Pre-interview sift questions

For this studentship competition, each project supervisor can nominate one candidate for interview by the DTP. To aid us in this process please could you answer for me the following questions and return to me by 11pm UK time on the **29th April**. Apologies for the short turn around we have very limited time to select a candidate. Note you are not expected to know all the answers, just do your best.

General Research

- Why do you want to do a PhD?

Through my extensive research experience across volunteer work in clinical and research labs and unpaid internships I have cultivated a strong passion for biomedical research and are keen to further my expertise. As a next career step after master's study, I am keen to progress to a PhD.

I want to become an independent researcher in the field of translational cancer research with an impeccable arsenal of bioinformatics skills. I am looking for the intellectual development to achieve this and the skillset to develop that I can transfer into insilico informed postulation of hypotheses, their wet-lab proof of concept and final insilico analysis of wetlab output to answer research questions. I realise this is also an avenue for technology transfer to underserved communities for example in Africa which, as reported in this project's background, is underrepresented in large population studies.

- Why are you interested in this particular project?

This project particularly builds up from my undergraduate and master's academic research projects on prostate cancer progression and drug discovery. It highlights the interplay of the microbiome, as development/progression of disease is multifunctional, in prostate cancer aggressiveness. My undergraduate study identified similar trends; high prostate cancer prevalence in African men, as the background of this study. I have relevant background knowledge on respective disease epidemiology and have done extra courses (Introduction to Linux, 16s rRNA sequencing, python programming) all directly relevant and align with the background experiences required for this project.

The opportunity to learn and attain skills for the development of new analytical techniques for big 'omics datasets more so prostate cancer perfectly aligns with my career ambitions.

Being a Ugandan, an African by decent, the project might leverage my background and experiences such as diet which affect the microbiome which could possibly now be a factor in prostate cancer aggressiveness.

- What research skills and training do you have which are relevant to this project?

I have hands-on lab experience with techniques like PCR, electrophoresis, ELISA, and nucleic acid extraction from internships in molecular biology techniques and cell culture. My bioinformatics skills covering primer design, molecular modelling with programmes like PyMol, and data analysis tools such as MiniTab, python programming and Linux terminal could aid the computational aspects of the project. I additionally have a background in coordinating both lab and field studies.

I have the experience of single authorship and paper correspondence through the publication process. This demonstrates my ability to take on constructive criticism, make rational discussion to satisfaction while reliably working as part of a team.

- What additional skills and training do you think you need?

I will need specialised training building up to genomic data/metagenomics and respective algorithm creation and data manipulation. Training in cutting-edge techniques like NanoString assays will also be valuable.

While I have experience with cell culture, I might need more specialised training in handling prostate cancer cell lines such as PC3, LNCaP, among others. More advanced bioinformatics skills in areas like metagenomics and microbial genomics will significantly be beneficial as they are expected to cover a significant part of the project objectives.

- Can you briefly discuss a recent piece of research that you have read and which you found of particular interest?

I have read the Gihawi *et al*. 2023 mBio paper (https://doi.org/10.1128/mbio.01607-23) which highlights issues of contamination and misclassification impacting cancer microbiome studies. The study reanalysed a previous widely-cited study by Poore et al. (2020) that reported strong associations between the presence of certain microbial signatures and different cancer types. The key findings and points of interest from this re-analysis were:
1. Major data analysis errors: Gihawi *et al*. identified two fundamental flaws in the original study's data analysis methods:
a. Overestimation of bacterial read counts: Due to contamination of bacterial genome databases with human sequences, millions of human reads were incorrectly classified as bacterial, leading to vastly inflated bacterial counts.
b. Artificial signal creation during normalisation: The normalisation process accidentally introduced distinct artificial signals for different cancer types, even when raw read counts were zero, allowing machine learning models to achieve high accuracy based on these artificial signals rather than true biological signals.

2. Lack of evidence for a cancer microbiome: By re-analysing the data using more robust methods, Gihawi *et al*. found little to no evidence for the existence of distinctive microbial signatures associated with different cancer types, contrary to the original study's claims.

3. Invalidation of subsequent studies: The authors highlight that over a dozen follow-up studies have relied on the flawed data from the original study, potentially rendering their findings invalid as well.

