# A Glossary of Terms for the Celera Assembler

## Basic Descriptive Terms:

Read:
A single sequence read produced by an ABI 3700 by our *internal* production pipeline.

Guide:
A read-sized sequence of the relevant genome supplied from an *external* data source, e.g. an STS marker, a BAC-end, or a fabricated piece of a known BAC.

Insert:
A segment of the target genome placed into a vector and ultimately end-sequenced by us. For example, we are currently planning on sequencing the ends of a 4/1 mix of 2Kbp and 10Kbp inserts.

“Fragment”:
Either a guide or a read. Unfortunately this term has a long history of different uses by different groups. In particularly, one may actually be talking about inserts. Usually the intended meaning is clear from context, but when it isn’t and its important to understand the precise meaning, be sure to ask for clarification.

Mate-Pair or Mates:
A pair of reads taken from the end of a given insert.

Overlap:
A pair of sequences, say A and B, overlap if there is an interval of A and an interval of B that match to within a user-specified level of similarity. If the sequencing error rate is less than 2% than a match with fewer than 4% differences constitutes an overlap. Typically, one is also implying that the segments involved constitute either a suffix/prefix pair (a "dovetail overlap") or all of one of the two sequences (a "containment overlap"). In pictures,

```
   A -------------------          or     A --------------------.
         ------------------- B                 ---------- B
```

Layout:
A layout is a (partial) positioning of a set of reads with respect to each other subject to the one constraint that every pair of reads that overlap in the layout do so as defined immediately above. The term layout is intended to specifically speak to the arrangement of the reads as opposed to their mutual connectivity (as in "contig" below) or the sequence(s) the set models (as in "consensus" below). A layout includes the orientation of the fragments and in the case that reads are mate-linked gives the estimated distance between contigs that contain each end of a mate pairing.

Contig:
A maximal set of reads in a layout which in aggregate cover a contiguous interval.

Scaffold:
A maximal set of contigs in a layout that are connected together by mate-links.

Consensus Sequence (or simply Consensus):
Given a collection of overlapping reads, that do not precisely match along their overlaps, a consensus sequence for the collection is, loosely speaking, one's best guess at the sequence the reads were sampled from. Often people mean something more precise: the mathematical definition of consensus sequence is one for which the sum of the differences between the consensus sequence and each one of the reads is minimal.

Multi Alignment:
A multi-alignment of a set of overlapping fragments is a matrix in which a row is a possibly empty prefix of blanks, followed by the sequence of a fragment interspersed with dashes, followed by a possibly empty suffix of blanks. One generally seeks the multi-alignment of the fragments that exposes their similarity and supports the evidence for a particular consensus sequence. Indeed, any computation that produces a consensus either implicitly or explicitly computes a multi-alignment of the underlying reads.

Assembly:
A layout and associated consensus sequence(s) and/or multi-alignment(s). In other words, we use this term to speak of a tentative reconstruction of segments of the target sequence and the locations from which the reads were sampled.

Assembly Snapshot/Release:
An assembly snapshot or release is a recording of the current state of knowledge about the potential assemblies of the set of data collected as of a given time. This term explicitly speaks to the incremental nature of the Celera assembler. A snapshot consists of a set of putative contigs, their consensus sequences, the potential orderings between them (based on mate information), validating and conflicting evidence, and partial maps of the anchor information used as input to the assembler. A snapshot thus models many potential assemblies that vary at the level of scaffolds.

Region, Interval:
A portion of a consensus sequence in an assembly snapshot defined by a pair of coordinates, contig identifier, and a release number

Feature:
A specific Region with certain characteristic(s) assigned to it.

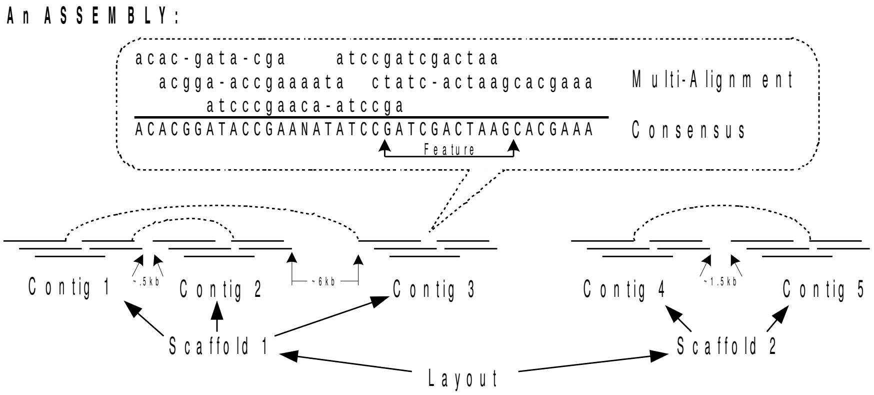

*Figure 1*: Illustration of Glossary Terms.

Figure 1 illustrates most of the terms above. One should note the following relationships between the terms:

- A snapshot models a set of possible assemblies.
- An assembly is a layout and associated consensus sequences/multi-alignments.
- A layout is an unordered set of scaffolds.
- A scaffold is an ordered, oriented, and distanced set of contigs.
- A contig is an overlapping set of fragments.

# Algorithm Specific Terms:

Unitig/Chunk:

A uniquely assembleable subset of overlapping fragments. A unitig and/or chunk is an assembly of fragments for which there are no competing choices in terms of internal overlaps. This means that a chunk is either a correctly assembled portion of a contig or it is an overcompressed assembly of several high-fidelity copies of a repeat. Every fragment belongs to one chunk.

Branch Point:

A branch point is a position on a fragment and/or chunk that is known to represent the boundary of a repetitive element. The inference one would like to make is that one side of the branchpoint is unique sequence and the other is repetitive, but internal repeat boundaries of micro- and mini-satellites are also detected as branchpoints.

U-unitig:

A unitig determined to represent unique sequence (between its branch points) by the assembler.
Also called discriminator-unique chunks.

Singleton unitig:

A unitig consisting of a single fragment.

Valid-unitig:

A unitig that involves only true overlaps, i.e., is correctly assembled. Hopefully, all U-unitigs are valid but this need not be so.

Invalid-unitig:
A unitig that involves one or more repeat-induced overalps, i.e., is an overcompressed incorrect subassembly.

## AUTHORS

Gene Myers

Created October 14, ’98

Last revised May 24, ‘99