1-1-2017

Development Of Student Data Visualization Tool, Adaption Of Clostridium Difficile Toxin A Into Protein Delivery Vehicle, And Elucidation Of Tcdc Mechanism Of Toxin Control

Adam Boyden
Wayne State University,

DEVELOPMENT OF STUDENT DATA VISUALIZATION TOOL, ADAPTION OF *CLOSTRIDIUM DIFFICILE* TOXIN A INTO PROTEIN DELIVERY VEHICLE, AND ELUCIDATION OF TCDC MECHANISM OF TOXIN CONTROL

by

ADAM BOYDEN

THESIS

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

In partial fulfillment of the requirements

for the degree of

MASTERS OF SCIENCE

2017

MAJOR: CHEMISTRY (Biochemistry)

Approved by:

_____

Advisor Date

DEDICATION

I would like to dedicate this work to my family, both blood and by marriage. I am grateful for the unwavering support and love from all of you. And most of all to my wife Kendra, who has supported and loved me more than I deserve.

ACKNOWLEDGEMENTS

I would first like to acknowledge Dr. Andrew Feig. He was very perceptive to my career goals and afforded me every opportunity to hone my skills as an instructor and continually work toward becoming a professor. He has also ingrained critical thinking and keeping the big picture in mind at all times. I would also like to acknowledge Dr. Tom Pentecost. He has continually inspired me by being a fun, entertaining, thoughtful, and thorough teacher and mentor. Your constant guidance has kept me on track when there was no light at the end of tunnel. I am also grateful to my colleagues Brianna Jackman and Amit Kumar. Amit spent extensive time showing me the way in lab and collaborating on the chimera project. Bri was a constant source of joy and humor as we reminded each other to have fun and enjoy the little things in life. Lastly, I would like to acknowledge my wife Kendra for unquestioning support and love throughout the graduate school process. Her faith in me as an educator and person has given me strength I did not know I had.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A_{495}: Absorbance at 495 nm
A_{280}: Absorbance at 280 nm
Antp: Antennapedia
BAB: Bax-TcdA-BoNT
Bax: Bcl-2-Associated X protein
BLAST: Basic Local Alignment Search Tool
BLyS: B Lymphocyte Stimulator
B. megaterium: *Bacillus megaterium*
***B. meg*:** *B. megaterium*
BoNT: Botulinum Nuero-Toxin A- Heavy Chain
BSA: Bovine Serum Albumin
$CaCl_2$: Calcium chloride
CAB: Casp9-TcdA-BoNT
CAE: Casp9-TcdA-EGF
Casp9: Caspase-9
CAV: Casp9-TcdA-VEGF
cDNA: complementary DNA
CPD: Cysteine protease domain
cPCR: Colony PCR
CPP: Cell penetrating peptide
CROP: C-terminal repeating oligopeptide
CV: column volume
DNA: Deoxyribonucleic acid
DTT: Dithiothreitol
EDTA: Ethylenediaminetetraacetic acid
EGF: Epidermal Growth Factor
Elong.: Elongation
ETC: 1-(3-Sulfopropyl)-2-(2-{[1-(3-sulfopropyl)naphtho[1,2-d]thiazol-2(1*H*)-ylidene]methyl}-1-butenyl)naphtho[1,2-d]thiazoliumhydroxide inner salt, triethylammonium salt
EtBr: Ethidium bromide
EtOH: Ethanol
FPLC: Fast performance liquid chromatography
FRET: Foerster resonance energy transfer
G4: G-quartet, G-quadruplex, G-tetrad
GGG-FITC: Glycine-glycine-glycine fluorescein isothiocyanate
GTD: Glucosyltransferase domain
HCl: Hydrochloric acid
HEPES: 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
His6: Six-histidine affinity purification tag
HRP: Horseradish peroxidase
HS GPA: High school grade point average
IP_6: Inositol hexakisphosphate
IPTG: Isopropyl-β-D-1-thiogalactopyranoside

