# *De novo* assembly using Newbler

Dag Ahrén
Microbial Ecology
Lund University
Sweden

# Data Analysis Suite

Data-Analysis software for processing, assembling and mapping the 454 reads:

**sffinfo** - extract fasta, quality and flowgrams as text from .sff files.
**sfffile** - join, split or trim sff files.

**gsAssembler (Newbler)** - to assembly reads into contigs/isotigs.
**gsMapper** - to map reads to a transcriptome or genome reference.
**gsAmplicon** – to analyse Variants in Amplicons.

# Data Analysis Suite

**Newbler**

# Newbler

- *De novo* assembly of:

  genomes

  transcriptomes

  amplicons (amplified fragments, e.g. with PCR)

# Newbler

- *De novo* assembly of:

  genomes

  transcriptomes

  amplicons (amplified fragments, e.g. with PCR)

# Why sequence transcriptomes?

- Requires less sequencing (cheaper) than sequencing an entire genome
- Provides more dynamic view of the activity in a cell.
- Gives relative expression levels for different cells under different conditions.
- Could identify alternate splicing, and fusion genes.
- Focuses on gene sequences, which are often the main research focus.
- Great complement to genome sequencing during gene prediction

# Challenges for assembly (1)

- **PCR artifacts (***eg.* **Chimeras** and **Mutations)**

- **Sequencing errors**, such as "***Homopolymer***" **errors** – when eg. 3+ run of same base.

- **MID**'s (multiplex indexes), **primers/adapters** (eg. *SMART* adapters used to synthesise cDNA) still in the raw reads.

- **Repeats** and **large** or **polyploid** genomes – repeated sequences in the transcriptome make assembly more difficult.

# Challenges for assembly (2)

- Extra sample preparation steps in cDNA synthesis - more risk of cloning errors or contamination, wider range of read lengths.

- Large expression level range - some transcripts have low read coverage and some very high coverage.

- Alternative splicing one gene gives different transcripts

# Pros and cons of Newbler

Cons:

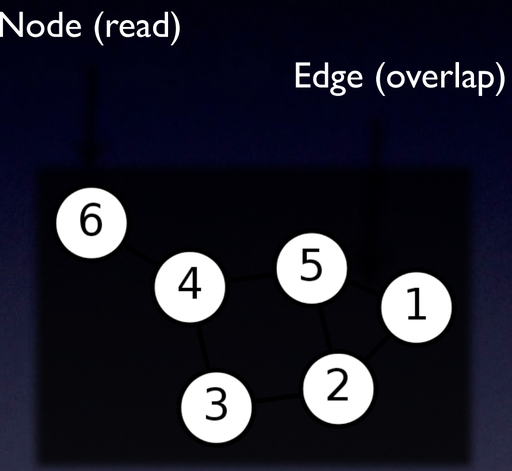
- Designed for 454 sequencing only (to some extent also Sanger)
- Proprietary software provided by Roche

# Pros and cons of Newbler

Pros:

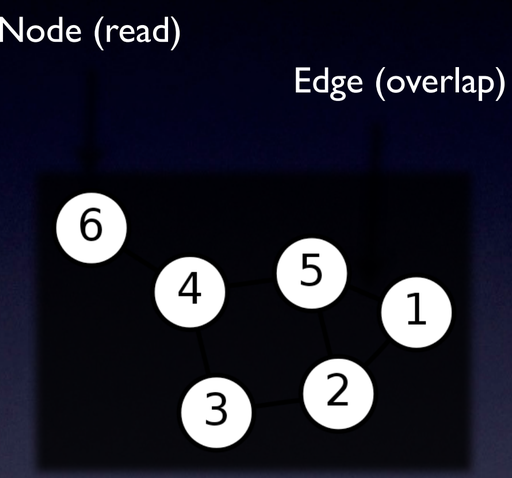
- *De facto* standard for 454 sequencing centers
- Relatively easy to use
- Graphics interface
- Roche support once you register as a user

# Overlap Layout Consensus (OLC) algorithm

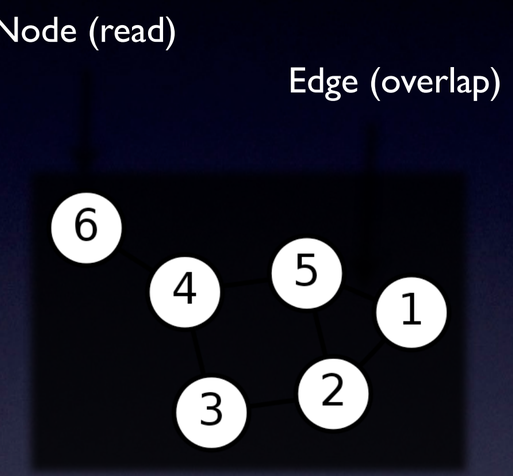


# Overlap Layout Consensus (OLC) algorithm

Construct a (directed) overlap graph, where nodes represent reads and edges represent overlap. Paths are contigs in this graph.



