From sequencing to analysis:

Building a comparative genomics tool for the biologist end-user

By

Paul Michael Mangiamele

A thesis submitted to the graduate faculty

In partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Human Computer Interaction

Program of Study Committee:

Lisa K. Nolan, Major Professor

Mark Ackermann

Jared Danielson

Chris Harding

Catherine M. Logue

Iowa State University

Ames, Iowa

2014

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to take this opportunity to extend my appreciation to those who assisted me in various aspects of this research endeavor. First to my major professor, Dr. Lisa K. Nolan, for her guidance and expertise and for her belief in my professional as well as personal growth. My committee members for their willingness to offer their time, insights, and encouragement. To the collaboration with Bryon Nicholson and Aaron West – showing me how much different skill sets can accomplish together. And lastly to my family and friends, who never doubted this wish would happen.

CHAPTER 1. GENERAL INTRODUCTION

1.1. Objectives

The objectives of this research are to (1) understand the current state of genomic next generation sequencing (NGS) by fully finishing multiple genomes, (2) create a framework for better methods of genomic finishing and comparison using Avian Pathogenic *Escherichia coli* (APEC), a common and economically important bacterial pathogen of poultry, as a test bed, and (3) implement a publically available program that assists the scientific community as a whole in NGS downstream analysis to visually compare genomes.

1.2. Rationale and Significance

Advancements in sequencing technology have driven an ever-growing body of genomic sequence data to new heights. Since publication of APEC O78 sequencing paper (Chapter 2) one year ago, sequencing technology has grown in leaps and bounds and has plummeted in price. The affordability of these systems and availability of sequencing services have made these technologies accessible to smaller laboratories, focusing on individual biological organisms and systems, such as our own lab with APEC. This 'perfect storm' has made it feasible to sequence several of the APEC isolates in our collection. Although data generation is only the beginning, two substantial NGS bottlenecks for many labs are (1) closing and finishing the genomes to high quality standards (1-3) and (2) taking the sequence data to biological insight and relevance, especially when the volume of data overwhelms paradigms for standard data analysis. This

dissertation explores the process of sequencing and closing a diverse set of APEC genomes to finished quality, then creates a framework to complete a new programmatic pipeline and visualization of comparative genomics.

1.3. Authors' Roles

The authors of Chapter 2, entitled "Complete Genome Sequence of the Avian Pathogenic *Escherichia coli* Strain APEC O78," were Paul Mangiamele, Bryon Nicholson, Yvonne Wannemuehler, Torsten Seemann, Catherine M. Logue, Ganwu Li, Kelly Tivendale, and Lisa K. Nolan. Mangiamele was the primary researcher and conducted all analyses and genomic work. Nicholson assisted in closing methods. Wannemuehler performed the polymerase chain reaction (PCR) to assist in closing gaps. Seemann performed the annotation. Logue completed the pulse field gel electrophoresis (PFGE) for confirmation of genomic sizing. Li and Tivendale had worked on the sequence previously. Nolan was a corresponding author who set the research objectives and played a major role in conducting the research.

The authors of Chapter 3, entitled "Toolbox for Exploring Avian Pathogenic *Escherichia coli* (APEC) Pathogenesis, Host Specificity, Evolution and Control," were Paul Mangiamele, Bryon Nicholson, Aaron West, Yvonne Wannemuehler, Torsten Seemann, Catherine M. Logue, Kelly Tivendale, Curt Doetkott, and Lisa K. Nolan. Mangiamele was the co-primary author and researcher responsible for sequence closing and finishing, and comparison methods and informatics programming. Nicholson was the co-primary author and researcher responsible for methodology, sequence analyses, and confirmed informatics approaches in a microbiology expertise. West programmed much of

the Perl scripting that encompasses the pipeline of cWGAP. Wannemuehler performed the PCR to assist in closing gaps. Seemann created Prokka and ran our annotations that were the basis for much of the analysis. Logue completed the PFGE for confirmation of genomic sizing. Tivendale sent in the genomes to be sequenced and performed animal experiments to confirm virulence findings. Doetkott performed the biostatistics on choosing what strains to sequence. Nolan set the research objectives and was the corresponding author for the manuscript.

The authors of Chapter 4, entitled "Comparative Whole Genomic Alignment Pipeline – cWGAP," were Paul Mangiamele, Bryon Nicholson, Aaron West, Torsten Seemann, and Lisa K. Nolan. Mangiamele was the primary author, who programed and created the interface and analysis methodology. Nicholson assisted with algorithmic creation and confirmed results. West was the primary Perl developer for file processing aspect of the program. Nolan consulted and was the corresponding author for the manuscript.

1.4. Dissertation Organization

This dissertation has five chapters including this general introduction (Chapter 1), three article chapters (Chapters 2-4), and a general conclusion (Chapter 5). Each of the three article chapters is a separate manuscript that is either published or soon to be submitted, excluding the general introduction (Chapter 1) and general conclusion (Chapter 5). Chapter 2 is a published Genome Announcement about the methods and details of our strain APEC O78. Chapter 3 is a genome-wide association study and analysis of four diverse strains of APEC, and the creation of a framework to display a new way to compare and visualize multiple genomes. Chapter 4 is a documentation and announcement for

public use of the tool outlined in Chapter 3 available for others in the community. Chapters 2 and 3 address Objective 1 of the dissertation. Chapter 3 addresses Objective 2 and 3, and presents a test case for using our program. Chapter 4 directly addresses Objective 3 and delves in depth on program creation in a piecewise manner. Chapter 5 is the general conclusion, giving a brief outline of the research as a whole.

CHAPTER 2. COMPLETE GENOME SEQUENCE OF THE AVIAN PATHOGENIC *ESCHERICHIA COLI* STRAIN APEC O78

A paper published in *American Society for Microbiology Genome Announcements*

Paul Mangiamele[1*], Bryon Nicholson[1], Yvonne Wannemuehler[1], Torsten Seemann[2], Catherine M. Logue[1], Ganwu Li[1], Kelly A. Tivendale[3] and Lisa K. Nolan[1§]

Abstract

Colibacillosis, caused by Avian Pathogenic *Escherichia coli*, is a disease with significant impact, causing extensive animal and financial losses globally. Though this disease is common and difficult to treat and manage, mechanistically, more knowledge is desired. Here, we present the fully closed genome sequence of a typical avian pathogenic *E. coli* strain belonging to the serogroup (O78).

---

[1] Department of Veterinary Microbiology and Preventive Medicine, College of Veterinary Medicine, 1802 University Blvd, VMRI 2, Iowa State University, Ames, Iowa 50011

[*] Primary researcher and author

[2] Victorian Bioinformatics Consortium, Monash University, Clayton, Victoria 3800, Australia

[3] Veterinary Science, The University of Melbourne, Parkville, Victoria 3010, Australia

[§] Corresponding Author

2.1. Genome Announcement

Colibacillosis, caused by avian pathogenic *Escherichia coli* (APEC), is one of the most significant infectious diseases affecting poultry (4-10). Poultry colibacillosis takes many forms, with systemic forms occurring most often (5). Collectively, these diseases result in annual multimillion-dollar losses due to mortality, decreased production, and condemnations (4, 6, 9). Indeed, colibacillosis poses a profound threat to one of humankind's cheapest sources of high-quality animal protein. Despite the importance of this disease, the mechanisms of APEC virulence largely remain unknown. Studies into APEC pathogenesis would be enhanced by public access to high quality genomic sequences. To date three APEC sequences are publically available. The sequence of APEC O1, an O1:K1:H7 strain isolated from the lung of a turkey, is fully closed (11). A draft sequence of a Brazilian APEC strain, SCI-07, a member of the O nontypeable:H31 serotype, from gelatinous edema lesions from a laying hen, is in 68 contigs (12), and a sequence of an O78 strain (χ7122) was recently released in four contigs (13). Here, we describe a fully closed and annotated sequence of another O78 strain with the idea that fully closed sequences representative of the most commonly isolated APEC serogroups, such as O1 and O78 strains, are needed to support future colibacillosis research (4).

APEC O78 is an O78 strain isolated from the lung of a turkey clinically diagnosed with colibacillosis. Genomic sequencing was performed using complementary sequencing technologies, combining results obtained with a Roche/454 FLX GS instrument and an Illumina Hi-Seq 2000. The following datasets were used in the final assembly: (i) GS-FLX, with 590,773 shotgun reads

totaling 237Mbp (~49-fold coverage); (ii) GS-FLX 8-kb mate-pair library with 474,583 shotgun reads totaling 168Mbp (~35-fold coverage) of which 330,857 were paired; and (iii) Illumina 100bp paired-end library with 27,389,600 reads totaling 2,587Mbp (~539-fold coverage). Both 454 read sets were assembled *de novo* using Newbler 2.6 (Roche), and Illumina reads were assembled separately with Velvet 1.1 (14) and Illumina's ELANDv2e assembler. The genome was closed using 454 assemblies as a 'reference' sequence and the Illumina data to add depth, correct errors, and close gaps. Whole-genome optical mapping (OpGen, Gaithersburg, MD) was used to validate scaffolds and contig order. The assembly was confirmed using PCR and Sanger sequencing and validated by consistency of paired-end evidence from 454 and Illumina reads.

Annotation was automated using NCBI Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP). The final version was checked against the previously completed Prokka 1.5.2 annotation.

The assembled genome consists of a single chromosome (4,798,435bp; 50.68 %GC content) and two plasmids, one 217.830kb and the other 113.260kb. The chromosome contains 4,696 protein-encoding genes, 88 tRNA-carrying genes, and 19 rRNA-carrying operons. The chromosome of APEC O78 is smaller than many other fully sequenced ExPEC genomes and its chromosomal structure appears different from other ExPEC. Assessment of the implications of these differences is ongoing, but addition of a genomic sequence of one of the commonly occurring serogroups among APEC significantly contributes to the toolset that can be used on studies of APEC pathogenesis and colibacillosis control.

2.2. Nucleotide Sequence Accession Number

Complete sequences of APEC O78 have been deposited in GenBank under accession no. CP004009.

2.3. Acknowledgements

This work was supported by grant USDA-NIFA award 0826675.

CHAPTER 3. TOOLBOX FOR EXPLORING AVIAN PATHOGENIC *ESCHERICHIA COLI* (APEC) PATHOGENESIS, HOST SPECIFICITY, EVOLUTION AND CONTROL

A paper to be submitted to *PLOS ONE*

Paul Mangiamele[1*], Bryon Nicholson[1*], Aaron West[2], Yvonne Wannemuehler[1], Torsten Seemann[3], Catherine M. Logue[1], Kelly A. Tivendale[4], Curt Doetkott[5], and Lisa K. Nolan[1§]

Abstract

Colibacillosis, caused by Avian Pathogenic *Escherichia coli* (APEC), is a significant disease causing extensive animal and financial losses globally. New colibacillosis control approaches, based on fundamental knowledge of APEC pathogenesis, are needed as current approaches are not fully effective. To date, few high-quality, finished APEC genomic sequences are available to adequately represent this diverse group of pathogens and support needed research. Here,

[1] Department of Veterinary Microbiology and Preventive Medicine, College of Veterinary Medicine, 1802 University Blvd, VMRI 2, Iowa State University, Ames, Iowa 50011

[*] Primary researcher and author

[2] Department of Chemistry, Iowa State University, Ames, Iowa 50010

[3] Victorian Bioinformatics Consortium, Monash University, Clayton, Victoria 3800, Australia

[4] Veterinary Science, The University of Melbourne, Parkville, Victoria 3010, Australia

[5] Information Technology Services, North Dakota State University, Fargo, ND 58105

[§] Corresponding Author

we describe and compare the genomic sequences of four APEC strains (two newly sequenced strains, APEC O2 and O18; one recently published but not yet fully described, O78; and one previously sequenced strain, APEC O1 that was re-annotated in the present study) that represent key groups of APEC in an effort to create a toolbox of strains for future study. Strain selection was based on analysis of over 452 APEC isolates for various traits with the intent of identifying strains that represent mainstream APEC but that differ in key traits in order to maximize the knowledge to be gleaned from their study. Comparative analysis of these four strains revealed that they harbor a common core of 108,471 base pairs (bp) that consists of 124 genes, with at least 8 islands with respective functionality. In addition, the data generated here were presented through a new comparative genomic framework, cWGAP, that visualizes the sequences of these strains, shows where they intersect to form a 'core APEC genome', draws attention to their conserved regions, and highlights their intergenic SNP regions through a heat map drawn across all fully sequenced and annotated genomes.

3.1. Introduction

Colibacillosis, caused by avian pathogenic *Escherichia coli* (APEC), is one of the most common infectious diseases affecting turkeys, layers, and broilers worldwide (5, 7-10, 15-17). This disease takes many forms, with the systemic infection occurring most often (5). Collectively, colibacillosis results in annual multimillion dollar losses due to mortality, lost production, and condemnations (5, 16, 17). Thus, this disease poses a profound threat to one of humankind's cheapest sources of high-quality protein. Despite the importance of

colibacillosis, the mechanisms of APEC virulence have largely remained a puzzle, hampering control efforts.

Much of what we do know about the molecular pathogenesis of APEC infection has relied on studies using indirect genome-wide approaches, such as suppression subtractive hybridization (18, 19), signature-tagged mutagenesis (20), and selective capture of transcribed sequences (21). Such indirect approaches were necessary since a complete APEC genome sequence only became available in late 2006 (11). With the passage of time, it has become clear that a single APEC sequence cannot adequately support state-of-the-art research into APEC pathogenesis, as APEC are very diverse (22), necessitating generation of multiple genomic sequences on which to base future advances in our understanding of APEC pathogenesis, evolution, host specificity, and control. Indeed, several groups have sought to fill this gap by describing draft APEC sequences (12, 13). Though these are very helpful, the nature of draft sequences precludes robust genomic comparisons that can provide critical clues as to mechanisms of disease, host specificity or disease control or insights into the evolution of virulence. Consequently, we believe that it is critical for future research that additional high quality and complete APEC genomic sequences be generated. Here, we seek to address this need by producing an "APEC toolbox", consisting of four APEC strains chosen for their ability to represent key groups of APEC, their complete genomic sequences, a thorough description of their relevant phenotypes, and their comparative genomic analysis.

3.2. Materials and Methods

3.2.1. Bacterial strains and serogrouping

Over 452 APEC isolates, originating from multiple geographic locations across the USA, from avian hosts with different forms of colibacillosis and various lesion types and hosts of all ages and types (layers, broilers, and turkeys) were considered for sequencing in this study. Isolates had been previously subjected to phylogenetic typing and virulence genotyping for over 200 genes, thought to be linked to APEC and/or human extraintestinal pathogenic *E. coli* (ExPEC) virulence (23, 24). In addition, these isolates have been assessed for resistance to 15 antimicrobials and content of plasmid replicons, and all were subjected to serogrouping through the Pennsylvania State *E. coli* Reference Center. Some of these isolates have been classified as to their virulence for chicks, chick embryos, rats, and mice and abilities to resist the effects of host complement (24). Since isolation, these strains have been stored at -80 C in Luria Broth (LB) with 20% glycerol.

3.2.2. Strategy employed to select APEC strains for sequencing

Strains were selected for sequencing using a multistep process. First, virulence genotyping data on the 452 APEC strains in our collection were subjected to cluster analysis so that strains representing different, but major clusters, could be included in our study. Also, because it is well ingrained in the literature that some of the more dominant APEC serogroups are O1, O2, and O78 (5, 7, 9, 10, 15-17), these serogroups were targeted for sequencing. Also included was an O18 strain, since there is interest in APEC's role in human diseases caused by such pathogens as neonatal meningitis *E. coli* (NMEC), which are often

O18 strains (22). Further, the strains selected for sequencing from each cluster were chosen to represent the typical phylogenetic group of that cluster and serogroup (Figure 1, Table 1 and 2 in Supplementary Data section). For instance, O78 strains tend to be assigned to phylogenetic group A by the older Clermont phylogenetic assay (25); thus, the O78 strain selected for sequencing was from the A phylogenetic group. Also, some attention was given to differences in other traits, in order to maximize what we could hope to learn from study of each strain individually and comparatively.

