# Gene Discovery and Annotation of *Gardnerella vaginalis*, a Bacterium Associated with Bacterial Vaginosis and Pre-term Labor

by

Marina Ibrahim

A thesis presented to the Honors College of Middle Tennessee State University in partial fulfillment of the requirements for graduation from the University Honors College

Spring 2019

Gene Discovery and Annotation of *Gardnerella vaginalis*,

a Bacterium Associated with Bacterial Vaginosis and Pre-term Labor

Marina Ibrahim

APPROVED:

_____

Dr. Rebecca Seipelt-Thiemann
Biology Department

_____

Dr. Lynn Boyd
Biology Department

_____

Dr. Mary Farone
Biology Department

_____

Dr. John R. Vile
University Honors College Dean

## ACKNOWLEDGEMENTS

I would first like to thank God, because it is through Him that all good things come, and if it wasn't for Him, I would not have been able to make progress with my project and have the people I had in my life support me throughout the entire process. I would also like to express my deepest gratitude to Dr. Seipelt-Thiemann, who always showed the greatest amount of care and attention to my project. If it wasn't for her patience and great knowledge, no progress would have been made with this project. I would like to thank Dr. Mary Farone as well for the advice and information she brought in order to further enhance my thesis. Lastly, I would like to thank my family's love, support, and patience as they listened to me talking non-stop about something they are not too interested in; I would not be where I am today if it wasn't for them.

## ABSTRACT

Human pregnancy lasts approximately 40 weeks, but genetic and environmental factors could contribute to premature birth. One factor is when other bacteria replace *Lactobacillus* spp. within the female urogenital tract. One bacterium found in patients diagnosed with bacterial vaginosis is *Gardnerella vaginalis*, which is complex in nature as some strains are pathogenic and others are not. To understand strain differences, a reference genome annotation was compared to transcriptome evidence within the Apollo Gene Annotation Platform. With approximately 30% of the annotations done, thirty-four gene models were re-annotated, while only two gene models did not require further annotation. Two polycistronic operons were identified, but no novel, putative genes were identified. Using this updated genome annotation will allow for a more informed comparison of genes and transcripts from pathogenic and non-pathogenic strains and may also allow identification of virulence factors that can be used as biomarkers and therapeutic targets.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF APPENDICES

# INTRODUCTION

Human pregnancy is considered a sensitive period of a woman's life as numerous genetic and environmental factors including genetic predisposition to premature birth, exposure to infectious agents, ingestion of potentially harmful chemical compounds, and changes in diet or behavior can have a tremendous impact on the mother and fetus. The process of gestation involves the development of the fetus through the physiological changes that the mother goes through (Soma-Pillay et al., 2016). Pregnancy lasts approximately 40 weeks with most single births occurring between 37 and 40 weeks of gestation. Preterm birth, that is birth at or before 37 weeks-gestation, occurs in 5-18% of total births (Romero et al., 2015) and 6-11% of live births (Zeitlin et al., 2013). The earlier the baby arrives, the greater health risks he or she has; some of these risks may be short term, such as staying in neonatal intensive care as they may be showing signs of infection and experiencing respiratory complications (Lawn et al., 2013). Other more serious risks include death, chronic respiratory illness, neurodevelopmental disorders, and long-term cognitive impairment (Gomez-Lopez et al., 2014). Not only is the baby affected by being born prematurely, but the mother also faces some physical and psychological complications, such as post-traumatic stress disorder symptoms of avoidance, hyperarousal, and intrusion (Ionio et al., 2016). Women who experienced preterm labor reported having higher than normal fatigue, anxiety, and flashbacks of their experience; in addition, these mothers had little to no early contact with their newborn, more postnatal health problems, and a significantly low positive feeling towards their baby (Henederson et al., 2016).

Preterm labor can result from a variety of different events, some of which include genetic factors, as well as maternal bacterial and viral infections. Preterm labor has been shown to occur in families and its occurrence rate differs by ethnicity, suggesting a genetic susceptibility component (Srinivas and Macones, 2005). Preterm premature rupture of membranes (PPROM), which occurs in at least one third of spontaneous preterm births (ACOG Practice Bulletin), has been linked to specific alleles of matrix metalloproteinase-1 (MMP1; Fujimoto et al., 2002), matrix metalloproteinase-8 (MMP8; Wang et al., 2008), or serpin peptidase inhibitor, clade H (SERPINH1; Wang et al., 2006). All three proteins are known to degrade extracellular matrix, so overproduction or enhanced activity has been suggested to be the mechanism causing membrane rupture (Maymon et al., 2000).

In addition to genetic susceptibility, pre-term labor and birth can also have environmental causes such as infection. It is important to note that the vaginal tract, similar to the digestive tract, is normally populated with different types of bacteria, known as the vaginal genomes of the microbiota. These bacteria form a mutualistic relationship with the female body. Bacteria gain nutrients from the environment, sloughed cells, and vaginal secretions. These bacteria continually re-colonize the area which keeps pathogenic bacteria from growing or over-growing into the same area. Furthermore, these bacteria not only do not harm the host, but actually produce compounds that help maintain the normal vaginal pH and vaginal health (Witkin et al., 2013). Normal pH ranges anywhere from 3.8-4.5 as it depends on the female's life stage. Most of the bacteria in the normal vaginal microbiome are of the *Lactobacillus* genus, specifically, *Lactobacillus crispatus, L. gasseri, L. iners, and L. jensenii* (Mendes-Soares

et al., 2014). Many physiological changes occur during pregnancy, including suppressive changes to the women's immune system in order to allow the growing fetus to be supported rather than seen as foreign by the mother's immune system. These changes can alter the vaginal microbiota and even allow shifts in microbial populations that can then lead to infection by pathogenic bacteria and pre-term labor (Silasi et al., 2015).

Bacterial vaginosis has been hypothesized to be one cause of or contributing factor to premature birth (Lata et al., 2010) and occurs when other bacteria replace *Lactobacillus* spp. within the female urogenital tract. It is present in ten to forty-one percent of women and has been linked to maternal and fetal low-health status; in addition, spontaneous abortion, amniotic fluid infection, endometritis, and post-cesarean wound infection occur more often due to bacterial vaginosis during pregnancy (McGreggor and French, 2000). Bacterial vaginosis is considered to be sexually transmitted and is diagnosed by the presence of indicator bacteria in addition to low vaginal acidity, or high vaginal pH, that results from infection (Leppaluoto, 2011). Bacterial vaginosis can be diagnosed using two methods, the Amsel criteria or the Nugent Gram stain scoring method. Three of the four criteria must be met in the Amsel criteria method for a positive diagnosis of bacterial vaginosis. Those criteria include a thin homogenous discharge, a pH greater than 4.5, an amine odor, and shed epithelial cells covered with bacteria, also called clue cells. The Nugent Gram stain scoring method sets out bacterial morphotypes that are associated with health, Gram-positive lactobacilli, and morphotypes associated with bacterial vaginosis, Gram-variable rods. If the Nugent score is between zero and three, it indicates a healthy microbiota; if the score is between seven and ten, it is a positive diagnosis for bacterial vaginosis (Aldunate et al., 2015). Females

who have a higher number of lifetime sexual partners, participated in their first intercourse at a young age, or practice regular douching are at a higher risk of obtaining bacterial vaginosis (Bautista et al., 2016).