kb: kilobase
KCl: Potassium chloride
Kcpm: kilo counts per minute
kDa: kilodalton
L: Liter
LAB: Luf-TcdA-BoNT
LAE: Luc-TcdA-EGF
LAB: Luc-TcdA-VEGF
LB: Lysogeny broth
Luc: Luciferase
M: Molar
mM: Millimolar
mL: Milliliter
min: Minute
MWCO: Molecular weight cutoff
NaCl: Sodium chloride
$NaPO_4$: Sodium phosphate
NEB: New England BioLabs
$NiSO_4$: Nickel (II) sulfate
nM: nanomolar
NP-40: Tergitol-type NP-40, nonyl phenoxypolyethoxylethanol
OD_{600}: Optical density at 600 nm
Oligo: Oligonucleotide
p35: Cyclin-Dependent Kinase 5 activator 1
PAB: p35-TcdA-BoNT
PAE: p35-TcdA-EGF
PAGE: Polyacrylamide gel electrophoresis
PAV: p35-TcdA-VEGF
PBS: Phosphate buffered saline
PCR: Polymerase Chain Reaction
Pfu: Polymerase from *P. furiosus*
PI: Phosphor imaging
PVDF: Polyvinylidene fluoride
QGRS: Quadruplex forming G-rich sequences
PNK: Phosphonucleotide kinase
RBD: Receptor binding domain
RhlB: ATP-dependent RNA helicase RhlB
RNA: Ribonucleic acid
SDS: Sodium dodecyl sulfate
Sfp: 4'-phosphopantetheinyl transferase
SrtA: Sortase A
T-10: TOP10 Electrocompetent cells
Tat: HIV trans-activator of transcription
Taq: Polymerase from *T. aquaticus*
TBE: Tris boric acid EDTA buffer
TcdA: Toxin A from *C. difficile*

TcdB: Toxin B from *C. difficile*
$TcdC^{C}$: TcdC from *C. difficile* with C-terminal His6
$TcdC^{152C}$: TcdC truncate starting at residue 152, with C-terminal His6
$TcdC^{152N}$: TcdC truncate starting at residue 152, with N-terminal His6
TCEP: *tris*(2-carboxyethyl)phosphine
TD: Translocation domain
Temp: Temperature
Tn: Transposon, Transposable element
tRNA: transfer RNA
Tris: Tris(hydroxymethyl)aminomethane
µM: micromolar
µm: micrometer
µg: microgram
µL: microliter
V: Volts
VEGF: Vascular Endothelial Growth Factor
XIAP: X-linked inhibitor of apoptosis protein

Introduction

This thesis is comprised of two primary areas of study; the first chapter focuses on pedagogical research and the following two on bench biochemistry research. At the beginning, work was focused solely on bench chemistry, developing cell specific protein delivery vehicles. Soon, work was split between the bench and developing a pedagogical tool for visualizing and analyzing longitudinal progression of student cohorts through various majors. As the chimera project began to stall, bench work became centered on elucidating a possible mechanism of toxin production within the pathogenic *Clostridium difficile*, while continuing pedagogical research. Eventually, frustrations working on bench-chemistry lead to focusing solely on the student data visualization tool and pedagogical research.

Chapter 1
Development of a New Student Data Visualization Tool: Changing the Paradigm of Data-Driven Decision Making

Introduction

In recent years, colleges and universities have invested significant resources towards improving student success.[1-3] When students succeed in college, they become productive members of their community and promote the advancement of their society. When students do not complete their courses of study, a portion of responsibility falls on the institution to assess the students' needs and supply the necessary resources for success. Student attrition leads to a loss of time and money invested for the student and the institution. Assessing student success has traditionally focused on academic output, but the issues are multi-dimensional. Researchers are beginning to expand their focus to psychosocial and financial aspects, and the role they play in student success.[4-6] Unfortunately, many of these investigations are carried out at the highest levels of an institution, and data rarely filter down to those faculty and staff closest to the issues. The goal of this work was to create a data visualization tool that can generate actionable outcomes from faculty and staff at all levels within an institution.

