# Project Assignment – Exploratory Genome Sequence Analysis

**Rakesh Santha Kumaran**

## Software/Tools Needed:

1. SINGULARITY: A Container allows your application portable, shareable and reproducible, as it allows you to stick an application and all of its dependencies into a single package. Singularity is a relatively new container software invented by Greg Kurtzer while at Lawrence Berkley National labs and is now developed by his company Sylabs. It was developed with security, scientific software, and HPC systems in mind.
   The detailed steps of installation of singularity and building a container as well as downloading a container are available at the link, singularity-tutorial.

2. WGSIM: It is useful for simulating the reads of a sequence from a reference genome. It allows us to parametrize the reads whether single or pairwise reads, required error rate, depth of reads, length of reads, rate of mutations allowed, and many other parameters. The details explanation of various parameters of wgsim tool is available at the link, simulating-sequence-reads-with-wgsim.
   After cloning the wgsim from the github link, https://github.com/lh3/wgsim, navigate to wgsim folder and execute the command, `gcc -g -O2 -Wall -o wgsim wgsim.c -lz -lm`.

3. SOAPDENEVO2: SOAPdenovo is a novel assembly tool that can build draft assembly for a given set of reads. After reads are generated from wgsim, this package is used for assembly.
   Use Git method to clone from the link, https://github.com/aquaskyline/SOAPdenovo2, and navigate to the folder and execute the command, sed -i -e "s/-lm/-lm -no-pie/" Makefile and then use make command.

4. ASSEMBLY-STATS: It takes a FASTA file as input and calculates both scaffold and contig statistics (N50, L50, etc.) from a scaffold FASTA file. It does this by breaking each scaffold wherever there is more than one N and then calculating statistics for both the scaffolds and contigs.
   Git clone from link, https://github.com/sanger-pathogens/assembly-stats.git and execute the cmake and make command and install the make file. The detailed steps are available at the singularity definition file.

5. SEQKIT: Providing statically linked executable binaries for multiple platforms.
   To install it, Use wget on the link,
   https://github.com/shenwei356/seqkit/releases/download/v2.2.0/seqkit_linux_amd64.tar.gz, and then below commands,

   ```
   tar -xzf seqkit_linux_amd64.tar.gz

   chmod a+x seqkit

   mv seqkit /opt/bin/

   \rm seqkit_linux_amd64.tar.gz
   ```

6. BLAST: It can rapidly align and compare a query DNA sequence with a database of sequences, which makes it a critical tool in ongoing genomic research.
   Use below commands to install it,

   ```
   wget
   https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ncbi-blast-2.13.0+-x64-linux.tar.gz

   tar -xzf ncbi-blast-2.13.0+-x64-linux.tar.gz

   \rm ncbi-blast-2.13.0+-x64-linux.tar.gz

   mv ncbi-blast-2.13.0+/bin/* /opt/bin
   ```

7. AUGUSTUS: It is a web-based gene prediction tool and give the related protein sequence for the given contig and chosen organism. It can be accessed via the link, https://bioinf.uni-greifswald.de/augustus/submission.php.

8. BLASTP: Blast tool for protein sequence, web based tool that can be accessed via https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

9. MEGAX11: It is used for aligning the sequence and also to construct the phylogenic tree. It needs to be installed via link, https://www.megasoftware.net/

## Experiment Steps:

After installing the tools, and downloading the locust genome, using seqkit extracted the contigs that are greater than 80000 nucleotides. Based on alphabetical order of first name, I have taken the 12$^{th}$ contig and have done the reads and assembly on that contig using wgsim and SOAPDenevo. And then, Experimented with different values of parameters for optimization.

Finally using AUGUSTUS, generated a related protein sequence for the 12$^{th}$ contig with reference to Fruit Fly (Drosophila melanogaster) organism. Then taken the longest protein sequence and with help of blastp identified the top 3 matching reference protein sequence and constructed a phylogenic tree on those 3 proteins using MEGA11 software.

## Experiment Results:

**Results for reads and assembly of sequence:**

Note: Observation has been made for the motive to have longest assembly sequence. But for the different motive like number of contigs to be high, the suggested observation may not be completely true. The observation is made to specific sample scenario so that it shows how we could interpret the data based on the results.