While bacterial vaginosis has a clinical definition, the exact infectious agent or agents that cause it remains somewhat unclear. One bacterium that has been found in women diagnosed with bacterial vaginosis is *Gardnerella vaginalis* (Figure 1).

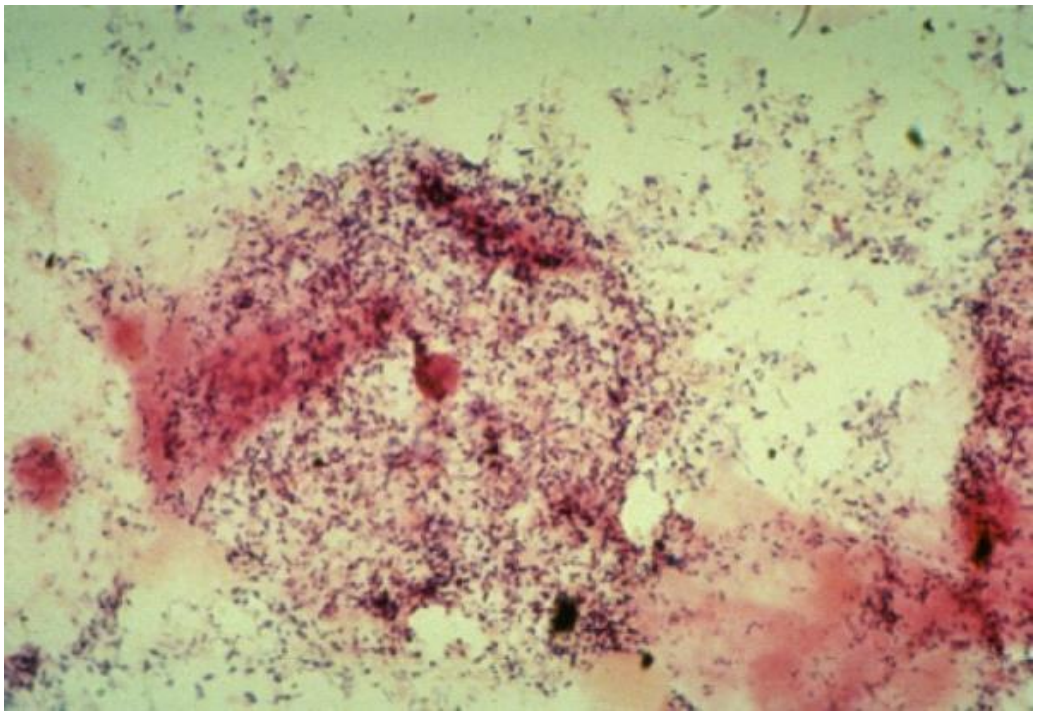

**Figure 1. *Gardnerella vaginalis* and Clue Cell from a Patient Sample.** Clue cell sample is provided by a physician in Murfreesboro, TN. This image is Gram stained and imaged at 1000X magnification on the Olympus BX60 microscope with an Olympus DIP72 camera and CellSens software.

*Gardnerella vaginalis* was formerly termed *Haemophilus vaginalis* and is a small, Gram-positive, rod-shaped bacterium. It was later classified as *Corynebacterium vaginale* before it obtained the name it has today. *G. vaginalis* is known for its ability to form a biofilm, which is a thin, slimy film of bacteria that sticks to a surface (Cornejo et al., 2018). One complicated issue regarding this bacterium is that some strains of *G. vaginalis* cause symptoms of bacterial vaginosis while other strains are found in normal and healthy women (Hickey and Forney, 2014). For this reason, it is important to determine how these strains differ.

To better differentiate between the pathogenic and non-pathogenic strains, the genetic and genomic differences between these strains must be examined. To make any meaningful comparisons among pathogenic and non-pathogenic strains, a complete reference genome that is fully annotated with experimentally verified genes must be generated. Currently, the *vaginalis* genomes of six strains have been fully sequenced (Yeoman et al., 2010; https://www.ncbi.nlm.nih.gov/genome/genomes/1967). Computational gene predictions using this sequence have also been performed using algorithms such as MAKER-P (Campbell et al., 2014). However, these computational predictions have not been validated using experimental RNA evidence. Therefore, this project aimed to use genomic and bioinformatics tools to more fully annotate a reference *G. vaginalis* genome using experimental RNA sequencing evidence. This analysis showed experimental evidence for many genes and allowed some operons to be identified. Furthermore, computational gene models were revised and refined using experimental evidence. This experimentally-verified genome annotation will allow a

better comparison between pathogenic and non-pathogenic strains to identify sources of virulence in pathogenic strains in the future.

## MATERIALS AND METHODS

### *Gardnerella vaginalis* Genome Data Resources

Genome sequence and resources for *Gardnerella vaginalis* (*Gv*) species were identified using the National Center for Biotechnology Information (NCBI). Sequences were available for 103 strains, but only six completed genomes were available. The completed genome that was chosen to act as the reference strain in this study was strain 14019 (NCBI ID: 171905 available at https://www.ncbi.nlm.nih.gov/genome/1967?genome_assembly_id=171905). This strain is most similar to the partially sequenced *Gv* strain studied at MTSU (strain 49145 NCBI ID: 297689 available at https://www.ncbi.nlm.nih.gov/genome/1967?genome_assembly_id=297689), and thus will be valuable in collaborative work. These two strains are 99.5% identical (https://www.ncbi.nlm.nih.gov/genome/1967?genome_assembly_id=297689). A computationally predicted genome annotation of this completed genome was also available (https://www.ncbi.nlm.nih.gov/assembly/GCF_000159155.2/). Furthermore, a next-generation RNA sequencing dataset utilizing this genome as the reference was also available at the following sites:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE80127

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL21713

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2113323

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2113322

**RNA Sequencing Data Alignment and Transcript Generation**

Paired RNA sequence fastq files were aligned to the reference genome using Bowtie2 version 2.2.6 (Langmeade et al., 2018) within the Cyverse Discovery Environment (Goff et al., 2011). The .sam files were converted to .bam files, which were then sorted and indexed using SamTools version 1.7 (Li et al., 2009). The output .bam files were then downloaded and used in the Apollo installation.

**Apollo Installation**

A web-based Apollo instance was set up with the kind help and perseverance of Mr. Tim Miller in the Computational Sciences Department following specifications for community-based annotation projects, as detailed in Lee et al. (2013).

**Apollo-Based Visualization of Operons and Revision to Gene Models**

Evaluation and revision of the computationally-predicted annotation for *G. vaginalis* was done using several data sources that were viewed and evaluated together using JBrowse Genome Viewer in combination with the web-based Apollo Gene Annotation Platform (Lee et al., 2013). The 14018 reference genome, 14018 reference genome annotation, aligned RNA sequencing reads (.bam output files), and the transcript index (.bai output files) were loaded into the platform and visualized. Within the Apollo Gene Annotation Platform, RNA sequencing reads were visually located along the reference genome sequence along with the computationally-predicted annotated genes. This allowed manual comparison of the annotated, computationally-predicted gene with experimentally-derived transcript information. First, some of the gene model required revision, that is the RNA evidence did not match the annotation exactly, while other gene models did not require revision because the RNA evidence did match the computational

annotation (see Annotation Revision). In addition, operons, which can have multiple genes encoded in a single RNA, were identified by the observation of contiguous genes with uniformly distributed RNA evidence that was not interrupted and also were transcribed in the same direction along the genome, as indicated by arrows in the genome annotation visualization.

