UNIVERSITY OF BIRMINGHAM

UNIVERSITY GRADUATE SCHOOL

2022-2023

PGT KEP PROJECT INDIVIDUAL ESSAY COVER SHEET

I confirm that I have read and understood the regulations on plagiarism* and have acknowledged the work of others that I have included in this dissertation.

Please read the following statement and **tick ONE box** regarding permission, or denial thereof, to view your dissertation by other students:

- **I AGREE** to allow my assignment to be seen by future students. ✅

By signing this form, I agree to allow access to future PGT KEP students, as part of the University of Birmingham, to view my assignment, or part thereof, for guidance as an example of good practice. For its part, the University will grant access to PGT KEP students as it deems appropriate, but in so doing forbids anyone to copy or use my assignment in any other way or for any other purpose.
I understand that my assignment will be available to view via Canvas and that any personal references will be anonymised.
I further understand that the University has no control over the actions of third parties, and should I have any concerns, my permission may be withdrawn, at any time, by advising the PGT KEP team in writing/via email.

- **I DO NOT AGREE** to allow my assignment to be seen by future students. ☐

Print Name: Frederick Fuller

Student ID: 2506189

Date: 23/08/23

---

*Plagiarism, in this context, is the reproduction of material from books and articles without acknowledgement. It is the act of passing off another person's work as your own, copying a fellow student's work or reproducing work submitted by a past student. Such actions are seen as a form of cheating and, as such, are penalised by examiners according to their extent and gravity.

You should not quote existing work without quotation marks and appropriate referencing. An attempt to present the work of someone else as your own may lead to your dissertation being awarded a mark of zero. You are required to state the full references of all sources that you use. If quotations are made, they must be explicitly and fully referenced, including stating the relevant page number(s). You will be penalised very severely if examiners find that you have presented a section of a book, an article or a paper without appropriate referencing. If you are not sure about how to quote an existing work, please ask for advice from your supervisor.

UNIVERSITY OF BIRMINGHAM

A BIOINFORMATICS DEVELOPMENT OF THE PROPOSED GBS DIAGNOSTIC EXPANSION OF LINEAR DIAGNOSTICS COMBINED WITH A DIAGNOSTIC ACCURACY IMPROVEMENT PIPELINE

This assignment is submitted as partial fulfilment of MSc Bioinformatics 2022-23.

Student Name: Frederick Fuller
Student ID Number: 2506189
Group/Individual Supervisor: Jean-Baptiste Cazier/ Mamunur Rashid
Word Count: 7751

1. Abstract

Linear Diagnostics is a company which specialises in sexually transmitted disease diagnostics. This individual essay encompasses two pieces of work. The first relates to the group project. As a group, we were tasked to develop a proposal for an expansion into group B streptococcus diagnostics. This involved the creation of an informative workflow pipeline, containing both code and pseudo code. The end result of this work is presented in the group report. However, it was not possible to present the vast majority of background information. Within this essay, the literature and methods that formulated the workings of the final outcome are explained and critiqued. This is of value to Linear Diagnostics as it projects any possible limitations and provides further explanatory detail to the group project.

The second piece of work is the individual project. I have been tasked by Linear Diagnostics to create a pipeline that will allow them to update their diagnostic systems thereby making them more precise. I have been provided past data to construct this pipeline, with the final output being sensitivity and specificity values. This pipeline uses the direct output from Linear Diagnostics diagnostic technology (fluorescence values from linear dichroism), in addition to cycle threshold values to achieve this task. To compliment the annotated R Markdown pipeline, I have provided an explanatory video of the code. As this is a tool created for Linear Diagnostics, their current knowledge of R studio language is unknown and this video serves as a helpful guideline. The video will also assist future recommendations (section 6), if Linear Diagnostics aims to pursue this further. The pipeline created is novel, as this technique has not been applied to EXPAR technology. This is due to the cycle threshold premise being applied to polymerase chain reaction assays, a similar, but not identical outcome. The accommodating R code and video can be found in the google drive in the appendices (section 8).

2. Introduction

In this introduction, the background of Linear Diagnostics is presented. This includes a summary of their past products and an introduction to *Streptococcus*. The outline of Linear Diagnostics current technology, and what differentiates this company from their competitors, is also provided. These two aspects are vital to understanding the following research and work provided for Linear Diagnostics.

2.1 Background

Linear Diagnostics (LD) is a company which specialises in next generation patient diagnostics of sexually-transmitted infections. LD utilises Exponential Isothermal Amplification (EXPAR) combined with linear dichroism. The resulting collaboration of these two techniques is a significantly quicker diagnostic system. The time taken to produce

a result using LD's technology is 20 minutes, whereas the closest competitor (Binx) takes 30 minutes. Within this tool, the sample undergoes three separate phases: sample processing, EXPAR amplification and finally detection. The sample undergoes chemical lysis and silica column binding with Guanidinium isothiocyanate, before being washed with a salt and elution concentration and a low salt buffer. Then the EXPAR reaction takes place, the DNA restriction enzyme is used to deconstruct the sequence at a particular location then the polymerase is used to extend the trigger sequence. This simple reaction creates a multitude of trigger sequences, which can be detected by fluorescence through linear dichroism. If the result shows the trigger sequence is present (positive diagnosis of an STI) then a significant earlier increase in fluorescence, compared to a sample without the trigger sequence, will be detected.

Linear dichroism takes advantage of the difference in light absorption based on the direction of light and the materials alignment. This technique is used to study the structure and properties of different molecules and materials, by observing the difference in their arrangement and how they interact with the light. In this case, this process is used to detect the trigger sequence. Understanding the technology that LD uses, and how the analysis can be improved, upon is vital to enhancing the business and bioinformatics capabilities. The output of this machine is then input into a pipeline based in Microsoft Excel to determine whether the sample is positive for the chosen trigger sequence or not. Deconstructing and improving this pipeline will form a basis of this individual project.

Currently, Linear Diagnostics provides a diagnosis service for *Chlamydia trachomatis* and *Neisseria gonorrhoeae* using CT/NG, which is a test for both afflictions, with only the use of one sample. The company is looking to expand into diagnostics for Group B Streptococcus (GBS), which will form the second basis of this individual project.

Streptococcus is caused by gram-positive bacteria, this is bacteria that gives positive results in the Gram stain test (Sizar et al., 2023). Streptococcus is distributed globally and has the ability to affect more than 27 species of fish living in fresh, brackish, and marine waters (Bowater et al., 2012, Choi et al., 2004, Keirstead et al., 2013). Among these bacteria, *Streptococcus agalactiae* is highlighted as not only being an invasive disease in fish, but is a zoonotic hazard (Yildirim‐Aksoy et al., 2018). Certain strains of this aquatic *S. agalactiae* have the capacity to transmit pathogens between these aquatic animals and humans (Delannoy et al., 2013). In humans, *S. agalactiae* is found to colonise the rectovaginal tract and can cause neonatal infectious diseases, this is known as Group B Streptococcus (GBS). The exact reasons behind the emergence of *S. agalactiae* in humans and other animals remains poorly understood (Yildirim‐Aksoy et al., 2018).