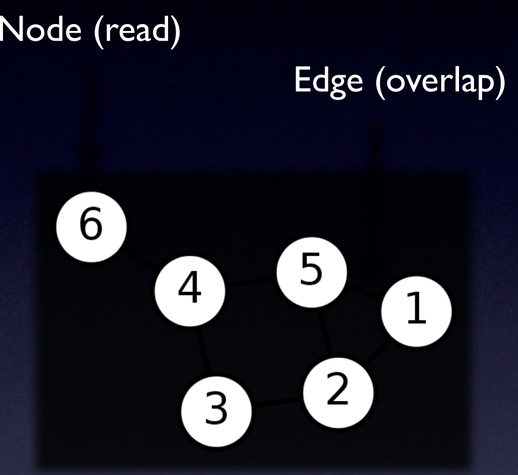
# Overlap Layout Consensus (OLC) algorithm




Construct a (directed) overlap graph, where nodes represent reads and edges represent overlap. Paths are contigs in this graph.

Problem: Find the consensus sequence by finding a path that visits all nodes in layout graph.

# Overlap Layout Consensus (OLC) algorithm

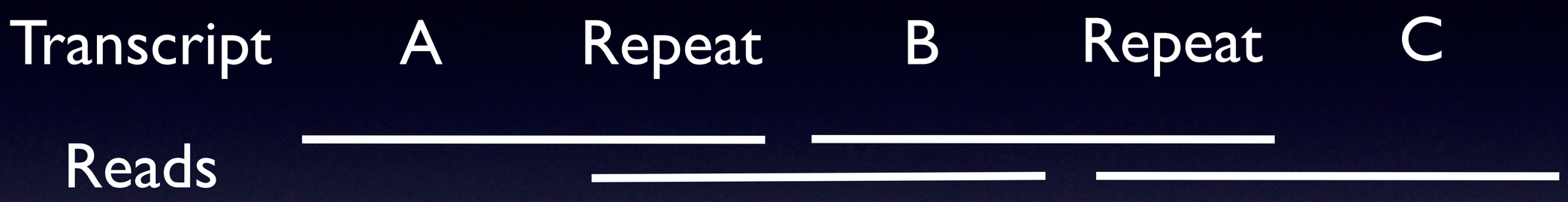



Construct a (directed) overlap graph, where nodes represent reads and edges represent overlap. Paths are contigs in this graph.

Problem: Find the consensus sequence by finding a path that visits all nodes in layout graph.

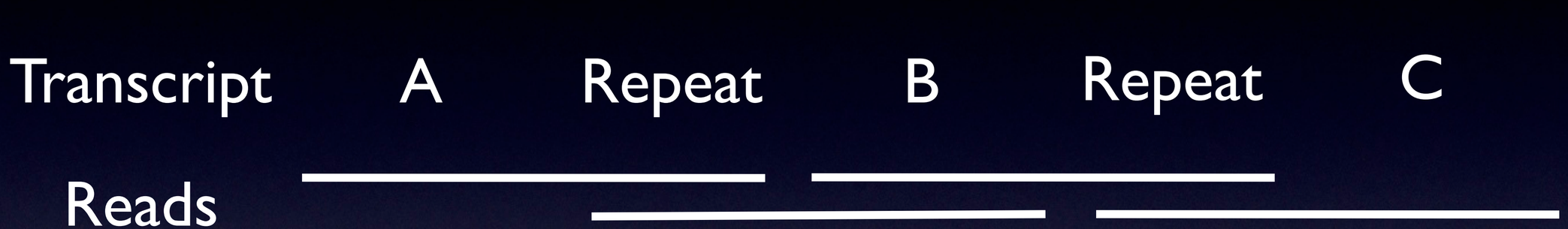
Note: Becomes complex since we need to take errors into account!