3.2.3. Antimicrobial susceptibility testing

Antimicrobial testing was performed on all isolates as per standards from the National Antimicrobial Resistance Monitoring System (NARMS) using broth microdilution. *E. coli* isolates were struck to tryptone soy agar (TSA) from frozen stock and incubated at 37 C for 18h. Colonies were selected using a sterile cotton swab and suspended in 5 ml of sterile water and adjusted to a 0.5 McFarland Standard using a nephelometer (TREK Diagnostics, Cleveland OH). Then, 10μl of the suspension was removed and added to 11 mls of Mueller Hinton (MH) broth with TES. The suspension was mixed using a vortex and added to the National Antimicrobial Resistance Monitoring System (NARMS) panel (CMV2AGNF; Trek) using an AIM Autoinoculator which dispensed 50μl of the broth suspension into the wells of each panel. All panels were sealed and incubated at 37°C for 18-20h. Following incubation, all plates were read using the Sensititre Autoreader and minimum inhibitory concentrations (MICs) recorded for each strain based on growth/ no growth in the wells of the plate. All MICs recorded were compared against the accepted breakpoints for E. coli recovered from

animals using the CLSI and NARMS criteria (see http://www.ars.usda.gov/Main/docs.htm?docid=6750&page=3). Antimicrobial resistance/ susceptibility was examined for the following antimicrobials: amikacin (0.5 - 64 μg/ml), ampicillin (1 - 32 μg/ml), amoxicillin/clavulanic acid (1/0.5 - 32/16 μg/ml), ceftriaxone (0.25 - 64 μg/ml), chloramphenicol (2 - 32 μg/ml), ciprofloxacin (0.015 - 4 μg/ml), trimethoprim/sulfamethoxazole (0.12/2.38 - 4/76 μg/ml), cefoxitin (0.5 - 32 μg/ml), gentamicin (0.25 - 16 μg/ml), kanamycin (8 - 64 μg/ml), nalidixic acid (0.5 - 32 μg/ml), sulfisoxazole (15-256 μg/ml), streptomycin (32 - 64 μg/ml), tetracycline (4 - 32 μg/ml), and ceftiofur (0.12 - 8 μg/ml). The minimum inhibitory concentration was calculated based on the well of least antimicrobial concentration showing no growth, and results were compared to NARMS established breakpoints to determine resistance.

3.2.4. Virulence genotyping and phylogenetic typing

Test and control organisms were examined for the presence of over 200 virulence genes and genomic islands known for their association with APEC or ExPEC chromosomal virulence. PCR was performed in multiplex using primers and protocol previously described (11, 22, 24).

Strains also were assigned to phylogenetic groups according to the PCR amplification methods described by Clermont *et al*. (26), (27). The first method assigns APEC to four groups (A, B1, B2, and D), while the more recently described method assigns strains to one of 12 groups (A, B1, B2, C, D, E, F, Clade I, II, III, IV, or V).

3.2.5. Genomic Sequencing

DNA preparation. APEC O2, O18, and O78 were all sequenced using analogous methods, with minor deviations in procedure between genomes. Whole genomic DNA was prepared using a Promega Wizard DNA Purification kit according to manufacturer's specifications.

Sequencing. Genomic DNA for each genome was purified and subjected to sequencing using a Life Sciences 454 FLX, generating both shotgun and mate-pair reads. *De novo* assembly was performed using Newbler 2.7 (Roche 454 Life Sciences). Then, using a complementary sequencing technology, Illumina, 100 bp paired-end libraries with insert sizes of 500 bp were generated at the University of Oregon Core Genomics Facility. Sequencing was performed on an Illumina HiSeq 2000. Base masking and de-multiplexing were performed using CASAVA 1.8.2 software. *De novo* assembly was performed using both Velvet 1.1 (14, 28) and ELANDv2e (Illumina). Previous 454 scaffolds generated by Newbler were used as scaffolding reference data, while Illumina data were added to supplement depth, correct errors, and close gaps. Final gaps were closed using primers designed to amplify out between contigs followed by Sanger sequencing. To assist with genome finishing, whole-genome optical restriction maps were generated for each genome using the restriction enzyme NcoI (OpGen, Gaithersburg, MD). MapSolver software was used to compare *in vitro* digestions to *in silico* digestions and confirm contig joins and orientation. Final single contigs were evaluated by Tablet (29) for consistent depth of coverage to scan for condensation or expansion sequencing errors.

3.2.6. Genomic annotation

Automated annotation was performed using Prokka 1.5.2 (30) with a custom database specifically set up for *E. coli* (Victorian Bioinformatics Consortium) for primary analysis. Though APEC O1 had previously been annotated (11), it was re-annotated in this study alongside the newly sequenced strains to ensure that the comparative analysis was up-to-date and consistent with other sequences. Protein coding regions were predicted using Prodigal (31), tRNA and tmRNA genes using ARAGORN (32), and rRNA genes using RNAmmer (33). Gene function was assigned primarily using BLASTp against the EcoCyc database (34, 35) and secondarily using HMMER3 (36) against Pfam-A 26.0 (37, 38). These GenBank files were used as the basis of our comparative genomic analysis and pipeline development using GenBank files as a medium.

Annotation, when submitted to NCBI, was automated using NCBI Prokaryotic Genome Annotation Pipeline (PGAP) to ensure consistency. While this pipeline is very similar to Prokka, it produced sufficiently different results. PGAP combines HMM-based gene prediction methods with a sequence similarity-based approach, which combines comparison of the predicted gene products to the non-redundant protein database, Entrez Protein Clusters, the Conserved Domain Database, and the COGs (Clusters of Orthologous Groups). Gene predictions were done using a combination of GeneMark and Glimmer (39-41). Ribosomal RNAs were predicted by sequence similarity searching using BLAST against an RNA sequence database and/or using Infernal and Rfam models. Transfer RNAs were predicted using tRNAscan-SE (42). In order to detect missing genes, a complete six-frame translation of the nucleotide sequence is done and predicted proteins (generated above) were masked. All predictions

were then searched using BLAST against all proteins from complete microbial genomes. Annotation was based on comparison to protein clusters and on the BLAST results. Conserved Domain Database and Cluster of Orthologous Group information was then added to the annotation. Frameshift detection and cleanup occurs and then the final output was sent back for final analysis.

3.2.7. Sequence analyses

Core and pan genome analysis. Core and whole genome alignments of APEC O1, O2, O18, O78 and the laboratory strain *E. coli* MG1655 (43, 44) were performed in progressive Mauve version 2.3.1 (45). APEC core regions were defined as contained in all APEC genomes and absent *E. coli* MG1655. Pan genomic data were defined as regions appearing in at least one APEC species but not in the non-pathogenic backbone. cWGAP (46) was used to separate core and pan genomic sequence alignments from the Mauve analysis, and back referenced Prokka annotations to provide genes contained within the core and pan APEC genome to generate abridged Genbank files.

Vaccine epitope analysis. Core APEC genes were submitted to the Vaxign vaccine prediction pipeline (47). The Vaxign pipeline uses open source programs to predict protein localization, transmembrane helices, and probable adhesins. Following prediction, the results were blasted for host genome similarity and epitope prediction. Results were filtered to exclude proteins with an adhesion probability below 0.25, localization in cytoplasm and inner membrane, or greater than two transmembrane helices.

Visualization comparison method. Using all information gathered we aimed to visualize pertinent information about each genome concisely and accurately, leveraging Mauve (45) and Circos via cWGAP (46).

Polymorphism and SNP analysis. SNP analysis was performed by Mauve 2.3.1 (45) SNPExporter, and filtering and parsing those results using Perl scripting. The data was filtered to only include majority consensus (75%) SNPs in APEC strains different from the reference strain MG1655. Polymorphic sites in each alignment were identified and listed by pattern. Data were then sorted and filtered for polymorphisms that were different or absent from *E. coli* MG1655. Datasets were reformatted for and visualized using Circos (48) using cWGAP and generating karyotype files (46). The SNP file generated from Mauve was filtered so that only SNPs where 3 or more of APECs were different from MG1655 were shown with the cWGAP scripts. Duplicate genes in the Circos file were removed by the gene filter script and genes multiple stop or rare codons were condensed (46). The percentages of SNPs appearing in each gene were calculated using SNPScript to generate a highlight file for Circos. All scripts were released and described in the cWGAP paper (46).

Genomic island (GI) identification. IslandViewer (49) is a web-based tool for identification of genomic islands in bacterial genomes, combining three methods of island identification. Sequence comparison using SIGI-HMM (50) was utilized to predict common genomic island characteristics using a Hidden Markov Model to identify codon patterns. Following, IslandPath-DIMOB (51) was then used to search the genome for common genomic characteristics of virulence islands. Finally genes were compared to a database of known antimicrobial and virulence genes (49, 52). Finished and annotated APEC and

MG1655 Genbank files were uploaded for analysis and a file containing the positions of all genomic islands was generated for genome visualizations.

Phylogeny construction. Phylogenic analysis was performed using MrBayes 3.2.2 (53, 54). Phylogenies were created using the genes *chuA*, *yjaA,* and the core APEC GIs 3, 5, 6, and 7. Genes were aligned using Clustal Omega and used to generate a nexus file. MrBayes was run using a general time reversible model with variation between sites described as an independent gamma rate model using MG1655 and DH1 as roots. The number of generations was set to 100,000 with 25,000 burn-in cycles, and posterior probability cutoff was set at 99%.

Comparative genomic analysis. In order to gain the most complete results from gene selection for vaccine development, a pipeline of tools was developed to reliably output information optimized for bacterial genomes. Much of this process has gone through significant rigor and can be extrapolated for other genomic sequences, including visualization. While Mauve has a very complete interactive visualization, expanding and visually presenting the core and pan genome was desired, while contrasting with other types of data like SNP and genomic island data. These requirements developed cWGAP (46).

Gene prevalence analysis. To determine the prevalence of core APEC epitopes identified using Vaxigen, multiplex PCR was used to amplify selected genes to determine the prevalence in the collection of 452 APEC and 200 avian fecal commensal *E. coli* (AFEC). PCR reactions were performed on whole DNA extractions from *E. coli*. PCR reactions were performed under the following conditions 95 C for 60 seconds, followed by 30 cycles of 95 C for 30 seconds, 65 C

for 30 seconds, and 72 C for 5 minutes, followed by a hold at 4 C using the following primers:

PillF: ATTATCCGGCAGCAGAGTGCC

PillR: CGACACTTGCAGATGGCACC

SorbF: TGTTGAGCAGACGAACCATCAGTAGC

SorbR: CGATGAAGTGGTATGGCCTACAGC

3.3. Results

3.3.1. Bacterial Strains, serogroups, genotyping, and phylogenetic typing

Establishing serogrouping (Table 2), phylogenetic analysis (Figure 2) and cluster analysis of virulence genotyping data (Figure 1), conveyed strains to sequence. Phylogenetic analysis of the sequenced strains revealed close relationships between the O1 and O18 strains and more distant relationships between the O78 and O2 strains (Figure 2). These data are corroborated by phylogenies in NCBI, but is based on regions determining phylotypes and hypothesized virulence regions and not full genome alignments (http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/). Of particular interest, strains O1 and O18 clustered closely together, while strains O2 and O78 in turn showed no difference.

3.3.2. Sequence analysis

Overview of APEC sequences. Completed genomic sequences were obtained for APEC O2, O18, O78 using similar methods. The visual flow and guide of this process is displayed in Figure 4. Details of the assemblies are shown in Table 6. The goal was to attain the best resolution of finishing possible for each

genome. A hybrid sequencing approach was used for each genome. Despite the fact that all the genomes being completed were APEC, they were each distinct enough to warrant a *de novo* sequencing and assembly approach due to large chromosomal rearrangements. Employing a reference-guided assembly resulted in significantly fragmented assemblies. Each genome is described in detail below.

APEC O1. APEC was completed in 2007 (11), and used different sequencing technology and closing methods than rest due to available technology of the time.

APEC O2. Due to a lack of relevant reference genomic sequences to guide its assembly, as well as repeated difficulties in spanning the gaps in sequence, APEC O2 was sequenced multiple times followed by *de novo* assembly. Four different sequencing technologies were used in an effort to bridge these gaps. These included (i) Roche/454 FLX Titanium GS, (ii) Illumina GAIIx, (iii) Illumina HiSeq2000, and (iv) Life Technologies Ion Torrent 316 chip. The following datasets were used in the final assembly: (i) GS-FLX, with 638,908 shotgun reads totaling 255.1 Mbp (48.9-fold coverage); GS-FLX 8-kb mate-pair library with 447,236 shotgun reads totaling 153.6 Mbp (30-fold coverage) of which 312,704 were paired. (ii) Illumina GAIIx with 862,731 paired reads totaling 67,223,839 (13.1 fold-coverage); (iii) Illumina HiSeq 2000 100 bp paired-end library with 9,614,323 paired reads totaling 803.22 Mbp (157.1 fold-coverage); and (iv) Life Technologies Ion Torrent 316 chip with 2,039,822 reads totaling 319.2 Mbp (62.4 fold-coverage). (Organized details are available in Table 6 in the 'Supplementary Data' section).

Results from all runs of these complementary sequencing technologies were combined, as described in the following. Both 454 read sets were assembled

de novo using Newbler 2.7 (Roche 454 Life Sciences). Illumina reads were assembled separately with Velvet 1.1 (14, 28) and Illumina's ELANDv2e assembler. The genome was brought down to two contigs using 454 assemblies as a 'reference' sequence with the Illumina data used to add depth, correct errors, and close gaps. Whole-genome optical mapping was used to validate the scaffolds and order the contigs. When the optical map showed that the gap was too large to span with PCR, we conducted additional Illumina and Ion Torrent runs to build varying-length reads. However, the data generated by these runs on assembly did not close this gap. Successful closure required using the two main scaffolds as reference, construction of an 'in house' BLAST database of assembled contigs from the Illumina and Ion Torrent contigs, and 'BLASTing' the pieces through an iterative process to find the connecting parts. Then, used the contig as a reference to guide all the reads together and correct any assembly issues, and the final contig assembly was confirmed using PCR, Sanger sequencing, and Whole-Genome Mapping, followed by validation by consistency of paired-end evidence from 454, Illumina, and Ion Torrent reads.

The assembled genome consists of a single chromosome (5,112,508 bp; 50.63 %GC content) and one plasmid, 199.734 kb, which was confirmed by pulse field gel electrophoresis (PFGE). The chromosome contains 4,784 protein-encoding genes, 89 tRNA-carrying genes, and 22 rRNA-carrying operons.

Nucleotide sequence accession numbers. Complete sequences of APEC O2 have been deposited in GenBank under accession no. CP006834.

APEC O18. Generation of a completed sequence for APEC O18 was relatively straightforward as compared to APEC O2 and O78. Employing a hybrid approach using both the Roche/454 FLX GS instrument and Illumina Hi-

Seq 2000. Final assembly of the following datasets were used: (i) GS-FLX, with 235,653 shotgun reads totaling 95.8 Mbp (~19.1-fold coverage); (ii) GS-FLX 8-kb mate-pair library with 219,416 shotgun reads totaling 67.6 Mbp (~13.5-fold coverage) of which 152,602 were paired; and (iii) Illumina 100 bp paired-end library with 14,386,242 reads totaling 1,358.8 Mbp (~274.4-fold coverage). Both 454 read sets were assembled *de novo* using Newbler 2.7 (Roche 454 Life Sciences), and Illumina reads were assembled separately with Velvet 1.1 (14, 28) and ELANDv2e (Illumina) assembler. The genome was closed using 454 assemblies as a 'reference' sequence, and the Illumina dataset was used to add depth, correct errors, and close gaps. Whole-genome optical mapping was used to validate scaffolds and contig order. The assembly was confirmed using PCR and Sanger sequencing and validated by consistency of paired-end evidence from 454 and Illumina reads. (Organized details are available in Table 6 in the 'Supplementary Data' section).

The assembled genome consists of a single chromosome (5,006,568bp; 51.73 %GC content) and three plasmids, (1) 131.266kb, (2) 110.346kb, and (3) 41.465kb, as confirmed by PFGE (Figure 6). The chromosome contains 4,581 protein-encoding genes, 84 tRNA-carrying genes, and 22 rRNA-carrying operons.

Nucleotide sequence accession numbers. The complete sequence of APEC O18 has been deposited in GenBank under accession no. CP006830.

APEC O78. Sequencing of APEC O78 was described previously (55). Organized details of relevant sequencing data are available in Table 6 in the 'Supplementary Data' section. The assembled genome consists of a single chromosome (4,798,435 bp; 50.68% GC content) and two plasmids, one 217.830

kb and the other 113.260 kb. The chromosome contains 4,696 protein-encoding genes, 88 tRNA-carrying genes, and 19 rRNA-carrying operons. The chromosome of APEC O78 is smaller than many other fully sequenced extraintestinal pathogenic *E. coli* (ExPEC) genomes, and its chromosomal structure appears different from those of other ExPEC genomes.

3.3.3. Visualization and analysis

Core and pan genome analysis. The pan-genome includes the "core genome" containing genes present in all strains, a "dispensable genome" containing genes present in one or more strains, and finally "unique genes" specific to single strains (56). Whole genome alignments of APEC O1, O2, O18, O78 and laboratory strain *E. coli* MG1655 were performed in progressive Mauve version 2.3.1 (Figure 3), using a seed weight of 17 and seed families. The seed size parameter sets the minimum weight of the seed pattern used to generate local multiple alignments during the first pass of anchoring the alignment (57). Core regions were defined as contained in all APEC (O1, O2, O18, O78) genomes and absent in *E. coli* MG1655 using genome subtraction. In-house Perl scripting filtered regions from the analysis, and back referenced annotations to provide genes contained within the core and pan APEC genome and generated abridged Genbank files (46). The output of this process created the core APEC genome, 108,471 bp that consists of 124 genes, with at least 8 islands with respective functionality (Figure 5, Table 3).