1. **Variation in the number of read pairs:**
   Pairwise reads with error rate 0.01 rate of mutation 0.001 and changing the number of read pair.
   >wgsim/wgsim -e 0.01 -**N 3000** -r 0.001 prjctseq.fa r1.fq r2.fq
   >wgsim/wgsim -e 0.01 -**N 5000** -r 0.001 prjctseq.fa r1.fq r2.fq

After analyzing the scaffold statistics overall the number of contigs in scaffolds have increased significantly. The length of scaffolds from N10 to N90 have significantly increased with increasing the N parameter value. The N10 means, A contig with size of the shortest contig such that the sum of contigs of that size or longer constitutes at least 10% of the total size of the assembled contigs. Similarly, for N90 and others. Also the longest sequence assembled is also higher with increasing the N value. But the composition of Nucleotides, A, C, G, T is similar and it is not showing significant variation with the change in parameter value.

**Observation: To have longer assembled sequence, Increasing the number of read pairs helps in increasing the scaffolds and help would help us in achieving the optimized condition.**

2. **Variation in error rate:**
   Pairwise read with, number of read pairs as 5000, rate of mutation as 0.001 and changing the error rates.
   > wgsim/wgsim **-e 0.1** -N 4000 -r 0.001 prjctseq.fa r1.fq r2.fq
   > wgsim/wgsim **-e 0.01** -N 4000 -r 0.001 prjctseq.fa r1.fq r2.fq

Increasing error rate decreases the longest length of the scaffolds, no of contigs in scaffolds, decreases length of N10 to N90 contigs

**Observation: To have longer assembled sequence, Overall having a minimum error can lead towards the optimized value.**

3. **Variation in length of reads:**
   Pairwise reads, with error rate 0.01, number of read pairs 5000, rate of mutation 0.01 and length of reads changed,
   > wgsim/wgsim **-1 500 -2 500** -e 0.01 -N 5000 -r 0.001 prjctseq.fa r1.fq r2.fq
   > wgsim/wgsim **-1 200 -2 200** -e 0.01 -N 5000 -r 0.001 prjctseq.fa r1.fq r2.fq

With decreasing the length of reads, longest assembled sequence and more precisely the average length of assembled sequence decreases. Also there is downward trend observed in length of N10 to N90 contigs.

**Observation: To have longer assembled sequence, increasing the length of reads would be helpful.**

4. **Variation between Single and Pairwise reads:**
   > wgsim/wgsim **-1 500 -2 500** -e 0.01 -N 5000 -r 0.001 prjctseq.fa r1.fq r2.fq
   > wgsim/wgsim **-1 500** -e 0.01 -N 5000 -r 0.001 prjctseq.fa r1.fq r2.fq

With single read the average length of assembled sequence is high and length of N10 to N90 contigs is also high for single read significantly. There is not much variation in composition of the A, C, G, T values. But number of contigs is significantly high for pairwise reads. Also Singleton_Num is very high for pairwise reads compared to single reads.