Emphasis on improving student success has lead to the emergence of learning analytics, focusing on aspects of students' lives and experiences, to find factors that can improve student learning and success.[7-9] Learning analytics has lead many to ask the question, "What leads students into academic trouble?" Answering that question requires parsing complex data on student performance, sometimes making it difficult to supply clear and actionable answers. Visualization tools help guide institutions or stakeholders through the complicated data, portray

it in an easy to follow manner, and allow for large amounts of information to be analyzed simultaneously to discern patterns. The ethics of student data analytics is highly discussed and it is important to note that for the privacy of faculty and students all courses, dates, and fields of study used herein have been de-identified.[10, 11]

The current paradigm of student data analytics relies heavily on predictive algorithms that analyze student demographic information and prior academic performance, or current data, to flag students at-risk for sub-optimal success in "real time".[12, 13] While these algorithms have been used towards many productive ends, they contain inherent limitations. These tools often fail to analyze the longitudinal progression of student cohorts. Also, many of the algorithms are end-point focused using graduation or a single class outcome as the measure of success, which may define the problem too narrowly. Some learning analytic tools, such as the Open Learning Initiative, have a wider definition of success and focus on student learning outcomes in individual courses, but tend to have microscopic focus within foundational courses prone to being a barrier to student success.[14] Most predictive algorithms do little to address the student deficit model, or the belief that students' own deficits lead to a lack of academic success, by not analyzing longitudinal effects of curriculum.[15-17] A truly effective analytic tool should focus on the curriculum and the student in parallel.

A class of visualization tools that depict student migration complement predictive algorithms.[18, 19] These tools allow administrators or faculty to determine populations of students who do not graduate with their initial major, or demonstrate how students migrate within a particular school of study. Changing majors clearly can affect time to degree and attrition from college but does not repair the more fundamental problems that might exist within the program. Migration plots are not sufficiently granular. They focus on starting and/or ending points without

addressing specific barriers that might catalyze student migration. Our tool uses a java generated user interface (UI) to present student data in a curriculum-centered fashion highlighting students' performance for each class in sequence, and not only identifying graduation as an endpoint (Figure 1.1). This tool lends itself to go beyond current student data, which most predictive

