Source: 454 Sequencing System Software Manual – GS De Novo Assembler

5. CDNA / TRANSCRIPTOME SEQUENCING APPENDIX

5.1 Introduction to cDNA Sequencing Analysis

cDNA is produced from the RNA (often, mRNA) in a sample. The cDNA is then sequenced resulting in reads that represent the original sample's mRNA. These reads can then be mapped against one or more references to gain information about the representation of individual mRNAs and genes from which the mRNAs are transcribed and their level of variation in the original sample. Alternatively, the reads can be assembled together to discover novel transcripts and splice variants for those mRNAs represented by multiple overlapping reads.

5.2 Transcriptome Assembly Concepts

5.2.1.1 Isogroup

An isogroup is a collection of contigs containing reads that imply connections between them. A discussion of the assembly process (see Section 1.1) explains how breaks can be introduced into the multiple alignments of overlapping reads, leading to branching structures between them. After attempting to resolve the branching structures, the Transcriptome Assembler groups all contigs whose branches could not be resolved into collections called isogroups. Using rules described in the following section, the assembler traverses the various paths through the contigs in an isogroup to produce the set of isotigs that gets reported. All possible paths through the contigs in an isogroup are traversed unless one or more thresholds is reached (see Section 5.2.2).

5.2.1.2 Isotig

An isotig is meant to be analogous to an individual transcript. Different isotigs from a given isogroup can be inferred splice-variants. The reported isotigs are the putative transcripts that can be constructed using overlapping reads provided as input to the assembler. Connections between contigs in an isogroup are represented by sequences (reads) that have alignments diverging consistently towards two or more different contigs (see Figure 103) or by a depth spike (Section 5.2.2.2). Traversal from the start contig to the end contig or from the end contig to the start contig should yield the same but reverse-complemented isotig sequence.

While many reads may contain poly-A tails, these tails are trimmed off prior to assembling the reads. Presently, the assembler ignores the fact that poly-A tails existed, so the orientation of reads in the assembly cannot be determined. Because of this lack of directionality, an isotig may be output as the reverse-complement of the biological transcript it represents. Contigs forming an isotig may be thought of as exons. This is not strictly correct, however, since untranslated regions (UTRs) and introns (in the case of primary transcripts) may exists in the reads generated from the sample.