December 2014

RNA-Seq de novo assembly training session Day 3 hands-on

Useful links:

SeqClean
A script for automated trimming and validation of ESTs or other DNA sequences by screening for various contaminants, low quality and low-complexity sequences.

CEGMA
CEGMA (Core Eukaryotic Genes Mapping Approach) is a computational method for building a highly reliable set of gene annotations in the absence of experimental data.

IGV
The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets.

Training session aims:
This training session provides you the post-processing of de novo assemblies.

Exercise n°4: Clean and correct assembly

Take back assemblies performed using Velvet/Oases and Trinity during day2. On each transcripts multi-fasta files, run:

- Seqclean to perform a simple cleaning step
- The chimera detection script to detect and correct potential chimeras
- The INDEL correction script to detect and correct potential insertions/deletions

Run these jobs on cluster without specific reservation.

Visualize some INDEL corrections using IGV.

What about the « efficiency » of each step on Velvet/Oases and Trinity assemblies?

Exercise n°5: ORF detection and annotation

Filter assemblies cleaned and corrected at the previous step and discard transcripts without ORF of at least 200 nucleotides.

Annotate the filtered sets of transcripts by similarity search against proteins coded by Zebrafish chr3 using Exonerate.

Give an estimation of the number of « full length transcripts » inside our assemblies.

Exercise n°6: Assembly using DRAP

Assemble reads from runs **ERR145651_t_norm**, **SRR748488_t_norm** and **SRR801554_t_norm** inside exercise_6 directory using DRAP.

Locate the output contigs fasta file (named **assembly_fpkm_1.fa** inside the output directories), blat contigs to `Danio rerio chr3` and extract the best blat hit (in psl format).

Start IGV and compare assemblies of the three runs. Load following files:

- Genome -> Load genome from file -> `Danio rerio chr3` fasta file
- File -> Load from file :
  - `Danio rerio chr3` GTF file
  - foreach `run`:
    - `run_vs_genome` BAM file
    - `run_vs_genome` TDF file
    - `run` best blat hits versus genome

igv_exercise_6.xml

Locate particular regions.

What is the origin zebrafish tissue of each run?
Which strategy should you adopt to build the most comprehensive transcriptome?

- sequencing deeply few samples or tissues
- sequencing lightly a large number of samples or tissues

Exercise n°7: Meta-assembly using DRAP

Build a comprehensive transcriptome using the runMeta tool of the DRAP package.
Meta-assemble the assemblies performed at exercice_6.

Locate the output contigs fasta file (named **assembly_fpkm_1.fa** inside the output directory), blat contigs to `Danio rerio chr3` and extract the best blat hit (in psl format).

Start IGV and compare tissues assemblies and meta-assembly. Load the previous session and add the track meta-assembly versus genome best blat hits.

Locate particular regions.
Did the meta-assembly compress efficiently individual tissues transcriptomes?
Compare tissues assemblies RMBT and meta-assembly RMBT and number of contigs.

Compare the number of zebra fish proteins which have an Exonerate best hit coverage greater than 90% and identity greater than 90% against contigs from tissues assemblies and meta-assembly.
Give some problems that still persist after the meta-assembly…

At the end of the session :

If you are using a training account do not forget to clean all the data and software packages you have generated, downloaded and installed.

For further questions :

- e-mail : support.genopole@toulouse.inra.fr .
- You can check the FAQ of the genotoul website: http://bioinfo.genotoul.fr/index.php?id=11 .
- Using the following link, you can have more information about the other training sessions provided by BIOINFO GENOTOUL: http://bioinfo.genotoul.fr/index.php?id=10.