**PROJECT 1 Proposal**

**Group: nkob**

**Members: Heather Baldwin, Ida Holaskova, Denzel Middleton**

**Date Sept 10, 2019**

**Whole Genome Sequencing of an Antibiotic Resistant Strain of *Neisseria gonorrhoeae***

**BACKGROUND**

In 2012, 78 million incidences of gonorrhea worldwide were reported to the World Health Organization (WHO) (*1*). Gonorrhea is a sexually transmitted disease (STD) transmitted by *Neisseria gonorrhoeae*. Widespread and improper use of antibiotics have advanced antibiotic resistance of *N. gonorrhoeae* to the extent of both multidrug-resistant and extensively drug-resistant strains being identified (*2*). *Neisseria gonorrhoeae* colonizes in warm, moist regions of the female reproductive tract as well as the urethra in both males and females. The spread of infection to the fallopian tubes and/or the uterus can lead to the pathogenesis of pelvic inflammatory disease (PID). Both infertility and ectopic pregnancy rates increase in patients with a history of PID.

*Neisseria gonorrhoeae* is constantly changing and new strains are evolving which has led to multidrug antibiotic resistance (*3*). Over the past decade, whole genome sequencing (WGS) has been utilized to examine the pathogenicity and antibiotic-resistant mechanisms of *N. gonorrhoeae* as WGS offers improved accuracy and finer resolution to previously utilized methods including multi-antigen sequence typing (*4*, *5*). In addition, WGS can aid in the identification of the presence of multiple infections, sample contaminants, and discrepant isolates. Once the WGS is obtained, data can be screened in the future for the presence of known antibiotic resistance gene(s). Of great significance, obtaining the WGS of various *N. gonorrhoeae* isolates is critical to the development effective treatment methods. The collected specimen, after verifying antibiotics resistance will be sequenced and compared to known resistant strains.

The specific aim of the research is to assemble the WGS of the GCWGS\_5192 (SRA number: SRX6631047) antibiotic resistant strain of *N. gonorrhoeae*.

**METHODS**

**Ethics Statement**

Prior to study initiation, this investigation will be reviewed by the institutional reviewboard at the West Virginia University.

***N. gonorrhoeae* Isolates Sampling**

Isolates of*N. gonorrhoeae* will be collected and cryopreserved according to protocol (*5*) in addition to clinical information about the antimicrobial susceptibility testing, epidemiological data (sex, age, sexual orientation), the anatomical site sample was collected from, previous gonorrhea diagnosis, other sexually transmitted infections diagnosed during the current gonorrhea infection, place of residence (zip code), type of clinic that the patient visited, HIV status, and probable state or country of infection. Bacterial isolates obtained from clinical specimens will be identified as *N. gonorrhoeae* by matrix-assisted laser desorption ionization - time of flight mass spectrometry (MALDI-TOF MS) using the BioTyper (Bruker Corporation) and frozen at 80°C in reconstituted powdered milk. Prior to antimicrobial susceptibility testing (AST) and sequencing, isolates will be thawed and subcultured twice on chocolate agar (Hardy Diagnostics) incubated at 35°C2°C with 5% CO2, and the organism identification will be confirmed with Vitek MS MALDI-TOF MS (bioMérieux).

