Methods: Assembly Markup

1. File Formats and Distribution

The markup that is distributed with an assembly contains both the overall assessment (i.e. predictions for where assembly errors might occur) as well as the raw underlying data (i.e. the individual pathologies).

This section describes the file formats only; for a detailed description of the content, see below.

The individual files are currently provided both in XML and GFF format:

Error predictions:
defects{.xml,.gff} – predicted locations and probabilities of errors

Raw pathologies:
hqd{.xml,.gff} – disagreements between reads and consensus
qual{.xml,.gff} – low base quality regions and suspicious contig ends
repeats{.xml,.gff} – repeats found after k-mer analysis with multiplicity (on a log scale)
matches{.xml,.gff} (optional) – repeats in the assembly with coordinates of alignments
stretch{.xml,.gff} – pathologies found from analyzing mate pair information

All markup files list intervals containing the following information:

- supercontig id (zero-based)
- contig id (zero-based)
- begin of interval (first base) on contig (zero-based)
- end of interval (last base) on contig (zero-based)
- type of interval (i.e. “possible_defect”)
- name of interval (optional, further information about the interval depending on the type)
- score (the meaning of which depends on the type)

The XML file is of the following form:

```
<!-- Broad Institute/MIT assembly markup file - DO NOT EDIT! -->

<assembly>

  <supercontig id="0"  len="3192231">
    <contig id="0">
      <coordinates start="1211"  stop="1211">
        <mark type="HQD_alt_seq"   name="G:T"  score="5"/>
      </coordinates>
    </contig>

    <contig id="84">
      <coordinates start="2989"  stop="2992">
        <mark type="HQD_alt_seq"   name="ACC:TACT"  score="5"/>
      </coordinates>
      <coordinates start="2990"  stop="2990">
        <mark type="unsupported_consensus"  score="1"/>
      </coordinates>
      <coordinates start="3853"  stop="3857">
        <mark type="HQD_alt_seq"   name="ACTTT:CTTGC"  score="3"/>
      </coordinates>
    </contig>
  </supercontig>

</assembly>
```

For each scaffold, it lists all intervals grouped by contigs. Unlike the scaffolds, contigs without intervals are not listed.

The GFF version is a simple tab-delimited flat file of the form

```
contig_1 ARACHNE qcmarkup   0     1797  1  . .  missing_unique_fosmid_coverage
contig_2 ARACHNE qcmarkup   37461 37761 0  . .  plasmid_compression:stretch(sd):-1.750000
contig_2 ARACHNE qcmarkup   68840 70796 2  . .  lonely_read
```

where the items are

1. contig id
2. “ARACHNE”
3. “qcmarkup”
4. start of interval
5. end of interval
6. score
7. “.”
8. “.”
9. Type, followed by “:”, followed by name.

Since the GFF version is simply a different format but contains the same data, we will refer only to the XML file version from now on.

2. Contents of Individual Files

a.) Overall Quality Assessment (defects.xml)

Each interval that contains a predicted probability to contain a defect has the following attributes:

- type: currently always “possible_defect”
- name: probabilities to find defects of three types
    - *misjoin*: Two adjacent regions in the assembly are in reality either more than 10,000 bases apart, are located on different chromosomes, or one of the regions is inverted relative to the “truth”.
    - *indel* (insertion or deletion): two regions in the assembly should, according to the “truth”, more than 100 bases but less than 10,000 bases closer to, or farther apart from each other.
    - *base errors*: consensus sequence labeled as having a quality score of 50 or more does not match the “truth”, including indels up to 100 bp.

The probabilities, which are a composite assessment based on all individual pathologies, are encoded in one single string in the form:

“p_misjoin:4.798377%;p_indel:6.921100%;p_baseerrors:87.313500%;”

In this example, there is a relatively low (but still higher than zero) probability of finding structural errors (misjoins, indels) in this region, but the probability of finding high-quality bases that are wrong is very high.

The entry for score is computed from the weighted number of individual pathologies, which are listed individually in the files described below.

b.) Read-to-consensus disagreements (hqd.xml)

The type "*HQD_alt_seq*" refers to intervals in which at least two reads disagree with the consensus, but agree with each other. In addition, disagreeing reads are required to match 48 bases on each end of the region in question. The score refers to the number of reads voting for alternative consensus sequence. The name lists both the alternative sequence and the sequence found in consensus (separated by a ":"), so, e.g.

```
<coordinates start="2989"  stop="2992">
  <mark type="HQD_alt_seq"   name="ACC:TACT"  score="5"/>
</coordinates>
```

means that there are five reads that suggest that the three bases "ACC", position 2989 through 2992, be replaced with the four bases "TACT". Note that this can be caused by assembly errors, sequencing errors or polymorphism.