**cDNA → Reads → Alignments → Contig graph → Final untangled assembly**



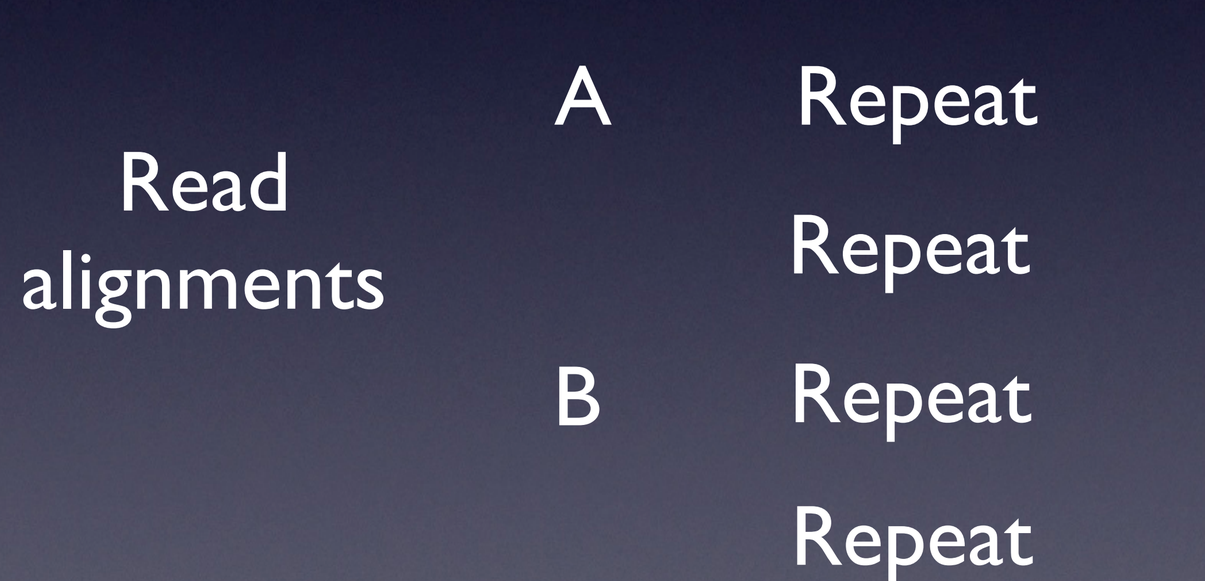
# How does Newbler work?