Genomic islands (GIs). A key part of the visualization was to guide the viewer's eye to potential areas of interest. In the present case, identification of putative virulence genes is of particular interest. In order to identify virulence

gene candidates, IslandViewer (49) was used to calculate GIs occurring in the APEC core genome followed by manual curation of these GIs. To accomplish this, the fully sequenced GenBank files from the Pokka annotation were fed into IslandViewer. Since it was the goal to compare internally with cWGAP, the GI output files were extracted from predicted GI coordinates and inserted into each track of data (Figure 5) rather than use IslandViewer's visualization extension. This data track leverages classification of virulence islands by GC% skew, codon usage, and mobile genetic elements. Regions positive for GIs were then compared to the Virulence Factor Database (58). The results of this prediction method are given in Table 3 and Figure 5.

Conserved regions. Examination of the core genome revealed large multi-gene clusters. BLAST searches were performed on these sections to ascribe a putative function label (Table 3 and Figure 5).

SNP analysis and visualization. Our SNP analysis was performed using Mauve's SNP calling from the progressive Mauve alignment backbone. The SNP comparisons were sequence-to-sequence differences, thus for every polymorphic site in an alignment, the SNP file records the nucleotides present in each genome at that site, along with the sequence coordinates of the site in each genome.

A Perl script was created and called SNPScript, part of cWGAP, to parse these data, assigning SNPs to their respective genes and intergenic regions (46). Genes less than 100 base pairs as well as genes containing SNP ratios in excess of 30% were manually examined for validity. Genes and polymorphisms were exported to a new track and visualized in Circos using a heat map scale. Genes containing the highest incidence of polymorphisms are labeled on the outside track. Multiple islands from the core APEC region (islands 2, 3, and 7) showed a

decreased incidence of polymorphisms, while islands 1,4, and 5 showed higher than average incidence of polymorphisms (Figure 5).

3.3.4. Vaxign analysis for vaccine development

The APEC core genome assembled from analysis of APEC O1, O2, O18 and O78 was analyzed via the Vaxign pipeline (47). This analysis revealed multiple potential targets for vaccine development. Vaxign filters out proteins predicted to occur in the cytoplasm or inner membrane, have greater than five trans-membrane helices, and have an adhesion probability of less than .025. From this analysis two operons, located within two operons (Sorbose and Fimbrial 2), were identified. Using PCR, the 452 APEC and 200 AFEC isolates of our collection were screened for these genes in an effort to determine the viability of these genes and the proteins they encode as likely candidates for vaccine targets. Though the genes of the Sorbose and Fim2 operons existed in greater proportions in APEC than AFEC, prevalence in non-pathogenic strains was high, suggesting that they might not be specific enough to the disease-causing strains to be useful vaccine targets.

3.4. Discussion

3.4.1. Strain selection for toolbox

Development of a set of representative genomes to support the study of important microbes has served the research community well, and availability of high quality reference genomes has greatly accelerated research on many fronts. Certainly, this has been true in pathogenic bacteriology--where once a genome was explored one mutated gene at a time, it is now possible to use pan-genomic

approaches to assess the activity of all the genes of an organism at the same time and under conditions of infection. Of course, the generalizability of the results of these studies depends on the representative nature of available genomic sequences. If they are not representative of the population of interest, the insights gained will be limited. Among homogeneous populations, this issue is not so concerning, but APEC are not homogeneous. They vary widely in serogroups, phylogenetic types, cluster types, virulence, host range and types of colibacillosis caused, making use of a few genomes on which to base future studies of APEC problematic. Here, we have sought to remedy the deficit of representative APEC genomes through generation of several high quality, finished genomes of 'mainstream' APEC that differ in certain key characteristics.

Based on phylogenetic typing (Figure 2), serogrouping (Table 2), and cluster analysis of virulence genotyping data for all the APEC strains in our collection (Figure 1), candidate APEC were identified for sequencing. Since APEC O1, O2 and O78 are considered to be among the most common serogroups causing disease in birds (59-61), making sure to include them in our final pool of sequencing candidates. In addition, we included an O18 strain, since such strains tend to bear much similarity to human Neonatal Meningitis *Escherichia coli* (NMEC), expanding our ability to study ExPEC host specificity. In addition, our data analysis revealed a tendency for O78 strains to be assigned to phylogenetic group A, while O1 and O18 tend to fall in phylogenetic group B2, and O2 strains occur in similar frequency in B2. Also, in a further effort to ensure the representative nature of the strains for study focus was given to strains lying within major APEC clusters, based on the statistical analysis of the virulence typing data. From this analysis it was found that certain serogroups tended to

fall in certain phylogenetic groups (Table 2). Thus, since O78 strains tend to fall into phylogenetic group A (according to Clermont's older scheme), an O78 strain from that phylogenetic group and from a major cluster was sequenced. A similar procedure was used to identify the O2 and O18 strains. From these, strains of different phylogenetic types were selected. This strategy ensured a genome of representative APEC from each of the major serogroups (O1, O2, and O78) and dominant phylogenetic types (A, B2, and D) and clusters occurring among APEC would be included. We believe this process, and the one undertaken to choose APEC O1, are among the most rigorous selection procedures ever used to choose ExPEC for sequencing. Consequently, we believe these sequences will underpin vital APEC research long into the future.

3.4.2. The core APEC region

A particular goal of this study was to determine what genes make an APEC, an APEC. That is, we sought to identify the core APEC genome. Identification of a core APEC genome would help focus future studies into the pathogenesis of colibacillosis and could serve as a basis to distinguish APEC from other ExPEC. In order to distinguish the APEC core genome from the *E. coli* backbone, we subtracted the sequence of *E. coli* MG1655, an avirulent, laboratory strain, from the genomes of APEC O1, O2, O18, and O78. Thus, the remaining 108 kb (108,471 bp) consisting of 124 genes constitutes the core APEC. Though addition of other APEC and non-pathogenic *E. coli* genomes to this analysis will likely refine this version of the core, it provides a starting point for future analysis.

3.4.3. Narrowing down the core APEC further

The core region, derived as described above, was further refined, as some genes present in the core, were found in other K-12 strains of *E. coli.* Further investigation into these K-12 genes revealed small percentage overlaps or improper gene annotation were the cause of aberrant annotation. Although, as cWGAP (46) became more mature and automated, additional functionality was built in to optimize parameters such a user specified overlap for gene detection as well as the ability to subtract an unlimited number of genomes. Utilizing and testing these methods, a new core was built by comparing APEC O1, O2, O18 and O78, then subtracting MG1655 as well as other K-12 such as DH1, MDS42, W3110, ATCC8739, BW2952, and DH10B. The output of this new core was 45,144 bp consisting of 52 genes. To optimize this further, an overlap percentage function was implemented and iterated for optimal values (as shown in Table 4). The K-12 genes decrease the overlap percentage stringency increased. After testing 0%-100%, a level of 20% was found to be optimal as it outputted 29,940 bp core consisting of 38 genes, of which 5 have K-12 labels. Increasing percentage overlaps induce more stringency to the analysis, but increases the chance of discarding useful data. Adjusting these parameters is a tradeoff, and the users are encouraged the limitation of their data. Details of this additional functionality in cWGAP is described (46).

3.4.4. Sequence accuracy and quality

"Finished sequence" refers to a region of DNA which has been closed to a point where there are no gaps or only well-characterized ones that cannot be resolved for biological reasons (http://www.genome.gov). According to the

human genome project (62), "finished sequence" must also be 99.99% accurate, containing error rates of less than 1 error in 10,000 assembled bases. This momentous endeavor created a gold-standard for accuracy of finished genomes (3). Subsequent projects that have similar high levels of support have also finished genomes to high standards, but many recently completed NGS genome projects, lacking the same financial resources, have generated low-quality drafts containing unresolved 'resolvable' gaps. Generating accurate genome sequences and genome annotation are important but time-consuming aspects of *de novo* genome sequencing projects. Since it was considered desirable for the APEC toolbox to contain high quality, finished genomic sequences, a significant amount of time and resources were expended to generate the best-finished quality sequences possible and make them available for future analysis.

3.4.5. Visualization and comparison method

The motivation behind building a new way to visualize and compare genomes (Figure 5) was to succinctly view a massive amount of genomic information and guide the investigator to genomic areas of interest. With this in mind, allowing flexibility of data tracks identifying interest areas is integral for future variation. Both Artemis and Mauve are very useful for comparative genomics, although the ability to add additional layers of information to their visual output that would aid in analysis is currently lacking. Although the reader is guided through the toolbox of analysis programs, their collective power is in their infinite flexibility with simple input by the addition of GenBank files. A user studying APEC can simply add another APEC GenBank file to gain a whole new set of comparison data, while another researcher could add a dataset of all

human ExPEC. Furthermore, adding new tracks of data such as global SNPs and other genomic islands can broaden analysis.

While Mauve does come with a visualization output, it was found to be insufficient for the specific end goals of data comparison. It uses the aforementioned MUSCLE alignment blocks and homologous region similarities to find areas of interest. Although this can be informative, many genomes are traditionally visualized circularly, as they are natively formed. Developing a method that can plug into the existing described pipeline using Circos (48) to create circular images is an optimal solution. Circos has enabled the user to create a base configuration file and base karyotype files developed from the outputted Genbank files from Prokka, the cWGAP Perl scripting (46), and external Perl scripts to create the initial circular diagram, found in many circular genomic maps. Circos additionally allows the user to create links based on the coordinates of the base karyotype files to show the relationship of the genes located in the karyotype files to each other, and most importantly: the core APEC genome karyotype. With the same data, a separate label file can be created that can label the genes and highlights and show the conserved regions of the genomes. In doing this, a large amount of data can be distilled to a single circular visualization that quickly shows the reader how all the genomes are interrelated though their core set of genes.

3.4.6. Summary

Elucidation of multiple genomes, as compared to single genomes, may be much more than additive, as it will enable the testing of many hypotheses that could not be evaluated with only a single APEC genome. For example, through

genomic comparisons of the strains described here, it may be possible to identify the mechanisms that are responsible for *E. coli*-caused respiratory disease, and septicemia, account for differences in the severity of colibacillosis associated with certain strains, or could be used to identify APEC control targets. Also, it may be possible through genomic comparison of all APEC and human ExPEC genomes to identify regions that are responsible for host specificity or that may be universal targets of future ExPEC vaccines. Similarly, such comparisons will enable identification of novel virulence genes and form the basis for future high-throughput comparative and functional analyses of the APEC genome. Many other beneficial outcomes of the proposed research to animal health are also expected. Additional analyses may uncover further insights into virulence. We also feel that this project has broader benefits that transcend just one group of production animals or one disease or even pathogenic bacteriology. For example, *E. coli* causes significant disease in all food animals (9); yet, the only genomes of pathogenic *E. coli* that exist do so because of their links to human disease. Certainly, this is true with APEC O1, the only fully sequenced pathogenic strain isolated from a non-human host. Although additional APEC genomes will better enable identification of unifying themes among APEC that can be exploited in disease control, they will also allow exploration of linkages between APEC and ExPEC of other food animals and human beings. Ultimately, these data may be able to determine if APEC are host specific, and if they are, determine what factors contribute to this tropism. If they are not host specific, then, these data will be helpful in determining if APEC are a threat to humans and other animals and if the ExPEC of humans and other animals are a threat to poultry. To better explore the possibilities of interspecies transmission of *E. coli*,

high throughput methods are needed to track strains in the production environment, through the food chain, and within the host environment. For example, with a multigenome APEC microarray, ramping up our tracking of APEC in the production environment and food chain from comparisons of isolates based on 200 or so genes to whole genome comparisons, targeting thousands of genes, giving great power to our observations. The success of such studies is totally dependent on the availability of high-quality, representative genomic sequences, making critical the work used here.

3.5. Acknowledgments

This work was supported by USDA-NIFA Award number: 0826675.

3.6. Supplementary Data Section

3.6.1. Figures

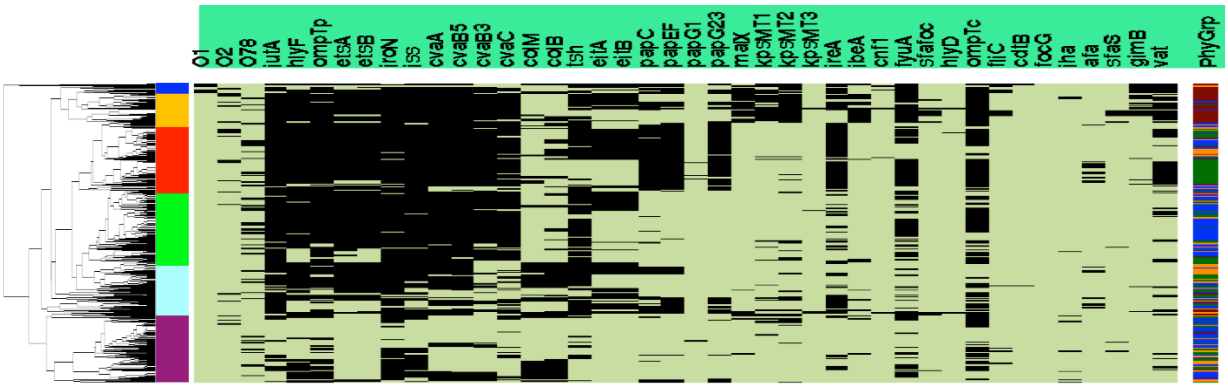

Figure 3.6-1 - 452 APEC were subjected to cluster analysis based on virulence genotype. To the right of the dendrogram is Column 1, highlighting each cluster with a different color (Cluster 1 = blue; 2 =mustard; 3 = red; 4 = green; 5 = light blue; and 6 = purple). Column 2 identifies all the O1 strains with a black bar; Column 3 = O2s; and Column 4 = O78s. The next 39 columns give the results for each isolate for each virulence gene, where black bar = gene is present; light green bar = gene is absent. Final Column = phylogenetic types with brown = B2; blue = A; orange = B1; and green = D. Note that all the O1 strains fall in the blue cluster (APEC O1 is the topmost isolate in the blue cluster) and lie in phylogenetic type B2. O2s are more variable and show some overlap with clusters containing O1 and O78 strains. Note, too, that there are several clusters of O78 strains in which no O1 or few O2 strains are found (green, light blue or purple clusters).

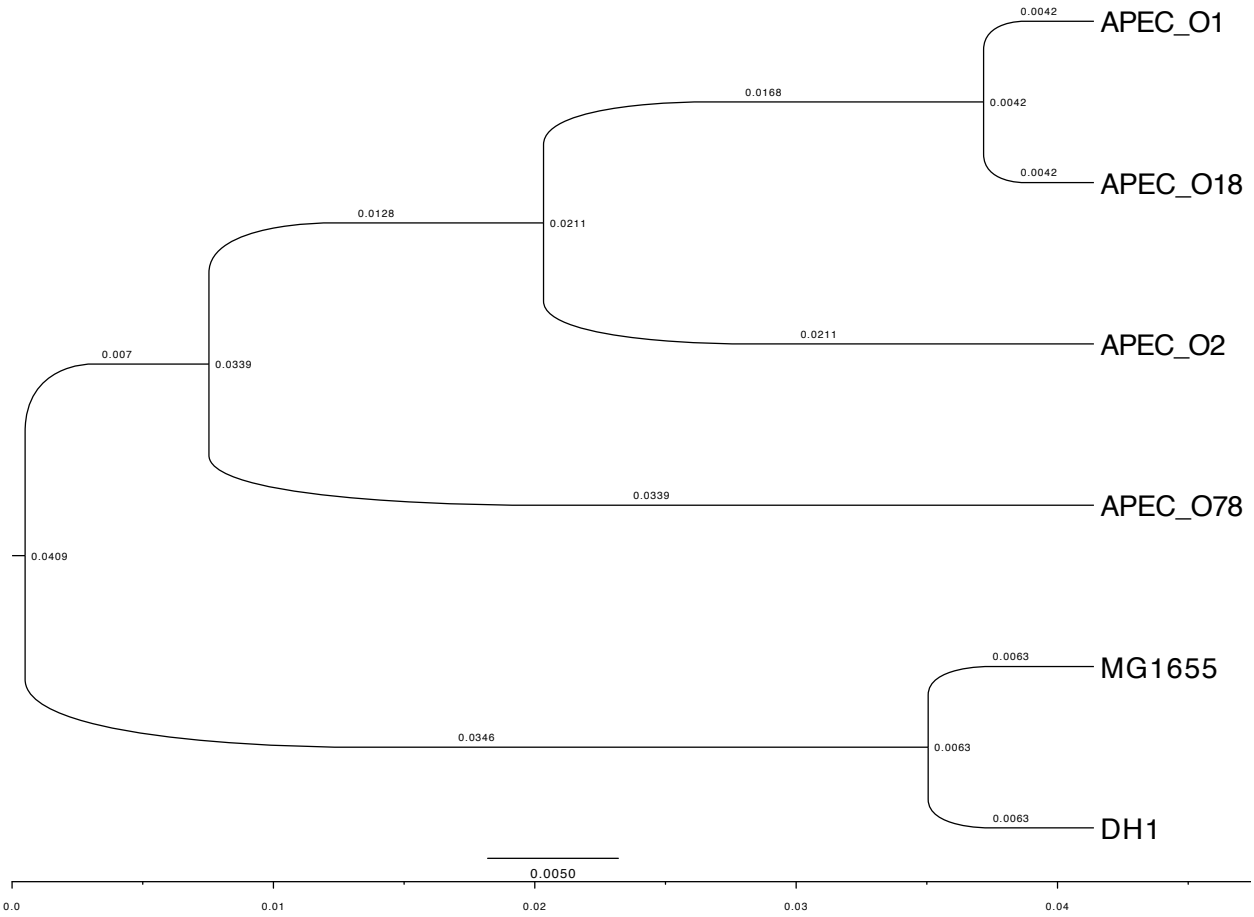

Figure 3.6-2 - Phylogenic analysis was performed using MrBayes 3.2.2 (53, 54). Phylogenies were created using the genes *chuA, yjaA,* and the core APEC genomic islands (GIs) is 3, 5, 6, and 7. Genes were aligned using Clustal Omega and used to generate a nexus file. MrBayes was run using a general time reversible model with variation between sites described as an independent gamma rate model using MG1655 and DH1 as roots. Number of generations was set to 100,000 with 25,000 burn-in cycles, and posterior probability cutoff was set at 99%.

