Shon Kurian George

Edinburgh, United Kingdom

+44 7778261592 | LinkedIn | Github | shon.kurian.george@gmail.com

EDUCATION

The University of Edinburgh, United Kingdom. Sep 2023 - Sep 2024.
Master of Science in Bioinformatics **Grade: Merit**

Relevant Courses**:** Bioinformatics Programming & System Management, Statistics & Data Analysis, Next Generation Genomics, Machine Learning In Python, Bioinformatics Algorithms.

Dr. D. Y. Patil Biotechnology and Bioinformatics Institute, Pune, India. Jun 2017 - Jul 2021.
Bachelor of Technology (Biotechnology) **CGPA: 8.79/10**

EXPERIENCE

R Package Developer working with **Jan Kückelhaus**, University Clinic Erlangen, Germany. Nov 2024 – Present.

- Currently developing the cypro R package, aimed at streamlining image-based cell profiling workflows by integrating data from diverse platforms like CellProfiler and CellTracker, hosted on GitHub for version control. Enhanced the package's usability by implementing intuitive Shiny-based interfaces, allowing detailed specification down to individual well-level configurations, significantly simplifying user interaction.
- Refined the S4 class objects within cypro to ensure accurate data integration from imaging platforms, enabling comprehensive cell movement analyses. Incorporated proactive user warnings and notifications within the Shiny application, minimizing user errors, safeguarding data integrity, and enhancing the user experience.

Postgraduate Student Researcher at **Abreu RNAlab,** University of Edinburgh, United Kingdom. Mar 2024 - Sep 2024.

"*Using spatial transcriptomics to evaluate if microRNAs regulate expression profiles at cell-type boundaries.*"

- Engineered a bespoke bioinformatics pipeline in R for 10x Genomics Visium Spatial Transcriptomics datasets, facilitating the identification of expression patterns of 5 tissue-specific microRNAs (miRNAs) across 8 datasets & 3 major tissue types (brain, heart, liver).
- Integrated multiple clustering methodologies (BayesSpace and K-means), achieving high tissue region segmentation accuracy as seen in the literature, directly contributing to high-fidelity statistical analysis. Achieved highly statistically significant results in miRNA signal validation using complex statistical models, delineating known regulatory roles in cellular boundary gene expression in 87% of datasets and 66% of tissue types evaluated.
- Spearheaded the development of an R Shiny application, enabling seamless & automated heatmap generation and dataset comparisons by reducing the manual processing time between datasets to under 15 seconds and making the comparison of the parameters of a specific dataset instantaneous.

PROJECTS

Detailed Collaborative Machine Learning Project on Modifiable Risk Factors Linked to Dementia. Leveraged the SHARE dataset to identify key modifiable dementia risk factors, using polynomial regression with GridSearchCV to determine third-degree polynomial as optimal based on lowest MSE and highest R^2; highlighted physical activity and social engagement as primary predictors. Contributed contextual framing and final interpretation, translating model outputs into actionable lifestyle and policy interventions informed by the 2017 & 2020 Lancet Commissions (Livingston et al.), while critically addressing limitations.

Analysis and Critique of *Automeris io* moth de novo Genome Assembly and Annotation by Skojec *et al*. and assembly using wtdbg2.
Executed a critical evaluation of the genome assembly workflow performed by Skojec *et al*., analysing 1.9 million next generation sequenced HiFi reads and comparing assembly quality between the Hifiasm (N50: 15.78 Mb) and wtdbg2 assemblers (N50: 1.1 Mb). Demonstrated the superior performance of Hifiasm in generating a high-quality assembly with 98.4% completeness and achieving a total assembly size of 490 Mb with only 600 contigs in the first draft compared to wtdbg2 which had 3,362 contigs.

Developed a Python script which enables users to query a taxonomic group & retrieve protein sequences via EDirect for analysis.
Developed a Python workflow for protein analysis, integrating NCBI and PROSITE databases to generate datasets with up to 1,000 sequences, automate alignments, and create customisable conservation plots, ensuring rigorous project and data management. Automated motif discovery and back-translation using EMBOSS tools, delivering motif frequency data & nucleotide sequences for further research insights.
(Refer Lucidchart workflows)

CORE COMPETENCIES

Programming Skills/Languages: R, Python, Bash/Shell Scripting, SQL (MySQL).
Operating Systems: Linux/UNIX, Windows. **Version Control & Containerisation:** Git and GitHub, Docker.
Bioinformatics Frameworks and Libraries:
R: Base R, tidyverse, pheatmap, ggplot2, shiny, Bioconductor, SPATA2, Seurat, BayesSpace, knitr, DESeq2. **CLI:** NCBI - EDirect and SRA Toolkit, FastQC and MultiQC, bowtie, Samtools, hisat2, bedtools, makeblastdb, Clustal Omega, EMBOSS suite, Cutadapt, NanoPlot, HiCanu, wtdbg2, QUAST, BlobTools, SPAdes, IGV. **Python:** pandas, numpy, matplotlib, seaborn, sklearn, sys, subprocess, time, shutil.
Software: Vim, OverLeaf, Zotero, Ollama, Open WebUI, MS Office Suite, Adobe Suite (Photoshop, Premiere Pro, Illustrator), Lucidchart.

INTERNSHIPS

Industrial Training Trainee at **Aurigene Pharmaceutical Services Limited (APSL),** Hyderabad, India. Feb 2021 – Jun 2021.

- Contributed to the In-Vivo Pharmacology Department, which managed 600+ rodents across 1,000+ efficacy studies annually, supporting drug discovery aligned with US FDA and DCGI standards, achieved through efficient time and project management. Gained exposure to multiple workflows including RNA isolation for 100+ tissue samples, ELISA with under 5% CV consistency across 96-well plates, and RT-qPCR used in regulatory submissions, developed through collaboration and innovative problem solving.

R & D Intern at **Serum Institute of India Private Limited,** Pune, India. Jun 2019 – Jul 2019.

- Gained exposure to state-of-the-art vaccine manufacturing processes in one of the world's largest vaccine producers, observing large-scale operations utilising fermenters and bioreactors with capacities up to 20,000 litres. Learned about Good Manufacturing Practices (GMP) and Good Laboratory Practices (GLP), critical to Serum producing 1.5 billion vaccine doses annually for global distribution.