, Inserts, Histograms, SNP Barcode, Features
  - Misassembly Walkthrough

- Demo
Outline

- AMOS Introduction
  - Getting Data into AMOS

- AMOS Validation Pipeline
  - Mate-Based Validation
    - C/E Statistic
  - Read Alignment Validation
  - Read Depth Validation

- AMOS Assembly Investigator
  - Contigs, Inserts, Histograms, SNP Barcode, Features
  - Misassembly Walkthrough

- Demo
AMOS Goals

- Open Source Assembly Package
- Modular design
- Flexibility in building “pipelines”
- Well defined input/output formats
- General use: does not depend on databases, proprietary data formats, specialized hardware, etc.
Modular Design

- Converters: Celera Assembler, .ACE, TIGR Assembler, Trace Archive
- Overlapper
- Contigger (Minimus)
- Consensus caller
- Comparative assembler (AMOScmp)
- Mate-pair based QC tool
- Viewer (Assembly Investigator)
- Pipeline executor
Assembly Data Conversions

CA Assembly w/ Surrogates to AMOS Message File (.asm, .frg)
$ toAmos -a prefix.asm -f prefix.frg -o prefix.afg -S

Finished Assembly to AMOS Message File (.contig, .frg)
$ toAmos -f prefix.frg -c prefix.contig -o prefix.afg

AMOS Message File to Bank
$ bank-transact -m prefix.afg -b prefix.bnk -c
**AMOS Validation Pipeline**

- Automatically scan an assembly to locate misassembly signatures for further analysis and correction

- **cavalidate prefix (.frg, .asm)**
  1. Load CA Assembly Data into Bank
  2. Evaluate Mate Pairs & Libraries
  3. Evaluate Read Alignments
  4. Analyze Depth of Coverage
  5. List Surrogates
  6. Load Misassembly Signatures into Bank

- **amosvalidate prefix (.afg)**
  - Same as cavalidate, except skips surrogates
Mate-Happiness: asmQC

- Evaluate mate “happiness” across assembly
  - Happy = Correct orientation and distance

- Finds regions with multiple:
  - Compressed Mates
  - Expanded Mates
  - Invalid same orientation (→ →)
  - Invalid outie orientation (← →)
  - Missing Mates
    - Linking mates (mate in a different scaffold)
    - Singleton mates (mate is not in any contig)

- Regions with high C/E statistic
Mate-Happiness: asmQC

- Excision: Skip reads between flanking repeats

- Truth

- Misassembly: Compressed Mates, Missing Mates
Mate-Happiness: asmQC

- Insertion: Additional reads between flanking repeats

- Truth

- Misassembly: Expanded Mates, Missing Mates
Mate-Happiness: asmQC

- Rearrangement: Reordering of reads
- Truth

Note: Unhappy mates may also occur for biological or technical reasons.
C/E Statistic

- The presence of individual compressed or expanded mates is rare but expected.

- Do the inserts spanning a given position differ from the rest of the library?
  - Flag large differences as potential misassemblies
  - Even if each individual mate is “happy”

- Compute the statistic at all positions
  - \((\text{Local Mean} - \text{Global Mean}) / \text{Scaling Factor}\)

- Introduced by Jim Yorke’s group at UMD
Sampling the Genome

Normal Library
Count=10000, Mean=4000, SD=400

0kb 2kb 4kb 6kb

8 inserts: 3kb-6kb
Local Mean: 4048
C/E Stat: \(\frac{4048-4000}{400/\sqrt{8}}\) = +0.33

Near 0 indicates overall happiness
C/E-Statistic: Expansion

Normal Library
Count=10000, Mean=4000, SD=400

8 inserts: 3.2kb-6kb
Local Mean: 4461
C/E Stat: \( \frac{(4461-4000)}{\sqrt{8}} = +3.26 \)

C/E Stat ≥ 3.0 indicates Expansion
C/E-Statistic: Compression

8 inserts: 3.2 kb-4.8kb
Local Mean: 3488
C/E Stat: \( \frac{(3488-4000)}{(400 / \sqrt{8})} = -3.62 \)

C/E Stat \( \leq -3.0 \) indicates Compression
Read Alignment

- Multiple reads with same conflicting base are unlikely
  - 1x QV 30: 1/1000 base calling error
  - 2x QV 30: 1/1,000,000 base calling error
  - 3x QV 30: 1/1,000,000,000 base calling error

- Regions of correlated SNPs are likely to be assembly errors or interesting biological events
  - Highly specific metric

- AMOS Tools: analyzeSNPs & clusterSNPs
  - Locate regions with high rate of correlated SNPs
  - Parameterized thresholds:
    - Multiple positions within 100bp sliding window
    - 2+ conflicting reads
    - Cumulative QV >= 40 (1/10000 base calling error)
Read Coverage