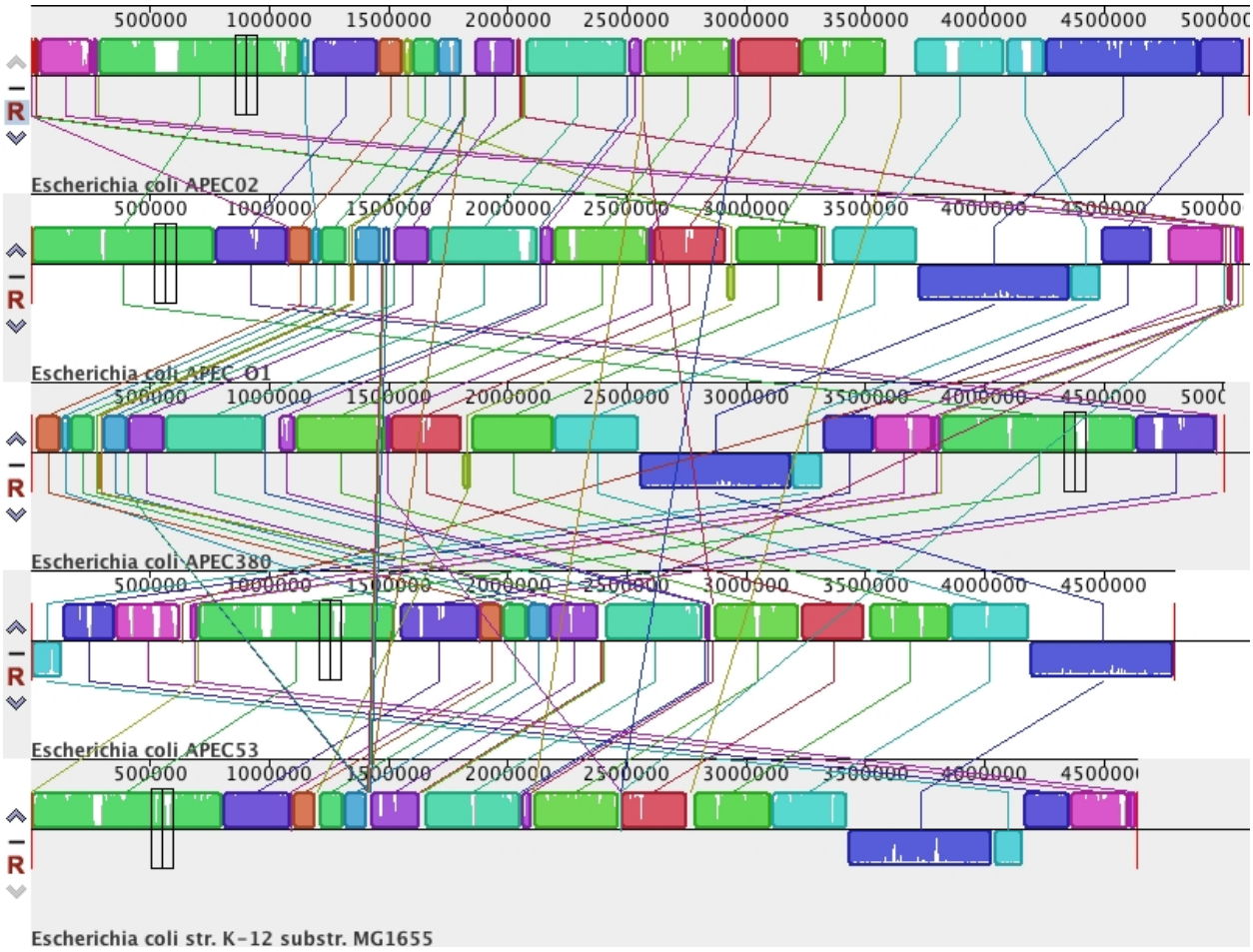

Figure 3.6-3 - Standard Mauve alignment visualization of all strains compared. APEC 380 = APEC O18, APEC53 = APEC O78.

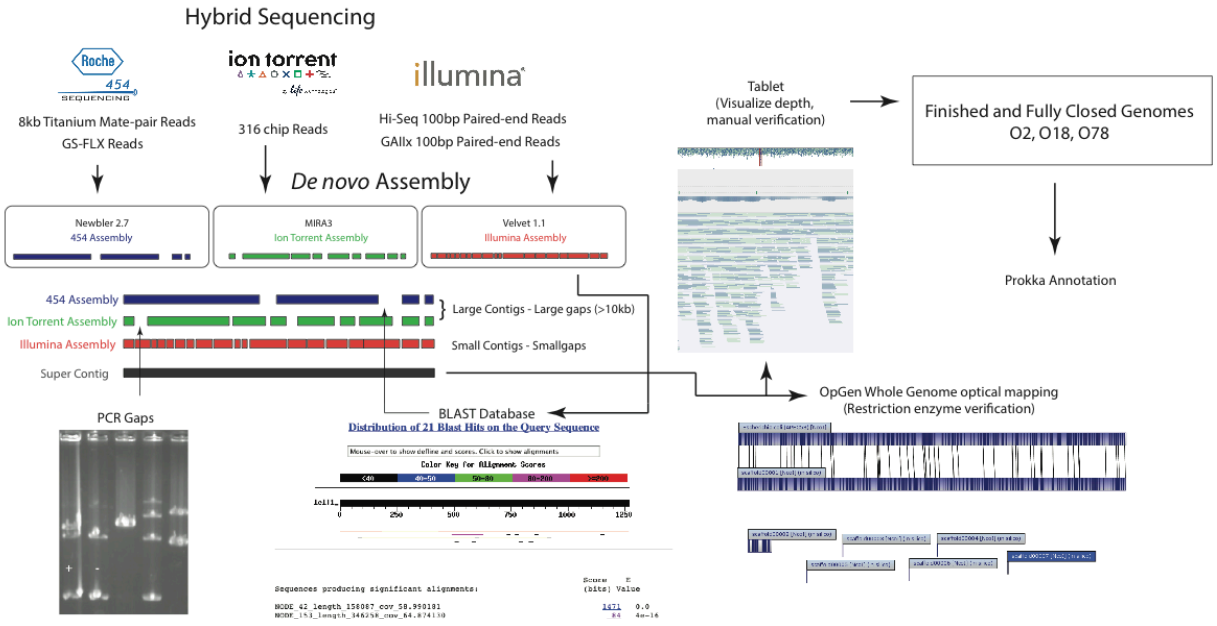

Figure 3.6-4 - A visual guide to how the entire APEC group was sequenced to such a high quality standard.

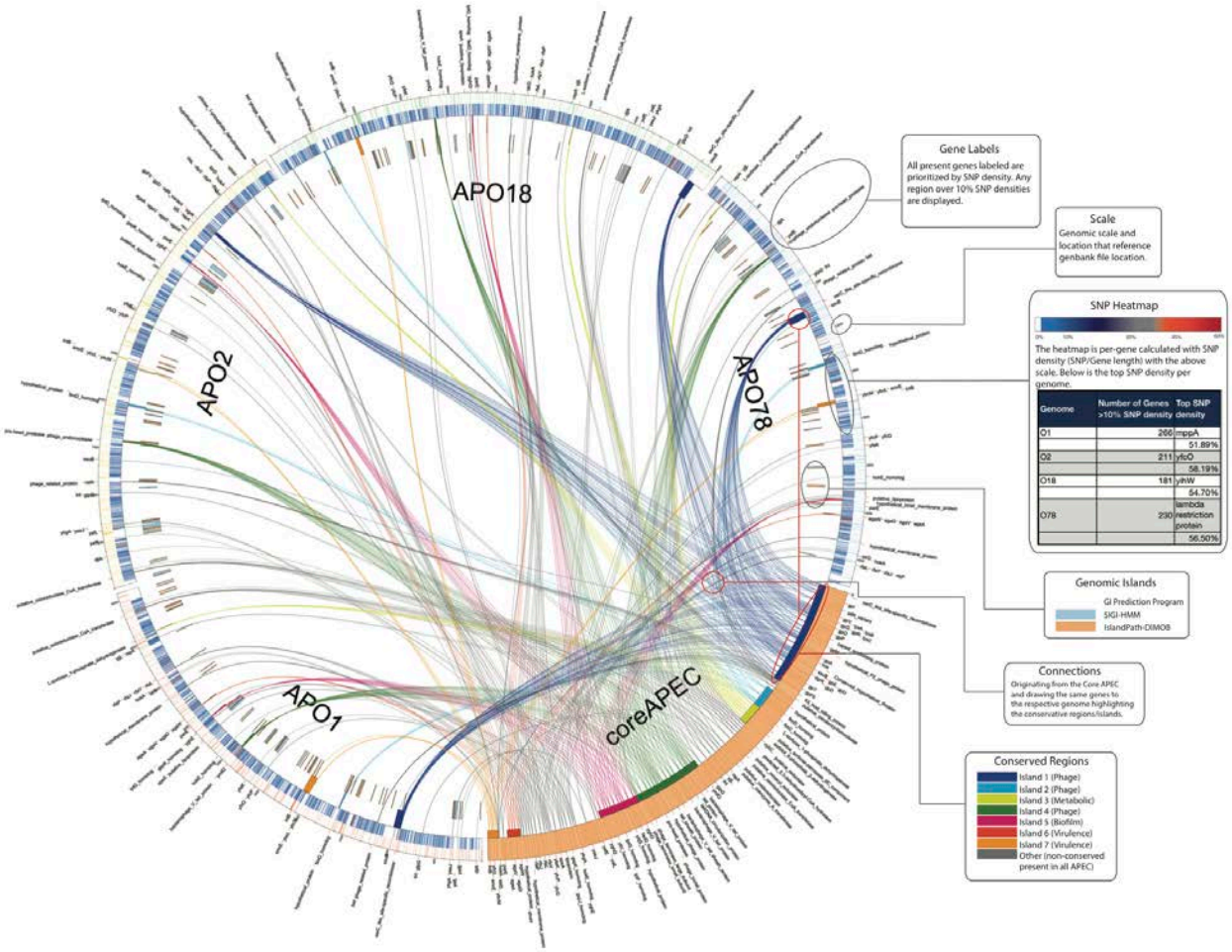

Figure 3.6-5 - Circos visualization output of cWGAP of representative APEC strains, with MG1655 subtracted. This figure is labeled for ease of identification.

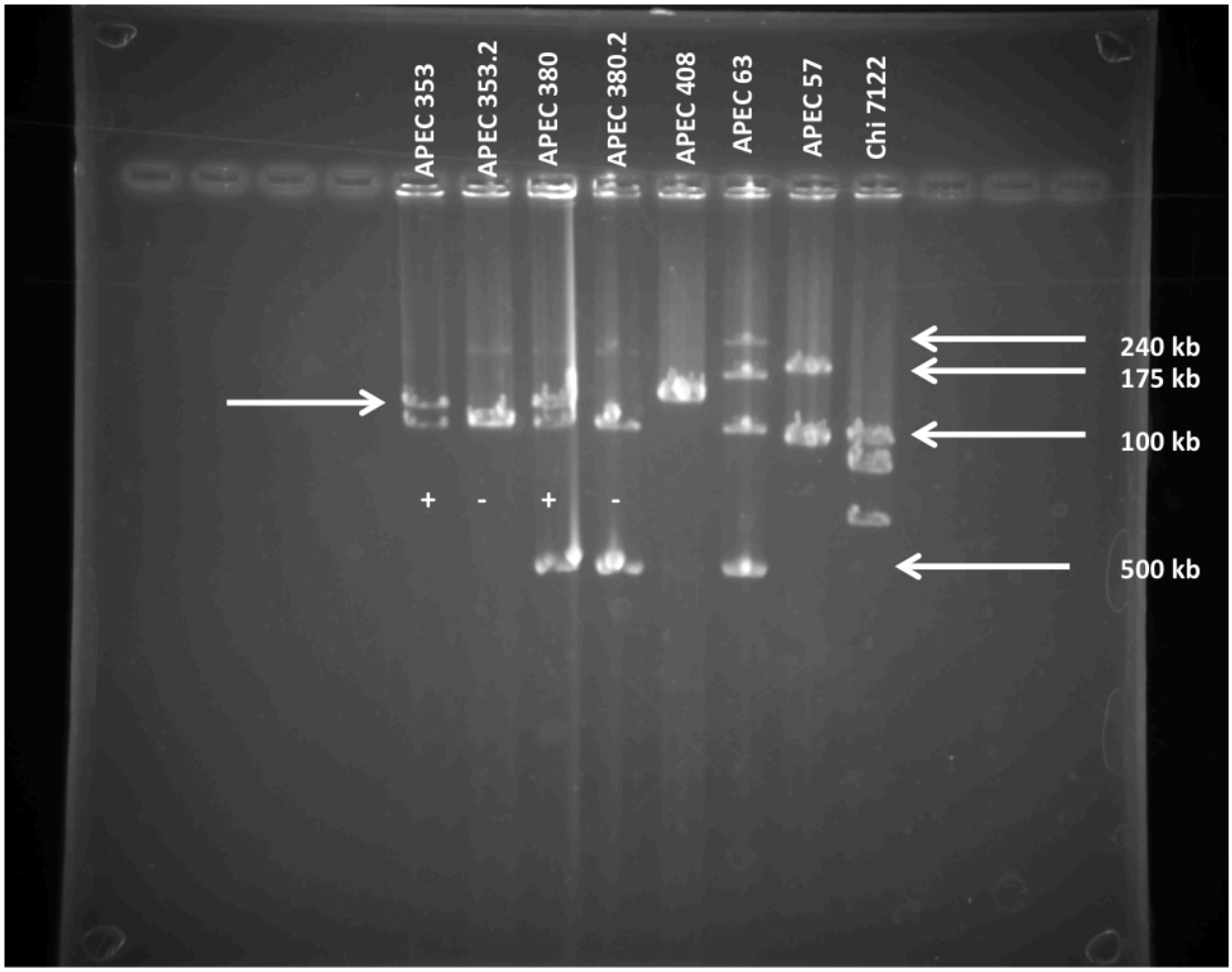

Figure 3.6-6 - APEC O18 (380) plasmid prep example showing sizes of plasmids isolated.

3.6.2. Tables

Table 3.6-1 - Characteristics of *E. coli* strains used in this study.

Strain	Isolation Source	Serogroup	K-1	Phylogenetic group	H-Type	Genome Length
APEC O1	The lung of a turkey	O1	+	B2	H7	5,082,025
APEC O2	Air sac of a chicken	O2	-	D	H4	5,112,508
APEC O18	Pericardium/lung chicken	O18	+	B2	H7	5,006,813
APEC O78	Lung of a turkey clinically diagnosed with colibacillosis	O78	-	A	H9	4,798,435

Table 3.6-2 - Phylogroup vs. serogroup analysis of all strains chosen.