2.2 Brief and Aims

This project report is divided into two sections. Firstly, as a group, we were tasked with creating an informative workflow-pipeline to assist in creating a proposal into the

diagnosis of GBS. While the results are presented in the group report, details on the methodology which are imperative to the development of this expansion, are elaborated in the later sections. This report will highlight the importance of this expansion into the wider field of STI diagnostics, as well as Linear Diagnostics. The background of these competitors will be highlighted to provide a deeper market overview, as well as providing significant literature to elucidate this case. Therefore, this will emphasise the convenience and value of this pipeline and further detail of its processes. The two main techniques that are used in the creation of this pipeline are Precision Weight Matrix (PWM) and Naive Bayes Classifier. These two approaches will be expanded on in the literature review. This code developed and represented in the group report relies on the understanding of research and results from "Sequence dependence of isothermal DNA amplification via EXPAR" by Qian *et al* (2012). To support this group project, further evaluation and explanation will be given to this companion paper.

The second section is based on my individual research and development, I have been tasked with creating a tool which Linear Diagnostics can use to compare and test different Ct thresholds in relation to sensitivity, specificity and accuracy. This will assist LD in creating more precise and therefore reliable diagnostics tests. This pipeline can therefore be used on both this pressing expansion into the diagnostics of GBS, and any STI. The balance of sensitivity vs specificity is vital for an efficient diagnostic test. Sensitivity is also referred to as the true positive rate of a model. This key term is used to correctly identify true positives in a diagnostic result. Comparatively, specificity measures the ability the model has to identify true negatives. If a pipeline were to have high sensitivity and low specificity, the result would capture as many true positives as possible in the results. The output of this would have most results be positive, increasing the number of false positives as a result. The opposite would be true if a pipeline were to be low in sensitivity and high in specificity. The pipeline would capture a large amount of true negatives, however would adversely increase the number of false negatives in the process. A balance between these two terms would hold the optimum number of both true positives and negatives. As this is a tool being provided for LD, a video has been created to accompany this individual essay. How much LD knows of R Studio script is not unknown so this video sets out to explain the commands used and the data flow processes. The video also allows LD to potentially change large sections of the pipeline and observe different samples or sample types.

3. Literature Review

Sexually Transmitted Infections (STIs) cause both a significant global health and economic challenge. There are over 35 bacterial, viral and parasitic pathogens that can be transmitted through sexual contact (Chesson, 2017). It was estimated, in 2008, that there were 489.9 million new cases of four curable STIs among adults aged 15 to 49. This represents a 11.4% increase from the 448.3 million cases reported in 2005 (WHO, 2012).

Specifically, of the cases in 2008, there were 105.7 million new instances of chlamydia, 106.1 million new cases of gonorrhoea, 10.6 million new cases of syphilis, and 276.4 million new cases of trichomoniasis (WHO, 2012). During 2008, it is estimated that 100.4 million adults were infected with chlamydia, 36.4 million with gonorrhoea, 36.4 million with syphilis, and 187 million with trichomoniasis (WHO, 2012). These significant statistics highlight the importance of education and diagnostics of STIs.

In addition to the lack of education, stigma is a large contributor to the increase of STI cases. Stigma refers to the negative perceptions society holds towards certain individuals or groups, leading to their isolation, rejection, and discrimination, Within the global arena of STIs, the existing social inequalities related to social class, race/ethnicity, immigration status, gender, gender expression and sexual orientation enhance and complicate stigmas, as each inequality carries different stigmas (Starrs et al., 2018).

Over the past half-century, significant advancements in basic, clinical, and translational sciences have revolutionised the public health's ability to address the various pathogens associated with STIs, including HIV (Garcia *et al*., 2021). However, despite these considerable achievements in understanding, detecting and treatments, STI cases continue to grow (Garcia *et al*., 2021). By developing and enhancing previous diagnostic measures, it is possible to substantially decrease the time taken from sampling to attaining a result, additionally increasing the number of patient output. As the diagnostics become increasingly accessible to all patients, these factors play an important role in reducing stigmas.

GBS remains a common cause of neonatal diseases, such as pneumonia, septicemia, and meningitis with its main form of transmission being from mothers to newborns during childbirth (Schrag et al., 2000). Although, its prevalence has declined in certain countries due to proactive preventative efforts. During the years 1998 to 2000 in the US, the overall early-onset disease rate was observed to be around 0.5 to 0.6 cases per 1,000 births, with variations caused by geography and race (Centers for Disease Control and Prevention A, 1998).Comparatively *Chlamydia trachomatis*, the most prevalent STI, had a rate of 2.5 cases per 1,000 people in the US (Koplan et al., 2001). However, it is worth noting that *Chlamydia trachomatis* is significantly more susceptible and widespread than GBS. In addition, Denmark reported a lower incidence, being 0.24 per 1,000 births confirmed GBS

infections, emphasising the impacts of geography and race can have on this condition (Carstensen et al., 1985). The transmission of this bacteria through childbirth is linked to the bacterial colonisation of the birth canal passage. As a result of this, the Centers for Disease Control now recommend intrapartum chemoprophylaxis for pregnant carriers as a strategy to reduce this neonatal GBS infection (Centers for Disease Control and Prevention B, 2002). Although this occurrence is lower compared to neonates, GBS is also responsible for invasive infections in non-pregnant adults (Bolaños et al., 2001).

Each year, more than 700,000 live births occur in the UK, and a significant concern within this context is GBS as it causes these early onset neonatal infections. It is estimated that around 50% of GBS-positive mothers can transmit the infection to their newborns. In the absence of intrapartum antibiotic prophylaxis (IAP) during labour, the incidence of GBS EOD in these cases is approximately 1-2%, highlighting the importance of diagnosis. Currently, the prevailing standard for detection is enriched culture medium (ECM) testing, this process requires 24-72 hours and is mainly provided to women that are deemed at a higher risk. This screening takes place around the 35-37 week mark of pregnancy. This factor, combined with the current long diagnosis time is flawed when accurately predicting GBS carriage during labour. LD seeks to break into the market of GBS diagnostics. The current opportunity is the development of a quick, cost effective and precise point-of-care (POC) diagnostic. This project's focal point involves the use of their current EXPAR and linear dichroism technology, ensuring a faster, and equally as accurate, test compared to the current competition.

In terms of market potential, an environment where GBS screening during labour becomes standard practice for all birthing mothers, could result in a market of £28 million in the UK alone, when compared with the current pricing of private testing options available . LD’s new technology has the potential to revolutionise the screening of GBS for women in labour.

3.1 Group B Streptococcus Expansion Review

The following section will be broken into two parts to provide a complete and comparative background. The first will discuss the literature and further reasoning from a business and biological perspective into this expansion, while the second will discuss the mechanics behind the pipeline requested by Linear Diagnostics.

3.1.1 Expansion Background

To fully comprehend the scope and attributes of the expansion into GBS diagnostics, an introspective review is required. The following questions have been sourced and will provide an estimated review of the many different aspects of this expansion and provide further reinforcement to the workflow pipeline.

How Does This Expansion Align With Competitor Companies?