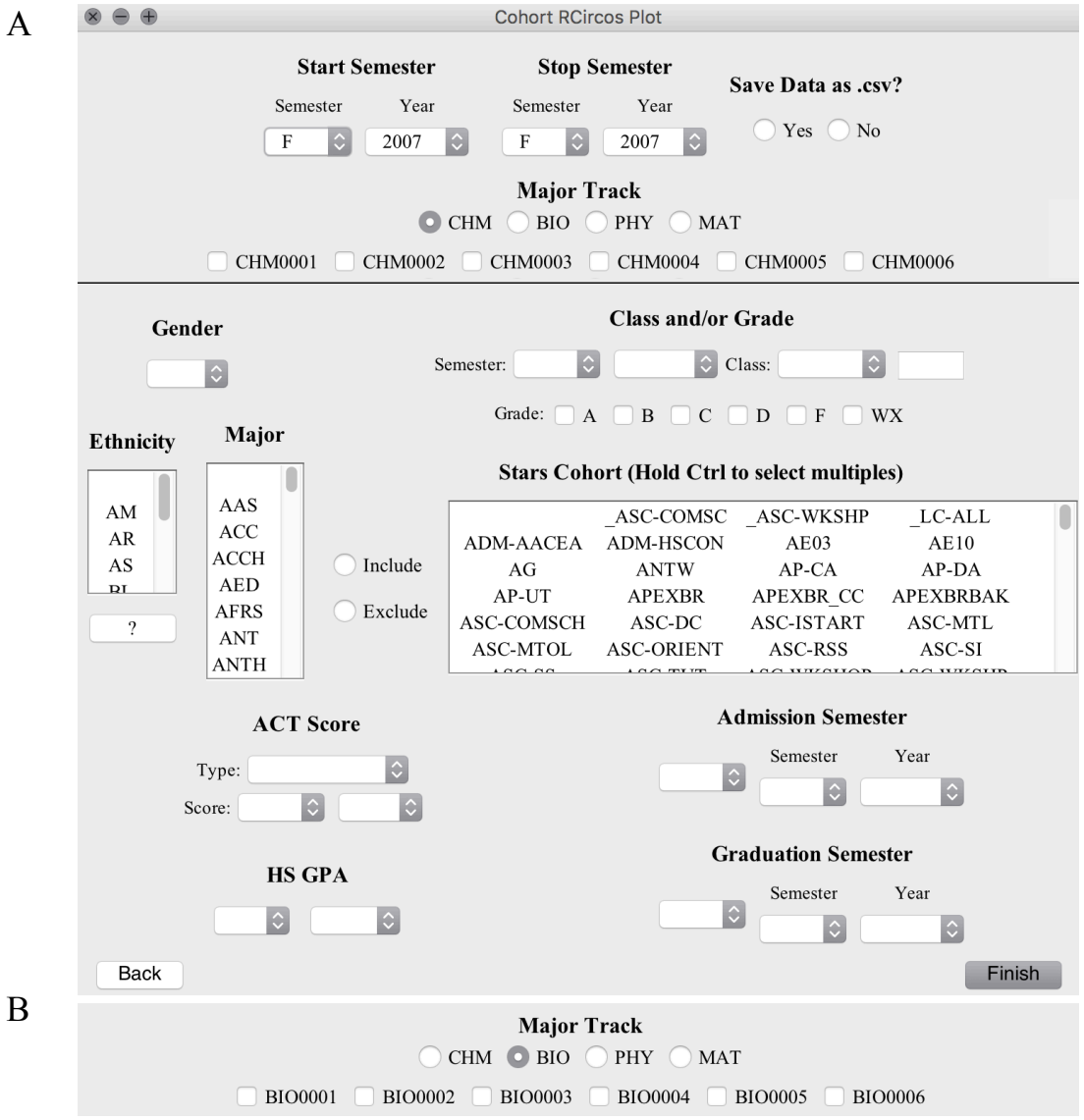


Figure 1.1: Java interface and demographic selectors. A) The java UI allows any user to generate Student Circos plots without prior experience with R syntax. The interface requires a range of semesters with the desired courses to build the required karyotype, and many optional demographic selectors by which to select a cohort. A cohort can also be selected based on a class selection with the option of specifying a particular semester and/or grade(s) earned in that class. **B)** The major track allows for easily switching between majors and will auto-update the window when selecting a new major. The user can select as many or as few of the supplied courses for each major track from which to build the karyotype.

algorithms use, and display historic student data for faculty and staff to draw conclusion about their programs as a whole. Student Circos plots allow the unique ability to track cohorts before and after a class of interest, making it a very robust tool for visualizing longitudinal student data.

Results

To probe student data effectively, the right people need the correct data and they have to ask the right questions. An investigator can generate a collection of questions and then easily become lost in a sea of data. Answering these questions by panning data for statistically significant trends can often be akin to taking a shot in the dark. This tool can help narrow the