**Annotation Revision**

Revision to a gene model involved moving the gene model to a new track in the Apollo workspace and zooming in to visualize RNA read, genome sequence information, and the gene model together. The gene model was then corrected based on the RNA sequencing information by adjusting the rectangular gene model "boxes" to be consistent in placement with the beginning and end of the transcriptional unit.

## RESULTS

*Gardnerella vaginalis* is a bacterium that is complex in nature as some of its strains are known to be disease-causing and other strains are known to be non-disease-causing and can potentially live harmlessly in a woman's urogenital tract (Hickey and Forney, 2014). Differences among these strains could be due to genetic differences in how these strains express their genes in response to particular stimuli in the environment. Therefore, to begin to understand how these strains truly differ it was important to establish which genes are present and expressed in a non-pathogenic strain to then compare to pathogenic strains.

The process of identifying and confirming which genes are present and expressed in an organism involves the genome annotation that generally begins with a sequenced genome and gene models that are predicted by computational algorithms based on known gene parameters. However, these predictions are often slightly to very inaccurate and require manual annotators to refine and correct gene models based on experimental evidence, such as known cDNA sequence, EST evidence, and RNA sequencing evidence. The goal of this project was to use the Apollo Gene Annotation Platform to annotate computationally predicted gene models based on experimental evidence. This included identifying polycistronic RNAs, verifying the 5' and 3' ends of RNAs, and identifying putative novel genes (Figure 2).

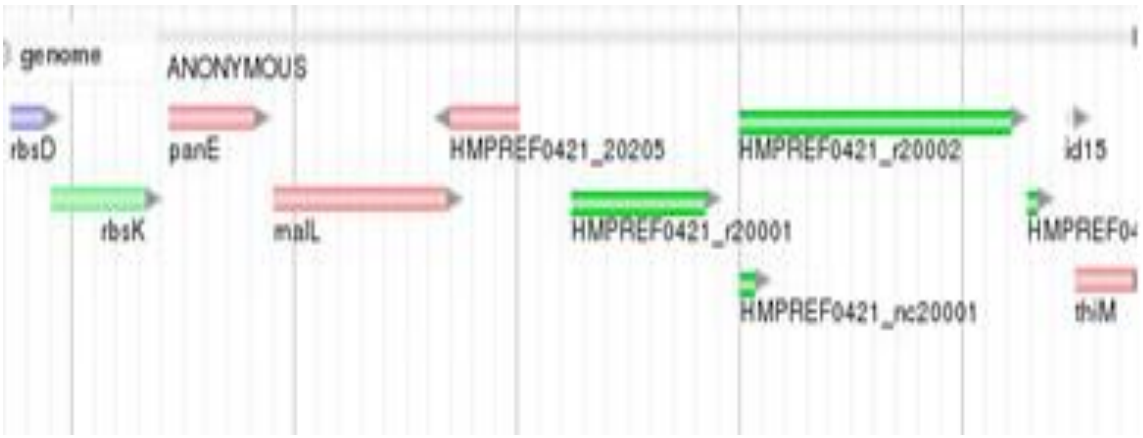


**Figure 2. Apollo Visualization Platform With Gene Models:** Arrows on the top panel indicate computationally predicted gene models. Each gene model is pointing in a certain direction, which indicates the direction of transcription. The tail of the arrow of the gene model denotes where the beginning is, and the arrow tip denotes the end of the gene. For example, *panE* is transcribed using the "top DNA strand" as the template, while *HMPREF0421_20205* is transcribed using the "bottom DNA strand" as the template.

Within this annotation platform, RNA evidence can be added to allow more accurate manual annotation. These are visualized as tiled rows located under the gene model (Figure 3). First, a strategy was developed to identify prokaryotic gene features expected in the RNA evidence, including 1) polycistronic RNAs where multiple genes are transcribed on a single RNA, 2) genes that were not transcribed under these specific conditions, 3) novel genes that were not computationally predicted, but have RNA evidence, 4) gene models that match the predicted annotation and therefore do not need revision, and 5) gene models whose 5’ and 3’ ends require revision based on RNA evidence. Representative examples of each follow and full lists can be found in the appendix.

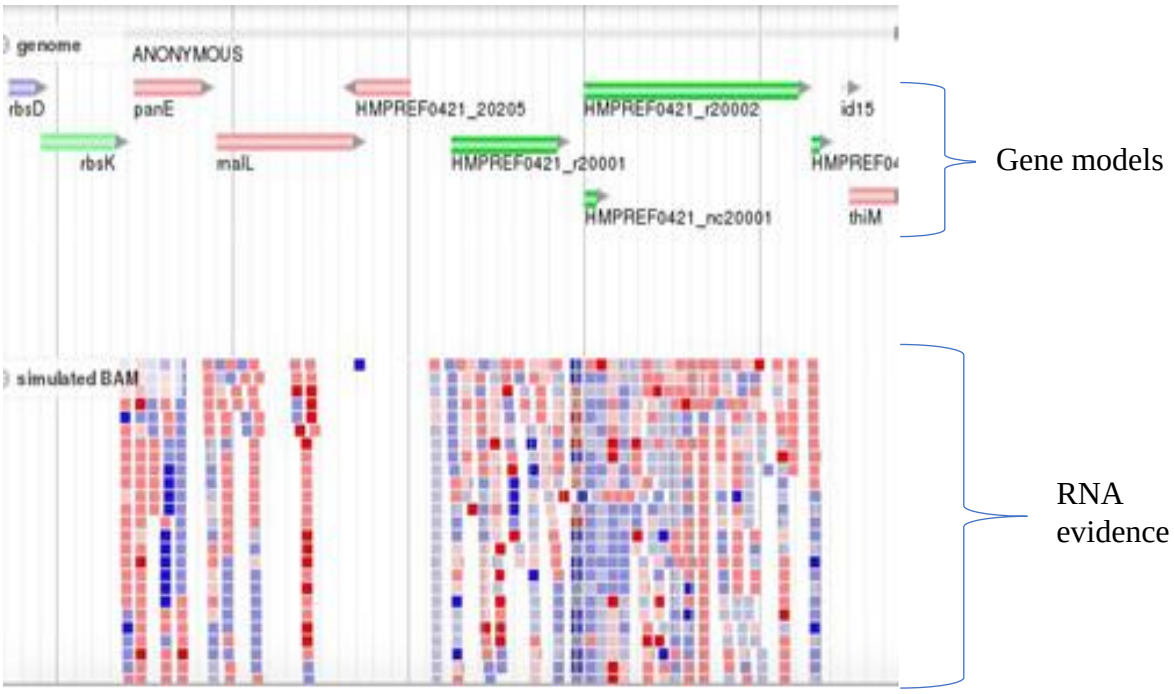


**Figure 3. Apollo Visualization Platform With Gene Model and RNA Evidence:** Arrows on the top panel indicate annotated genes with direction of transcription noted. The tiled rows in the lower panel represent RNA sequencing aligned to that region of the genome.

First, operons were identified by visual identification of RNA sequencing evidence that was contiguous along several gene models that had the same direction of transcription (Figure 4). This showed evidence that a simple polycistronic RNA was utilized by the organism to encode multiple genes or proteins. The given gene models are transcribed in the same direction and there appears to be RNA evidence under each one of the gene models. A full list of other polycistronic operons can be found in the appendix.