Currently, there is a lack of a cost-effective point-of-care (POC) diagnostic tool for Group B Streptococcus during labour in the UK. Annually, over 700,000 babies are born in the UK, yet 35% of expectant mothers have routine NHS GBS tests, this has been confirmed by LD. This current approach to GBS diagnostics subjects 65%-85% of GBS-negative women to unnecessary antibiotics (Saari et al., 2015). To overcome this currently unsuccessful and wasteful practice, a swift and highly sensitive POC intrapartum diagnostic test needs to be present in maternity wards, which LD is aiming to satisfy.

This expansion will include development that will encompass not only the assay, reagents and other diagnostic technology, but a custom-designed disposable assay cartridge. This cartridge will incorporate essential elements of the diagnostic test, for example, sample concentration, DNA amplification, temperature and the Linear Dichroism detection technology. This cartridge is also protected as the intellectual property of Linear Diagnostics Limited (LDL), resulting in the reinforcement of this technology's longevity.

What is the Healthcare Need, Technological Challenge and Market Opportunity Behind this Expansion?

As stated in the answer to the previous question, there is a dire need for diagnostic technology that prevents the use of giving patients unnecessary intrapartum antibiotic prophylaxis (IAP). Concerns are raised by Seedat *et al* (2019) about the use of universal screenings and how these dated practices lead to increased adverse effects due to the risk of antibiotic resistance. In addition to this risk, this unnecessary service provided by the hospital increases their expenditure, further emphasising the cost effective benefits of this new technology in the event they were implemented in hospitals. Using a sample of 313 women, it was proven by Picchiassi *et al* (2019) that the use of IAP also increased the length of time mothers spent in hospital. This factor not only further increases hospital workload, but negatively impacts the stigma of GBS diagnostics.

Finally, a study by Zimmerman and Curtis (2019), found a correlation between the use of IAP and a lower abundance of Actinobacteria, especially *Bifidobacteriaceae* in the intestinal microbiota of the affected infants of the mothers. This important influence on the infants during this crucial 'critical window' stage may have an adverse affect on their immune development. All these factors combined underline the importance of this expansion and relevant pipeline.

Commercially, there are limited alternatives on the market which do not require overnight culturing. The primary competitor is BD GeneOhm test, this product provides a sub-60-processing time. However, this test presents drawbacks in POC settings due to the need for specialised equipment that is uncommon in maternity wards, an example of this is a vortex and microcentrifuge. Furthermore, its execution mandates the presence of skilled laboratory staff, given that precise laboratory methodology profoundly influences its capabilities. This key limitation of the technology used by competitors constrains their efficiency to produce timely and accurate results. This expansion of LD aims to provide a true POC product with the assistance of this developed pipeline.

What is the Underpinning Scientific Innovation for this Expansion?

There are two main technologies that revolutionise the diagnostics industry that Linear Diagnostics utilises. The first of these is the EXPonential Amplification Reaction (EXPAR). This utilises isothermal DNA amplification. This technology is used to replicate DNA sequences without the need for thermal cycling, as required in techniques like PCR which would significantly lengthen the diagnostic test. This reaction is coupled with a bioreagent, after this the sample undergoes linear dichroism. The use of linear dichroism ensures a fluorescence output of gene presence without the need of fluorescent enzymes (which would require further heating or cooling). Linear dichroism is a characteristic displayed by certain molecular arrangements, causing uneven absorption of linearly polarised light across two perpendicular axes.

Within biomolecules, an array of chromophores exists (these being molecules with the ability to selectively absorb light). The process of measurement mandates alignment of chromophores, whether this alignment is complete or partial, in relation to the incoming light. This alignment is facilitated by shear forces generated as the sample flows under precisely defined conditions. Notable responsive shifts in molecular alignment, linear dichroism also proves receptive to the creation of complexes, particularly when intertwined with oligonucleotides or antibodies. Comparatively, the output from linear dichroism is similar to that of qPCR results. Applying this technology to the diagnosis of further STI's (namely GBS) could revolutionise the STI diagnosis market.

The intellectual property conceived by Dr.Matt Hicks and Professor Tim Daffrom at the University of Birmingham outlines how linear dichroism can be harnessed to detect particular molecules within a given subject. The method involves modifying extended scaffold molecules to bind with the target, thereby producing a distinctive signal at a precise wavelength. This wavelength precision allows for simultaneous detection of numerous targets within the same sample.

EXPAR stands as an emerging and swift DNA amplification technique employed for enhancing short oligonucleotides in the context of molecular diagnostics. This technology offers a host of advantages: its uncomplicated nature, cost-effectiveness, rapidity (yielding detection in less than 5 minutes), adaptability, capacity for optimisation and resilience against sample cross-contamination.

3.2 Pipeline Mechanics

The mechanics of this pipeline are influenced heavily by: "Sequence dependence of isothermal DNA amplification via EXPAR" by Qian *et al* (2012). In summary, this paper uses precision weight matrices (PWM) and a Naive Bayes Classifier (NBC) to assess template performance. This template performance can then be evaluated and compared with other templates in order to select which sequence is most efficient for diagnostics. However, there are potential limitations and critiques on this study. For this expansion, the CDS2 gene is used when searching for a trigger template and its design.

3.2.1 Design and Methodology

EXPAR Template Design Based on Thermodynamic Criteria

In this study, a total of 384 template sequences were used, compromising 64 sequences with known characteristics and 320 novel designs. Sequences 2-11 were derived from a previously published study (Van Ness *et al*., 2003), introducing single-nucleotide changes in specific positions. For the remaining 320 sequences, a randomised distribution of bases were assigned to 14 variable positions within the trigger complement and the nicking enzyme post-cut site. Additionally, an extra thymine (T) was appended to position 20 (D) of each template. An example of this is shown below in Figure 1:

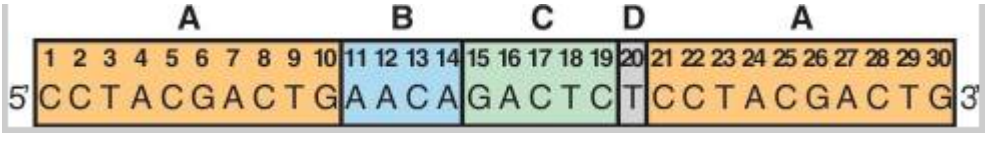

Figure 1: Example template sequence. A is the trigger complement which is duplicate for positions 1-10 and 21-30. B is the nicking enzyme post cut-site. C is the complement of the nicking enzyme recognition site. D is the additional T at the end of the trigger-binding site. Taken from Qian et al (2012).

EXPAR Performance Screening

The EXPAR reactions were executed using the designed template sequences along with their corresponding trigger sequences. The reaction mixture consisted of the key components essential for this reaction to take place. Real-time fluorescence amplification curves were generated utilising a Bio-Rad Opticon real-time thermocycler, which differs from Linear Diagnostics' linear dichroism technology. To ensure quality control, rigorous replication of each template sequence was conducted to ensure data robustness.

Data Analysis and Classification

The analysis of the real-time fluorescence curves involved the development of the EXPAR Data Analysis Tool (EDAT) using the MATLAB programming language. This tool facilitated the fitting of the experimental data to sigmoidal curves, and significant parameters, including P10 (point of 10%%), P90 (point of 90%), N10 (nadir of 10%) and N90 (nadir of 90), were extracted from these curves. These features of this model provide the key aspects, insights and behaviours of this model.