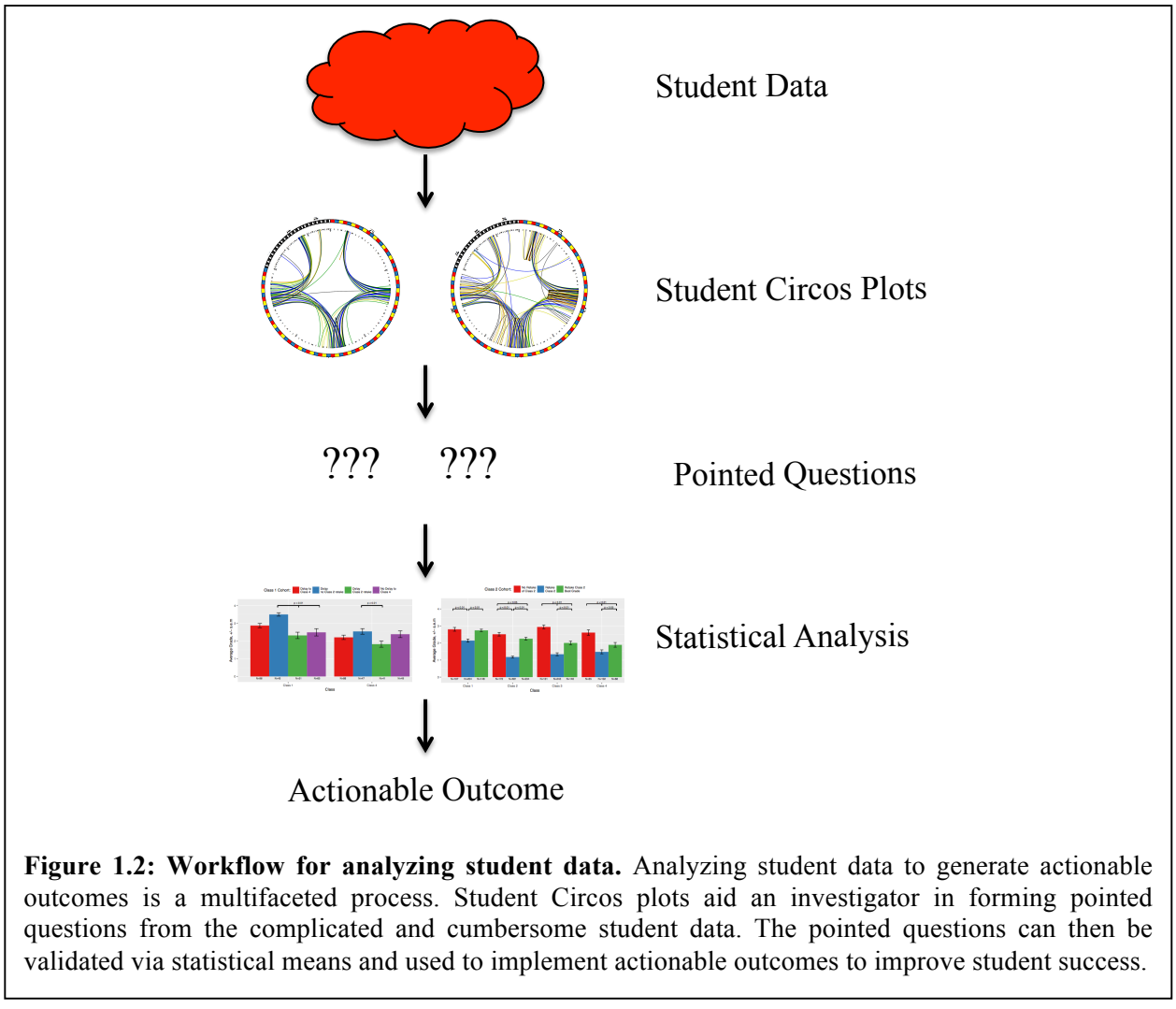


Figure 1.2: Workflow for analyzing student data. Analyzing student data to generate actionable outcomes is a multifaceted process. Student Circos plots aid an investigator in forming pointed questions from the complicated and cumbersome student data. The pointed questions can then be validated via statistical means and used to implement actionable outcomes to improve student success.

focus leading them to a quicker understanding of longitudinal cohort progressions (Figure 1.2). Using the Student Circos plots allows swift visualization of many cohorts to quickly analyze if the proposed questions have merit. Then, more pointed questions can be generated and the data table output can be used to analyze for statistically significant relationships. The investigator can then use their own experience to filter the results and propose an actionable outcome, working to improve student success.

Reading Student Circos Plots

Circos is a program that was designed for visualizing data, most often of genomic origins, in a circular layout maximizing the data-to-ink ratio.[20] Student Circos plots are built off of the parent Circos program because the circular layout allows the longitudinal analysis of student progression through majors. Because this tool is based off of Circos, aspects have adopted nomenclature such as "karyotype". The karyotype is the outside circle of class names and color boxes. The colored boxes denote semesters where red, blue, and yellow indicate fall, winter, and spring/summer semesters respectively. The colored lines signify the grade a student earned in the class where: A=green, B=blue, C=yellow, D=orange, WF=black. Students' progress through a major via three main paths ending in one of three possible outcomes: graduation, still currently enrolled, or stopping out (Figure 1.3). Other endpoints exist (such as transfer out) which could also be added if needed by an institution, but these three encompass the majority of students. Plotting full cohorts together, one can gather information on longitudinal student progression. Visualizing cohorts over time can identify important trends that warrant further analysis, such as statistical tests between cohorts.

Proof of Concept For Student Circos Plots

To validate the ability to draw conclusions using Student Circos plots, the tool was used to identify the semester a curriculum change was implemented within the first class of a major

