**Statistical analysis of sequins with Anaquin**

Tim Mercer, Wendy Chen, Ted Wong

Garvan Institute of Medical Research

**INTRODUCTION**

Anaquin is a C++ statistical analysis tool for sequin data-analysis. It covers RNA, DNA and metagenomics. In this tutorial, we will demonstrate the usage of the software on simulated data-sets and real data-sets.

**1. Downloading and installing software**

1.1 | Create a directory to store all of the executable programs used in this protocol:

$ mkdir Spike

1.2 | To install Anaquin, download the latest binary package (<https://dl.dropboxusercontent.com/u/7376968/anaquin.tar>) and extract the archive into the new directory:

$ cd Spike

$ wget https://dl.dropboxusercontent.com/u/7376968/anaquin.tar

$ tar -xf anaquin.tar

$ cd anaquin

1.3 | Please check if you can see the version number.

$ ./anaquin -v

Version 1.0. Garvan Institute, copyright 2015.

1.4 | To test the installation:

$ ./anaquins -t

**===================================**

**All tests passed** (260 assertions in 7 test cases)

The -t option runs internal test cases shipped with the software. All test cases must pass. Please contact Garvan Institute if any of the test fails.

**2 RNA PROCEDURE**

In this section, we will use R\_K562 data-set. The data-set contains reads aligned only to the in-silico chromosome. The other reads have been filtered out to save file size.

2.1 | Download the data-set

$ wget https://dl.dropboxusercontent.com/u/7376968/R\_K562.zip

$ unzip R\_K562.zip

You should find three BAM files in the zip-file, each corresponds to a biological replicate.

$ ls R\_K562

RK1A\_filtered.bam RK2A\_filtered.bam RK3A\_filtered.bam

2.2 | Convert the first replicate for mixture A to SAM

$ samtools view -h -o R\_K562/RK1A\_filtered.sam R\_K562/RK1A\_filtered.bam

2.3 | Examine the SAM file

$ grep -v @ R\_K562/RK1A\_filtered.sam | head -n 2

13892:20439 163 **chrT** 388560 50 67M58378N59M

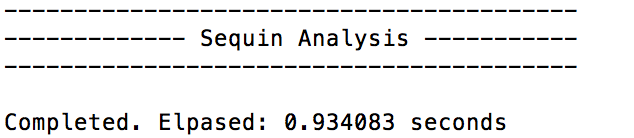
3931:76483 163 **chrT** 388571 50 56M58378N64M33114N6M

The reads are mapped to the in-silico (chrT) chromosome.

2.4 | Calculate the alignment statistic

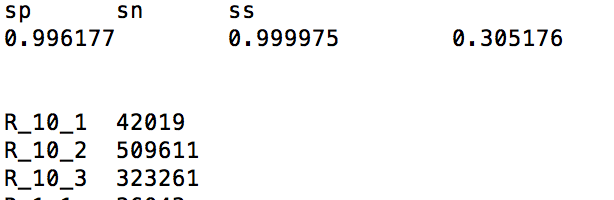
$ ./anaquin rna -al R\_K562/RK1A\_filtered.bam

You should see the execution time.



The output files are saved in the spike\_out folder. You should see ralign\_base.stats*,* ralign\_exon.stats and ralign\_intron.stats.

$ head spike\_out/ralign\_exon.stats



The file lists the specificity (sn) and sensitivity (sp) at the exon level. The limit of sensitivity is about 0.31, this is the lowest quantity of abundance that can still be detected. The file also lists number of counts for each sequin.

Similarly for ralign\_base.stats and rliagn\_intron.stats.