Precision Weight Matrix (PWM) Approach

A comprehensive 10-fold cross-validation method was used. Precision Weight Matrices were generated from templates that were in the lowest 30% of the P90 range as well as templates that were in the highest 40% of the Diff value range. A support vector machine (SVM) algorithm was used using these P90 and Diff values as features to determine a classification boundary between these two scores.

Naive Bayes Classification (NBC) Approach

A position motif count matrix was developed using different templates. Utilising this approach, these templates were categorised into performance classes. The 10-fold cross-validation strategy was consistently applied to ensure the reliability of predictions.

Development of Tool and Performance

To streamline these computational methods the EXPAR Template Sequence analysis tool (ETSeq) was created using Python programming. This tool integrates both the PWM and NBC approach and accepts user-defined sequences. The predicted classifications are generated into a Microsoft Excel file. This tool was validated through extensive analysis, however for this diagnostic expansion task, the use of this tool is not necessary.

3.2.2 A Critique of the Methodologies

Understanding the methodologies behind this paper is crucial to forming the backbone behind the computational aspect of this expansion. Knowing this, it is equally as important to critique the methods to fully evaluate other possible solutions if an error were to arise in the future.

Thermodynamic Criteria for Template Design

This study would benefit from further explanation of the underlying assumptions and potential complexities associated with the thermodynamic criteria. For instance, the influence of other factors, such as local sequence context and template-template interactions, might impact the stability and performance of the templates. A more detailed exploration of these variables could enhance the robustness of this design. Likely by incorporating experimental validation and computational simulations can help overcome these limitations.

Experimental Replication and Data Quality

Replication of experiments are outlined in this study, However, it is worth considering the potential impacts of factors like pipetting errors and batch variations that could affect consistency across replicates. Moreover, the study lacks a detailed discussion of detailed steps taken to minimise these potential sources of variation. Providing insights into the experimental controls and techniques employed to mitigate such issues would bolster the reliability and replicability of this study.

Classification Approaches and Predictive Power

This study introduces PWM and NBC. These techniques offer insights into the template performance prediction, however they may inherit limitations in dealing with the complex sequence-structure relationships in this biological system. Having a broader dataset or external ‘real life’ data could demonstrate the generalizability of these classification methods.

Possible Limited Scope

The thermodynamic criteria plays a pivotal role in template design, it may not encompass all factors that influence EXPAR performance. The study predominantly focuses on thermodynamics and sequence motifs, potentially overlooking other elements that contribute to amplification efficiency, such as DNA secondary structures and interactions within the polymerase enzyme. Expanding the scope to encompass these additional considerations could offer a more holistic understanding of template performance. Additionally, instead of using the Bio-Rad Opticon real-time thermocycler, linear dichroism could be used in order to provide a similar output to the diagnostic results and potentially reveal an additional feature to this aspect.

3.2.3 Comparison Index Analysis

Similar to the previous section, the results of this paper will be summarised in a comprehensive view, highlighting the important factors that influenced the GBS expansion development. For example, one of the outcomes of this paper is the development of the computational software they use to perform some of these computational methods, for this expansion the use of this computational tool is not necessary. The main output that this expansion utilises is the significant position motifs, the stages of this creation will be explained.

Position-Dependent Nucleotide Frequency Maps

To gain insights into the sequence patterns of the EXPAR templates, the researchers generated position-dependent nucleotide frequency maps. The result of this is shown below:

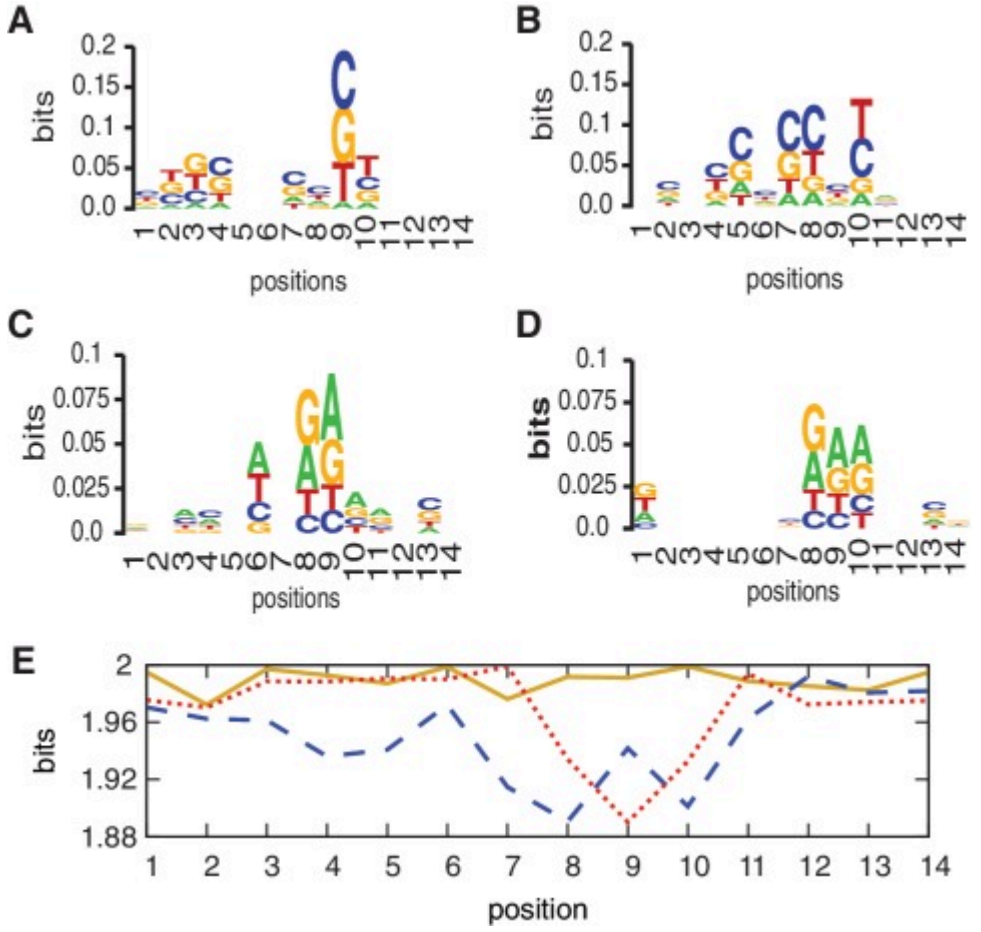

Figure 2: Position-dependent nucleotide frequency maps (sequence logos) A: Low P90 (fast amplification). B: Large positive Diff. C: High P90 (slow amplification). D: Large negative Diff (poor temporal separation between specific and non-specific amplification). E: Shannon entropy for each variable position. Blue: Class I templates. Red: Class II templates. Orange: Class I and Class II combined. Taken from Qian et al (2012).

Shannon Entropy Calculation

Shannon entropy, a measure of randomness or uncertainty in the nucleotide distribution at variable positions was calculated. This technique pinpointed the positions within the template sequences where nucleotide composition varied significantly. These positions highlighted contribute towards the final significant position motif output.

Machine Learning-Based Motif Identification

To identify the characteristics associated with performance, the NBC approach was used. NBC was used to identify the significant sequence motifs linked to well-performing (Class I) and poorly-performing (Class II) templates. These motifs represented patterns are indicative of either fast amplification with good specificity or slow amplification with poor specificity.