**cDNA → Reads → Alignments → Contig graph → Final untangled assembly**



# How does Newbler work?

**cDNA → Reads → Alignments → Contig graph → Final untangled assembly**

# How does Newbler work?

**cDNA → Reads → Alignments → Contig graph → Final untangled assembly**

Transcript A Repeat B Repeat C

Reads

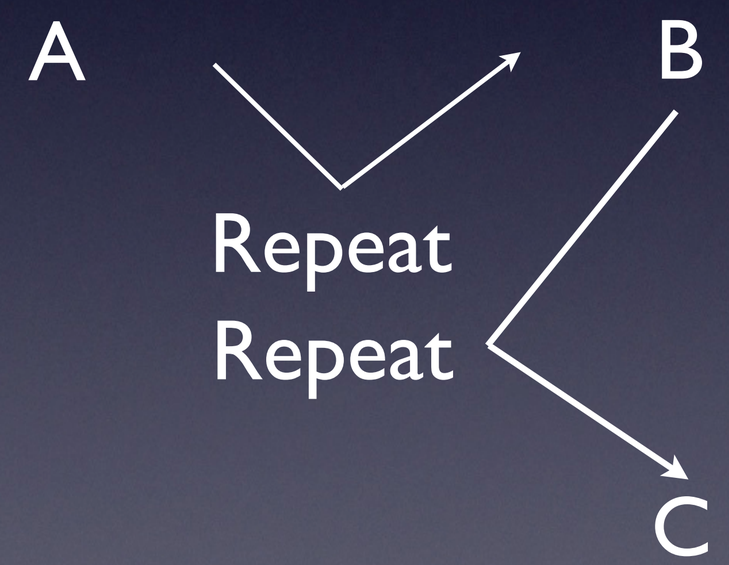
Read alignments

A Repeat

Repeat B

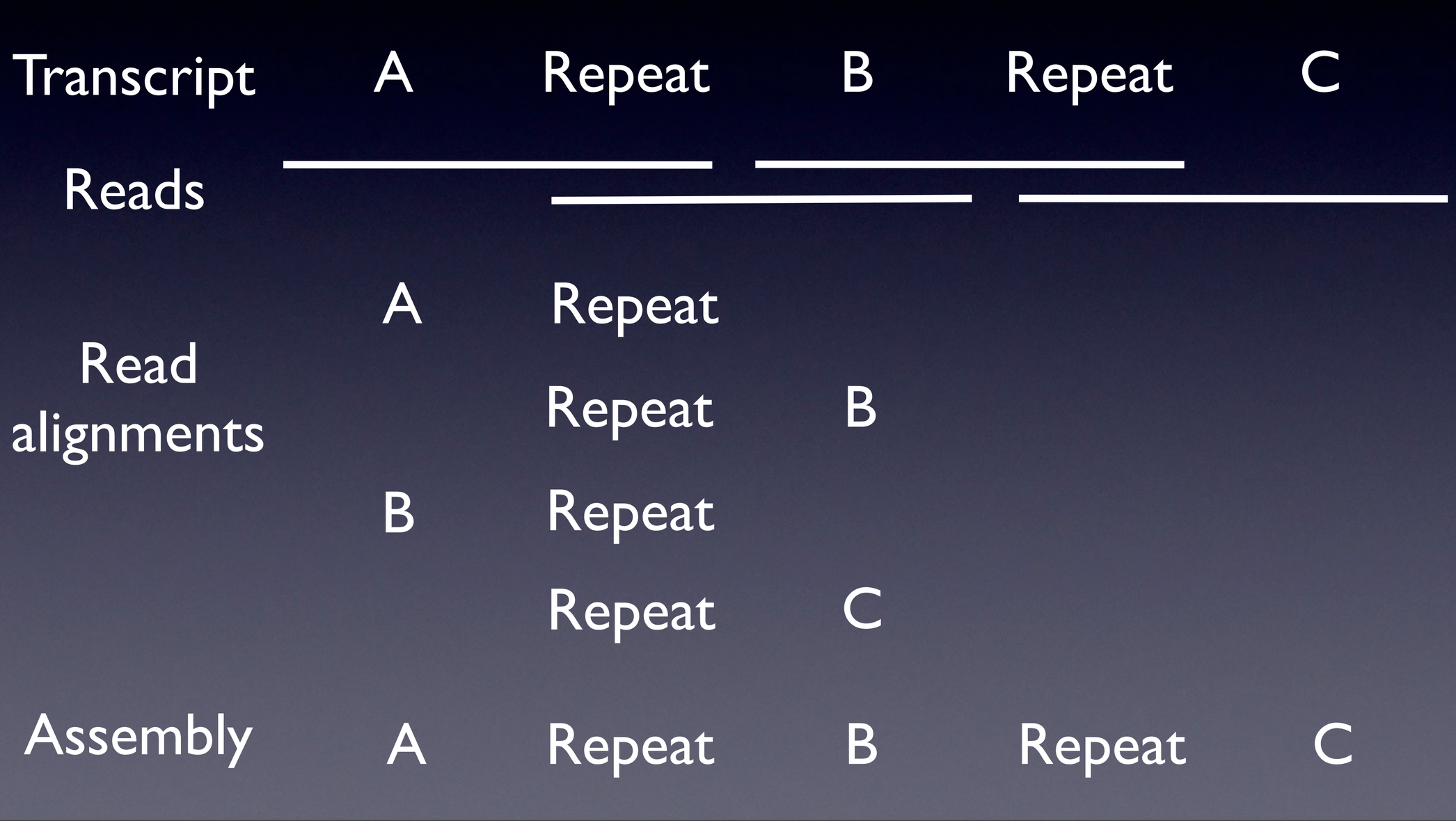
B Repeat

Repeat C



# How does Newbler work?

**cDNA → Reads → Alignments → Contig graph → Final untangled assembly**



# Output files from a Newbler transcriptome assembly

A directory is generated and given the project name

All output files are located in this directory

Fasta and quality files

Ace files for viewing etc

Statistics and log files of the assembly

# 454NewblerProgress.txt

Generating project

reading sequence files (SFF)

Indexing reads

Overlap detection

Aligning seeds

Building tree/Computing alignments

Detangling alignments

Building contigs/scaffolds

Generating output

# 454NewblerProgress.txt

Generating project

reading sequence files (SFF)