**Cell culture and Antimicrobial Resistance Testing**

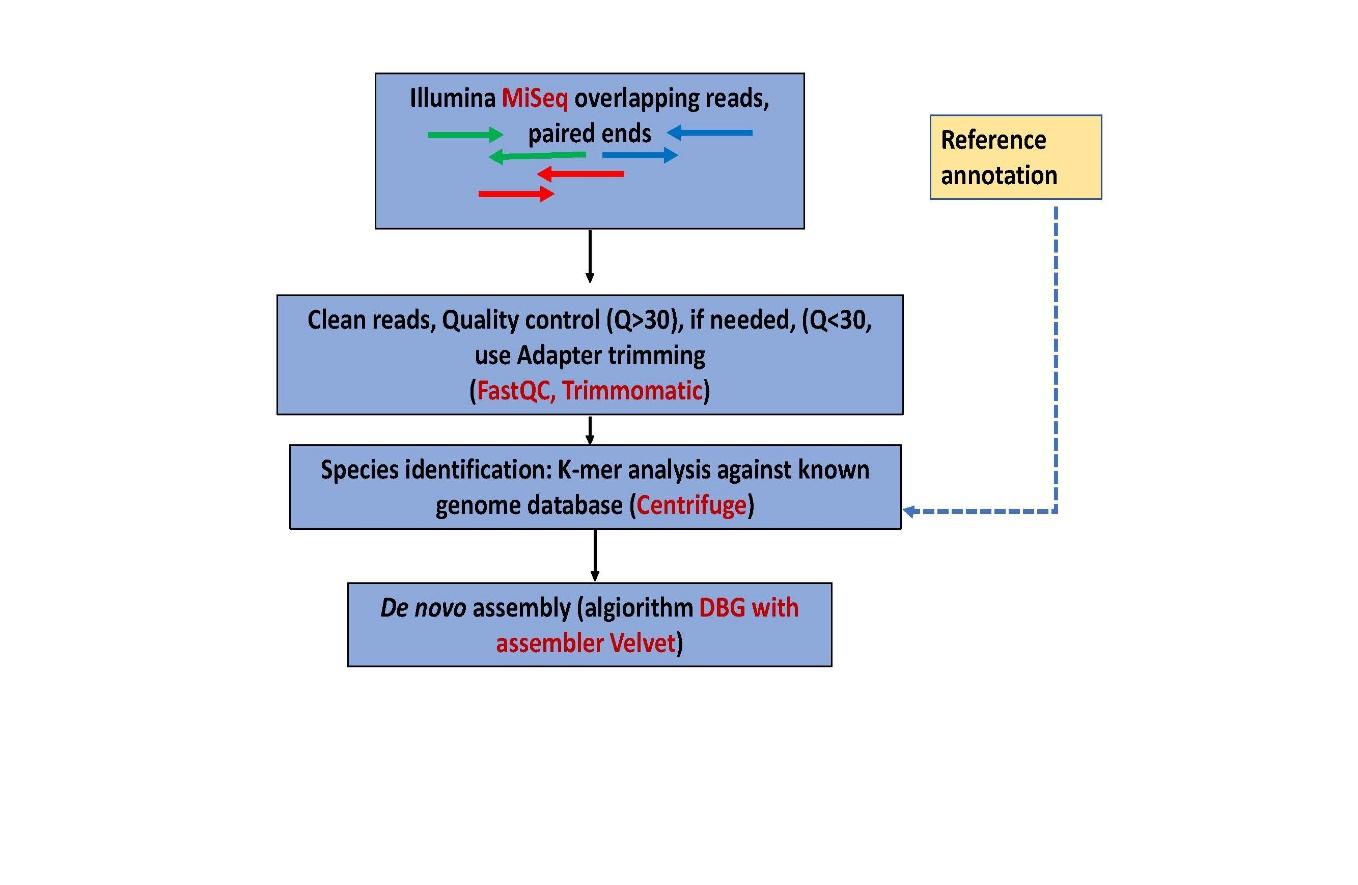
The gonococcal isolates will be cultured (*5*) and antimicrobial susceptibility testing will be done to determine the minimum inhibitory concentration (MIC) with Etests (bioMérieux; Marcy-l'Étoile, France) for cefixime and ceftriaxone, and an agar dilution breakpoint method or Etests for ciprofloxacin, azithromycin, and spectinomycin (*5*).