	A	B1	B2	D	Totals
O1	0	0	6	1	7
O2	6	4	18	27	55
O18	1	1	10	1	13
O78	93	6	3	1	103
Totals	100	11	37	30	**178**

Table 3.6-3 – Core APEC identified islands - Examination of the core genome revealed large multi-gene clusters. BLAST searches were performed on these sections to ascribe a putative function label. The following image is a list of the genes by the putative function label and corresponding color to Figure 5's "Conserved Regions."

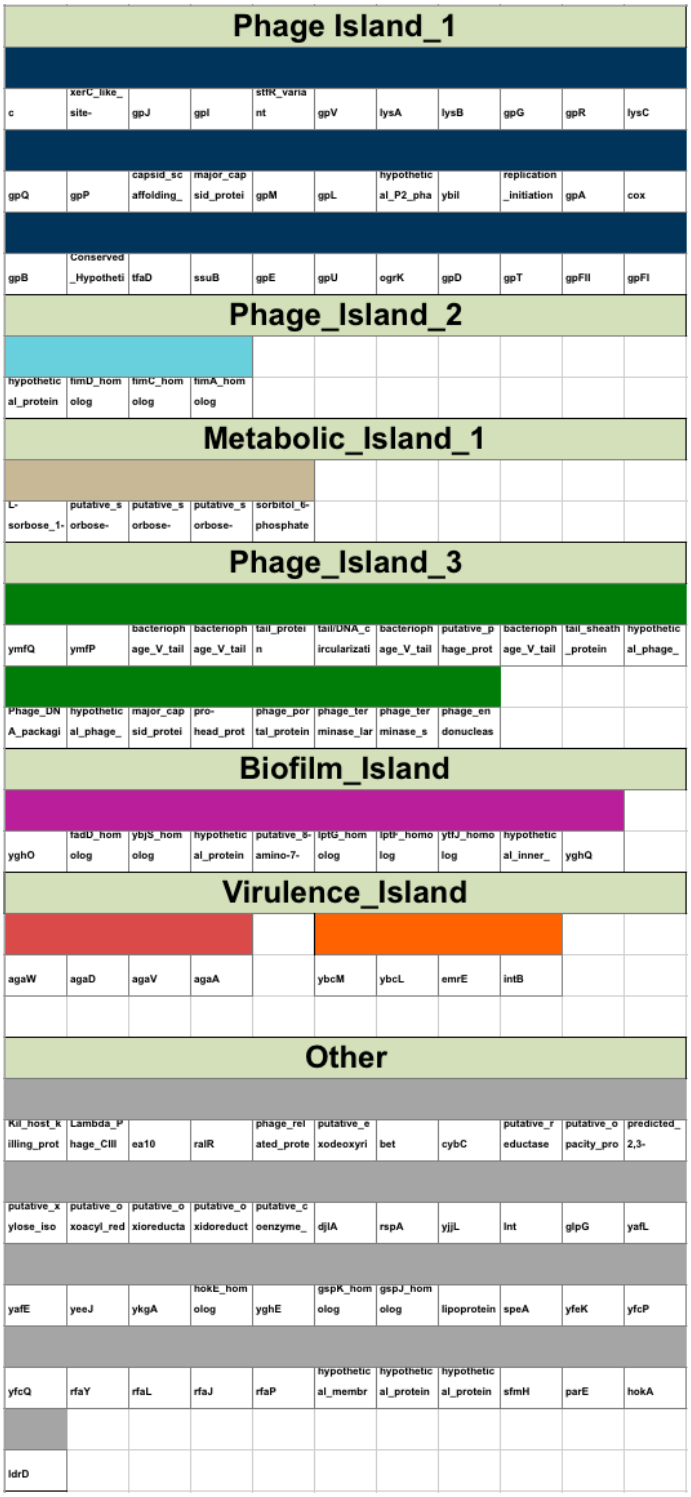

Table 3.6-4 – Leveraging new functions of cWGAP, this is a narrowing list of core APEC first using additional *K-12* subtraction (*K-12* genes highlighted in red) to arrive at a new core subset. Then creating subsequent subsets increasing gene overlap stringency. This table is highlights the optimized core at 20%.

0%	10%	20%	30%
APECO2 4411969 4412859 pstA			
APECO2 515192 515962 ycgJ	APECO2 515192 515962 ycgJ	APECO2 515192 515962 ycgJ	
APECO2 926261 927643 cusC			
APECO2 622253 623011 yafL_1			
APECO2 2010972 2012390 yedQ_1	APECO2 2010972 2012390 yedQ_1	APECO2 2010972 2012390 yedQ_1	APECO2 2010972 2012390 yedQ_1
APECO2 4764586 4765296 hypothetical_protein	APECO2 4764586 4765296 hypothetical_protein	APECO2 4764586 4765296 hypothetical_protein	APECO2 4764586 4765296 hypothetical_protein
APECO2 4765299 4765931 hypothetical_protein	APECO2 4765299 4765931 hypothetical_protein	APECO2 4765299 4765931 hypothetical_protein	APECO2 4765299 4765931 hypothetical_protein
APECO2 4766077 4766682 hypothetical_protein	APECO2 4766077 4766682 hypothetical_protein	APECO2 4766077 4766682 hypothetical_protein	APECO2 4766077 4766682 hypothetical_protein
APECO2 4766736 4768241 glpD			
APECO2 4600409 4600621 small_toxic_polypeptide	APECO2 4600409 4600621 small_toxic_polypeptide		
APECO2 4600754 4601181 cspA	APECO2 4600754 4601181 cspA		
APECO2 456830 458110 hemL			
APECO2 1373602 1374369 ssuB			
APECO2 3337404 3337556 small_toxic_polypeptide	APECO2 3337404 3337556 small_toxic_polypeptide		
APECO2 3337541 3337698 sok	APECO2 3337541 3337698 sok	APECO2 3337404 3337556 small_toxic_polypeptide	APECO2 3337404 3337556 small_toxic_polypeptide
APECO2 4963539 4964486 sorC	APECO2 4963539 4964486 sorC	APECO2 3337541 3337698 sok	APECO2 3337541 3337698 sok
APECO2 4964834 4965706 rluF			
APECO2 4097784 4099067 yjjL_3	APECO2 4097784 4099067 yjjL_3	APECO2 4097784 4099067 yjjL_3	APECO2 4097784 4099067 yjjL_3
APECO2 4099134 4100348 rspA_2	APECO2 4099134 4100348 rspA_2	APECO2 4099134 4100348 rspA_2	APECO2 4099134 4100348 rspA_2
APECO2 4100846 4102219 argH			
APECO2 3802107 3803999 parE			
APECO2 1581639 1583372 Phage_terminase-like_protein,_large_subunit	APECO2 1581639 1583372 Phage_terminase-like_protein,_large_subunit	APECO2 1581639 1583372 Phage_terminase-like_protein,_large_subunit	APECO2 1581639 1583372 Phage_terminase-like_protein,_large_subunit
APECO2 1583384 1583566 hypothetical_protein	APECO2 1583384 1583566 hypothetical_protein	APECO2 1583384 1583566 hypothetical_protein	APECO2 1583384 1583566 hypothetical_protein
APECO2 1583566 1584807 phage_portal_protein,_HK97_family	APECO2 1583566 1584807 phage_portal_protein,_HK97_family	APECO2 1583566 1584807 phage_portal_protein,_HK97_family	APECO2 1583566 1584807 phage_portal_protein,_HK97_family
APECO2 1584785 1585435 phage_prohead_protease,_HK97_family	APECO2 1584785 1585435 phage_prohead_protease,_HK97_family	APECO2 1584785 1585435 phage_prohead_protease,_HK97_family	APECO2 1584785 1585435 phage_prohead_protease,_HK97_family
APECO2 1585450 1586655 phage_major_capsid_protein,_HK97_family	APECO2 1585450 1586655 phage_major_capsid_protein,_HK97_family	APECO2 1585450 1586655 phage_major_capsid_protein,_HK97_family	APECO2 1585450 1586655 phage_major_capsid_protein,_HK97_family
APECO2 1586704 1586904 hypothetical_protein	APECO2 1586704 1586904 hypothetical_protein	APECO2 1586704 1586904 hypothetical_protein	APECO2 1586704 1586904 hypothetical_protein
APECO2 1586907 1587230 putative_phage_protein_(possible_DNA_packaging)	APECO2 1586907 1587230 putative_phage_protein_(possible_DNA_packaging)	APECO2 1586907 1587230 putative_phage_protein_(possible_DNA_packaging)	APECO2 1586907 1587230 putative_phage_protein_(possible_DNA_packaging)
APECO2 1587227 1587637 putative_phage_head-tail_adaptor	APECO2 1587227 1587637 putative_phage_head-tail_adaptor	APECO2 1587227 1587637 putative_phage_head-tail_adaptor	APECO2 1587227 1587637 putative_phage_head-tail_adaptor
APECO2 1587612 1588118 hypothetical_protein	APECO2 1587612 1588118 hypothetical_protein	APECO2 1587612 1588118 hypothetical_protein	APECO2 1587612 1588118 hypothetical_protein
APECO2 1989573 1992224 fimD_2	APECO2 1989573 1992224 fimD_2	APECO2 1989573 1992224 fimD_2	APECO2 1989573 1992224 fimD_2
APECO2 1992266 1992976 focC_1	APECO2 1992266 1992976 focC_1	APECO2 1992266 1992976 focC_1	APECO2 1992266 1992976 focC_1
APECO2 1993338 1993901 fimA_3	APECO2 1993338 1993901 fimA_3	APECO2 1993338 1993901 fimA_3	APECO2 1993338 1993901 fimA_3
APECO2 2428209 2429006 ureR	APECO2 2428209 2429006 ureR	APECO2 2428209 2429006 ureR	APECO2 2428209 2429006 ureR
APECO2 2429016 2429567 putative_kinase_inhibitor	APECO2 2429016 2429567 putative_kinase_inhibitor	APECO2 2429016 2429567 putative_kinase_inhibitor	APECO2 2429016 2429567 putative_kinase_inhibitor
APECO2 2429736 2430068 emrE	APECO2 2429736 2430068 emrE	APECO2 2429736 2430068 emrE	APECO2 2429736 2430068 emrE
APECO2 3557107 3557838 hypothetical_protein	APECO2 3557107 3557838 hypothetical_protein	APECO2 3557107 3557838 hypothetical_protein	APECO2 3557107 3557838 hypothetical_protein
APECO2 3557984 3559960 speA			
APECO2 1587612 1588118 hypothetical_protein	APECO2 1587612 1588118 hypothetical_protein	APECO2 1587612 1588118 hypothetical_protein	APECO2 1587612 1588118 hypothetical_protein
APECO2 1588115 1588675 hypothetical_protein	APECO2 1588115 1588675 hypothetical_protein	APECO2 1588115 1588675 hypothetical_protein	APECO2 1588115 1588675 hypothetical_protein
APECO2 1588684 1588854 hypothetical_protein	APECO2 1588684 1588854 hypothetical_protein	APECO2 1588684 1588854 hypothetical_protein	APECO2 1588684 1588854 hypothetical_protein
APECO2 1588838 1590334 Mu-like_prophage_tail_sheath_protein_gpL	APECO2 1588838 1590334 Mu-like_prophage_tail_sheath_protein_gpL	APECO2 1588838 1590334 Mu-like_prophage_tail_sheath_protein_gpL	APECO2 1588838 1590334 Mu-like_prophage_tail_sheath_protein_gpL
APECO2 1590334 1590690 Phage_tail_tube_protein	APECO2 1590334 1590690 Phage_tail_tube_protein	APECO2 1590334 1590690 Phage_tail_tube_protein	APECO2 1590334 1590690 Phage_tail_tube_protein
APECO2 1590690 1590959 hypothetical_protein	APECO2 1590690 1590959 hypothetical_protein	APECO2 1590690 1590959 hypothetical_protein	APECO2 1590690 1590959 hypothetical_protein
APECO2 1591101 1592936 phage_tail_tape_measure_protein,_TP901_family,_core_	APECO2 1591101 1592936 phage_tail_tape_measure_protein,_TP901_family,_c	APECO2 1591101 1592936 phage_tail_tape_measure_protein,_TP901_family,_c	APECO2 1591101 1592936 phage_tail_tape_measure_protein,_TP901_family,_
APECO2 1592997 1594325 Mu-like_prophage_DNA_circulation_protein	APECO2 1592997 1594325 Mu-like_prophage_DNA_circulation_protein	APECO2 1592997 1594325 Mu-like_prophage_DNA_circulation_protein	APECO2 1592997 1594325 Mu-like_prophage_DNA_circulation_protein
APECO2 1594322 1595401 Mu-like_prophage_tail_protein_gpP	APECO2 1594322 1595401 Mu-like_prophage_tail_protein_gpP	APECO2 1594322 1595401 Mu-like_prophage_tail_protein_gpP	APECO2 1594322 1595401 Mu-like_prophage_tail_protein_gpP
APECO2 1595401 1595949 phage_baseplate_assembly_protein_V	APECO2 1595401 1595949 phage_baseplate_assembly_protein_V	APECO2 1595401 1595949 phage_baseplate_assembly_protein_V	APECO2 1595401 1595949 phage_baseplate_assembly_protein_V
APECO2 1595949 1596374 Phage_protein_GP46	APECO2 1595949 1596374 Phage_protein_GP46	APECO2 1595949 1596374 Phage_protein_GP46	APECO2 1595949 1596374 Phage_protein_GP46
APECO2 1596361 1597419 hypothetical_protein	APECO2 1596361 1597419 hypothetical_protein	APECO2 1596361 1597419 hypothetical_protein	APECO2 1596361 1597419 hypothetical_protein
APECO2 1597410 1597994 hypothetical_protein	APECO2 1597410 1597994 hypothetical_protein	APECO2 1597410 1597994 hypothetical_protein	APECO2 1597410 1597994 hypothetical_protein
APECO2 1597998 1598921 hypothetical_protein	APECO2 1597998 1598921 hypothetical_protein		
K-12 Genes: 12	K-12 Genes: 7	K-12 Genes: 5	K-12 Genes: 5
Total Genes: 52	Total Genes: 42	Total Genes: 38	Total Genes: 37

Table 3.6-5 - Characteristics of *E. coli* strains used in this study.

Strain	Source	Plasmids (size(s) kb)	Serogroup	MLST	Genes associated with the conserved virulence region of APEC plasmids							
					*iut*A	*sit*A	RepFIB	*hly*F	*omp*T	*ets*AB	*iss*	*iro*N
DH5α[a]	-	0	NT	ST1060	-	-	-	-	-	-	-	-
APEC O1	(11)	4 (241,174,101,49)	O1	ST95	+	+	+	+	+	+	+	+
APEC O2	White	1 (199)	O2	ST117	+	+	+	+	+	+	+	+
APEC O78	Arkansas	2 (218,113)	O78	ST23	+	+	+	+	+	+	+	+
APEC O18	Nebraska	3 (131,110,41)	O18	ST95	+	+	+	+	+	-	+	+
APEC χ7122[c]	(13)	3 (103,90,60)	O78	ST23	+	+	+	+	+	+	+	+
APEC O2[d]	(63)	2 (180, 101)	O2	ST135	+	+	+	+	+	+	+	+

[a] Negative control for rat neonatal meningitis model and ELA
[b] Postive control for rat neonatal meningitis model
[c] Positive control for chick colisepticemia model
[d] Positive control for ELA

Table 3.6-6 – Details of APEC sequence assembly**.**

APEC O2	454 FLX Titanium GS			Illumina GAIIx			Illumina HiSeq2000			Ion Torrent		
	GS-FLX			100bp paired-end			100bp paired-end			316 chip		
	Reads	*Totaling*	*Fold*	*Reads*	*Totaling*	*Fold*	*Reads*	*Totaling*	*Fold*	*Reads*	*Totaling*	*Fold*
	638,908	255.1 Mbp	48.9	862,731	67,223,839	13.1	9,614,323	803.22 Mbp	157.1	2,039,822	319.2Mbp	62.4
	GS-FLX 8-kb mate-pair											
	447,236	153.6 Mbp	30									

APEC O18	454 FLX Titanium GS			Illumina HiSeq2000		
	GS-FLX			100bp paired-end		
	Reads	*Totaling*	*Fold*	*Reads*	*Totaling*	*Fold*
	235,653	95.8 Mbp	19.1	14,386,242	1,358.8 Mbp	274.4
	GS-FLX 8-kb mate-pair					
	219,416	67.6Mbp	13.5			

APEC O78	454 FLX Titanium GS			Illumina HiSeq2000		
	GS-FLX			100bp paired-end		
	Reads	*Totaling*	*Fold*	*Reads*	*Totaling*	*Fold*
	590,77	237Mbp	49	27,389,600	2,587Mbp	539
	GS-FLX 8-kb mate-pair					
	474,583	168Mbp	35			

CHAPTER 4. COMPARATIVE WHOLE GENOMIC ALIGNMENT PIPELINE – CWGAP

A paper to be submitted to *Nature Biotechnology*

Paul Mangiamele[1*], Bryon Nicholson[1], Aaron West[2], Torsten Seemann[3], and Lisa K. Nolan[1§]

4.1. Abstract and Introduction

Advancements in sequencing technology have driven an ever-growing body of genomic sequence data to new heights. Sequencing projects across all organisms are growing exponentially due to the feasibility afforded by next-generation sequencing (NGS) technology. This trend is accelerating read length per run, as well as significantly decreasing the cost per run, subsequently outpacing Moore's Law in the cost per genome for the past 6 years (64). The affordability of these systems and availability of sequencing services have made these technologies accessible to smaller laboratories focusing on individual biological organisms and systems. However, if the purpose of these projects is to answer research questions, data generation is only the beginning. A substantial

[1] Department of Veterinary Microbiology and Preventive Medicine, College of Veterinary Medicine, 1802 University Blvd, VMRI 2, Iowa State University, Ames, Iowa 50011

[*] Primary researcher and author

[2] Department of Chemistry, Iowa State University, Ames, Iowa 50010

[3] Victorian Bioinformatics Consortium, Monash University, Clayton, Victoria 3800, Australia

[§] Corresponding Author

bottleneck for many labs is the next steps taking the sequence data to biological insight, especially when the volume of data overwhelms paradigms for standard data analysis. Here, we present our tool, the comparative Whole Genomic Alignment Pipeline (cWGAP), which addresses this bottleneck by extending the functionality and visual aspects of comparative genomics programs, while focusing on making the process iterative and easy to use for the biologist end-user.