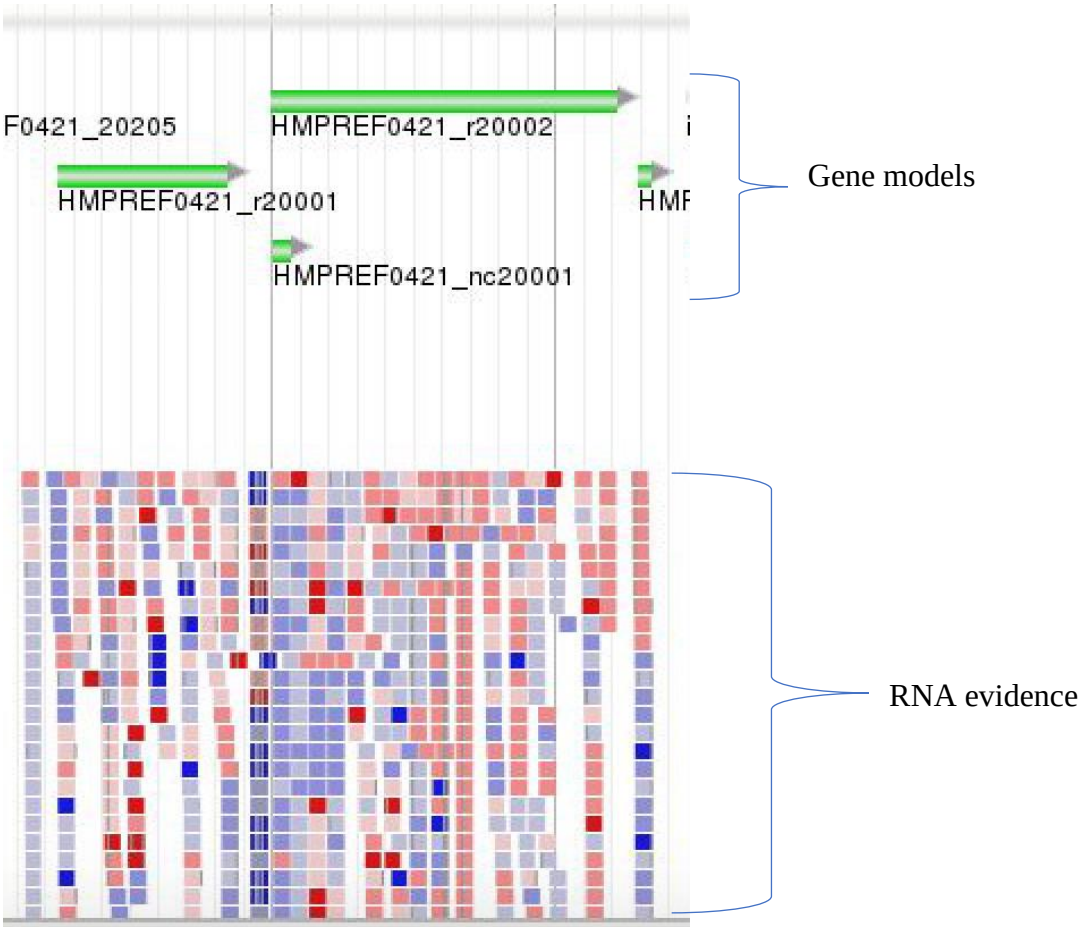


**Figure 4. Gene Model Example of Polycistronic Operon Annotation.** A polycistronic operon is shown where several genes transcribed in the same direction with non-interrupted RNA evidence (HMPREF0421_r20001, HMPREF 0421_nc20001, and HMPREF0421_r20002). Arrows on the top panel indicate annotated genes with direction of transcription noted. The tiled rows in the lower panel represent RNA reads.

Some gene models were present without RNA evidence accompanying them, which meant that the program could not read any evidence for that particular gene which could be due to issues with multiple factors. The most likely explanation is that the genes were not transcribed under conditions used in the RNA sequencing experiment (Figure 5).

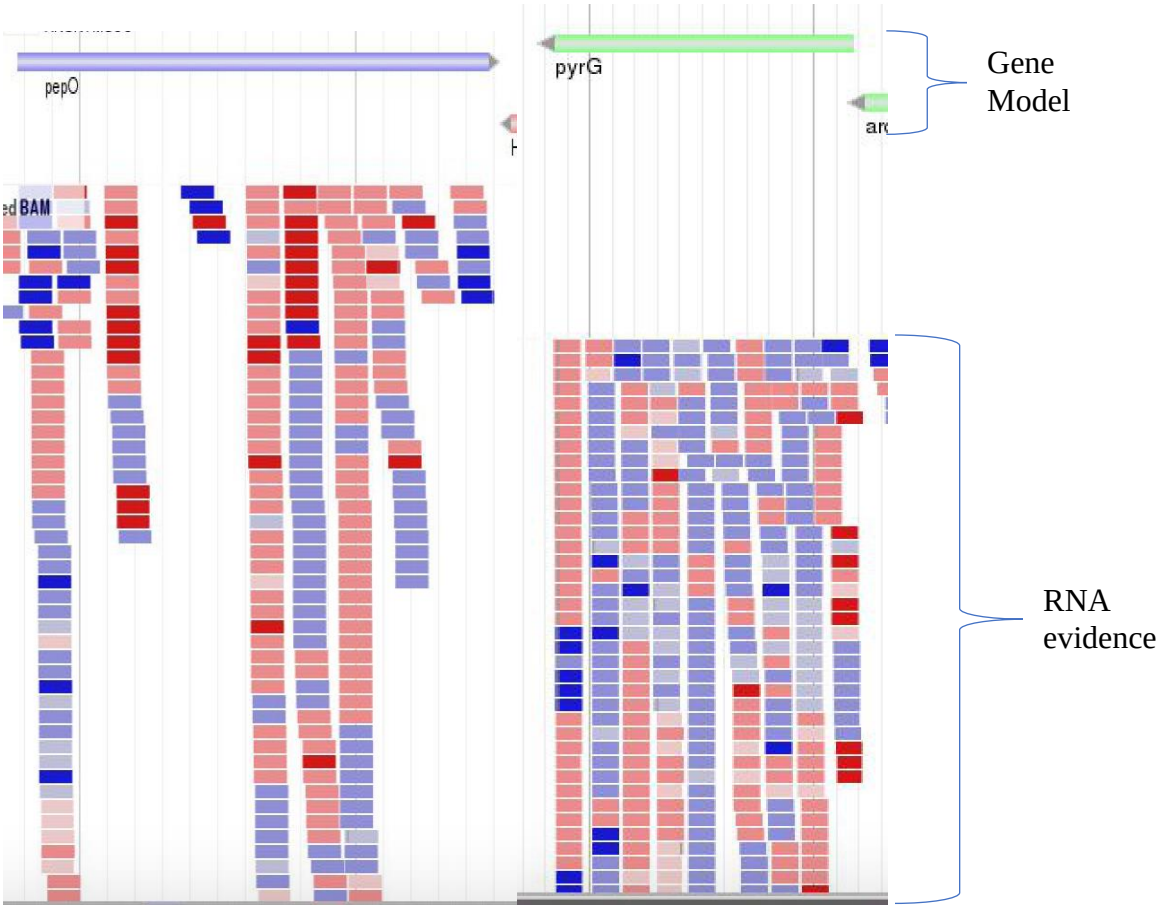


**Figure 5. Non-transcribed Genes:** Some regions do not show RNA reads across the entire gene (*pepO*) as would be expected (*pyrG*). These may be issues with annotation, the RNA sequence experiment, or alignment failures, but likely this is due to lack of transcription under these conditions.

