Comparative Genomics 2018

Practical 2: Gene Prediction

Group 11
Shuhan Xu, Fuqi Xu, Milda Valiukonyte

Summary
In this practical, we first used Glimmer3 to predict the genes for our five genomes. For each genome, Glimmer3 first searched for long open reading frames (ORF) and used it to train a interpolated context model (ICM). The model is used to predict the genes in the genome. Glimmer3 predicted much fewer genes for the eukaryote genome than the prokaryote genomes as it does not account for introns which disrupt the coding sequence. From the genes obtained, we extracted the proteins sequences used the script provided. We then used GENSCAN with the parameters in Humanlso.smat to predict the genes for our eukaryote genome. Using our own written script, we created separate files for amino acid and nucleotide sequences from the output of GENSCAN. We also produced a graphical diagram for the location of the predicted genes and used blast to searched for the protein names of the first two sequences.

Exercise 1

1. tigr-glimmer long-orfs -n -t 1.15 01.fa 01.long-orf-coords

long-orfs:	program which finds long non-overlapping ORF from a DNA sequence file
-n:	no header information in the output file
-t 1.15:	consider only genes with entropy distance score lower than 1.15 in the program
0.1.fa:	sequence file

0.1.long-orf-coords: output file containing the coordinates of the long ORF

2. tigr-glimmer extract -t 01.fa 01.long-orf-coords > 01.longorf

extract:	program which reads the sequence file, searches for the region defined by the coordinates in the coordinate file, and outputs the region in a fasta file.
-t:	exclude the last three characters of the output string
01.fa:	sequence file

01.long-orf-coords: coordinates of the long ORF

01.longorf:	fasta file of the long ORF

3. tigr-glimmer build-icm -r 01.icm < 01.longorf

build-icm:	program which builds an interpolated context model (ICM) from a training set of sequences
-r:	use the reverse of the input sequence to build the ICM
01.icm:	the output file for the built ICM
01.longorf:	fasta file of the long ORF

4. tigr-glimmer glimmer3 -o50 -g110 -t30 01.fa 01.icm 01.glimmer

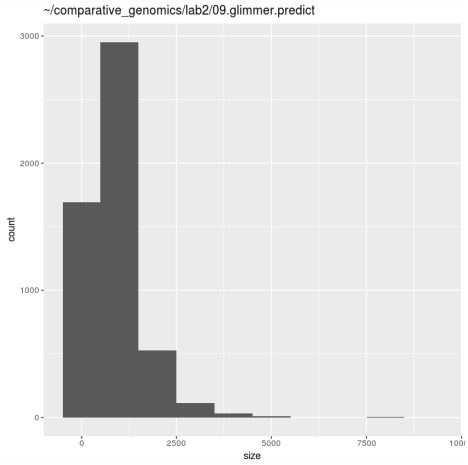
glimmer3:	program which makes gene prediciton
-o50:	maximum overlap length of 50 for genes
-g110:	minimum gene length of 110
-t30:	threshold score of 30 for genes
01.fa:	sequence file
01.icm:	ICM file
01.glimmer:	tag for the two output file 01.glimmer.detail and 01.glimmer.predict

5. One can train using known genes from the genome or genes from an closely related species or strain

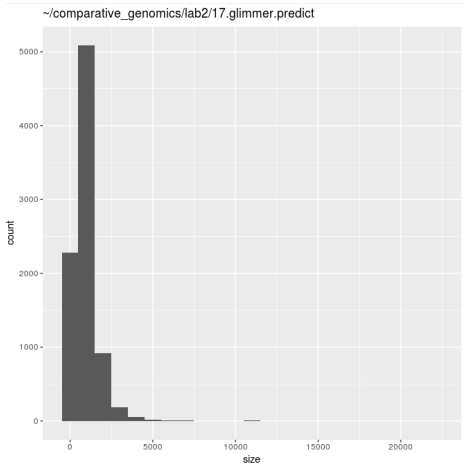
6. Glimmer is not suitable for eukaryotic genomes. It does not consider the presence of introns. In the No.24 genome, it's prediction gives few ORFs.

7.

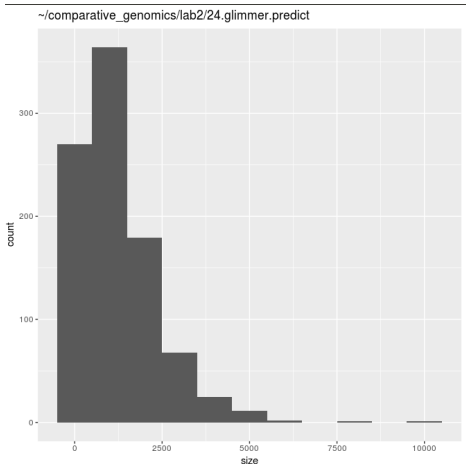
09.fa.txt



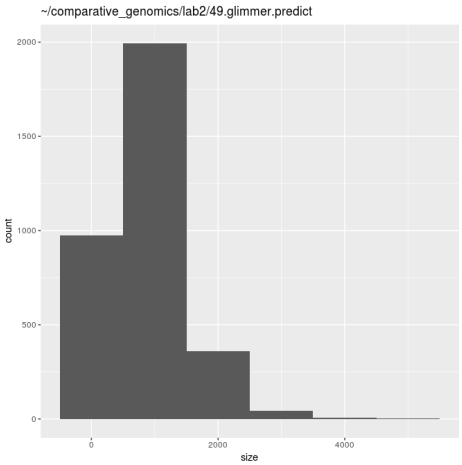
17.fa.txt



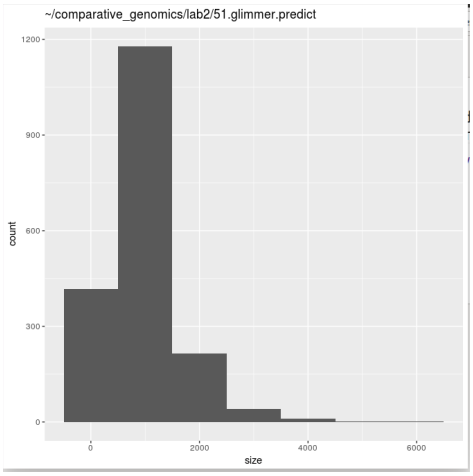
24.fa.txt



49.fa.txt



51.fa.txt



8. 24.fa.txt has much fewer genes compared with the other genomes. For instance, the count for genes of size 1000bp is only 350 for 24.fa.txt as compared to 1000-5000 for other genomes. This is due to the fact that 24.fa.txt is a eukaryote genome and glimmer3 cannot account for introns which disrupt the gene. In general, the shape of the distribution is the same for all genomes with the highest count at 1000.

9.

protein sequences for 09.fa.txt: see attachment 09.fa.txt.pfa
protein sequences for 17.fa.txt: see attachment 17.fa.txt.pfa
protein sequences for 24.fa.txt: see attachment 24.fa.txt.pfa
protein sequences for 49.fa.txt: see attachment 49.fa.txt.pfa
protein sequences for 51.fa.txt: see attachment 51.fa.txt.pfa

Exercise 2

1.
output for GENSCAN: see attachment 24.out
python script to separate protein and nucleotide sequences: see attachment separate.py
file for amino acid sequences: see attachment predicted_protein_sequences.py
file for nucleotide sequences: see attachment predicted_nucleotide_sequences.py

2. see attachment 24.ps

3.
first nucleotide sequence: Cos7p (COS7)
Second nucleotide sequence: alpha-glucoside permease (MPH2)