Genome Assembly Forensics: Finding the Elusive Mis-assembly

Michael Schatz, Adam Phillippy, and Mihai Pop

Center for Bioinformatics and Computation Biology, University of Maryland, College Park, MD, USA

http://amos.sourceforge.net/forensics

Genome Biology 2008, 9:R55

Abstract

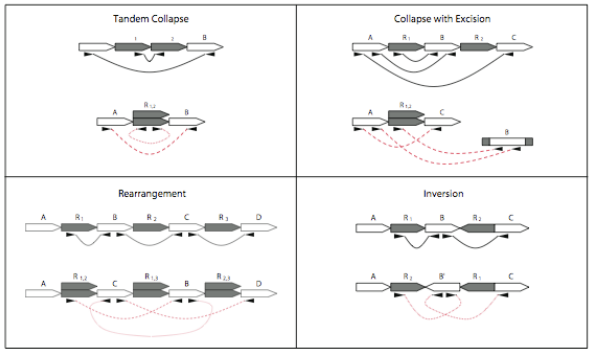
Since the initial "draft" sequence of the human genome was released in 2001, it has become clear that it was not an entirely accurate reconstruction of the genome. Despite significant advances in sequencing and assembly since then, genome sequencing continues to be an inexact process. Genome finishing and validation have remained a largely manual and expensive process, and consequently, many genomes are presented as draft assemblies. Draft assemblies are of unknown quality and potentially contain significant mis-assemblies, such as collapsed repeats, sequence excision, or artificial rearrangements. Too often these assemblies are judged only by contig size, with larger contigs preferred without regard to quality, because it has been difficult to gauge large scale assembly quality.

Our new automated software pipeline, *amosvalidate*, addresses this deficiency and automatically detects mis-assemblies using a battery of known and novel assembly quality metrics. Instead of focusing on a single assembly characteristic as other validation approaches have tried, the power of our approach comes from leveraging multiple sources of evidence. *amosvalidate* statistically analyzes mate-pair orientations and separations, repeat content, depth-of-coverage, correlated polymorphisms in the read alignments, and read alignment breakpoints to identify structurally suspicious regions of the assembly. The suspicious regions identified by individual metrics are then clustered and combined to identify (with high confidence) regions that are mis-assembled. This approach is necessary for accurately detecting mis-assemblies because each of the individual characteristics has unavoidable natural variation, but, when considered together, have greatly increased analysis power. Furthermore, our pipeline can easily be adjusted to analyze assemblies utilizing new sequencing technologies where some metrics are unreliable or not available, such as base pair quality or mate pairs.

Our validation pipeline provides a robust measure of assembly quality that goes beyond the simple measures commonly reported. Evaluation of the pipeline has shown it to be highly sensitive for mis-assembly detection, and has revealed mis-assemblies in both draft and finished genomes. This is particularly troubling as scientists move away from the "gene by gene" paradigm and attempt to understand the global organization of genomes. Without a correct genome sequence or even a clear understanding of the errors present, such studies may draw incorrect conclusions. Our goals are to help scientists locate mis-assembled regions of an assembly and help them correct those regions by focusing their efforts where it is needed most. *amosvalidate* is compatible with many common assembly formats and is released open-source at http://amos.sourceforge.net.

Genome Mis-assemblies

Mis-assemblies almost always occur from complications related to repeated sequences in the genome. For example, a common mis-assembly is for the assembler to mis-compute the number of occurrences of a repeat, and either include too few or too many. The first type, called a collapse, can also excise other unique portions of the genome between the repeat instances. Repeats can also cause the assembler to rearrange the unique portions of the genome between instances of a repeat, or even invert the orientation of those sequences. Recognizing these mistakes requires a careful analysis of all of the assembly data for various assembly signatures.



Mis-assembly Signatures

1. Mate-Pair Validation

Are the mate-pairs correctly oriented and their separation within the library distribution?

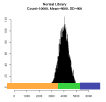
CE Statistic

$CE = (\mu_l - \mu_g) / (\sigma_g / \sqrt{N})$

Local Mean: 4048
C/E Stat: +0.33
Near 0 indicates overall happiness

Local Mean: 4461
C/E Stat: +3.26
CE Stat ≥ 3 indicates expansion

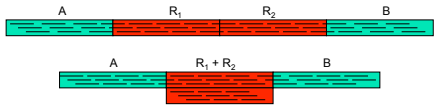
Local Mean: 3488
C/E Stat: -3.61
CE Stat ≤ 3 indicates compression



2. Coverage Analysis

Is the depth of coverage higher than usual?

Compute the depth of coverage at each position in the assembly



3. Repeat Analysis

Does the repeat occur enough in enough copies?

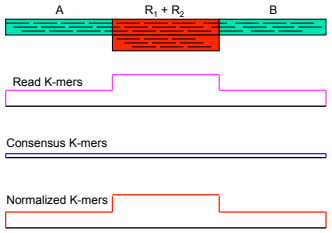
Normalized K-mers

$K^* = K_R / K_C$

K_R = K-mer frequency computed in the reads

K_C = K-mer frequency computed in the consensus

The sequence of an N copy repeat should occur N times in the consensus.