For biologists with limited bioinformatics skills, cWGAP provides an easy to use web-based interface that allows users to upload and compare GenBank files. The backend of cWGAP synchronizes many of the predefined views, options, and paradigms of the datasets entered. cWGAP visually presents the user an ever-growing selection of data output such as core and pan genomic data, SNP data, intergenic link data, and cluster analysis.

cWGAP was designed around genomic studies of *Escherichia coli*, an organism for which there is substantial gene content variability among individual isolates (65-67). In particular, a previous genomic project (65) describes the process and goals to find and visualize the pan and core genome through comparative genomics. Other comparative genomic tools did not have the specified functionality desired, necessitating the creation of our own tool that encompasses other sequencing functionalities such as identification of virulence and genomics islands, and provides expandability for future functionality as needed. cWGAP provides a broad range of functions and data flexibility not provided by publically available comparative genomic systems, including Artemis Comparison Tool (68, 69), VISTA (70), Ensembl Compara (71), BRIG

(72), CGAT (73), and UCSC Browser (74) (see Table 1 for comparison). Testing had also has shown cWGAP to be useful beyond our initial dataset.

A key feature of cWGAP is its visual comparison of multiple diverse genomes, allowing the user to quickly and easily extract a set of common genes. Additional functionality was built in by subtracting common genes from the comparison for more succinct results, as well the flexibility to add additional supplementary data tracks (i.e., SNP data, IslandViewer data) that the user requires.

Our Avian Pathogenic *E. coli* (APEC) comparison project (55) will be used as a case study to illustrate how cWGAP can be used to further the state-of-knowledge of a set of genomic data and how the results are used to generate an 'all-encompassing' genomic analysis figure (Figure 2).

cWGAP is available for use: http://cwgap.it

4.2. Methods

Each one of the following is a breakdown of an individual section of the cWGAP pipeline. A high-level flow diagram of cWGAP is shown in Figure 4.2-1. A visual breakdown of all individual parts is available in Figure 2. The examples presented in this chapter come from our APEC analysis (65). These data are available for download and further analysis on our server (http://ecoli.cvm.iastate.edu).

4.2.1. Overall cWGAP flow

The cWGAP pipeline is made up of many moving parts, although it can be broken down into the major features as shown in Figure 1. A user arriving at the website can drag and upload annotated GenBank files, specify the options

desired for analysis, and instruct the program to compare the genomes. This process initiates the cWGAP pipeline. The pipeline starts by processing the uploaded files by organizing and passing them through progressive Mauve for alignment (75). This results in an alignment backbone, which contains all aligned sequences that become the basis for the rest of the analysis. From here, the cWGAP 'Rosetta Stone' scripting aligns nucleotide regions, which are compared or subtracted based on user requirements to derive a core and pan genome, which is then referenced to the GenBank files to obtain a human-readable list of genes. These gene lists, corresponding to core and pan genomic data, are formatted for analysis in Circos (48). The translation results in three files: (1) a karyotype file, (2) a corresponding gene label file, and (3) a global links file for each genome. If the user chooses to visualize SNP data, the Mauve SNP script can be employed to build a SNP backbone, and a cWGAP extension (SNPScript) that builds heat maps into Circos format, aligning with the karyotype base pair locations. The end result is then displayed to the user via the web interface, and all pertinent data can be downloaded locally.

4.2.2. Sequence annotation pre-processing

The Prokka annotation pipeline (76) was used for consistent and cohesive automatic annotations before submitting them to cWGAP for comparative analysis. Prokka is optimized for bacterial genomes, offers enhanced annotation accuracy through the addition of custom databases, and provides fast annotations for easy iteration. Furthermore, being Perl-based, this annotation pipeline can easily be integrated into the cWGAP pipeline and can be fully downloaded and run locally. Though other automatic annotations like PGAP for

NCBI (PGAP), RAST (77), xBASE2 (78) will generate valid GenBank files for cWGAP, all genomes should be annotated using the same pipeline for consistent, coherent results. Progressive Mauve uses the FASTA sequence when developing the backbone alignment, thus annotation errors from other automatic annotation programs will not interfere with alignments but will appear in the human readable gene lists.

4.2.3. Mauve and the Backbone

Mauve is well described (57, 75, 79, 80) and has significant functionality beyond its visual interface implementation for which Mauve is most commonly used (See Figure 4). cWGAP leverages progressive Mauve (75) using positional homology multiple genome alignments to extend their previous method (57) to aligning regions conserved in subsets of the genomes. The progressive Mauve aligner offers a platform on which to base study of the combined effects of gene gain, loss, and rearrangement in microbial species and excels at aligning rearranged genomes with different gene content (75). Thus, progressive Mauve is ideal for processing genomic information into the cWGAP pipeline.

One of the primary output files of the Mauve alignment process is the backbone file. Progressive Mauve utilizes a revised backbone from the original Mauve backbone that observes alignment regions conserved among subsets of the genomes (75). The following is a breakdown of an interpretation of the Mauve backbone file. A snippet and description of our APEC alignment is shown in Table 1. These data are passed to the cWGAP 'Rosetta Stone' scripts for additional analysis. Lines two, four, and six are "core genome" lines since they

have homologous sequences among all the “compare” genomes (all APEC sequences) and lack homology with the “subtract” genome (MG1655).

4.2.4. The karyotype

Data visualization begins with translating the GenBank file into a Circos-readable base pair location file, hereby referred to as the karyotype file. Creating a translation from the GenBank format to a karyotype for each genome was a pivotal part of the visualization. Since the goal was not to re-invent the wheel for something as standard as GenBank files, the Genbank2Circosk.pl script from Texas A&M portal CLI Portal project (https://cpt.tamu.edu/cpt-software/portal/genbank2circosk.pl), written by Eric Rasche, was implemented. The script was neither fully automated, nor wrote the exact information needed for a Circos visualization; however, it accomplished much of what was needed. Modifying the code and releasing a new version of the script through the cWGAP program was necessary, and it is called as a dependency of the cWGAP scripts that can stand-alone.

The karyotype files, which emerge from the new genbank2circosk.pl script, represent each base pair index for the expressed gene from annotation. Circos performs a visualization of the base file and data ranges in circular form. The Circos chromosome definitions are formatted as follows in the example snippet of the APECO2 karyotype file:

```
BAND ID GENE_NUMBER GENE_NAME START END COLOR

band APECO2 1 hypothetical_protein 325 942 red
band APECO2 2 putative_transcriptional_regulator 966 1199 red
band APECO2 3 hypothetical_protein 1741 2025 red
band APECO2 4 hypothetical_protein 2361 2552 red
```

4.2.5. SNP analysis and visualization

The SNP analysis is performed using Mauve's SNP-calling function on the progressive Mauve alignment backbone data. In creating cWGAP, the Java package from the Mauve.jar (org.gel.mauve.analysis.SnpExporter) was extracted and ran independently via the pipeline. The SNP comparisons are sequence-to-sequence differences, thus for every polymorphic site in a genome alignment, the SNP file records the nucleotides present in each genome at that site, along with the sequence coordinates of the site in each genome. A sample output file snippet is listed in Table 4. This file format closely follows the backbone file, although outputs a line for every polymorphic site in an alignment. Each line shows the nucleotides present and sequence coordinates in each genome at that site. The SNP pattern displayed with sequences are ordered the same as when input for alignment, similar to the backbone.

The file by itself may be used for analysis, although it can be parsed further and format each result into a correlative visual format. SNPScript, a separate Perl script, was created to accomplish this by assigning SNPs to their respective genes and regions, and calculates the percentage SNPs within the gene compared to other genomes. Genes with fewer than 100 base pairs and genes containing SNP ratios in excess of 30% were manually examined for validity. Genes and polymorphisms are exported to a new track and visualized in Circos using a specified heat map scale. Genes containing the highest incidence of polymorphisms were labeled in the outside track for the APEC example. The following is an example snippet of the SNPScript processed file for APECO2:

```
chr start end SNP_Percentage

APECO2 69206 69281 0
APECO2 69317 69392 0
APECO2 69662 70801 0.0412642669007902
APECO2 70815 72347 0.0359007832898172
APECO2 72319 72780 0.017353579175705
```

4.2.6. Links and relationships

Link files contain base pair location coordinates for each individual genome within the core genome. These links make up the core functionality of Circos, as they show the relationship between each compared genome visually. Inherent in the subset of data from the backbone file, a map is generated where the core genome connects to every base pair range from its respective genome. This set of coordinates is processed into a Circos links file for each genome, and retain these values for other parts of the analysis. This part of the visualization not only creates cohesiveness – it also shows the user clusters of genomic data to easily identify genomic islands and other patterns within the data. The following is an example snippet of the APECO2 links file:

chr1	start1	end1	chr2	start2	end2
coregenome	1	1419	APECO2	2010972	2012390
coregenome	1420	4071	APECO2	1989573	1992224
coregenome	4072	4782	APECO2	1992266	1992976
coregenome	4783	5346	APECO2	1993338	1993901
coregenome	5347	6144	APECO2	2428209	2429006

4.2.7. cWGAP 'Rosetta Stone' Perl scripts

The cWGAP 'Rosetta Stone' Perl scripts are the heart of file processing in the pipeline. Here the Mauve backbone is taken and the data is separated into core and pan genomic groups. Much of the information of this nature is contained within the Mauve files, although it takes significant data manipulation and manual searching to pull the data into this format and casting them into a

Circos-based visualization. The goal was to automate many of these steps. Since much of the comparative genomics algorithmic work is handled by progressive Mauve, the main idea was to focus on putting that data into a visual format easily understood by biologist end-users. Here, the Perl scripts are described in detail, although there is no substitute for understanding how the scripts work through reading the Perl code. The cWGAP script can be viewed, downloaded, and run locally via the author's Github under an MIT use license (https://github.com/paulmm/cWGAP).

The script begins by using the karyotype axes from the original GenBank files; it builds the translation matrix and brings data together for comparison. The script parses data ranges available from each respective GenBank file into either a "compare" and "subtract" category, sets backbone ranges of each respective genome, and then pulls the ranges from both and holds the data that appear within those ranges. In addition, since there will be overlap with varying levels of similarity, a percentage overlap function was added so that the user can specify how much overlap in the genes and backbone is acceptable, i.e., the user sets the desired stringency for calling overlapping data. This step is iterative allowing the user to adjust the analysis to a particular dataset, beginning with the default setting of 0% overlap.

The scripts then create all permutations of all the genomes to create the pan and core genomic sets. In the case of the four APEC "compare" genomes (G1-G4) and one *E. coli* MG1655 "subtract" genome (sG1), the permutation looks like the following (1 being a positive match, 0 being a negative match):

G1	G2	G3	G4	sG1
1	1	1	1	0
1	1	1	0	0
1	1	0	1	0
1	0	1	1	0
0	1	1	1	0
0	1	1	0	0
		•••		
0	0	0	0	0

With all the pertinent data saved and compared, it is possible to output the relevant data for Circos to process visual data natively. The script takes care of casting the data output, and the shell script organizes the files into the project files to feed and run into Circos.

4.2.8. Web interface

Running cWGAP scripts is simplified through the shell scripting and is how the user sends their data through the entire pipeline. Since many biologists are not comfortable running Unix command line programs (81-83), an easy-to-use web interface was created that runs the pipeline automatically and eliminates the large list of dependencies needed to make the pipeline executable (See Table 3). Thus, using cWGAP allows the biologist end-user to focus on data output and not the time-consuming casting of data into the correct format for analysis and visualization.

In order to accomplish this task, modern web technologies were leveraged such as Ruby on Rails, HTML5, jQuery, and JavaScript. We also performed

iterations of user testing in order to streamline programmatic flow, followed by fine-tuning of the program based on user feedback to ensure the appropriateness of the interface to the biologist. It was a sincere intent to build a simple, yet powerful and effective interface enabling the user to accurately interpret and visualize their data and identify the core and pan genome of their genome comparison sets.

4.3. Results and Discussion

4.3.1. cWGAP motivation

The generation of genomic sequences of representative APEC was motivated by a desire to create a toolbox of genomic tools for community study of this important pathogen. In order to facilitate and direct future study of APEC, the project sought to identify a core set of genes (core APEC) that made an APEC an APEC. Here, the core APEC was defined as, genes found in all completely sequenced APEC (i.e., APEC O1, O2, O18 and O78), including those present in *E. coli* K-12 strains such as *E. coli* MG1565. In addition, the study sought to define the core pathogenome of APEC, i.e., all the shared genes in these strains minus the *E. coli* backbone as represented by *E. coli* MG1565. Unfortunately, no single tool was available that would accomplish these end-goals (Table 2). Mauve was the closest tool that showed reliable and pertinent data through comparative genomics, although the visual aspect was static in the way in which it displayed data. Consequently, the progressive Mauve algorithm was used to create a new visualization with the pan and core genomes requiring significant effort to extract. Since bacterial genomes are circular in nature, this project desired to visualize the pertinent data in a circular fashion, such as with

Circos. Therefore, a method was created to exchange information between progressive Mauve and Circos, while parsing through data, adding pertinent data tracks, and optimizing user flexibility. cWGAP accomplishes all of these goals and allows others to use other genome datasets to extract core and pan genomes.

4.3.2. Interface

It was of primary importance to create an interface that would be easy to use for biologists, while still outputting data of publication quality. In addition, since this process is computationally intensive, it needs to inform the user in real-time of the progress of the analysis and errors that arise along the way due to input or formatting complications. After significant iterative development and user testing we have arrived at cWGAP. Using the latest web development methods and technologies, we wanted to convey a modern web look and feel, as well as give users options to tailor analysis to their needs.

4.3.3. Privacy

cWGAP allows users to keep their work private, as the entire system can be used locally. A secure login and authentication was implemented for our web system to keep users safe and secure. Although, inherently using anything on the web in insecure (84). Thus, if privacy is of chief concern, a user can download the pipeline and run it locally behind firewalls in a secure environment for maximum privacy.

4.3.4. Local running of cWGAP and dependencies

To run cWGAP, Perl and a Linux/BSD OS are needed. Perl, like languages such as Python or Ruby, is an interpreted language. This means that the user does not need to compile the cWGAP code — it is read in by the Perl executable, which in turn interprets, compiles and runs the code. Table 3 gives a list of dependencies to download and compile on a user's local machine to execute the pipeline reliably. This installation can be time consuming; to reduce this unwieldy aspect of cWGAP, the web application version was created. We recommend using the web implementation for most projects, as the web version will consistently run with the latest updates.

4.3.5. Diagram visualization

Use of visual methods to process, organize, and make data accessible is innate to human understanding (85, 86). Representing genomic output data of assembled, annotated genomes with links and "attention areas" in visual form allows for a quick "birds-eye" view of massive amounts of data for rapid discovery. Leveraging human cognition pattern matching (87) over computational-only approaches to post-genomic assembly can facilitate confirmation of assemblies and annotations in comparative genomics (88, 89). Displaying output as visual maps of core and pan genomes will allow the biologists to employ their skill-set more effectively. Utilizing human originality and spatial intuition with a hybrid human–computer optimization framework will be necessary to enable the finishing process of data generated by NGS to keep pace with the progression of genomic sequencing technology.