This re-analysis is of particular interest to researchers in the fields of microbiome research, cancer biology, and bioinformatics, as it challenges a widely-cited study and highlights the importance of rigorous data analysis and interpretation, especially in the context of low-biomass samples and potential biases introduced by computational methods. The findings emphasised the need for caution when interpreting microbiome-based associations and the importance of robust validation and replication studies.

- What do you know of the development opportunities of the MMB DTP PhD programme?

The MMB (Molecules, Microbes and Materials) Doctoral Training Partnership (DTP) PhD programme at the University of East Anglia (UEA) offers the following development opportunities for students:

1. Interdisciplinary research environment: As a student of Norwich Medical school, basing on my supervisor's affiliation, I will be part of one of three Graduate Schools (Norwich Bioscience Institutes, Medicine and Health Sciences, or Science). This will foster interdisciplinary research and collaboration within the Norwich Research Park.

2. Microbial bioinformatics training: The programme aims to provide specialised training in microbial bioinformatics, equipping students with the skills to run sophisticated computer analyses alongside laboratory work. This training is delivered through problem-based learning, training events, monthly master classes, and annual retreats.

3. Professional Placements: All students are required to undertake a Professional Placement lasting up to 12 weeks, or for iCASE (industrial Cooperative Awards in Science & Technology) students, a placement with a non-academic partner lasting 3-18 months. These placements provide opportunities to engage with businesses and gain industry experience.

4. Transferable skills development: The Graduate Schools offer training programmes that provide a solid foundation in transferable skills, nurturing students' knowledge and abilities to support their

career progression.

5. Career support: As a student, I will have access to career advice and support from UEA Career Central, including dedicated career advisers for PhD students, to help me manage my career and develop the confidence and skills to pursue careers in various sectors.

6. Student support services: As a student, I can benefit from a wide range of support services offered by UEA Student Services, ensuring a well-rounded and supportive environment during their PhD studies.

Summarily, the MMB DTP PhD programme at UEA provides development opportunities through interdisciplinary research, specialised training in microbial bioinformatics, professional placements, transferable skills development, career support, and comprehensive student support services.

Laboratory

- If possible, please summarise a project with lab work elements in that you have been involved with.

I was involved in a molecular epidemiology project determining the circulating pathogens in a cattle keeping community. The project involved field/ animal blood sample collection and preservation on fast technology for analysis of nucleic acids (FTA) cards as well as whole blood for serum. The serum was used for ELISA disease screening and positive samples'(diseased) FTA cards were washed, elution of DNA followed by nested PCR to identify first the genera of suspected pathogens (*Anaplasma species, Bebesia species, Theileria species* among others) and then the species and endemic subspecies respectively.
I was involved in the field sample collection, PCR optimisation, gel electrophoresis, and data analysis (partly) of this project as I was one of the contracted field/ lab technologists.

- Pick an experimental protocol (e.g. RNA extraction or RT-PCR) and explain what it does in simple terms.

In order to assess gene expression in a cell/group of cells, we determine the amount of the genes expressed as mRNA but since this is unstable, we reverse transcribe it to DNA. The DNA is then amplified and quantified in real time to measure respective gene expression with respective primers.

- How would you go about monitoring an experiment where you have to handle a large number of samples?

I would run the samples in batches with controls first to ensure precision and accuracy and secondly to optimise the carrying capacity of available equipment. I would include more stepwise quality control measures to track progress.

- What are some common pipette handling errors?

Pipetting different samples with the same pipette tip which causes cross contamination.
Not using the right pipette tip for the respective pipette.
Dispensing with excess pressure that causes a flush or the sample/liquid/reagent.
Not returning pipettes to maximum volume after use which weakens the pipette spring over time.
Resting the pipette horizontally rather than vertically which causes liquid content to flow back beyond the tip and potentially causing cross contamination to the next content.
Infrequent /irregular pipette calibration which creates variability in pipetting volumes hence errors.