4. Micro-Heterogeneity

Do the reads agree with each other?

Multiple reads with same conflicting base are unlikely

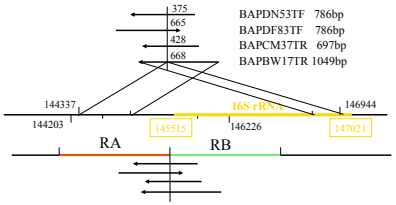
- 1x QV 30: 1/1000 base calling error
- 2x QV 30: 1/1,000,000 base calling error
- 3x QV 30: 1/1,000,000,000 base calling error

Regions of correlated SNPs are likely to be assembly errors or interesting biological events

5.Read Breakpoints

Do the singleton reads align to the assembly?

A consistent breakpoint shared by multiple reads can indicate a collapsed repeat.

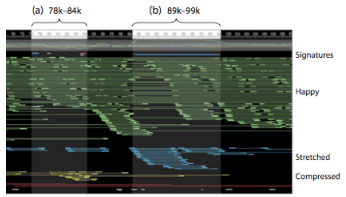


Automatic Validation Pipeline

Our validation pipeline, amosvalidate, scans the assembly for mis-assembly signatures. Independently, each signature type may miss certain mis-assemblies or report false-positives. Many false positive signature derive from statistical variation or otherwise innocuous events. For example, a large variance in a library's mate-pair size distribution can cause clusters of overlapping stretched or compressed mate-pairs. However, combining multiple mis-assembly signatures increases the likelihood that all mis-assemblies will be detected, and the tagged regions identify are true errors in the assembly. For example, a region with a largely negative CE value, and a high rate of correlated SNPs is very likely to truly be a collapsed repeat. This particular combination is especially strong, since mate-pair and sequence data are independent sources.

Mis-assembly signatures are combined by merging signatures that co-occur within a small window (2 Kbp by default). If multiple signatures of at least two different evidence types occur within this window, the region is flagged as 'suspicious'. Each such region is reported along with detailed information about the individual signatures, and forms the initial focus for subsequent validation and correction efforts. For manual analysis, these regions, along with the individual mis-assembly features, can be viewed alongside the assembly data in the AMOS assembly viewer, Hawkeye.

An example D. virilis mis-assembly shown in Hawkeye. Sequencing reads are represented as thick boxes connected to their mate by thin lines. Correctly sized (happy) mates are shown in green, stretched in blue, and compressed in yellow. A CE statistic plot is given at the top, with mis-assembly signatures plotted directly below as intervals. (a) The amosvalidate region that suggests a compression mis-assembly. (b) The amosvalidate region that suggests an expansion mis-assembly.



				Mis-assembly signatures			Suspicious regions		
Species	Len	Ctgs	Errs	Num	Valid	Sens	Num	Valid	Sens
B. anthracis	5.2	87	2	1,336	21	100.0	127	2	100.0
B. suis	3.4	120	10	1,047	30	80.0	158	9	90.0
C. burnetii	2.0	55	22	1,375	70	100.0	124	19	100.0
C. caviae	1.4	270	12	625	16	83.3	50	8	66.7
C. jejuni	1.8	53	5	290	11	80.0	61	3	60.0
D. ethenogenes	1.8	632	12	688	22	91.7	88	9	100.0
F. succinogenes	4.0	455	21	1,670	27	95.2	266	14	66.7
L. monocytogenes	2.9	172	1	1,381	5	100.0	201	1	100.0
M. capricolum	1.0	17	3	83	0	0.0	16	0	0.0
N. sennetsu	0.9	16	0	91	0	NA	13	0	NA
P. intermedia	2.7	243	21	1,655	57	100.0	201	20	100.0
P. syringae	6.4	274	64	2,841	200	98.4	366	55	98.4
S. agalactiae	2.1	127	21	687	53	95.2	112	18	85.7
S. aureus	2.8	824	41	1,850	69	97.6	227	18	75.6
W. pipientis	3.3	2017	31	761	92	100.0	132	30	100.0
X. oryzae	5.0	50	151	2,569	379	100.0	100	69	100.0
Totals	46.8	5412	417	18,949	1,052	96.9	2,242	275	92.6

In the systematic evaluation of 16 Phrap assemblies using amosvalidate, we found the sensitivity of our methods is quite good; 96.9% of known mis-assemblies are identified by one or more amosvalidate signatures, and 92.6% are identified by one or more amosvalidate suspicious regions. The over-prediction of mis-assembly signatures can be mostly ignored, because most follow up analysis will only use the more confident suspicious regions. The over-prediction of suspicious regions appears to indicate a limitation of our methods when used with Phrap, since we found Phrap generates otherwise correct consensus sequence, but with many incorrectly place reads into the wrong repeat copies. These mis-placed reads will often create false-positive signatures. We argue that this is the correct behavior for amosvalidate to ensure every true mis-assembly is detected..