Multi-Base Motif Identification

Beyond singular nucleotides, multi-base motifs (patterns of consecutive nucleotides at specific positions) were identified. Some of these multi-base motifs occurred repeatedly at various positions within the template sequences and were strongly associated with template performance.

Visualisation

The identified motifs and their relative importance were visually represented using sequence logos and motif plots. These visualisations make it easy to interpret and comprehend the significance of specific sequence patterns. The result of this is shown on the following page::

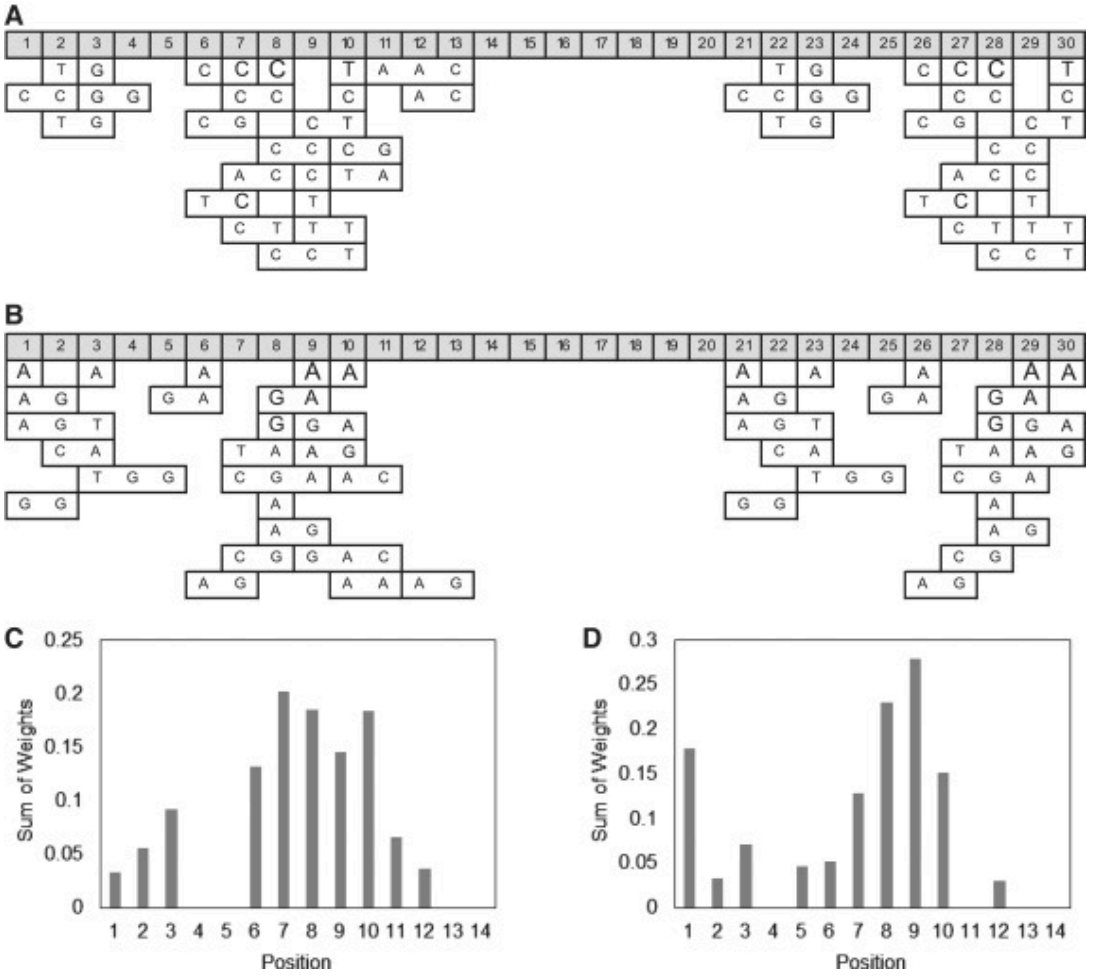

Figure3: Visual representation of the significant position motifs. Shows the highest ranking position motifs. A: Well-performing Class I templates. B: Poorly performing Class II templates. To show the significance in motifs, the font size was used and correlated to each motif's weight. Also shows the attribute evaluation. C: Well performing Class I templates. D: Poorly performing Class II templates.

In addition to this visual output, this analysis revealed important biological insights. For instance, it was observed that the size and hydrophobicity of nucleotide bases played a role in template performance. Certain nucleotides, such as cytidine and thymidine, were associated with well-performing templates, while guanine and adenine were linked to poor performance.

In conclusion, this final output can be used to find the best possible sequence that can be used in the diagnosis for GBS. Using the output in figure 3, the sequence that will be isolated for the diagnosis (linear dichroism mechanism) should fit as much as possible with the well-performing Class I templates, as these hold the most significant and concurrent motifs.

4. Methodology

As the methodology and results for the GBS expansion project was depicted in the group report, this and the proceeding sections will be focussed on the individual project. Referring back to the brief and aims, the task was set by Linear Diagnostics to create a tool which could output the sensitivity and specificity values when changing the diagnostic fluorescence threshold. To do this, the premise of cycle thresholds were used. I was provided by Linear Diagnostics with past Chlamydia linear dichroism output results in the form of a Microsoft Excel Spreadsheet. To manage the sample size manageable whilst retaining a reasonable number, only the urine samples were used.

The output of the linear dichroism technology is very similar to qPCR results, therefore the pipeline utilises the technique of Ct values and Ct thresholds that will be explained in further detail in the Results (section 5). The pipeline is developed using RStudio as this language is suited to handling data in this format and size whilst also providing all the applicable tools. Within the Microsoft Excel Spreadsheet provided by Linear Diagnostics, I was tasked to test this tool only on the urine samples. Not only does this reduce the samples into a much more manageable data set but also increases the speed of the pipeline considerably, allowing many more tests without computational delay.

5. Results

This results section will start by evaluating the background behind the Ct value and Ct threshold. This is required to later understand the mechanics of the pipeline and how the Ct value is acquired for each sample. The calculations for both sensitivity, specificity and accuracy will be included accommodating a background into how each term is defined. To fully understand each step of the pipeline, every step will be broken down into individual processes. Not only will this clarify the mechanics of the pipeline but it will also act as a guide for Linear Diagnostics for any future tampering.

5.1 Cycle Threshold Values Assignment

Cycle Threshold (Ct) values are important parameters in molecular biology. These values are most commonly used in techniques like polymerase chain reaction (PCR) and quantitative real-time PCR (qPCR). In qPCR reactions, genetic material is repeatedly replicated through a temperature cycle. During each cycle, the genetic material is heated to denature it, then cooled to allow short DNA primers to bind to complementary sequences on the target material, and finally, the temperature is raised again to allow DNA polymerase to synthesise new DNA strands using the primers as ‘starting points’. A fluorescent dye or a probe is used in this reaction and binds to the newly synthesised DNA strands. As the genetic material gets amplified, the amount of fluorescence also increases.