Indexing reads

Overlap detection

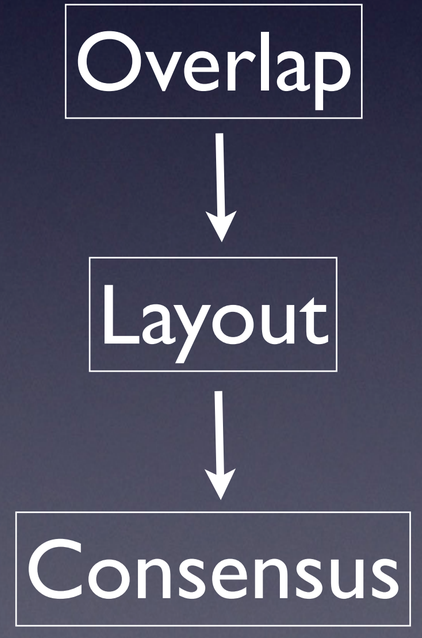
Aligning seeds

Building tree/Computing alignments

Detangling alignments

Building contigs/scaffolds

Generating output



# Isogroups, Isotigs, Contigs ?

Some definitions to understand Newbler output:

- An **Isogroup**:represent a **gene**

  - collection of isotigs containing reads that imply connections between the isotigs.

- An **Isotig**: represents a **transcript**.

  - different isotigs from a given isogroup can be inferred splice-variants.

**Contigs**: contigs forming an isotig may be thought of as **exons**.

- this is not strictly correct, as untranslated regions (UTRs) and introns (in the case of primary transcripts) may exist in the reads generated from the sample.

# Transcriptome options

Newbler collects into “Isogroups”, then creates “Isotigs”

New options for transcriptomes:

-cdna = for transcriptome (cDNA assembly)

-ig = max contigs in an isogroup (default 500 contigs)

-it = max number of isotigs in an isogroup (default 100)

-icc = maximum number of contigs in one isotig (default 100 contigs)

-icl = isotig contig length threshold, below which traversal stops (default 3 base pairs)

Pages 160-171 of Part C of the Roche Assembly manual gives a good table of all the options.

# Example Transcriptome output files

In the Assembly subdirectory:

**454Isotigs.fna** fasta file of all Isotigs, and single contigs.

**454Isotigs.qual** quality scores (Phred-based) for each base in '454Isotigs.fna' file. (eg: 20 = 1 in 100 probability of incorrect base call; 50 = 1 in 100,000)

**454Contigs.fna** fasta file of all contigs, which are used to create the Isotigs.

**454Contigs.qual** quality scores for each base.

**454NewblerMetrics.txt** statistics of the assembly such as # of reads and bases aligned, # of overlaps found, mean contig sizes,

**454ReadStatus.txt** status of each read in assembly (Assembled, PartiallyAssembled, Singleton, TooShort, Outlier), and alignment 3' and 5' positions within contig.

**454TrimStatus.txt** each read's original and revised trim-points used in the assembly.

# View the assembly

```
isotig00013 isogroup00007   ACAACACCTTCTGTTTCATTGACAGTCTCATGACTTTTGCTGCCATTAGTTGTTAAATGCGAAATGACAGTGCATAGCGAATCTCGTCCACCATGTA
EZAKA7J02HVX44              ACAACACCTTCTGTTTCATTGACAGTCT
EZAKA7J02I9H0R              ACAACACCTTCTGTTTCATTGACAGTCTCATGACTTTTGCTGCCATTAGTTGTTAAATGCGAAATGACAGTGCATAGCGAATCTCGTCCACCATGTA
EZAKA7J02GE27X              ACAACACCTTCTGTTTCATTGACAGTCTCATGACTTTTGCTGCCATTAGTTGTTAAATGCGAAATGACAGTGCATAGCGAATCTCGTCCACCATGTA
EZAKA7J02HK063              ACAACACCTTCTGTTTCATTGACAGTCTCATGACTTTTGCTGCCATTAGTTGTTAAATGCGAAATGACAGTGCATAGCGAATCTCGTCCACCATGTA
EZAKA7J02FUAKD                          ATTGACAGTCTCATGACTTTTGCTGCCATTAGTTGTTAAATGCGAAATGACAGTGCATAGCGAATCTCGTCCACCATGTA
EZAKA7J02I1D19                                      TGACTTTTGCTGCCATTAGTTGTTAAATGCGAAATGACAGTGCATAGCGAATCTCGTCCACCATGTA
```

# Exercises in the learning activities

- sfftools
- Newbler using graphics
- Newbler using commandline