A key part of the visualization is to guide the viewer's eyes to potential areas of interest. In the comparison of APEC, the goal was to identify putative virulence genes on the chromosome. Thus, focus was given to the flexible aspect of adding information to the visualization – in this case two specific tracks of data, calculating genomic islands through IslandViewer (49) and manually curating and determining islands via the core APEC subset of genes and looking for patterns, and how the overlap and intersection tracks with all relevant data.

4.3.6. Adding additional analysis into the visualization

Once a user creates a baseline cWGAP comparison, the sky is the limit for additional analysis. After the data are converted to a Circos format, it is easy to add karyotype and highlight ranges through Circos on the existing visualization. To illustrate the utility of this aspect, virulence data from IslandViewer (49) was added, manually curated the core APEC genome for genes, then highlighted these clusters and displayed them visually. The overlap of these areas was observed for identifying genes for further study (Figure 2) (65).

4.3.6.1. IslandViewer data track

Fully sequenced genomes and finished GenBank files from the Prokka annotation (76) were fed into the IslandViewer (49) input to predict genomic islands. Although IslandViewer has its own visualization extension, we wanted to compare and contrast against our own data. To do so, the respective CSV files were downloaded and extracted the predicted island coordinates into each track of data. This data track identifies genomic islands by GC% skew, codon usage, and occurrence of mobile genetic elements. Regions positive for genomic islands were compared to the Virulence Factor Database (58) to assess the likelihood that

these genomic islands are pathogenicity islands. The raw table of values by prediction method is listed in Table 10 in the Supplementary Data section, and visualized in the Figure 4 above under "Genomic Islands".

4.3.6.2. Handpicked islands

As the core genome was extracted, several gene clusters became identifiable. Using a handful of tools, including PortEco and NCBI Blast, the ability to find functions of each of the conserved regions is made easier. The next step was to build a table looking at each cutoff for locations and function to determine what is neighboring and separate within each genome. The final result is in Figure 2, clearly identifying conserved regions in the core and how they translate to the respective genomes.

Both of these data tracks may be integrated into the cWGAP pipeline in the future if users find the data helpful. Since many of these programs are Unix-based it would be simple to add their processes into the pipeline and add the options to the web interface. Utilization of the MIT license, users are encouraged to fork the project on Github to make these improvements themselves, which we can integrate into the global cWGAP changes.

4.3.7. License for use

An MIT License is required for use of this software (http://opensource.org/licenses/MIT) (https://github.com/paulmm/cWGAP). While this pipeline will work for many genomic sequences, some genomic data may work better than others. Thus, we encourage feedback in order to improve this pipeline, expand its applicability to different datasets, add functionality, and address any problems.

4.4. Supplementary Data section

4.4.1. Tables

Table 4.4-1 - Progressive Mauve backbone data - The first row labels the information contained by each column. The order in which GenBank files are added to progressive Mauve correlates to the backbone order. The following comparison backbone file is generated with APEC data: seq0 is APECO2, seq1 is APECO1, seq2 is APECO18, seq3 is APECO78 and seq4 is MG1655 (our "subtract" genome in this case). Each sequence contains a pair of columns, which denote the base pair location of the beginning (leftend) and end (rightend) of each homologous match. Each subsequent row below the label corresponds to a segment of DNA conserved among all five uploaded genomes. Thus, the second line indicates that the segment between coordinates 2011957-2012457 in the first genome is homologous to the segment between coordinates 1650339-1650839, 539454–539954, and 2354011-2354510 of the second, third and fourth genomes, respectively (all 500 base pairs in length). The zeros indicate a negative match, in this case *E. coli* MG1655 (a non-APEC *E. coli* strain), since the alignment is not present. Similarly, the third row indicates that the segment [3922319-3923236] in APECO1 is homologous to [2756638-2757555] in APECO18.

seq0_leftend	seq0_rightend	seq1_leftend	seq1_rightend	seq2_leftend	seq2_rightend	seq3_leftend	seq3_rightend	seq4_leftend	seq4_rightend
2011957	2012457	1650339	1650839	539454	539954	2354011	2354510	0	0
0	0	3922319	3923236	2756638	2757555	0	0	0	0
1990798	1994081	1638620	1641903	527735	531018	2328934	2332215	0	0
0	0	3918597	3922174	2752916	2756493	0	0	0	0
1798304	1798306	1453878	1453880	393995	393997	2162824	2162826	0	0
0	0	3918272	3918391	2752591	2752710	0	0	0	0
-4714886	-4715067	3907767	3907947	2742086	2742266	0	0	0	0
0	0	3893856	3898983	2728183	2733302	0	0	0	0

Table 4.4-2 - Comparative genomics comparison table.

	cWGAP	Mauve	Artemis Comparison Tool	VISTA	Ensembl Compara	BRIG	CGAT	MapView	UCSC Browser
Data Visualization Flexibility									
Core Genome Visualization	+	-	-	-	-	-	-	-	-
Pan Genome Visualization	+	-	-	-	-	-	-	-	-
Adding Visual data beyond program	+	-	-	-	-	-	-	-	-
Rapid data processing (> 15 min)	+	+			-				-
Shows links and rearrangement data	+	+	+	+	-	+	+	+	+
Circular comparisons	+	-	-	-	-	+	-	-	-
Fast interactive performance			+	+	-			+	+
Data Processing capacity									
Core Genome Computation	+	+	-	-	-	-	-	-	-
Pan Genome Computation	+	+	-	-	-	-	-	-	-
GenBank compatible	+	+	+	+	-	+	-	-	-
SAM/BAM Compatible	-	-	-	-	-		-	-	-
Pairwise comparisons	+	+	-	-	+	+	-	-	-
Subtract genome	+		-	-	-	-	-	-	-
No File Size Limit (Large genomes)	+	+	-	-	-	+	-	-	
Data hosting									
Host public datasets	+		+	+	-	-	-	+	+
Upload user data	+	+	+	-	-	-	-	-	-
Secure/Private	+	+		-	-	-		-	-
Local hosting	+	+	+	-	-	+	-	-	-
Remote hosting	+		+	+	-	-	+	-	+
Downloadable data files	+			+	-	-		-	+
Interface									
Modern Look and Feel	+	+	-	-	-	-	-	-	-
Command line	+	+	-	-	+			-	-
Desktop application	+	+	+	-	-	+	+	-	-
Web-based interface	+	+	+	+	-		-	+	+
OpenID/OAuth login	+		-	-	-			-	-
Easy to use	+	+	-	-	-	+		-	-
Biologist-centric	+	+	-	-	-	-	-	-	-
Unix OS	+	+	+	+	+	+	+	+	+
Windows OS	+	+	+	+	-	+	+	+	+
Mac OS	+	+	+	+	-	+	+	+	+
Other OS	+	-	-	+	-	-	-	+	+

+ Positive
Not Applicable
- Negative

Table 4.4-3 - List of program dependencies.

Global	gbk2circosk	Circos	
Linux/BSD OS	common::sense	Carp	List::Util
At least Perl v5.14.2	YAML::XS	Clone	Math::Bezier
	JSON::XS	Config::General	Math::BigFloat
	Scalar::Util::Numeric	Cwd	Math::Round
	Bio::SeqIO	Data::Dumper	Math::VecStat
	file::Slurp	Digest::MD5	Memoize
	Moose	File::Basename	POSIX
		File::Spec::Functions	Params::Validate
		File::Temp	Pod::Usage
		FindBin	Readonly
		Font::TTF::Font	Regexp::Common
		GD	Set::IntSpan
		GD::Image	Storable
		GD::Polyline	Sys::Hostname
		Getopt::Long	Text::Balanced
		IO::File	Text::Format
		List::MoreUtils	Time::HiRes

Table 4.4-4 - SnpExporter snippet – this file format closely follows the backbone file (Table 1), although outputs a line for every polymorphic site in an alignment. Each line shows the nucleotides present and sequence coordinates in each genome at that site. The SNP pattern displayed with sequences are ordered the same as when input for alignment, similar to the backbone.

SNP pattern	sequence_1_Contig	sequence_1_PosInContg	sequence_1_GenWidePos1	sequence_2_Contig	sequence_2_PosInContg	sequence_2_GenWidePos2	sequence_3_Contig	sequence_3_PosInContg	sequence_3_GenWidePos3	sequence_4_Contig	sequence_4_PosInContg	sequence_4_GenWidePos4	sequence_5_Contig	sequence_5_PosInContg	sequence_5_GenWidePos5
aca--	APECO2	4252531	4252531	APECO1	4359006	4359006	APECO18	3192252	3192252	null	0	0	null	0	0
gag-g	APECO2	4252527	4252527	APECO1	4359010	4359010	APECO18	3192256	3192256	null	0	0	MG1655	4038106	4038106
cct-t	APECO2	4252519	4252519	APECO1	4359018	4359018	APECO18	3192264	3192264	null	0	0	MG1655	4038114	4038114
gaa-a	APECO2	4252518	4252518	APECO1	4359019	4359019	APECO18	3192265	3192265	null	0	0	MG1655	4038115	4038115
Tatgg	APECO2	4252514	4252514	APECO1	4359023	4359023	APECO18	3192269	3192269	APECO78	6	6	MG1655	4038119	4038119
Tcttt	APECO2	4252511	4252511	APECO1	4359026	4359026	APECO18	3192272	3192272	APECO78	9	9	MG1655	4038122	4038122
Aaatt	APECO2	4252507	4252507	APECO1	4359030	4359030	APECO18	3192276	3192276	APECO78	14	14	MG1655	4038127	4038127

4.4.2. Figures

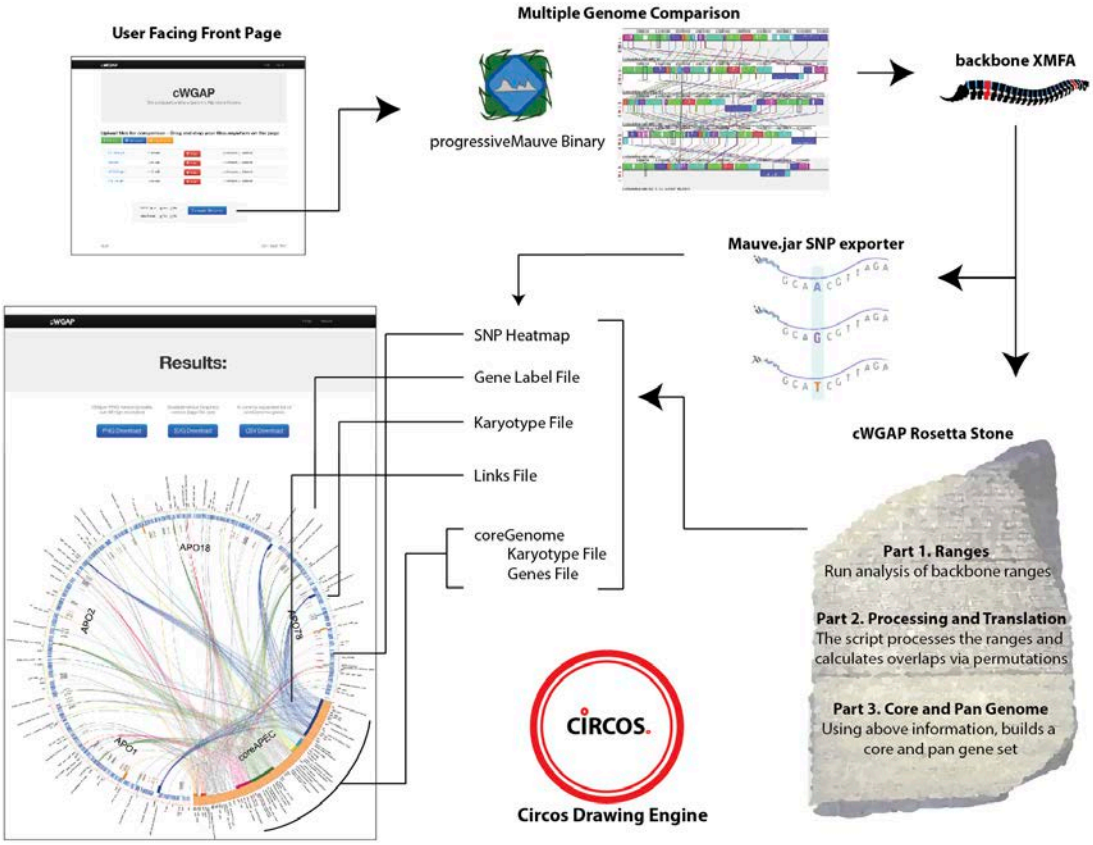

Figure 4.4-1 - Programmatic flow of cWGAP in action.

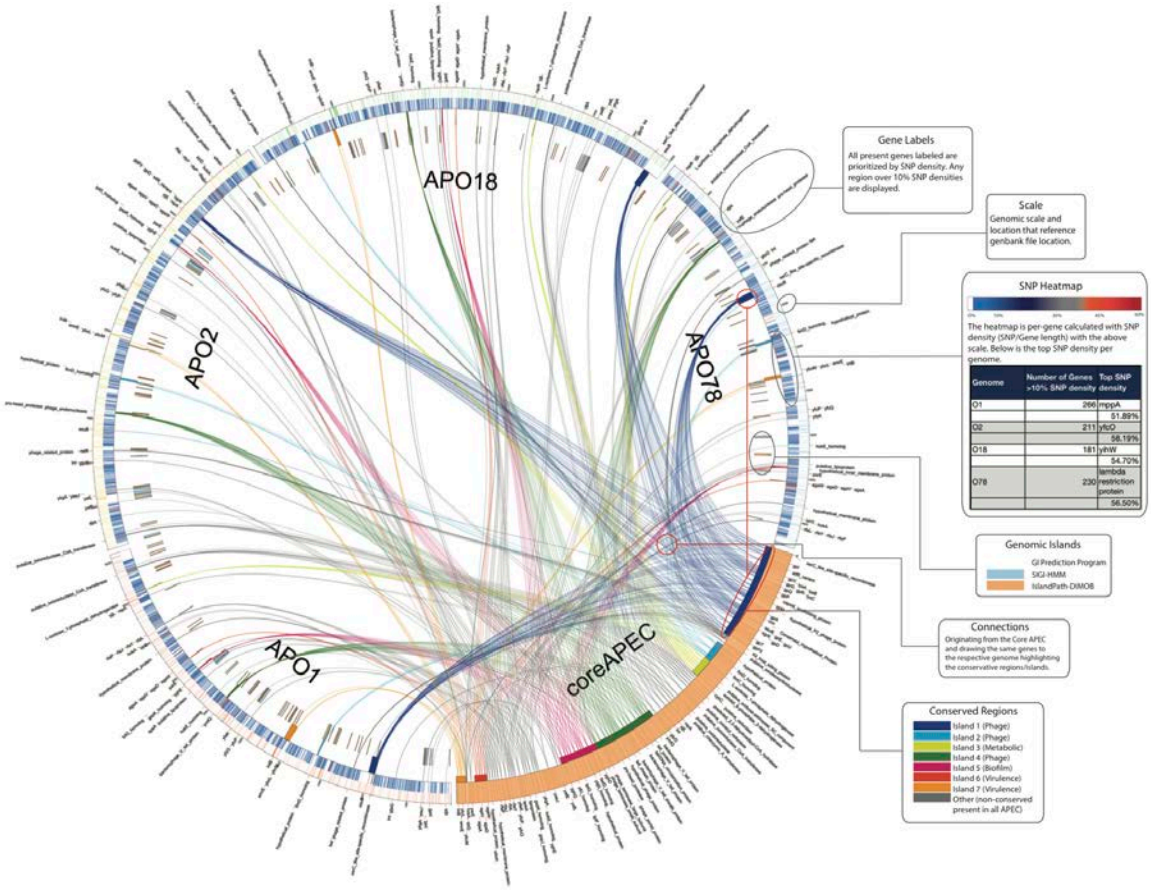

Figure 4.4-2 - Circos visualization of representative APEC strains.