**Observation: To have longest assembled sequence, having a single read would be helpful.**

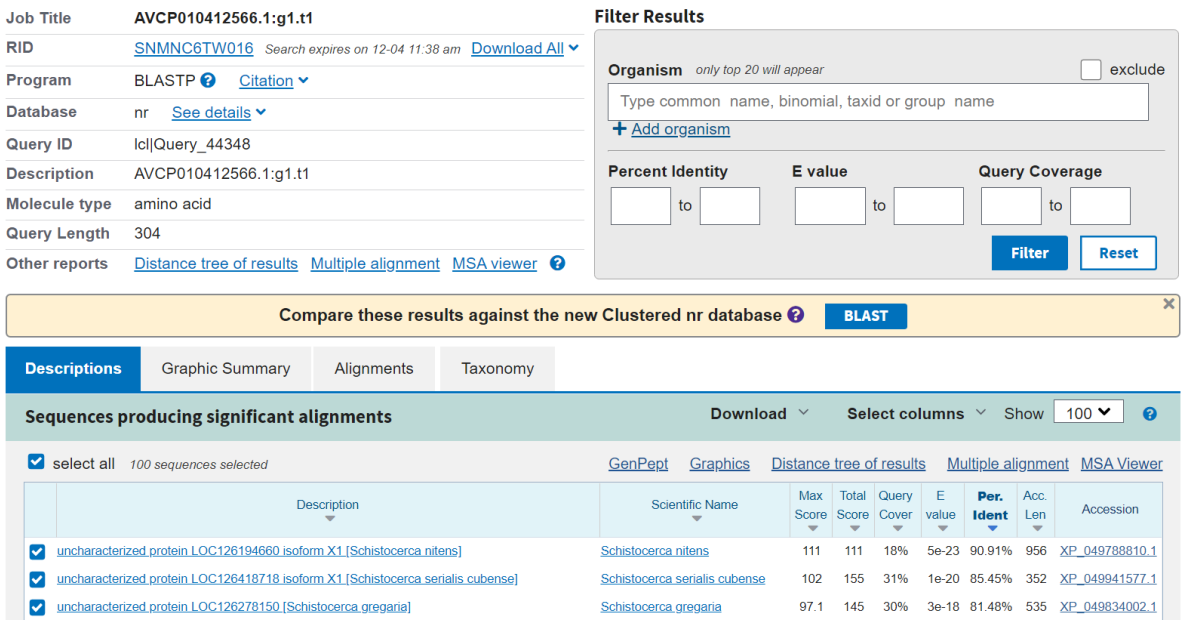
**Constructing the phylogenic tree using AUGUSTUS, BLAST, MEGA:**

1. <u>Determining the longest protein sequence:</u>
   Copy the 12th contig and choose the reference organism as Fruit fly(Drosophila melanogaster)

Longest protein sequence determined:

```
>AVCP010412566.1:g1.t1
MFSHHLSSHNIQKIFFLFFSFKKNKIWKGHIQNKGVSAHSLIETQVKHQLRGGPIYATEGFIVAIRETRRENEESEDE
EEEEEKKFSDDEWSASQENETEIKDITGPTATTLYSRLFISIPPFPAYLLPFSLPASCPIPRKYPVTKGALLGYFGGRFWIWRDEEVALKQITRATSS
SLQIQNLPPKYPSSAPFVTGYFLGMGQEAGSEKGRSPVTQKVYTIKGRATCNSTHMIYQLTCLHCEAFYVGMTSNKLSIRMNGHRQTVLVVLHHPGYL
GYFPPTPTFQNSRDGNSPPIVHLVSTSAKF
```

2. Determining the 3 best reference proteins for the obtained protein sequence using Blastp:



| | |
|---|---|
| Job Title | AVCP010412566.1:g1.t1 |
| RID | SNMNC6TW016 *Search expires on 12-04 11:38 am* Download All |
| Program | BLASTP Citation |
| Database | nr See details |
| Query ID | lcl\|Query_44348 |
| Description | AVCP010412566.1:g1.t1 |
| Molecule type | amino acid |
| Query Length | 304 |
| Other reports | Distance tree of results Multiple alignment MSA viewer |

**Filter Results**

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to | E value to | Query Coverage to

Filter Reset

Compare these results against the new Clustered nr database BLAST

Descriptions | Graphic Summary | Alignments | Taxonomy

**Sequences producing significant alignments** Download Select columns Show 100

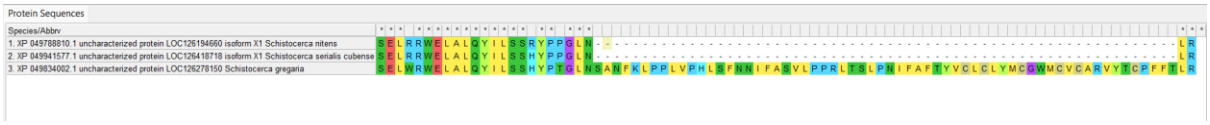
select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