Novel genes are found when RNA evidence exists, but the annotation platform does not match this RNA evidence to a gene model. New genes, or novel genes, did not exist within the Apollo Gene Annotation Platform.

With the genome organization completed, the next step was to identify and correct gene models based on RNA sequencing evidence, that is, to correct any apparent discrepancies in the 5' and 3' ends of transcripts. Some gene models were accurate, as shown by the representative gene *HMPREF0421_r20002* (Figure 6). This RNA evidence shows a correct 5' and 3' end for the close up view of *HMPREF0421_r20002* transcript (Figure 7). All gene models that were correct are listed in Appendix A. However, not all predicted gene models were supported by the RNA evidence. Gene *malL* represents a gene model that required revision based on experimental evidence (Figure 8). To correct, or annotate the computationally predicted gene model, the gene model had to be dragged to the annotation track found above the gene model and RNA evidence. Once the model was dragged to the track, the focus was then zoomed in all of the way to the point that you see the RNA sequence and the RNA evidence as long lines inside of the blocks that were previously viewed. The gene model then gets corrected by dragging or trimming the start and stop codon based on where the majority of the RNA evidence is found. For example, gene model *atpC* (Figure 9) needs to have its stop codon extended more towards to the left because there is more RNA evidence towards the left that the computationally predicted gene model is not including. Once that is corrected, the rest of the gene model could be examined to determine if there needs to be any more cutting or extending of the gene model.

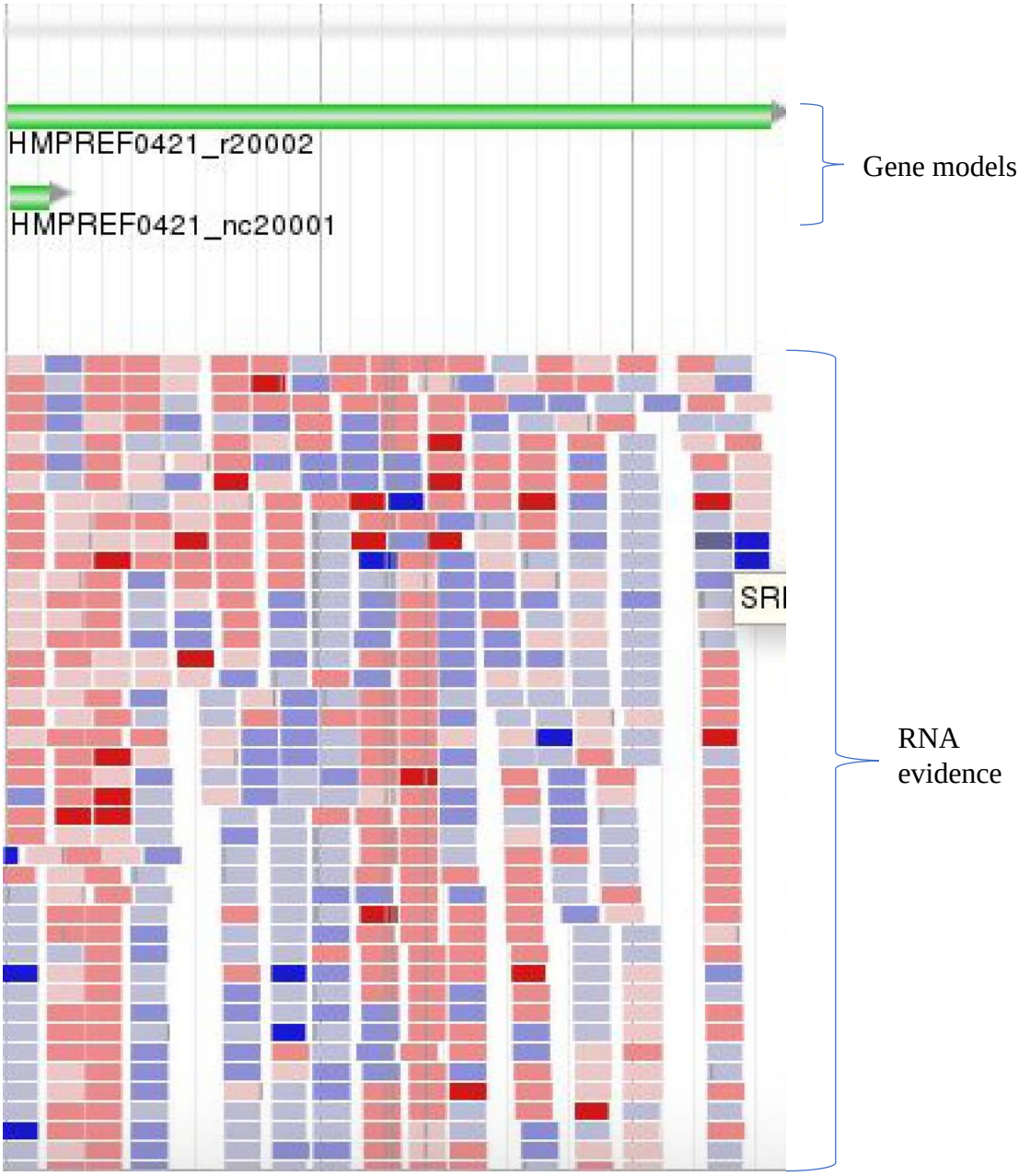


**Figure 6. Gene Model Example Requiring No Revision:** An example of an annotated gene that did not require revision (*HMPREF0421_r20002*). Arrows on the top panel indicate annotated genes with direction of transcription noted. The tiled rows in the lower panel represent RNA reads.