**WGS Procedures and Analyses**

The DNA will be extracted from a single isolate with a Wizard Genomic DNA Purification kit (Promega; Madison, WI, USA). Genomic DNA obtained from pure *N. gonorrhoeae* culture (0.5 ng) will be used to create sequencing libraries with the Nextera kit (Illumina, San Diego, CA, USA) (*6*) and sequenced on the Illumina MiSeq platform (SanDiego, CA, USA) at the Genomic Core at West Virginia University with collaboration with Genomics Core Facility at Marshall University (MU) in WV. Next generation sequencing (NGS) is the cutting-edge technology and provides more accuracy and speed than previously known Sanger sequencing (*7*). Paired-end sequencing will be utilized to over single-end sequences due to higher quality and accuracy-alignable sequence data (*8*).

The workflow of WGS raw data processing will be carried out via High Performance Computing center (Spruce Knob). The experimental software pipeline is demonstrated in Figure 1 below, modified from https://github.com/tseemann/nullarbor.Specifically, FastQC will be used for quality check (*9*) of the raw WGS data. If the reads have Q score below 30, the Trimmomatic will be used because it is specifically designed for paired-end data in addition to being flexible and efficient (*10*). Even though the purity of the culture (*N. gonorrhoeae*) should be guaranteed using MALDI-TOF MS; as a second precaution to exact species identification, the Centrifuge will be utilized in the pipeline for identification of other possible - contaminant prokaryotic species. Centrifuge requires a relatively small index and classifies sequences at a very high speed allowing it to process millions of reads from a typical high-throughput DNA sequencing run within minutes (*11*). Samples will have less than 70% identity to *N. gonorrhoeae* will be unacceptable and will be excluded. Finally, de Bruijn graph (DBG) assembler algorithm, based on the k-mer approach, will be employed for it is faster and better suited for short reads (*12*). Specifically, Velvet, one of the fastest suitable assembler for paired-end data (*12*, *13*) will be used for the genome assembly. Assemblathon script on contig files will be used to calculate N50 contig length for data reporting *(12).* The assembled genome will be compared to the known reference genome available on NCBI.

[The sample analyzed will be referred as GCWGS\_5192 (SRA number [**SRX6631047**](https://www.ncbi.nlm.nih.gov/sra/SRX6631047%5Baccn%5D)). The coverage is 107% (2,295,960 pb /2,153,922 bp) compared to reference genome *N. gonorrhoeae* (NCBI Accession NC\_002946).]



**Figure 1.** An overview of the genome assembly protocol.

**Data availability**

All genomes used in this study will be deposited into the NCBI Whole Genome Database.

**Grants**

WV-INBRE grant (P20GM103434), the COBRE ACCORD grant (P20GM121299) and the West Virginia Clinical and Translational Science Institute (WV-CTSI) grant (U54GM104942).

**Facilities**

The MU Genomics Core Facility provides Next Generation Sequencing (NGS) services to investigators at universities, and in government and industry. The Genomics Core is equipped with an Illumina MiSeq sequencer with the capability to perform whole genome sequencing.

**Division of Labor**

HB researched and wrote the background and checked the quality of the dataset, IH and HB searched the methods, IH wrote the methods and the workflow and DM predicted and wrote the expected outcomes and contributed to experimental design, pipeline components identification and selection. HB will write code for quality control using FastQC and Trimmomatic, if need be. IH will write code for Centrifuge. DM will code for the assembling of the genome using DBG and Velvet. HB and IH will contribute to the interpretation of the results. DM, IH, and HB will contribute equally to the writing of the manuscript. If conflict arises among investigators, including lack of contribution, the issue will be discussed with the course instructor.

**EXPECTED OUTCOME**

An assembled whole genome *N. gonorrhea* sequence is expected at the completion of experimentation. This new knowledge will aid in identifying novel genes involved in antibiotic resistance phenotypes of *N. gonorrhoeae* being reported. Alternatives to the pipeline outlined above would be to use the ABySS assembler instead of Velvet to assemble the *N. gonorrhoeae* genome due to mishaps in k-mer development or inaccurate matches in contig formation. The challenges that may arise include, data with low quality, k-mer development, repeats, and sample contamination.