The Ct value is the cycle number at which the fluorescence signal emitted by the amplification process crosses a certain predefined threshold. This threshold is chosen to be identifiable at each position of a fluorescent curve. Ct values are inversely proportional to the initial amount of target genetic material in the sample: the higher the initial amount, the fewer cycles needed to cross the threshold. The background behind this term is relevant to understand the later mechanics of the tool.

For this project, I determined the Ct value to be the time of which the delta value of the fluorescence output to be the largest. This ensures that for each sample the Ct value will be located in the same position, to ensure there is no variability. A representation of the Ct value assignment and the accommodating threshold is shown below in Figure 2:

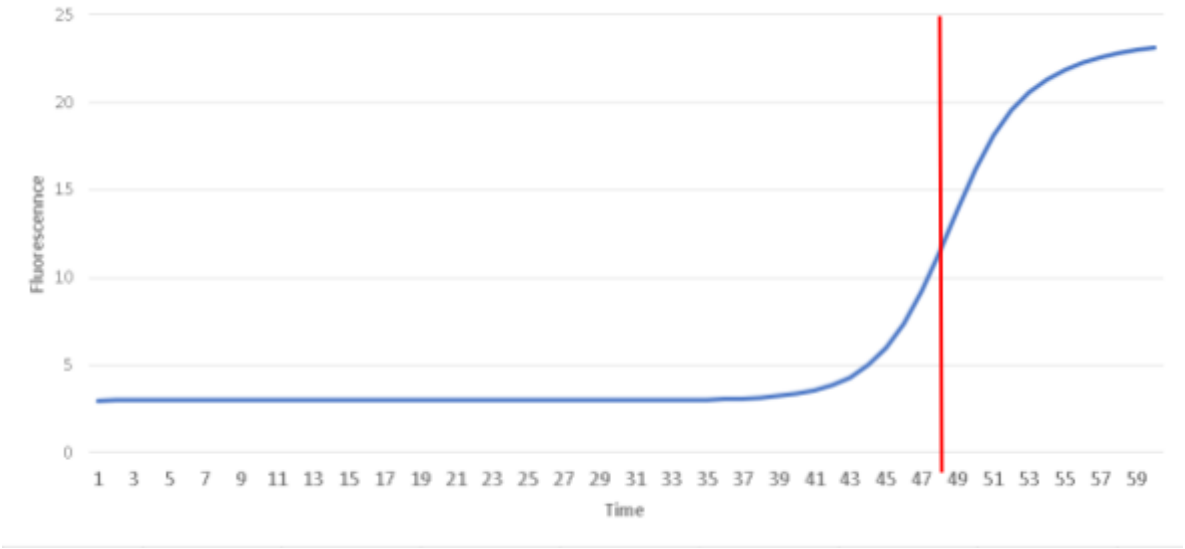

Figure 4: A line chart created from using fluorescence output from linear dichroism. The x axis represents time (seconds). The red line represents the Ct value this sample would have (48 seconds), if the designated spot was the location of highest delta.

Using Figure 2 as a reference, if the Ct threshold was designated at 51 seconds, this sample would be diagnosed as positive as the Ct value is prior to the threshold.

5.2 Sensitivity vs Specificity Calculation

Using the output from the previous section, the 'positive' or 'negative' binary classification can be tied to each individual sample. The data provided by Linear Diagnostics also contains the true diagnostic outcome whether the patient was positive for Chlamydia or not. Using this information, type I and type II errors can be inferred as well as the true positives and true negatives.

- True Positive (TP): This occurs when the diagnostic tool diagnoses the sample positive, matching the true outcome.
- True Negative (TN): This result is when the diagnosis is negative, which matches the negative outcome of the patient.
- False Positive (FP): This outcome is when the diagnostic tool result is positive, when in reality the sample should be negative.
- False Negative (FN): This final result occurs when the diagnostic tool outcome is negative when in reality the sample is positive.

Once these outcomes can be allocated to each sample, the overall sensitivity and specificity can be calculated for all samples. This is done using these equations:

$$Sensitivity = TP/(TN + FN)$$
$$Specificity = TN/(TN + FP)$$

Additionally, accuracy can also be calculated from the given output. Accuracy measures the proportion of correctly classified samples (both true positives and true negatives) out of all samples. The equation is shown here:

$$Accuracy = (TP + TN)/(TP + FP + TN + FN)$$

5.3 Pipeline Flow

As this pipeline is building off of previously provided data. It is necessary to include a flow-chart to visually depict all the stages concurrently. As each section is reliant on multiple separate processes, they are each depicted below:

1. Installing and Loading Packages and Loading the Document

Firstly, the packages used for this tool are installed and loaded into the environment. The packages used are: "openxlsx" and "dplyr" which are essential for the later data manipulation and analysis. Using the "read.xlsx" function provided by the "openxlsx" package, individual sheets can be isolated and loaded into different objects. The sheet "Sequence 10 Data Hood" is the location of the raw linear dichroism output, therefore it is read into "CTNGdata". The "Sequence 10 Analysis Hood" is where the true diagnostic

result is located, therefore is read into the “CTNGanalysis” object which will be used further down the pipeline.

2. Isolating the Urine Samples, Removing Retests and Cleaning the Samples

Within the Excel data sheet that was provided, under the “Sequence 10 Analysis Hood”, samples can be filtered by sample type. As tasked, samples were filtered by “Urine” and the identification digits were used to isolate the corresponding sample data in the “Sequence 10 Data Hood”. However, these isolated samples still contain those designated with “Retest”. Within the Excel spreadsheet cell formula, the “Retest” row investigates the delta values of the fluorescence curve of the sample. If the delta values present an abnormal pattern that has been designated by Linear Diagnostics, “Retest” appears in this row. This is a reliable form of quality control that the parent company has enforced, and gives reasoning as to why these samples need to be removed before continuing this analysis.

From here, the data is reformatted into a more suitable structure, removing unnecessary information (for example: date, gender and other blank columns and rows). As some of the samples have fluorescent periods longer than others, the otherwise blank cells labelled as “Na” will be replaced with 0. This won't affect the following delta value functions as the Ct value would be in the time before the 0 values start.

3. Creating the Delta Values and Ct Values

To calculate the delta values, the following created function “calculate_delta” will be used. In summary, this function calculates the difference between each of the fluorescence values. This function then repeats this difference calculation for all samples. As this section was created as a function, a new data frame object can be used for the results output, allowing the raw data to be preserved and not overwritten.

Once this new data frame has been created, the function “calculate_ct” is created to find the determinant Ct value for each sample. This function locates the row of which the highest delta value is present, then using this information determines the time of which this highest value occurred using the row number. Once this has been achieved this process is repeated for all samples in the input data frame.This final output of the function will contain the sample id number and the corresponding Ct value. For the same reason as the “calculate_delta” function creation, this process is also a function, allowing the possibility if there is any future error for the user to view past data objects.

4. Ct Threshold Diagnostic

At the start of this next stage, the Ct threshold will be designated by the user. This was created into its own individual object as to increase the ease of access for change and manipulation. Firstly, using the earlier input “Sequence 10 Analysis Hood” Excel sheet in the object “CTNGanalysis”, the true diagnostic results of the samples are isolated and then input into the samples corresponding column.