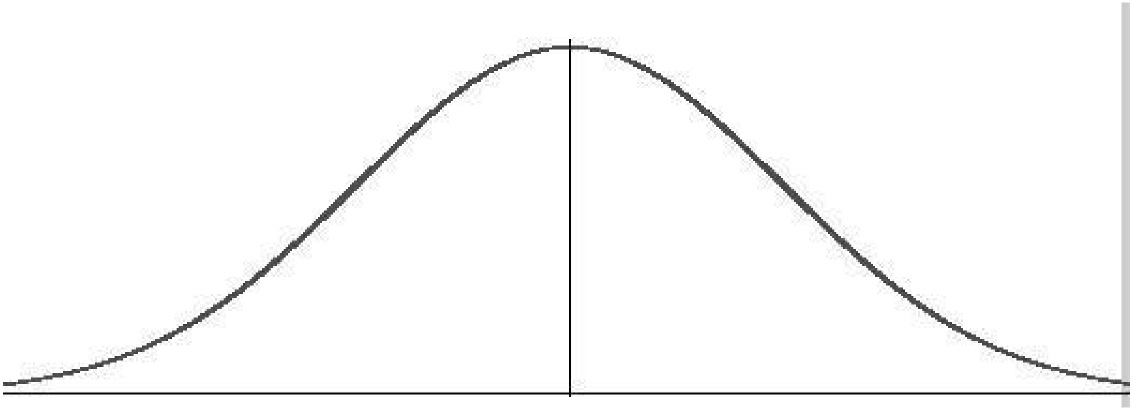
Maths/statistics

- What mathematics and statistics courses have you completed or what experience do you have?

As part of my academic studies, I have been taught data analysis with Minitab, Graphpad Prism and R as part as Research Methods. I have additionally learnt R from YouTube tutorials and Coursera. I have attached my Coursera courses as proof.

- What distribution is this?

Normal distribution



- What is the difference between a Student's and Welch's t-test?

Student's t-test assumes equal variances between the two groups being compared, while Welch's t-test does not make this assumption.

- When do you use non-parametric and parametric tests? Give an example of both in biological research.

When the data is not normally distributed following a normality test or when data is skewed, maltimordal especially in small samples. When data is ordinal or nominal, rather than continuous or when there are outliers in the data that cannot be removed as part of the analysis. When the median is chosen as the measure of central tendency instead of the mean. In all these instances, we use non-parametric tests. An example is when comparing pain levels (none, mild, moderate, severe) between two patient groups. Since the pain data is ordinal, a non-parametric test like the Mann-Whitney U test would be appropriate.

When the data follows a normal distribution, especially with larger sample sizes, when the assumptions for parametric tests are met, such as equal variances. Parametric tests like the t-test are more powerful . An example could be when comparing the mean Gleason scores between two groups of patients. If the Gleason scores are normally distributed, a parametric two-sample t-test would be appropriate.

- Look at this plot. Don't worry about the axes or scale; for all intents and purposes, it's just X and Y.

    1. What are some of the interesting features of this plot?

The data is not distributed to fit a line of best fit. Consequently, it appears to be non-linear with a sigmoidal curve.

There is a maximum value that Y approaches at higher values of X.

There is clustering of datapoint when X and Y are still low

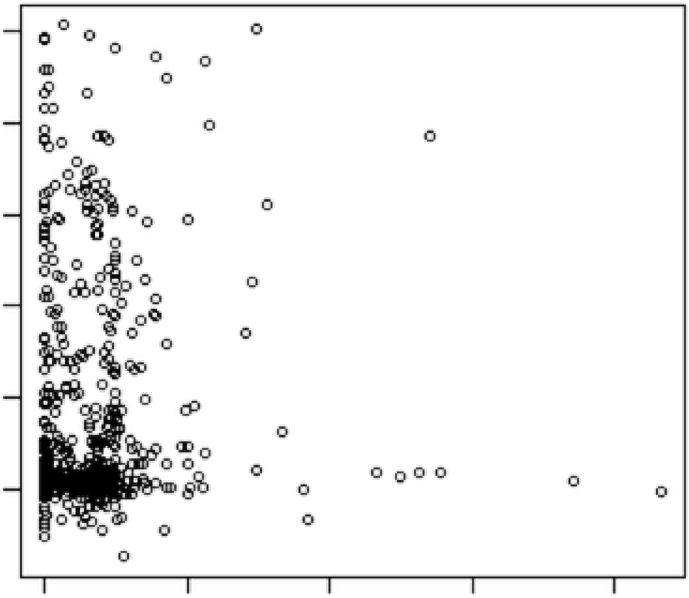
There is a baseline value where Y remains relatively constant initially as X increases.