In the event of low-quality data, NCBI has 2130 matches to *N. gonorrhoeae* isolates in which the dataset can be changed to use an alternate raw data set to carry out the experiments. K-mer size is important in the assembling process; a k-mer too small will allow repeats to occur. Repeats will ruin the assembled genome. Contamination of host, other bacteria, or viral components can affect the assembling process of a whole genome sequence of *N. gonorrhoeae*.

**REFERENCES**

1. L. Newman, J. Rowley, S. V. Hoorn, N. S. Wijesooriya, M. Unemo, N. Low, G. Stevens, S. Gottlieb, J. Kiarie, M. Temmerman, Global Estimates of the Prevalence and Incidence of Four Curable Sexually Transmitted Infections in 2012 Based on Systematic Review and Global Reporting. *PLOS ONE*. **10**, e0143304 (2015).

2. J. W. Tapsall, F. Ndowa, D. A. Lewis, M. Unemo, Meeting the public health challenge of multidrug- and extensively drug-resistant Neisseria gonorrhoeae. *Expert Review of Anti-Infective Therapy; London*. **7**, 821–34 (2009).

3. S. R. Harris, M. J. Cole, G. Spiteri, L. Sánchez-Busó, D. Golparian, S. Jacobsson, R. Goater, K. Abudahab, C. A. Yeats, B. Bercot, M. J. Borrego, B. Crowley, P. Stefanelli, F. Tripodo, R. Abad, D. M. Aanensen, M. Unemo, Public health surveillance of multidrug-resistant clones of Neisseria gonorrhoeae in Europe: a genomic survey. *Lancet Infect Dis*. **18**, 758–768 (2018).

4. R. S. Lee, T. Seemann, H. Heffernan, J. C. Kwong, A. Gonçalves da Silva, G. P. Carter, R. Woodhouse, K. H. Dyet, D. M. Bulach, T. P. Stinear, B. P. Howden, D. A. Williamson, Genomic epidemiology and antimicrobial resistance of Neisseria gonorrhoeae in New Zealand. *Journal of Antimicrobial Chemotherapy*. **73**, 353–364 (2018).

5. A. L. Bailey, R. F. Potter, M. A. Wallace, C. Johnson, G. Dantas, C.-A. D. Burnham, Genotypic and Phenotypic Characterization of Antimicrobial Resistance in Neisseria gonorrhoeae: a Cross-Sectional Study of Isolates Recovered from Routine Urine Cultures in a High-Incidence Setting. *mSphere*. **4** (2019), doi:10.1128/mSphere.00373-19.

6. M. Baym, S. Kryazhimskiy, T. D. Lieberman, H. Chung, M. M. Desai, R. Kishony, Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes. *PLOS ONE*, 15 (2015).

7. H. P. J. Buermans, J. T. den Dunnen, Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. **1842**, 1932–1941 (2014).

8. T. Nakazato, T. Ohta, H. Bono, Experimental Design-Based Functional Mining and Characterization of High-Throughput Sequencing Data in the Sequence Read Archive. *PLoS ONE*. **8**, e77910 (2013).

9. J. Brown, M. Pirrung, L. A. McCue, FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics*. **33**, 3137–3139 (2017).

10. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 2114–2120 (2014).

11. D. Kim, L. Song, F. P. Breitwieser, S. L. Salzberg, Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* (2016), doi:10.1101/gr.210641.116.

12. A. R. Khan, M. T. Pervez, M. E. Babar, N. Naveed, M. Shoaib, A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective. *Evol Bioinform Online*. **14**, 117693431875865 (2018).

13. D. R. Zerbino, *Curr Protoc Bioinformatics*, in press, doi:10.1002/0471250953.bi1105s31.