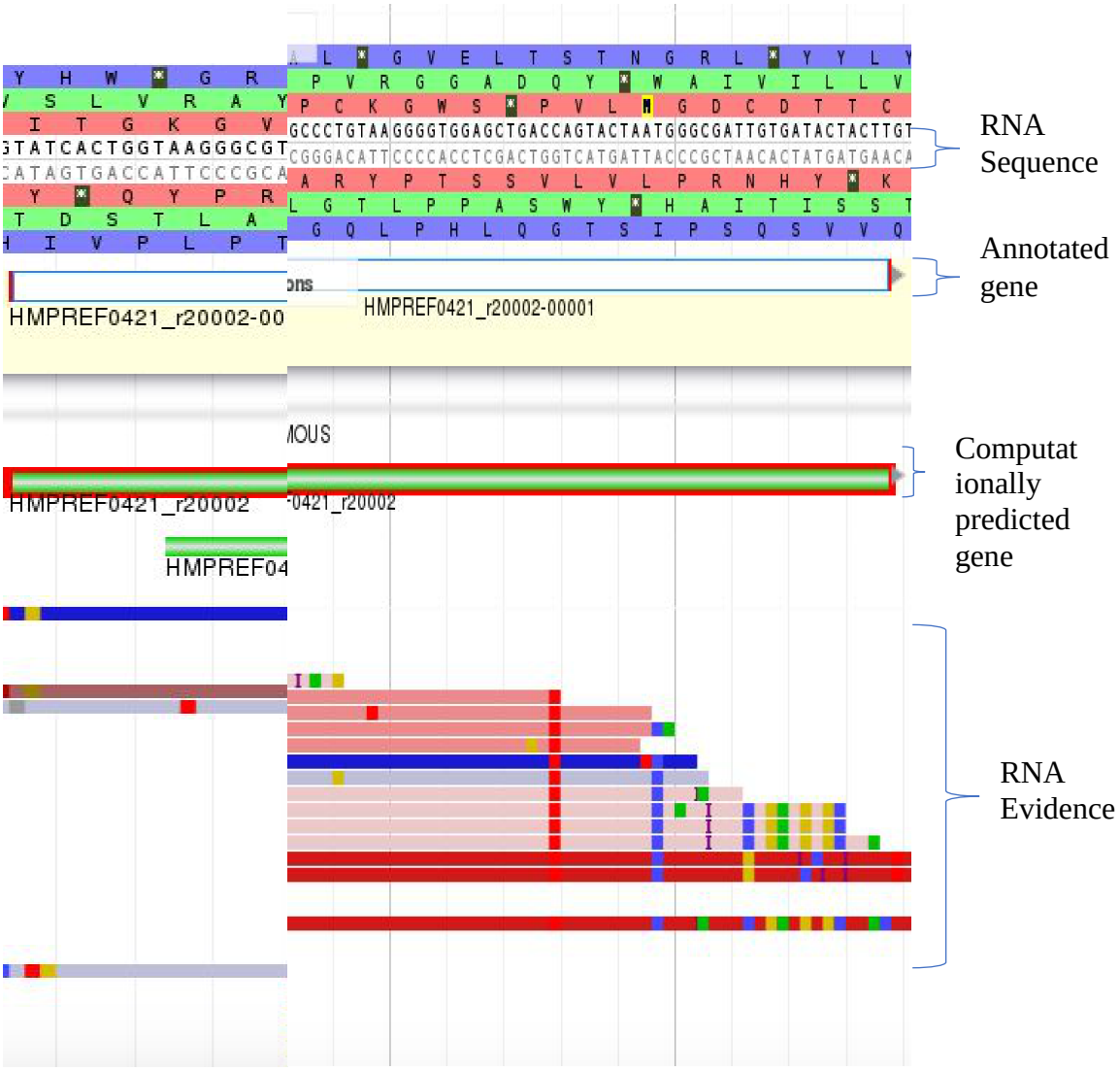


**Figure 7. Close Up View of Gene Model *HMPREF0421_r20002* Requiring No Revision:** Throughout the entire gene model, there was RNA evidence provided throughout the entire model. From the start codon (left) all through the stop codon (right), there was contiguous RNA evidence suggesting there is no revision required.

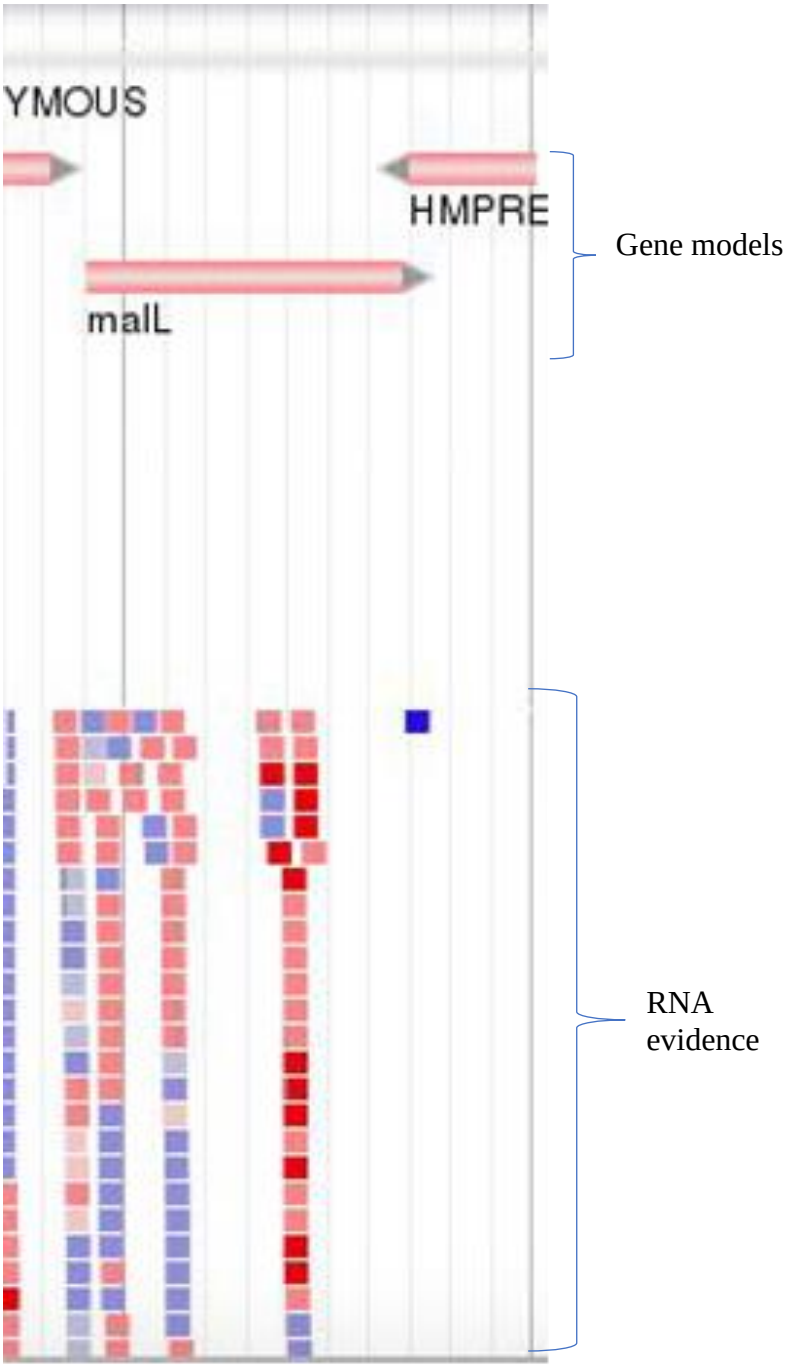


**Figure 8. Gene Model Example Requiring Revision.** An example of an annotated gene that did require revision (*malL*). Arrows on the top panel indicate annotated genes with direction of transcription noted. The tiled rows in the lower panel represent RNA reads.