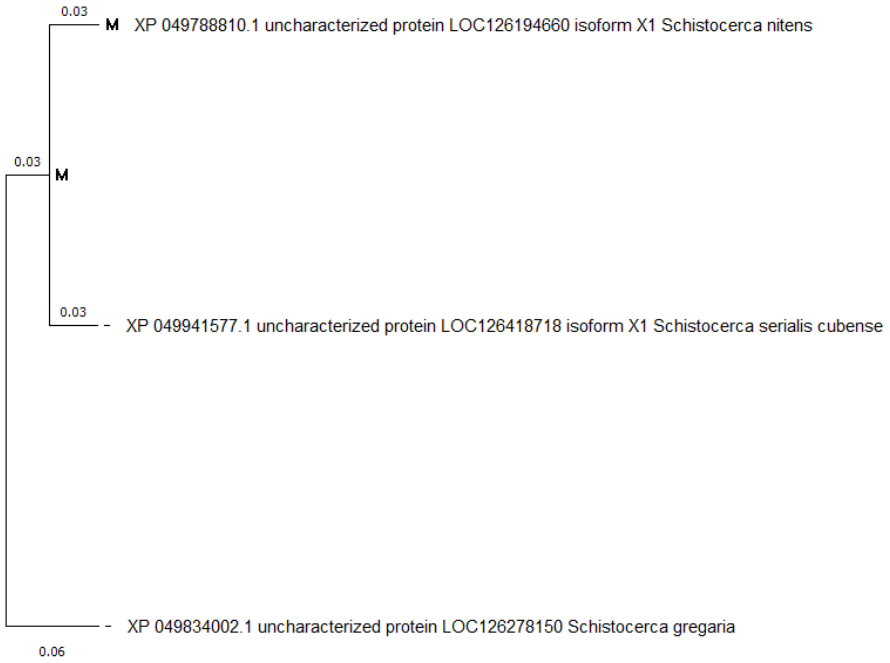
| Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|---|---|---|---|---|---|---|---|---|
| uncharacterized protein LOC126194660 isoform X1 [Schistocerca nitens] | Schistocerca nitens | 111 | 111 | 18% | 5e-23 | 90.91% | 956 | XP_049788810.1 |
| uncharacterized protein LOC126418718 isoform X1 [Schistocerca serialis cubense] | Schistocerca serialis cubense | 102 | 155 | 31% | 1e-20 | 85.45% | 352 | XP_049941577.1 |
| uncharacterized protein LOC126278150 [Schistocerca gregaria] | Schistocerca gregaria | 97.1 | 145 | 30% | 3e-18 | 81.48% | 535 | XP_049834002.1 |