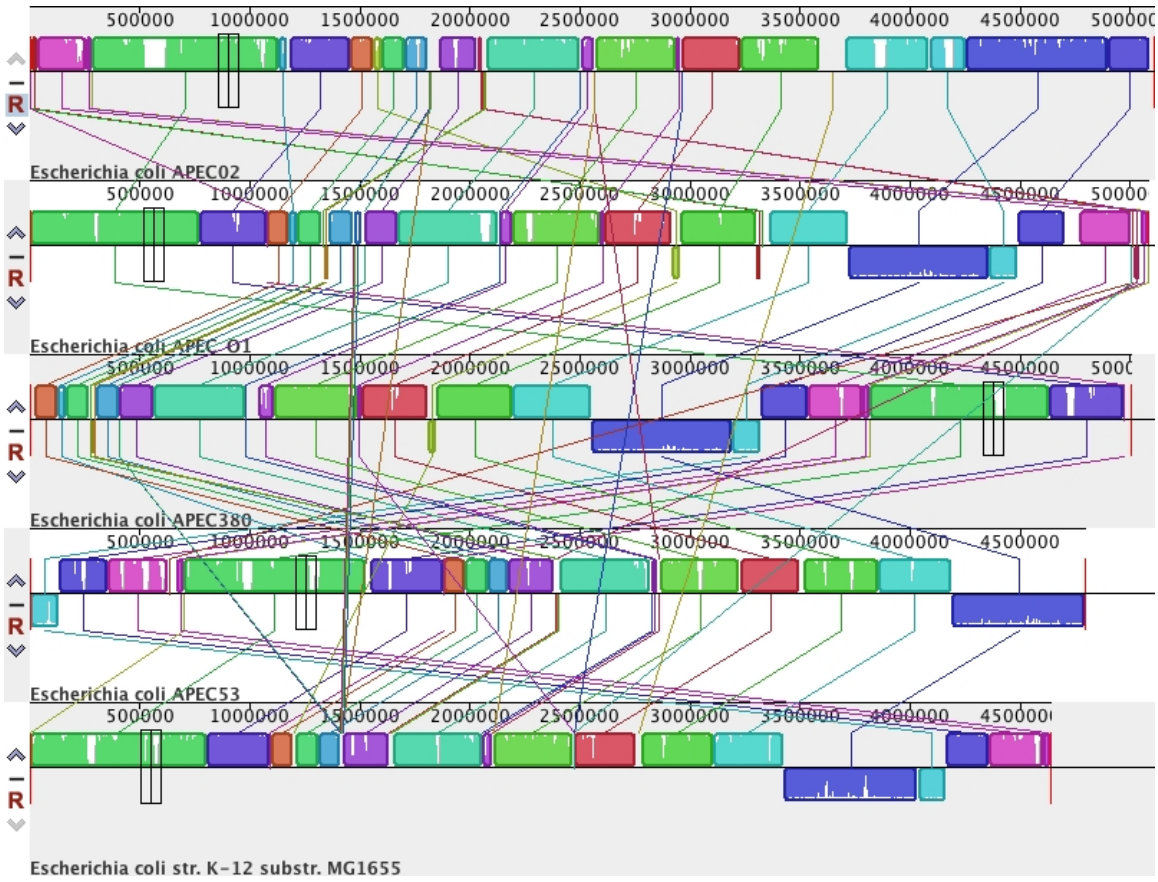

Figure 4.4-3 - Mauve alignment of all strains.

CHAPTER 5. GENERAL CONCLUSIONS

5.1. Summary

Having one foot planted in next generation biological sequencing, and the other firmly planted in human computer interaction, it is the author's goal to create useful tools geared for the biologist end-user. Spending the last six years exploring how the user interacts with bioinformatics data, the author has leveraged human computer interaction practices, user-centered design, and user experience and a unique viewpoint was gained for the creation of bioinformatics tools which are truly useful for their audience. The papers presented are publication breadcrumbs for how this has been accomplished.

Chapter 2 introduced APEC O78, a genome on which significantly improved assembly methods created a refined final genomic resolution. Leveraging multiple sequencing runs for respective long and short read technologies, the process merged the runs together with our published process.

Chapter 3 investigated a similar process to the assembly methods outlined in Chapter 2 on multiple finished genomes. Expanding this into a genome-wide association analysis to include sequence analysis for each genome with respect to: comparative genomic analysis, core and pan genome analysis, vaccine epitope analysis, polymorphism and SNP analysis, phylogeny, genomic island identification, gene prevalence analysis, and visualization of all genomes compared. Thus, a truly two pronged toolbox for (1) exploring APEC pathogenesis representing the most diverse set of APEC sequenced yet and (2)

creating a methodology for a new way to use comparative genomics programmatically and visually.

Chapter 4 fully documented and described the methods for the creation of our new visual comparative genomics program, cWGAP, which was outlined in Chapter 3, and describes how it can be used in other areas of genomics and the greater NGS scientific community. The paper is an announcement for use of the web interface, program, and code for full public use and improvement, creating a group effort to become the definitive process of comparative genomics.

This dissertation completes a circle of research that states the objectives, designs and implements research to address the stated objectives, and finally disseminates the results through publishing and releasing the tools and code for public use to benefit the scientific community as a whole. The implications of this entire research effort benefit (1) the community of APEC researchers by fully exploring the most diverse set of APEC sequences that underpin vital research long into the future, thus safeguarding the public's cheapest source of high-quality protein for the future, and (2) the bioinformatic and genomics research community by creating a new platform of visualizing the core and pan genome while being flexible to add other pertinent data to the study, thus expanding to other genomes of human health importance.

BIBLIOGRAPHY

1. GOLD (Genomes OnLine Database).

2. Liolios K, Mavromatis K, Tavernarakis N, & Kyrpides NC (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 36(Database issue):D475-479.

3. Chain PS, *et al.* (2009) Genomics. Genome project standards in a new era of sequencing. *Science* 326(5950):236-237.

4. Barnes HJ, Nolan LK, & Vaillancourt J-P eds (2008) *Diseases of Poultry* (Iowa State University, Ames, IA), 12 Ed, pp 691-716.

5. Gross WB (1984) *Colibacillosis* (Ames, IA) 8th Ed.

6. Gross WG (1994) *Escherichia coli in Domestic Animals and Humans* (CAB International, Wallingford, U.K).

7. Gyles CL (1994) Escherichia coli in Domestic Animals and Humans. (CAB International, Wallingford, Oxon, UK).

8. Gyles CL, Prescott, J.F., Songer, J.G., and Thoen, C.O., eds (2004) Escherichia coli. Pathogenesis of Bacterial Infections in Animals, 3rd ed. *Iowa State University Press*.

9. Gyles GL (1993) *Escherichia coli. In: Pathogenesis of Bacterial Infections in Animals* (Iowa State University Press, Ames, IA) 2nd Ed.

10. Morris M (1989) Poultry Health Issue. *Poultry Times* July 3:11.

11. Johnson TJ, *et al.* (2007) The Genome Sequence of Avian Pathogenic *Escherichia coli* Strain O1:K1:H7 Shares Strong Similarities with Human Extraintestinal Pathogenic *E. coli* Genomes. *Journal of Bacteriology* 189(8):3228-3236.

12. Rojas TC, *et al.* (2012) Draft genome of a Brazilian avian-pathogenic Escherichia coli strain and in silico characterization of virulence-related genes. *J Bacteriol* 194(11):3023.

13. Dziva F, *et al.* (2013) Sequencing and functional annotation of avian pathogenic Escherichia coli serogroup O78 strains reveal the evolution of E. coli lineages pathogenic for poultry via distinct mechanisms. *Infect Immun* 81(3):838-849.

14. Zerbino DR & Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18(5):821-829.

15. Barnes HJ, Vaillancourt, J.-P., and and Gross, W.B. (2003) *Diseases of Poultry* (Iowa State University Press, Ames, IA) 11th Ed.

16. Barnes HJ & Gross WB (1997) Colibacillosis. *Disease of Poultry 10th ed.*), pp 131-141.

17. Gross WG & Gyles CL (1994) Diseases due to Escherichia coli in Poultry. *Escherichia coli in domestic animals and humans*, (CAB International, Wallingford), pp 237-259.

18. Lymberopoulos MH, *et al.* (2006) Characterization of Stg fimbriae from an avian pathogenic *Escherichia coli* O78:K80 strain and assessment of their contribution to colonization of the chicken respiratory tract. *J. Bacteriol.* 188(18):6449-6459.

19. Stocki SL, Babiuk LA, Rawlyk NA, Potter AA, & Allan BJ (2002) Identification of genomic differences between *Escherichia coli* strains pathogenic for poultry and *E. coli* K-12 MG1655 using suppression subtractive hybridization analysis. *Microb. Pathog.* 33(6):289-298.

20. Li G, Laturnus C, Ewers C, & Wieler LH (2005) Identification of genes required for avian Escherichia coli septicemia by signature-tagged mutagenesis. *Infect Immun* 73(5):2818-2827.

21. Morrow BJ, Graham JE, & Curtiss R, 3rd (1999) Genomic subtractive hybridization and selective capture of transcribed sequences identify a novel Salmonella typhimurium fimbrial operon and putative transcriptional regulator that are absent from the Salmonella typhi genome. *Infect Immun* 67(10):5106-5116.

22. Johnson TJ, *et al.* (2008) Comparison of extraintestinal pathogenic Escherichia coli strains from human and avian sources reveals a mixed subset representing potential zoonotic pathogens. *Appl Environ Microbiol* 74(22):7043-7050.

23. Mora A, *et al.* (2013) Poultry as reservoir for extraintestinal pathogenic Escherichia coli O45:K1:H7-B2-ST95 in humans. *Vet Microbiol* 167(3-4):506-512.

24. Tivendale KA, *et al.* (2010) Avian-pathogenic Escherichia coli strains are similar to neonatal meningitis E. coli strains and are able to cause meningitis in the rat model of human disease. *Infect Immun* 78(8):3412-3419.

25. Gordon DM, Clermont O, Tolley H, & Denamur E (2008) Assigning Escherichia coli strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environ Microbiol* 10(10):2484-2496.

26. Clermont O, Bonacorsi S, & Bingen E (2000) Rapid and simple determination of the Escherichia coli phylogenetic group. *Appl.Environ.Microbiol.* 66(10):4555-4558.

27. Clermont O, Christenson JK, Denamur E, & Gordon DM (2013) The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ Microbiol Rep* 5(1):58-65.

28. Zerbino DR (2010) Using the Velvet de novo assembler for short-read sequencing technologies. *Curr Protoc Bioinformatics* Chapter 11:Unit 11 15.

29. Milne I, *et al.* (2013) Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform* 14(2):193-202.

30. Seemann T (2013) Prokka.

31. Hyatt D, *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.

32. Laslett D & Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* 32(1):11-16.

33. Lagesen K, *et al.* (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 35(9):3100-3108.

34. Keseler IM, *et al.* (2013) EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res* 41(Database issue):D605-612.

35. Keseler IM, *et al.* (2011) EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Res* 39(Database issue):D583-590.

36. Finn RD, Clements J, & Eddy SR (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39(Web Server issue):W29-37.

37. Finn RD, *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res* 38(Database issue):D211-222.

38. Punta M, *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res* 40(Database issue):D290-301.

39. Borodovsky M (1993) GeneMark: parallel gene recognition for both DNA strands. *Computers & Chemistry* 17(19):123-133

40. Delcher AL, Harmon D, Kasif S, White O, & Salzberg SL (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 27(23):4636-4641.

41. Lukashin AV & Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* 26(4):1107-1115.

42. Lowe TM & Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25(5):955-964.

43. Riley M, *et al.* (2006) Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. *Nucleic Acids Res* 34(1):1-9.

44. Blattner FR, *et al.* (1997) The complete genome sequence of Escherichia coli K-12. *Science* 277(5331):1453-1462.

45. Darling AC, Mau B, Blattner FR, & Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14(7):1394-1403.

46. Mangiamele P, Nicholson B, West A, Seemann T, & Nolan LK (2014) Comparative Whole Genomic Alignment Pipeline - cWGAP. *Unpublished*.

47. He Y, Xiang Z, & Mobley HL (2010) Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. *J Biomed Biotechnol* 2010:297505.

48. Krzywinski M, *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19(9):1639-1645.

49. Langille MG & Brinkman FS (2009) IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* 25(5):664-665.

50. Waack S, *et al.* (2006) Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics* 7:142.

51. Hsiao W, Wan I, Jones SJ, & Brinkman FS (2003) IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics* 19(3):418-420.

52. Dhillon BK, Chiu TA, Laird MR, Langille MG, & Brinkman FS (2013) IslandViewer update: Improved genomic island discovery and visualization. *Nucleic Acids Res* 41(Web Server issue):W129-132.

53. Ronquist F, *et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 61(3):539-542.

54. Huelsenbeck JP & Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17(8):754-755.

55. Mangiamele P, *et al.* (2013) Complete genome sequence of the avian pathogenic Escherichia coli strain APEC O78. *Genome announcements* 1(2):e0002613.

56. Medini D, Donati C, Tettelin H, Masignani V, & Rappuoli R (2005) The microbial pan-genome. *Curr Opin Genet Dev* 15(6):589-594.

57. Darling AC, Mau B, Blattner FR, & Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 14(7):1394-1403.

58. Chen L, *et al.* (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res* 33(Database issue):D325-328.

59. Rodriguez-Siek KE, *et al.* (2005) Comparison of *Escherichia coli* Isolates Implicated in Human Urinary Tract Infection and Avian Colibacillosis. *Microbiology* 151(Pt 6):2097-2110.

60. Rodriguez-Siek KE, Giddings CW, Doetkott C, Johnson TJ, & Nolan LK (2005) Characterizing the APEC pathotype. *Vet. Res.* 36(2):241-256.

61. Sojka WJ & Carnaghan BA (1961) Escherichia coli Infection in Poultry. *Res.Vet.Sci.* 2:340-352.

62. International Human Genome Sequencing C (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431(7011):931-945.

63. Johnson TJ, Johnson SJ, & Nolan LK (2006) Complete DNA Sequence of a ColBM Plasmid from Avian Pathogenic *Escherichia coli* Suggests that it Evolved from Closely Related *ColV* Virulence Plasmids. *Journal of Bacteriology* 188(16):5975-5983.

64. National Human Genome Research Institute (NHGRI) Genome Sequencing Program (GSP) N (2014) DNA Sequencing Costs.

65. Mangiamele P, *et al.* (2014) Toolbox for Exploring Avian Pathogenic Escherichia coli (APEC) Pathogenesis, Host Specificity, Evolution and Control. *PLoS One*.

66. Perna NT, *et al.* (2001) Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature* 409(6819):529-533.

67. Welch RA, *et al.* (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proc Natl Acad Sci U S A* 99(26):17020-17024.

68. Abbott JC, Aanensen DM, Rutherford K, Butcher S, & Spratt BG (2005) WebACT--an online companion for the Artemis Comparison Tool. *Bioinformatics* 21(18):3665-3666.

69. Carver TJ, *et al.* (2005) ACT: the Artemis Comparison Tool. *Bioinformatics* 21(16):3422-3423.

70. Mayor C, *et al.* (2000) VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* 16(11):1046-1047.

71. Vilella AJ, *et al.* (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* 19(2):327-335.

72. Alikhan NF, Petty NK, Ben Zakour NL, & Beatson SA (2011) BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402.

73. Uchiyama I, Higuchi T, & Kobayashi I (2006) CGAT: a comparative genome analysis tool for visualizing alignments in the analysis of complex evolutionary changes between closely related genomes. *BMC Bioinformatics* 7:472.

74. Miller W, *et al.* (2007) 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* 17(12):1797-1808.

75. Darling AE, Mau B, & Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5(6):e11147.

76. Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*.

77. Aziz RK, *et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75.

78. Chaudhuri RR, *et al.* (2008) xBASE2: a comprehensive resource for comparative bacterial genomics. *Nucleic Acids Res* 36(Database issue):D543-546.

79. Darling AE, Treangen TJ, Messeguer X, & Perna NT (2007) Analyzing patterns of microbial evolution using the mauve genome alignment system. *Methods Mol Biol* 396:135-152.

80. Darling AE, Tritt A, Eisen JA, & Facciotti MT (2011) Mauve assembly metrics. *Bioinformatics* 27(19):2756-2757.

81. Paul Mangiamele AP, Bryon Nicholson, and Lisa K. Nolan (2013) Development and Evaluation of a Web-Based Comparative Genomics Database for Biologists: BioComb. *In prep.*

82. Powell DR & Seemann T (2013) VAGUE: a graphical user interface for the Velvet assembler. *Bioinformatics* 29(2):264-265.

83. Blankenberg D, *et al.* (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* Chapter 19:Unit 19 10 11-21.

84. Aspnes J, Feigenbaum J, Mitzenmacher M, & Parkes D (2003) Towards Better Denitions and Measures of Internet Security. *EECS Harvard*.

85. Miller G (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The psychological review* 63:81-97.

86. Glicksohn A & Cohen A (2011) The role of Gestalt grouping principles in visual statistical learning. *Atten Percept Psychophys* 73(3):708-713.

87. Eiben CB, *et al.* (2012) Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nat Biotechnol* 30(2):190-192.

88. Nagarajan N, Read TD, & Pop M (2008) Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics* 24(10):1229-1235.

89. Valouev A, Zhang Y, Schwartz DC, & Waterman MS (2006) Refinement of optical map assemblies. *Bioinformatics* 22(10):1217-1224.