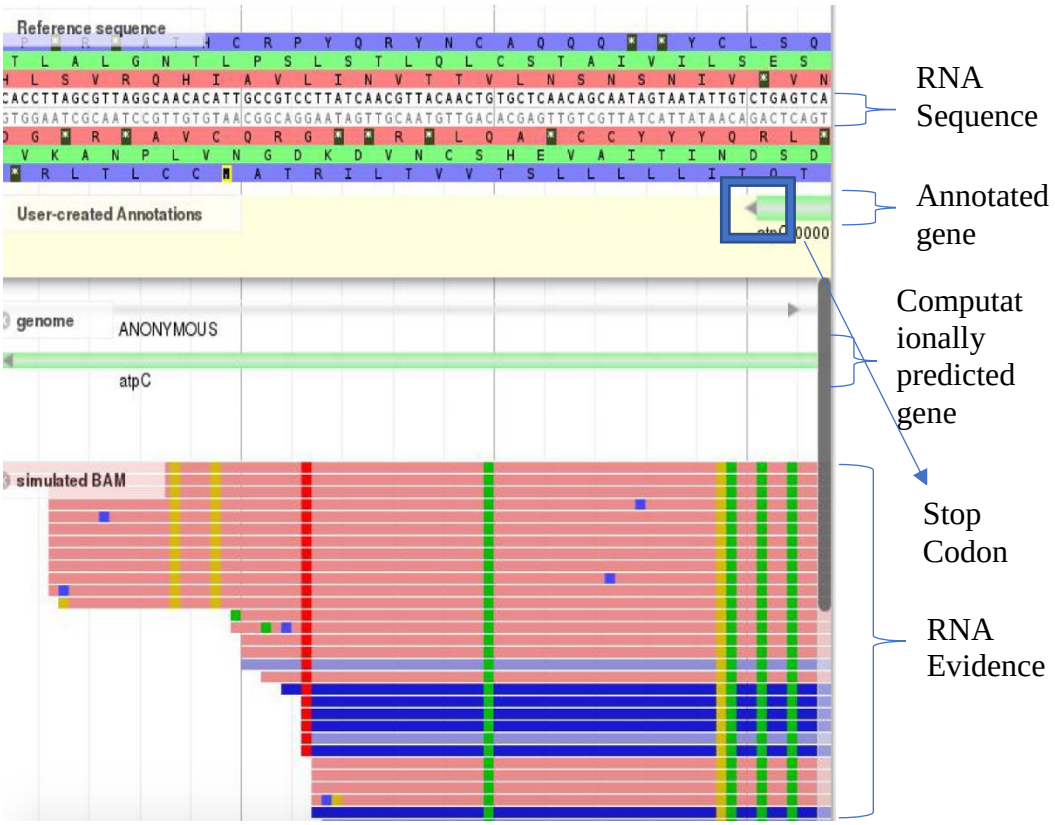


**Figure 9. Example Annotation for Correcting a Gene Model at the 5' End of a Gene:** The computationally predicted gene is labeled *atpC*, which shows a different 5' end than the actual data.

## CONCLUSION

*Gardnerella vaginalis* is a bacterium that can be categorized as two types, 1) pathogenic, which is disease-causing and 2) non-pathogenic, which is non-disease-causing (Hickey and Forney, 2014). Those differences may be due to genetic variability or to specific gene-environment interactions. To understand the nature of these strains and how they differ, it is first important to have an accurate and experimentally-confirmed genome annotation. The aim of this project was to use experimental evidence from an RNA sequencing experiment to confirm, revise, and identify genes and operon structures in a non-pathogenic strain using the Apollo Gene Annotation Platform.

First, polycistronic operons, which have contiguous RNA evidence for multiple genes transcribed using the same DNA strand template, were identified. These polycistronic RNAs, as well as monocistronic RNAs, which had been computationally predicted were next examined for accuracy of the 5' and 3' ends of RNAs. Experimental RNA sequencing evidence was again used to visualize the correct ends to revise each gene model. With thirty percent of the annotations completed, a total number of thirty-four gene models were corrected (Appendix A). Finally, the process of curating the genome and gene models also included the search for novel genes that had not been predicted computationally. This involved RNA evidence being present without a corresponding gene model, which indicated that there were transcripts present that the computational prediction annotation platform did not detect. No novel genes were identified within the Apollo Gene Annotation Platform.

The conclusions drawn from this strain of *G. vaginalis* can be compared to those of strains 317, 594, and 409-05. All strains had the potential of being pathogenic, possess

a large number of certain genes that promote their ability to compete and exclude other bacteria colonies found within the vaginal tract, and all strains encoded bactericidal toxins. Of those bactericidal toxins, two were uniquely produced by strain 409-05 that allowed for mucin degradation, which is a characteristic common with bacterial vaginosis. Because of this, the detection of *G. vaginalis* in the vaginal tract does not yield adequate information as to the physiological potential of the bacterium (Yeoman et al., 2010).

In all, these curation efforts have produced a more accurate genome annotation with thirty-four total revisions based on experimental evidence to guide future research into pathogenic and non-pathogenic differences. Next steps would include using RNA sequence for the pathogenic strain to identify "pathogenic-specific" novel genes and other genome features. Pan-Genome analysis could be used to increase our understanding of the completed annotations. To perform the Pan-Genome analysis, the PanSeq web-based software (Rouli et al., 2015) could be used to examine the *G. vaginalis* strains to analyze which strains are classified as pathogenic, which are classified as non-pathogenic, and which are common to both. Such analyses would not only expand current knowledge about this bacterium, thus improving outcomes for women and pregnancy, but would also contribute to the development of personalized medicine for infectious disease management.

## REFERENCES

ACOG Practice Bulletin. 1998. Premature rupture of membranes: clinical management guidelines for obstetrician-gynecologists. International Journal of Gynaecology and Obstetrics. 63:75-84.

Aldunate M, Srbinovski D, Hearps AC, Latham CF, Ramsland PA, Gugasyan R, Cone RA, Tachedjian G. 2015. Antimicrobial and immune modulatory effects of lactic acid and short chain fatty acids produced by vaginal microbiota associated with eubiosis and bacterial vaginosis. Frontiers in Physiology. 6:164.

Bautista CT, Wurapa E, Sateren WB, Morris S, Hollingsworth B, Sanchez JL. 2016. Bacterial vaginosis: a synthesis of the literature on etiology, prevalence, risk factors, and relationship with chlamydia and gonorrhea infections. Military Medical Research. 3:4.

Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, Lei J, Achawanantakun R, Jiao D, Lawrence CJ, Ware D, Shiu SH, Childs KL, Sun Y, Jiang N, Yandell M. 2014. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiology. 164:513-524.

Cornejo OE, Hickey RJ, Suzuki H, Forney LJ. 2017. Focusing the diversity of Gardnerella vaginalis through the lens of ecotypes. Evolutionary Applications. 11:312–324.

Fujimoto, T., Parry, S., Urbanek, M., Sammel, M., Macones, G., Kuivaniemi, H., Romero, R., Strauss, J. F., III. 2002. A single nucleotide polymorphism in the matrix metalloproteinase-1 (MMP-1) promoter influences amnion cell MMP-1 expression

and risk for preterm premature rupture of the fetal membranes. The Journal of Biological Chemistry. 277:6296-6302.

Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, Matasci N, Wang L, Hanlon M, Lenards A, Muir A, Merchant N, Lowry S, Mock S, Helmke M, Kubach A, Narro M, Hopkins N, Micklos D, Hilgert U, Gonzales M, Jordan C, Skidmore E, Dooley R, Cazes J, McLay R, Lu Z, Pasternak S, Koesterke L, Piel WH, Grene R, Noutsos C, Gendler K, Feng X, Tang C, Lent M, Kim SJ, Kvilekval K, Manjunath BS, Tannen V, Stamatakis A, Sanderson M, Welch SM, Cranston KA, Soltis P, Soltis D, O'Meara B, Ane C, Brutnell T, Kleibenstein DJ, White JW, Leebens-Mack J, Donoghue MJ, Spalding EP, Vision TJ, Myers CR, Lowenthal D, Enquist BJ, Boyle B, Akoglu A, Andrews G, Ram S, Ware D, Stein L, Stanzione D. 2011. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. Frontiers in Plant Science. 2:34.