The curve has a point of inflection or region of maximum rate of change in Y with respect to X.

The data points are more variable in the middle range of X values compared to the lower and higher ends.

    2. How would you predict Y given X?

I would use a non-linear regression model as the sigmoidal shape suggests a logistic regression model.



Programming/computing

- Can you describe your interest in computing?

My interest in computing stems from its power to accelerate research and discovery in the life sciences in terms of big data which is a current trend in forming current clinical decisions. I am fascinated by how computational approaches can complement wet-lab techniques, enabling one to analyse large datasets, model biological systems, and gain deeper insights into complex phenomena like cancer progression.

I have developed a keen interest in computing and bioinformatics tools through my academic and research experiences. During my Bachelor's and Master's programs, I gained hands-on experience with various bioinformatics software and databases, such as PyMol, cBioPortal, UCSF Chimera, and AutoDock Vina. I have used these tools for tasks like molecular modelling, structure visualisation, and virtual screening of potential drug candidates.

Additionally, in my research internships, I have worked extensively with data analysis software like Excel, Minitab, and GraphPad Prism to process and interpret experimental data. I also have some experience with command-line tools and bioinformatics pipelines, such as primer design and the 16S rRNA intermediate bioinformatics course through H3AfricaBioNet.

Moving forward, I am eager to further develop my bioinformatics skills and apply them to novel cutting-edge research projects, such as the proposed microbiome role in aggressiveness of prostate cancer among African men. I believe that a strong foundation in computing and bioinformatics will be invaluable in my future career as a translational biomedical researcher.

- Do you have any Linux experience or other command line experience?

Yes, I have used Linux (Ubuntu) as a main operating system since 2017 for 3d modelling with in Blender and academically as part of my undergraduate research project while running autodock vina , open bubble, shell scripts and anaconda for molecular docking file preparation, docking via commandline and filtering data output
As I venture into bioinformatics, I have realised that many programmes have been written to run in Linux, as such I have had to learn to use basic commands such as executing commands, creating and changing directories , changing permissions.
I have additionally used Linux to fetch data from online biological data repositories such as Sequence Read Achieve(SRA).

- You have some data on your laptop and you need to transfer it to a Linux server for processing, what protocol or program would you use to do that?

I would use the SCP (Secure Copy) protocol. First, I would ensure that the server is running the

OpenSSH server software and that the firewall allows SSH traffic. I have done this while on the 16s rRNA intermediate Bioinformatics course where we were using an institution's server to run experimental data locally on personal computers.

- What programming experience do you have?

I do not have specific programming experience but have attended Python programming courses (Fundamental skills in Bioinformatics and Programming in Python) on Coursera. From this, I am familiar with data types, calculations, loops and other fundamental python knowledge. I have used these experimentally to organise files on my laptop. I have additionally written Linux shell scripts to create files, change directories and permissions while attending an Introduction to Linux course.

- Have you done any web development or made an android/ios app or anything like that?

No, I have not.

- Write a program (in any language or pseudo language) that prints the numbers from 1 to 100. But for multiples of three print "Fizz" instead of the number and for the multiples of five print "Buzz". For numbers which are multiples of both three and five print "FizzBuzz".

Using the for loop in python;

```
for num in range(1, 101):
    if num % 3 == 0:
        print("Fizz")
    elif num % 5 == 0:
        print("Buzz")
    elif num % 3 == 0 and num % 5 == 0:
        print("FizzBuzz")
    else:
        print(num)
```

- You have a directory named `sequences` in your home folder. Within a terminal window, what command would you use to print out a list of fasta (ending with extension .fa) files in that directory

```
ls ~/sequences/*.fa
```

Bioinformatics

- If possible, please summarise a project with bioinformatic elements in that you have been involved with.