- Find regions of contigs where the depth of coverage is unusually high

- Collapsed Repeat Signature
  - Can detect collapse of 100% identical repeats

- AMOS Tool: analyzeReadDepth
  - 2.5x mean coverage
Assembly Investigator
Assembly Investigator Goals

Interactively explore and analyze

- **Libraries**
  - Insert Sizes, Read Length, Inserts

- **Scaffolds & Contigs**
  - Sizes, Composition, Sequence, Multiple Alignment, SNP Barcode

- **Inserts**
  - Happiness, Coverage, CE Statistic

- **Reads**
  - Clear Range, Quality Values, Chromatograms

- **Features**
  - Arbitrary regions of interest
  - Including Misassembly Signatures!!!
Main Window: Contig View
Contig View Expanded

- Quality Values
- Normalized Chromatogram
- No size restrictions
Chromatograms are loaded from specified directories, or on demand from Trace Archive.
Main Window: Contig View

Display Inserts
Insert View
Insert View

Toolbar
Position
Insert and Read Coverage
Scaffold Features
Inserts
Details

Current Contig Position
Standard Feature Types

[B] **Breakpoint**
Alignment ends at this position

[C] **Coverage**
Location of unusual mate coverage (asmQC)

[S] **SNPs**
Location of Correlated SNPs

[U] **Unitig**
Used to report location of surrogate unitigs in CA assemblies

[X] **Other**
All other Features

---

Loading Features:
$ loadFeatures bankname featfile

Featfile format:
Contigid type end5 end3 comment
**Insert Happiness**

- **Happy**
  - Oriented Correctly &&
  - \(|\text{Insert Size} - \text{Library.mean}| \leq \text{Happy-Distance} \times \text{Library.sd}\)

- **Stretched**
  - Oriented Correctly &&
  - \(\text{Insert Size} > \text{Library.mean} + \text{Happy-Distance} \times \text{Library.sd}\)

- **Compressed**
  - Oriented Correctly &&
  - \(\text{Insert Size} < \text{Library.mean} - \text{Happy-Distance} \times \text{Library.sd}\)

- **Misoriented**
  - Same or Outies

---

- **Linking**
  - Read’s mate is in some other scaffold

- **Singleton**
  - Read’s mate is a singleton

- **Unmated**
  - No mate was provided for read

---

Both mates present

Only 1 read present
Histographs & Statistics

- Insert Size
- Read Length
- GC Content
- Overall Statistics

- Bird’s eye view of data and assembly quality
# Assembly Reports

- **Contigs**
  - Select from 172 contigs in xo4.bin

- **Features**
  - Select from 372 features

- **Reads**
  - Contig information

- **Scaffolds**
  - Select from 10 scaffolds in xo4.bin

- **Full Integration:** “Double click takes you there”
Assembly Reports

- Contigs
- Reads
- Features
- Scaffolds

Misassembly Walkthrough: Correlated SNPs

- Full Integration: “Double click takes you there”
SNP View

Zoom Out

SNP Sorted Reads

Polymorphism View
SNP Barcode

Colored Rectangle indicate the positions and composition of the SNPs
SNP Barcode

Colored Rectangle indicate the positions and composition of the SNPs
Collapsed Repeat

-5.5 CE Dip

Individual Compressed Mates

Read Coverage Spike

68 Correlated SNPs
Confirmed Misassembly

Collapsed repeat
- Compressed mates (-5.5 CE Stat)
- Correlated SNPs (68 Positions within 1400bp)
- Spike in Read Coverage
Fixing collapsed repeats with AMOS

1. Select reads and mates in region of collapse.
   - AMOS: findMissingMates, select-reads

2. Reassemble those reads with stricter parameters.
   - AMOS: minimus

3. Inspect new assembly to ensure misassembly was corrected.
   - AMOS: amosvalidate, Assembly Investigator

4. Patch the collapsed region of the original assembly with corrected version.
   - AMOS: stitchContigs
Replace the reads between the stitch reads in the original contig with corresponding region in the patch contig.

Can also close gaps or fix contig ends.
Current Research

- **Misassembly signature detection**
  - Read alignment breaks
  - Singleton / Missing mate analysis
  - Integrated & Dynamic Thresholds of detection

- **Automated assembly improvement**
  - Automatic contig patching
  - Automatic repeat separation
  - Automatic parameter tuning

- **Exotic Assembly**
  - Multiple haplotypes
  - Metagenomic assembly
  - 454 & Sanger Sequencing Hybrids
More Information

- Contact AMOS
  - amos-help [ at ] lists.sourceforge.net

- AMOS Team
  - Art Delcher
  - Adam Phillippy
  - Mihai Pop
  - Steven Salzberg
  - Michael Schatz
  - Dan Sommer