Gomez-Lopez N, Stlouis D, Lehr MA, Sanchez-Rodriguez EN, Arenas-Hernandez M. 2014. Immune cells in term and preterm labor. Cellular and Molecular Immunology. 11:571–581.

Henderson J, Carson C, Redshaw M. 2016. Impact of preterm birth on maternal well-being and women's perceptions of their baby: a population-based survey. BMJ Open. 6:e012676.

Hickey RJ, Forney LJ. 2014. Gardnerella vaginalis does not always cause bacterial vaginosis. Journal of Infectious Diseases. 210:1682-1683.

Ionio C, Colombo C, Brazzoduro V, Mascheroni E, Confalonieri E, Castoldi F, Lista G. 2016. Mothers and Fathers in NICU: The Impact of Preterm Birth on Parental Distress. Europe's Journal of Psychology. 12:604-621.

Laing C, Buchanan C, Taboada EN, Zhang Y, Kropinski A, Villegas A, Thomas JE, Gannon VP. 2010. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinformatics. 11:461.

Langmead B, Wilks C, Antonescu V, Charles R. 2019. Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics. 35:421-432.

Lata I, Pradeep Y, Sujata, Jain A. 2010. Estimation of the incidence of bacterial vaginosis and other vaginal infections and its consequences on maternal/fetal outcome in pregnant women attending an antenatal clinic in a tertiary care hospital in North India. Indian Journal of Community Medicine. 35:285.

Lawn JE, Davidge R, Paul VK, Xylander SV, Johnson JG, Costello A, Kinney MV, Segre J, Molyneux L. 2013. Born Too Soon: Care for the preterm baby. Reproductive Health. 10:S5.

Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG, Lewis SE. 2013. Web Apollo: a web-based genomic annotation editing platform. Genome Biology. 14:R93.

Leppäluoto PA. 2011. Bacterial vaginosis: what is physiological in vaginal bacteriology? An update and opinion. Acta Obstetricia et Gynecologica Scandinavica. 90:1302–1306.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 25:2078-9.

Maymon, E., Romero, R., Pacora, P., Gervasi, M.-T., Bianco, K., Ghezzi, F., Yoon, B. H. 2000. Evidence for the participation of interstitial collagenase (matrix

metalloproteinase 1) in preterm premature rupture of membranes. American Journal of Obstetrics and Gynecology. 183:914-920.

McGreggor JA and French JI. 2000. Bacterial vaginosis in pregnancy. Obstetrics Gynecology Survery. 55:S1-19.

Mendes-Soares H, Suzuki H, Hickey RJ, Forney LJ. 2014. Comparative Functional Genomics of Lactobacillus spp. Reveals Possible Mechanisms for Specialization of Vaginal Lactobacilli to Their Environment. Journal of Bacteriology. 196:1458–1470.

Michelle Silasi, Ingrid Cardenas, Karen Racicot, Ja-Young Kwon, Paula Aldo, Gil Mor. 2015. Viral infections during pregnancy. American Journal of Reproductive Immunology. 73:199-213.

National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/

Pederson Pedersen TL, Nookaew I, Ussery DW, Månsson M. 2017. PanViz: interactive visualization of the structure of functionally annotated pangenomes. Bioinformatics. 33:1081-1082.

Priya Soma-Pillay, Nelson-Piercy C., Heli T., Alexandre M. 2016. Physiological changes in pregnancy. Cardiovascular Journal of Africa. 27:89-94.

Romero R, Dey SK, Fisher SJ. 2014. Preterm labor: One syndrome, many causes. Science. 345:760–765.

Rouli L, Merhej V, Fournier P-E, Raoult D. 2015. The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbes and New Infections. 7:72–85.

Srinivas, S. K., Macones, G. A. 2005. Preterm premature rupture of the fetal membranes: current concepts. Minerva Ginecologica. 57:389-396.

Wang, H., Parry, S., Macones, G., Sammel, M. D., Ferrand, P. E., Kuivaniemi, H., Tromp, G., Halder, I., Shriver, M. D., Romero, R., Strauss, J. F., III. 2004. Functionally significant SNP MMP8 promoter haplotypes and preterm premature rupture of membranes (PPROM). Human Molecular Genetics. 13:2659-2669.

Wang, H., Parry, S., Macones, G., Sammel, M. D., Kuivaniemi, H., Tromp, G., Argyropoulos, G., Halder, I., Shriver, M. D., Romero, R., Strauss, J. F., III. 2006. A functional SNP in the promoter of the SERPINH1 gene increases risk of preterm premature rupture of membranes in African Americans. Proceedings of the National Academy of Sciences of the United States of America. 103:13463-13467. Note: Erratum: Proceedings of the National Academy of Sciences of the United States of America. 103:19212 only.

Witkin SS, Mendes-Soares H, Linhares IM, Jayaram A, Ledger WJ, Forney LJ. 2014. Influence of Vaginal Bacteria and D- and L-Lactic Acid Isomers on Vaginal Extracellular Matrix Metalloproteinase Inducer: Implications for Protection against Upper Genital Tract Infections. mBio. 4:e00460-13. Note: Erratum: mBio. 5:e00874-14.

Yeoman CJ, Yildrim S, Thomas SM, Durkin AS, Torralba M, Sutton G, Buhay CJ, Ding Y, Dugan-Rocha SP, Muzny DM, Qin X, Gibbs RA, Leigh SR, Stumpf R, White BA, Highlander SK, Nelso KE, Wilson BA. 2010. Comparative Genomics of *Gardnerella vaginalis* Strains Reveals Substantial Differences in Metabolic and Virulence Potential. Plos one. 5:e12411.

Zeitlin J, Szamotulska K, Drewniak N, Mohangoo A, Chalmers J, Sakkeus L, Irgens L, Gatt M, Gissler M, Blondel B. 2013. Preterm birth time trends in Europe: a study of

19 countries. BJOG: An International Journal of Obstetrics and Gynaecology. 120:1356–1365.

**Appendix A**

| Gene Models | Require Revision | Didn't Require Revision | Polycistronic |
|---|---|---|---|
| *atpD* | X | | |
| *atpG* | X | | |
| *atpH* | X | | |
| *atpF* | X | | |
| *atpB* | X | | |
| *HMPREF0421_t20001* | X | | |
| *lysC* | X | | |
| *HMPREF0421_20060* | X | | |
| *gatC* | X | | |
| *pfo* | X | | |
| *HMPREF0421_20067* | X | | |
| *HMPREF0421_20068* | X | | |
| *fahA* | X | | |
| *yggS* | X | | |
| *gluO* | X | | |
| *glpE* | X | | |
| *ppk* | X | | |
| *mutT* | X | | |
| *HMPREF0421_20087* | X | | |
| *upp* | X | | |
| *HMPREF0421_20090* | X | | |
| *HMPREF0421_20105* | X | | |
| *rnpA* | X | | |
| *rpmH* | X | | |
| *gyrB* | X | | |
| *gyrA* | X | | |
| *HMPREF0421_20140* | X | | |
| *pknA* | X | | |
| *HMPREF0421_20149* | X | | |
| *HMPREF0421_20152* | X | | |
| *ppc* | X | | |
| *HMPREF0421_20172* | X | | |
| *atp2Cl* | X | | |
| *atpE* | X | | |
| *clpB* | | X | |
| *glgP* | | X | |
| *gatB, gatA, gatC* | | | X |
| *spoU, HMPREF0421_20070, HMPREF0421_20071, rpsI* | | | X |