Using the set Ct threshold, the samples that contain a Ct value that is lower (the samples that have fluorescence rising earlier) than this value are distinguished as “Pos”. The samples that have Ct values higher than this threshold are then appropriately marked “Neg”. The final output of this section is a data frame that contains three rows. The first is the earlier Ct value, the second is the “CTNGanalysis” true diagnostic result and the third is the new ‘test’ Ct threshold result. Sequentially, these rows are labelled: “result”, “CTres” and “ThreshRes”. Visually, this output makes it very easy for the user to understand and analyse.

5. Confusion Matrix

Using the previous output, the following true positives, true negative, false positives and false negatives can be assigned to each sample. The output is bound onto the previous data frame with an additional row assigned “ConfMatrix”. Using this result, the sum of the true positives, true negatives, false positives and false negatives are calculated and output. Not only is this used in the next calculation but for the user this is an important statistic to present.

Finally, using these sums, the sensitivity, specificity and accuracy are calculated using the formula earlier in the results section.

5.4 Pipeline Summary

To visually represent the pipeline flow as well as the data inputs and outputs, a flow chart has been created and is shown below as Figure 3:

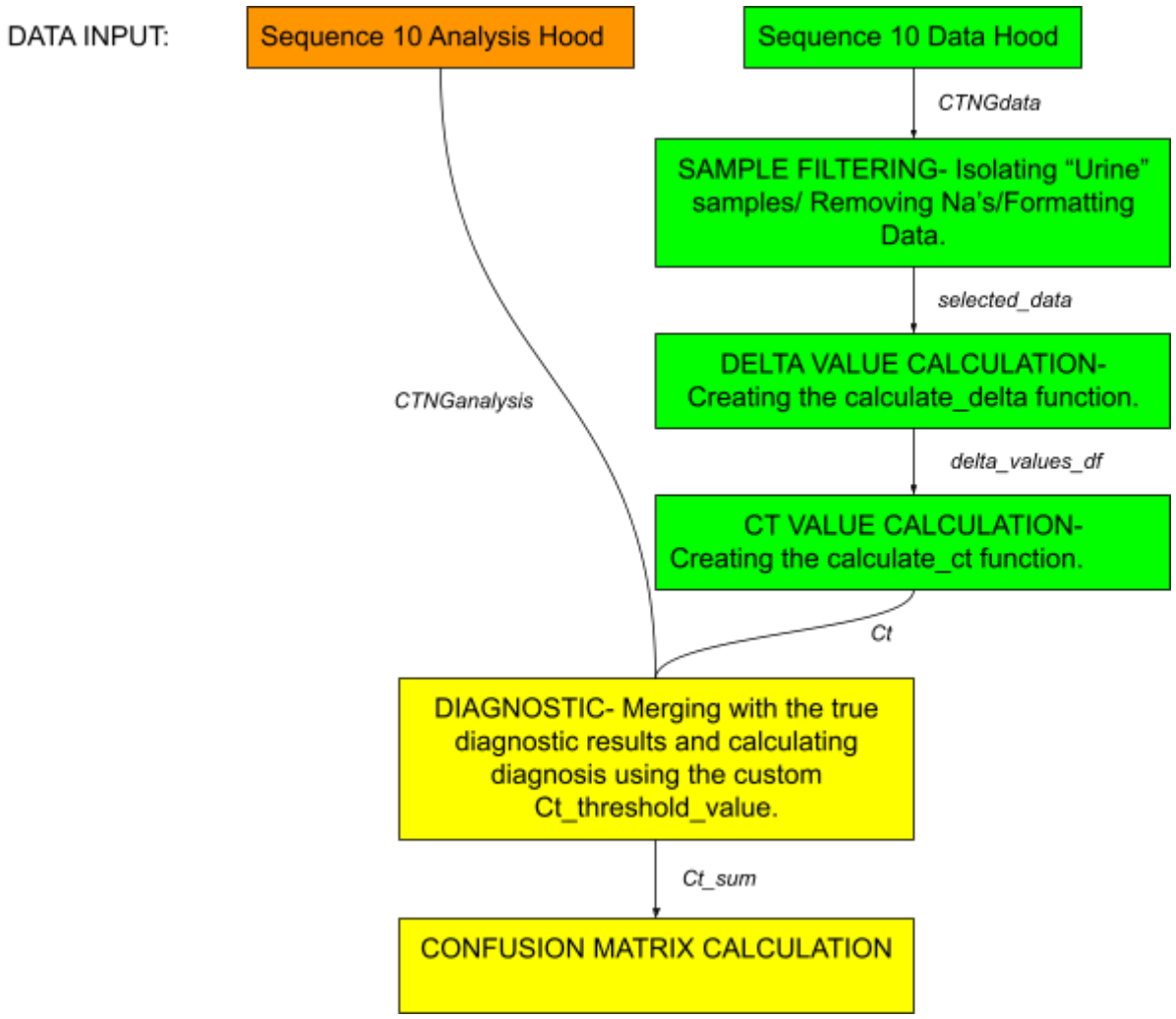

Figure 5: A flow chart to show the processes and data flow of the pipeline. Noting the object outputs of each process.

6. Future Recommendations and Conclusion

In summary, I have created a pipeline that can produce accurate sensitivity, specificity and accuracy that is determined by the custom Ct threshold set by the user. This pipeline is functional and all the different techniques and commands used have been explained through the annotation on the code itself. However, there are a few aspects of this pipeline that can be further improved upon, or introduced to improve its capability and accessibility. These are expanded upon and developed in the following sections.

6.1 R Shiny Creation

R shiny is a web application for R studio, the coding language that this pipeline was written in. There are two essential components of R shiny that define its use and importance. Firstly, there is the User Interface (UI). This is the front-end of the application. This is how the application looks to the user. In this case, the UI would present a few select options that the user can change easily. These would likely be: the Ct threshold and the samples intended to be used (either filtered by sample type or id). The second component is the server. This is the back-end of the application which contains the pipeline itself. By isolating the pipeline at the back-end it preserves the code, removing the possibility of it being tampered with, breaking essential pieces.

Once the desired values have been input, the pipeline will run in the back end. Once complete, the outputs will display on the UI. This will also give a timeline of how the data is transformed through the individual processes of the pipeline, identical to the different chunk outputs in the R Markdown File provided. Another benefit that using R shiny provides is that it is usable for anywhere as it is a web application. This minimises the risk of different versions of code being created and therefore incorrect results being produced.

6.2 Sample Size Increase and Linear Dichroism Output

Currently, the pipeline uses a satisfactory number of samples to provide adequate sensitivity, specificity and accuracy values. However, if the sample size were to be increased, the number of true positives, true negatives, false positives and false negatives would increase. This would provide a more reliable and applicable output. To implement this however, the pipeline would need to be rewritten to a certain extent. Either including a system that does not require sample id to function or uses samples based off of sample type.

Another aspect that can be included is the overall data input and its formatting. This pipeline formats the data in a way that is reliant on the Microsoft Excel spreadsheet being consistent. The code can be updated in an attempt to accept wider varieties of formatting.