  My recent MSc project aimed to identify genes that are increasingly mutated with progressing prostate cancer (PCa) by analysing mRNA sequencing data from a cohort of 491 PCa patients. The researcher found that the genes ENOX1, CCDC122, and LACC1 had increasing deep deletions associated with increasing age at diagnosis which in-turn positively correlated with disease severity measured by histological Gleason scores.

  The bioinformatics analysis involved retrieving mRNA sequencing data from the cBioPortal database for the Prostate Adenocarcinoma (TCGA Firehose Legacy) cohort.

  Analysing genomic alterations, specifically deep deletions, of ENOX1, CCDC122, and LACC1 across the patient samples using the cBioPortal tools.

  Examining the effect of deep deletions of these three genes on overall survival using Kaplan-Meier plots.

  Performing protein-protein interaction analysis for ENOX1, CCDC122, and LACC1 using the STRING database.

Conducting pathway enrichment analysis using Enrichr and Reactome to identify pathways enriched by genes upregulated following the deep deletion of the three genes.

Notable findings were - Deep deletions of ENOX1, CCDC122, and LACC1 simultaneously co-occurred in 16% of the patient samples.

The top enriched pathways following the deep deletion of these genes were estrogen biosynthesis, KSRP signalling, omega-3 and omega-6 fatty acid metabolism, and Rap1 signalling.

The research suggested that these pathways could be potential druggable targets in PCa patients with deep deletions of ENOX1, CCDC122, and LACC1. They proposed future work involving wet lab experiments using PCa cell lines to validate the findings.

Link to preprint: doi.org/10.1101/2023.10.12.23296974

My undergraduate research project "A Molecular Docking Study of Human STEAP2 for the Discovery of New Potential Anti-Prostate Cancer Chemotherapeutic Candidates", aimed to use the STEAP2 protein, which is a prostate cancer-specific biomarker, as a target for discovering potential new drugs for treating prostate cancer through computational molecular docking approaches.

The project involved retrieval of human STEAP2 sequence from UniProt, BLAST -ing it to find similar sequences from different species for phylogenetic analysis using MEGAX of conserved motifs. Homology modelling (HHpred, AlphaFold, RaptorX,SwissModel) was used to predict its 3d structure which was then used for molecular docking in Autodock Vina with a set of FDA approved drugs from the drugbank to identify potential drug candidates that can bind STEAP2. The top scoring complexes, by binding energy, were shortlisted as promising drug candidates such as Triptorelin and Leuprolide and mechanism of action were postulated as potent inhibitors of STEAP2 activity.

Where applicable validation was done with independent homology modelling engines, 3d structure analysis, independent 3d docking software and binding site prediction tools.

The study demonstrated how integrating various bioinformatics techniques like sequence analysis, modelling, docking, and structural analysis could aid in rational drug discovery targeting a specific protein implicated in cancer progression. The promising results suggested further investigation into developing STEAP2-targeted therapies for prostate cancer treatment.

Link to paper: doi: 10.3389/fbinf.2022.869375

- Pick a bioinformatic algorithm and explain what it does in simple terms.

The BLAST (Basic Local Alignment Search Tool) algorithm is used to compare a query sequence of nucleotides or amino acids (DNA, RNA, or protein) against a database of sequences to find regions of similarity.

It works by aligning short,high scoring sub sections of a query sequence instead of aligning the entire length of the query sequence. It then compares these subsections of the sequence against a database of sequences of the same sequence type. It scores any matches and reports those exceeding a given threshold as results/hits.

- What is the difference between technical and biological replicates?

Technical replicates are repeated measurements/tests on the same sample done to assess the technical reliability of the measurements such as its precision and accuracy, while biological replicates which are done in parallel to different samples, evaluate the biological relevance and reproducibility of the experimental findings across different biological systems or samples.

- Give us a command to find, if a given sequence name (`eg. tubulin`) is present in a given file (`eg. gene_seqs.fa`)

grep 'tubulin' gene_seqs.fa

- Write down best pairwise alignment possible for following two DNA sequences ATGCTACGCGTAGCAT and TGCTTAGTACC

```
ATGCTACGCGTAGCAT
-TGCTTAGTACC----
```