Best reference proteins identified,

```
>XP_049788810.1 uncharacterized protein LOC126194660 isoform X1 [Schistocerca nitens]
MLHFRPSSTMSPAQHPHNDPIKFYLHSLRRHAFALARLRSHILFSQACLTFGITPKGLTLKVPISGCNPSFHQSLYQFQT
EQSIALTHLILHLHHINSANEHTRQLLSLIKVLNLSSPTSTPAVQSILLQANRKLEQHATLHLKKLSNLLVSHLRKGNSLT
LHNLSSKPQPPLIAHKPSLSHLLNLPLPAPLLPKPQNSFQHNLEPQHPNSVVNLSSKPLSQSEPLSYPKASPSAPLPDST
KQPSSKIYCPTPVLSAGNITLPRRKMILILLLMIQLPKTPSKLNPAWNSSVLRHSGTPPPLPQNHPLQTFQEFLTSSLAS
QSFLKTLNPTPNITTAEAQAIRDLKADRSIVILPADKGSTTVVLDRREYVAEGLRQLSDNTTYKVCQGNPIPDVQAELQG
ILRTLGPLQNLSPDSINLLTPPTPRTPNFYLLPKIHKPNHPGRPIVADYQAPTERISAYVDQHLQPITCSLPSFIKDTNH
FLERLESLPNLLPPETILVTIDATSLYTNIPHVQGLAAMEHFLSRRSPATLPKTSFLITLASFILTHNFFTFEGQTYQQL
KGTAMGTRMAPSYANLFMGRLEEAFLVTQVCQPKVWYRFIDDIIMIWTHSEEELQNFLSNLNSFGSIRFTWSYSKSHATF
LDVDLHLSNGQLHMSVHIKPTNKQQYLHYDSCHPFHIKRSLPYSLGLRGKRICSSPESLKHYTNNLTTAFASRNYPPGLV
QKQITRATSSSPQTQNPPQKNHKSAPLVTGYFPGLDQTLNVALQQGYDFLKSCPEMRSILHEILPTPPRVSFRRPPNLRN
LLVHPYEIPKPPSLPSGSYPCNRPRCKTCPMHPPTTTHSSPVTRKVYTIKGRATCESTHVIYQLTCLHCDAFYVGMTSNK
LSIRMNGHRQTVFVGNEDHPVAKHALVHSQHILAQCYTVRVIWILPTNTNLSELRRWELALQYILSSRYPPGLNLR
>XP_049941577.1 uncharacterized protein LOC126418718 isoform X1 [Schistocerca serialis
cubense]
MIWTHSEEELQNFLSNLNSFGSIRFTWSYSKSHATFLDVDLHLSNGQLHTSVHIKPTNKQQYLHYDSCHPFHIKRSLSYS
LGLCGKRICSNPESLNHYTNNLKTAFTSRNYPPDLVQKQIARATSSSPQTQNLPQKNPKSAPLVTGYFPGLDQTLNVALQ
QGYDFLKSCPEMRSILHEILPTPPIVSFRRPPNLRNLLVHPYEIPKPSSLPSGSYPCNRPRCKTCPMQPPTTTYSSPVTR
KVYTIKGRAMCESTHVIYQLTCLHCDALYVGMTSNKLSIHMNGHRQTVSVGNEDHPVAKHALVHGQHILAQCYTVRVIWI
LPTNTNLSELRRWELALQYILSSHYPPGLNLR
>XP_049834002.1 uncharacterized protein LOC126278150 [Schistocerca gregaria]
MQSPSLPNLLPLETILVTTDATSLYTNILYVQGLAVMEHFLSCRSPATLPKTSFLITLASFILTHNFFTFEGQTYQQLKG
TAMGTRMAPSYPSLFMGRLEEAFLVSQACQSKVWYKFNDDIFMIWTHSKELQNFLSNLNSFGSIRFTWSYSKSHATFLDV
DLHLFNGQLHTFVHIKPINKQQYLHYDSCHPFHIKQPLPYSLGLRGKRICSSPESLNHYTNNLKTALASCNYPPDLVQKQ
IARATSSSPQTQNLSQKNPKSAPLVTGYFPGLDQTLHVALQQGYNFLKSCPEMRSILHEILPTPPRVSFRRLPNLRSLLV
HPYEIPKPPSLPSGSYPCNHPRCKTCPMHPPTTTYSSPVIRKVYMIKGRATCESSHVIYQLTCLHCEAFYVGMTSNKRSI
HMNGHRQAVFDGNEDHPVAKHALVHSQHILTQCYTVRVRWILPTDTNLSELWRWELALQYILSSHYPTGLNSANFKLPPL
VPHLSFNNIFASVLPPRLTSLPNIFAFTYVCLCLYMCGWMCVCARVYTCPFFTLR
```

3. Aligning the sequence and forming the phylogenic tree:



The tree clearly shows that Schistocerca serialis cubense is closely related to chistocerca nitens than to Schistocerca gregaria. It also shows how the three protein sequence are related and the branch length between them.

Definition file:
Prepared the singularity definition file and then prepared the separate file for script to be run after the environment is made.
Sample screenshot:

```
bootstrap: docker
From: ubuntu:20.04

%post
    apt -y update
    apt -y install bc zip unzip wget tzdata
    apt -y install build-essential git cmake zlib1g zlib1g-dev


    cd /opt
    mkdir bin

    git clone https://github.com/lh3/wgsim
    cd wgsim
    gcc -g -O2 -Wall -o wgsim wgsim.c -lz -lm
    cd ..

    git clone https://github.com/aquaskyline/SOAPdenovo2
    cd SOAPdenovo2/
    sed -i -e "s/-lm/-lm -no-pie/" Makefile
    make
    cd ..

    git clone https://github.com/sanger-pathogens/assembly-stats.git
    cd assembly-stats/
    mkdir build
    cd build
    cmake ..
    make
    make install
    cd ../..

    wget
    https://github.com/shenwei356/seqkit/releases/download/v2.2.0/seqkit_linux_amd64.t
    ar.gz
    tar -xzf seqkit_linux_amd64.tar.gz
    chmod a+x seqkit
    mv seqkit /opt/bin/
    \rm seqkit_linux_amd64.tar.gz
```

**Conclusion:**
**Based on results of assembly:**
Analyzing the parameters of the tools used to maximize our understanding of the tools and apply to the different requirements of analysis of genomes.
**Based on Results of phylogenic tree:**
It helps in understanding how different protein sequence are related to each other.