Another type, "*unsupported_consensus*", indicates regions in which there is not a single read that matches exactly with consensus for at least 48 bases on either side. This is not necessarily a consensus problem but can be caused by low quality reads and/or haplotype differences. The score refers to the length of the interval.

c.) Base Quality (qual.xml)

The type "*low_qual*" is computed by smoothing consensus bases and marking regions in which the base quality score is lower than 30. The score is always 0.

The type "*single_read_hang*" flags possibly chimeric reads at the beginnings or ends of contigs. The score is the length of the interval.

d.) Repetitive k-mers (repeats.xml)

The type is always "*repeat-48-mer*", the score gives the multiplicity (i.e., how many times this k-mer is found in the assembly consensus sequence) on a logarithmic scale. The scale (in part) is:

Mult	=>	Score
2	=>	2
3	=>	3
4	=>	3
5	=>	4
6	=>	5
7	=>	6
8	=>	6
9	=>	7
10	=>	7
30	=>	15
50	=>	20
70	=>	23
90	=>	26

```
100  =>    27
300  =>    39
500  =>    45
700  =>    49
900  =>    52
```

In this example, a region is a 480 base pair long copy of a repeat (base 663 through 1143), with a total of 6 copies (see log scale) in the assembly. One copy has a single base difference at base 1091:

```
<coordinates start="663"  stop="1090">
  <mark type="repeat-48-mer"  score="5"/>
</coordinates>
<coordinates start="1091"  stop="1091">
  <mark type="repeat-48-mer"  score="4"/>
</coordinates>
<coordinates start="1092"  stop="1143">
  <mark type="repeat-48-mer"  score="5"/>
</coordinates>
```

This annotation reflects properties of the underlying genome more than the assembly, but repetitive regions are more apt to contain assembly errors.

e.) Repeats (matches.xml)

Based on k-mer matches, this file is a more detailed annotation of repeats. Here, we list the corresponding regions of a repeat, i.e. where the copies are located. The repeat has to be at least 128 bases long; isolated differences are allowed but there have to be sufficient perfect 48 base pair matches.

The location of matching regions is encoded into the type, e.g.

match_s1_c14-[37808-37989][1621318-1621499]+

means that one other copy is in scaffold 1, contig 14 (both zero based), from base 37808 through 37989 in contig coordinates, from base 1621318 through 1621499 in Arachne-internal supercontig coordinates[1].

f.) Mate-pair pathologies (stretch.xml)

This file lists a number of different pathologies based on mate-pair link analysis.
The type “*illogical_read*” describes clusters of read pairs (a minimum of two is required), for which both end reads point in the same direction. The score is the number of reads found in the interval.

The type “*lonely_read*” marks regions in which there are clusters (a minimum of two) of reads, for which the mates have been assembled in different scaffolds. The score is the number of reads found in the interval.

Note that in this file, different types can be combined when sharing the same interval, e.g. the type “illogical_reads.lonely_reads” means the occurrence of both events.

[1] Caution: these coordinates DO NOT correspond the coordinates in the agp file or chromosme.fasta or supercontigs.fasta.

The types “*fosmid_compression*”, “*fosmid_expansion*”, “*plasmid_compression*” and “*plasmid_expansion*” are the results of statistical mate pair linking analysis, broken down by insert sizes. There has to be a minimum of two inserts crossing a region. Then, the mean deviation for all inserts spanning the region from the library size is reported, e.g.

```
<coordinates start="758"  stop="11758">
  <mark type="fosmid_compression" name="stretch(sd):-2.410000 "  score="1"/>
</coordinates>
```

indicates that the average stretchiness of all fosmids crossing the region is 2.4 standard deviations below the mean, thus making this spot a candidate for a possible deletion of sequence.

The types “*missing_unique_fosmid_coverage*” and “*missing_unique_plasmid_coverage*” refer to regions in the assembly where there is either a lack of insert coverage, broken down by insert library sizes, or there is a repetitive region longer than an insert library size. In other words, if both end reads of an insert are in repeat regions, this insert gets discarded as indicator of assembly correctness. A complete lack of inserts with at least one end read anchored in unique sequence may signal a misassembled region. Of course, it may also just reflect complexity of the actual genome or problems that arose during the sequencing process.