7. References

Bolaños, M., Cañas, A., Santana, O., Pérez-Arellano, J., de Miguel, I., Martín-Sánchez, A., 2001. Invasive Group B Streptococcal Disease in Nonpregnant Adults. European Journal of Clinical Microbiology & Infectious Diseases 20, 837–839. https://doi.org/10.1007/s100960100612

Bowater, R.O., Forbes-Faulkner, J., Anderson, I.G., Condon, K., Robinson, B., Kong, F., Gilbert, G.L., Reynolds, A., Hyland, S., McPherson, G., Brien, J.O., Blyde, D., 2012. Natural outbreak of Streptococcus agalactiae (GBS) infection in wild giant Queensland grouper, Epinephelus lanceolatus (Bloch), and other wild fish in northern Queensland, Australia. Journal of Fish Diseases 35, 173–186. https://doi.org/10.1111/j.1365-2761.2011.01332.x

Carstensen, H., Henrichsen, J., Jepsen, O.B., 1985. A National Survey of Severe Group B Streptococcal Infections in Neonates and Young Infants in Denmark, 1978–83. Acta Paediatrica 74, 934–941. https://doi.org/10.1111/j.1651-2227.1985.tb10060.x

Centers for Disease Control and Prevention A, 1998. Adoption of Hospital Policies for Prevention of Perinatal Group B Streptococcal Disease—United States, 1997. Jama 280, 958. https://doi.org/10.1001/jama.280.11.958

Centers for Disease Control and Prevention B, 2002. Revision of Guidelines for the Prevention of Perinatal Group B Streptococcal Disease. Jama 287, 1106. https://doi.org/10.1001/jama.287.9.1106-jwr0306-2-1

Chesson, H., 2017. Sexually Transmitted Infections: Impact and Cost-Effectiveness of Prevention, in: Disease Control Priorities, Third Edition (Volume 6): Major Infectious Diseases. World Bank Publications.

Choi, Y.J., Kim, E.J., Piao, Z., Yun, Y.C., Shin, Y.C., 2004. Purification and Characterization of Chitosanase from Bacillus sp. Strain KCTC 0377BP and Its Application for the Production of Chitosan Oligosaccharides. Applied and

Environmental Microbiology 70, 4522–4531. https://doi.org/10.1128/aem.70.8.4522-4531.2004

Delannoy, C.M., Crumlish, M., Fontaine, M.C., Pollock, J., Foster, G., Dagleish, M.P., Turnbull, J.F., Zadoks, R.N., 2013. Human Streptococcus agalactiae strains in aquatic mammals and fish. BMC Microbiology 13. https://doi.org/10.1186/1471-2180-13-41

Garcia, P.J., Miranda, A.E., Gupta, S., Garland, S.M., Escobar, M.E., Fortenberry, J.D., 2021. The role of sexually transmitted infections (STI) prevention and control programs in reducing gender, sexual and STI-related stigma. EClinicalMedicine 33, 100764. https://doi.org/10.1016/j.eclinm.2021.100764

Keirstead, N.D., Brake, J.W., Griffin, M.J., Halliday-Simmonds, I., Thrall, M.A., Soto, E., 2013. Fatal Septicemia Caused by the Zoonotic Bacterium Streptococcus iniae During an Outbreak in Caribbean Reef Fish. Veterinary Pathology 51, 1035–1041. https://doi.org/10.1177/0300985813505876

Koplan, J., Jaffe, H., Spencer, J., Devine, O., Flock, M., Berman, S., Weinstock, H., Centers for Disease Control and Prevention, Department of Health and Human Services, 2001. Sexually Transmitted Disease Surveillance 2000. Division of STD Prevention.

Picchiassi, E., Coata, G., Babucci, G., Giardina, I., Summa, V., Tarquini, F., Centra, M., Bini, V., Cappuccini, B., Di Renzo, G.C., 2019. Intrapartum Test for Detection of Group B Streptococcus Colonization During Labor. Obstetrical & Gynecological Survey 74, 269–270. https://doi.org/10.1097/01.ogx.0000557707.43337.84

Qian, J., Ferguson, T.M., Shinde, D.N., Ramírez-Borrero, A.J., Hintze, A., Adami, C., Niemz, A., 2012. Sequence dependence of isothermal DNA amplification via EXPAR. Nucleic Acids Research 40, e87–e87. https://doi.org/10.1093/nar/gks230

Saari, A., Virta, L.J., Sankilampi, U., Dunkel, L., Saxen, H., 2015. Antibiotic Exposure in Infancy and Risk of Being Overweight in the First 24 Months of Life. Pediatrics 135, 617–626. https://doi.org/10.1542/peds.2014-3407

Schrag, S.J., Zywicki, S., Farley, M.M., Reingold, A.L., Harrison, L.H., Lefkowitz, L.B., Hadler, J.L., Danila, R., Cieslak, P.R., Schuchat, A., 2000. Group B Streptococcal Disease in the Era of Intrapartum Antibiotic Prophylaxis. New England Journal of Medicine 342, 15–20. https://doi.org/10.1056/nejm200001063420103

Seedat, F., Geppert, J., Stinton, C., Patterson, J., Freeman, K., Johnson, S.A., Fraser, H., Brown, C.S., Uthman, O.A., Tan, B., Robinson, E.R., McCarthy, N.D., Clarke, A., Marshall, J., Visintin, C., Mackie, A., Taylor-Phillips, S., 2019. Universal antenatal screening for group B streptococcus may cause more harm than good. BMJ l463. https://doi.org/10.1136/bmj.l463

Sizar, O., Leslie, S.W., Unakal, C.G., 2023. Gram-Positive Bacteria [WWW Document]. NCBI Bookshelf. URL https://www.ncbi.nlm.nih.gov/books/NBK470553/ (accessed 7.29.23).

Starrs, A.M., Ezeh, A.C., Barker, G., Basu, A., Bertrand, J.T., Blum, R., Coll-Seck, A.M., Grover, A., Laski, L., Roa, M., Sathar, Z.A., Say, L., Serour, G.I., Singh, S., Stenberg, K., Temmerman, M., Biddlecom, A., Popinchalk, A., Summers, C., Ashford, L.S., 2018. Accelerate progress—sexual and reproductive health and rights for all: report of the Guttmacher– Lancet Commission. The Lancet 391, 2642–2692. https://doi.org/10.1016/s0140-6736(18)30293-9

Van Ness, J., Van Ness, L.K., Galas, D.J., 2003. Isothermal reactions for the amplification of oligonucleotides. Proceedings of the National Academy of Sciences 100, 4504–4509. https://doi.org/10.1073/pnas.0730811100

WHO, 2012. Global incidence and prevalence of selected curable sexually transmitted infections: 2008. Reproductive Health Matters 20, 207–208. https://doi.org/10.1016/s0968-8080(12)40660-7

Yildirim‐Aksoy, M., Beck, B.H., Zhang, D., 2018. Examining the interplay betweenStreptococcus agalactiae,the biopolymer chitin and its derivative. MicrobiologyOpen 8, e00733. https://doi.org/10.1002/mbo3.733

Zimmermann, P., Curtis, N., 2019. Effect of intrapartum antibiotics on the intestinal microbiota of infants: a systematic review. Archives of Disease in Childhood - Fetal and Neonatal Edition 105, 201–208. https://doi.org/10.1136/archdischild-2018-316659

8. Appendices

Link to accommodating google drive, containing R Markdown file and Tutorial Video: https://drive.google.com/drive/folders/1q7xFovretmgohY_4QF0dvWWirhs5A8zf?usp=drive_link