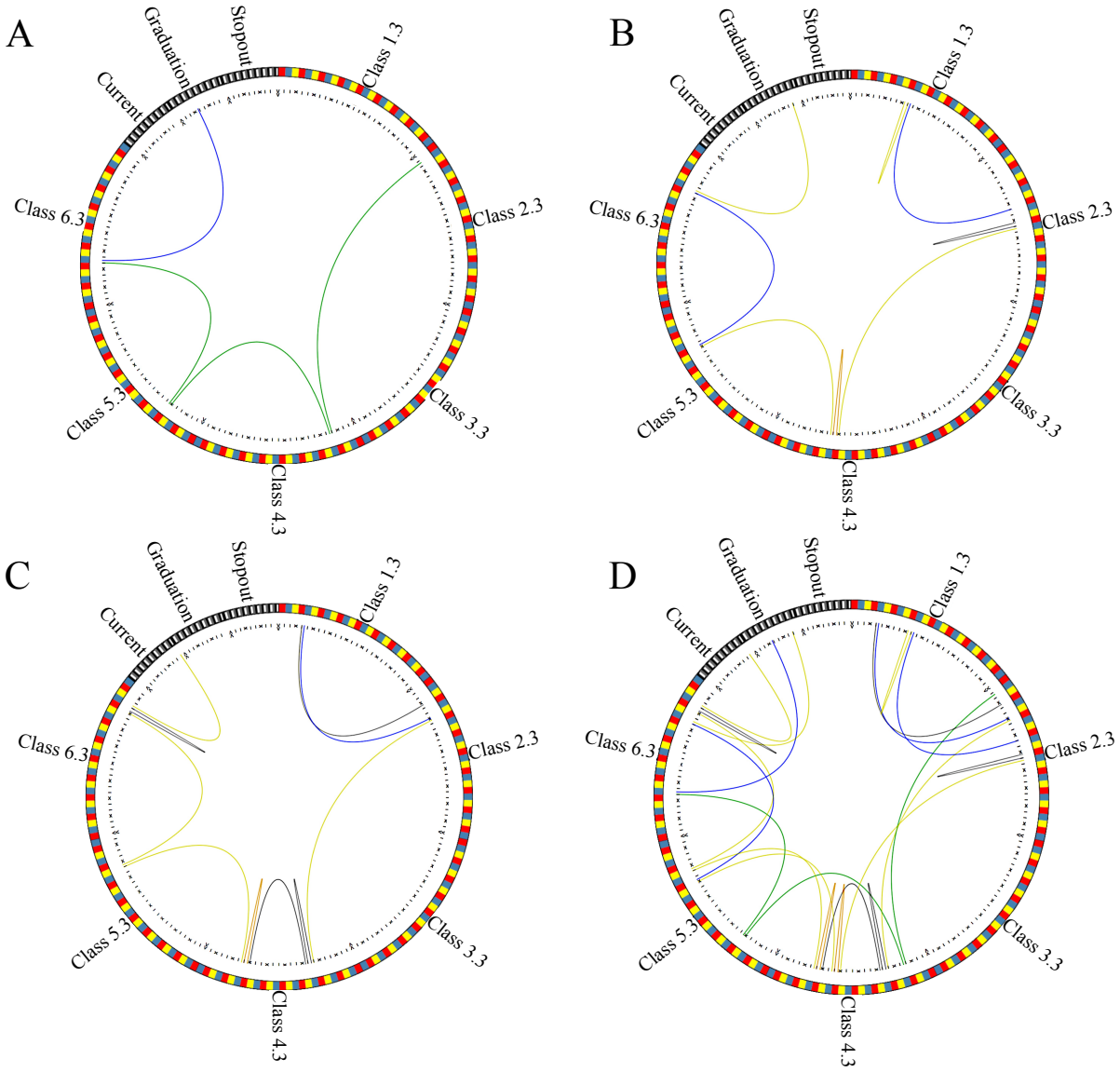


Figure 1.3: Reading a Student Circos plot. The karyotype displays the semesters and courses selected by a user. Red boxes symbolize fall semesters, blue winter, and yellow spring/summer. The interior of the karyotype has a series of tick marks for easily identifying when a semester begins (-), when a class begins (<<), or when the academic year begins (X). These plots show three students with expected paths through a given major. **A)** Student tracking through without delay and with good grades to graduation **B)** Student struggled through the track with repeated attempts at courses, but still graduated **C)** Student struggled through the track and has not yet reached graduation. **D)** Overlay of all three students.

(Figure 1.4). A series of plots were generated using winter and fall semesters (spring/summer semesters were ignored because they are not part of a normal progression and have low levels of enrollment) covering a four-year span. Before Semester 4, each cohort took the second class in this progression over a wide range of semesters indicating the cohorts are not proceeding

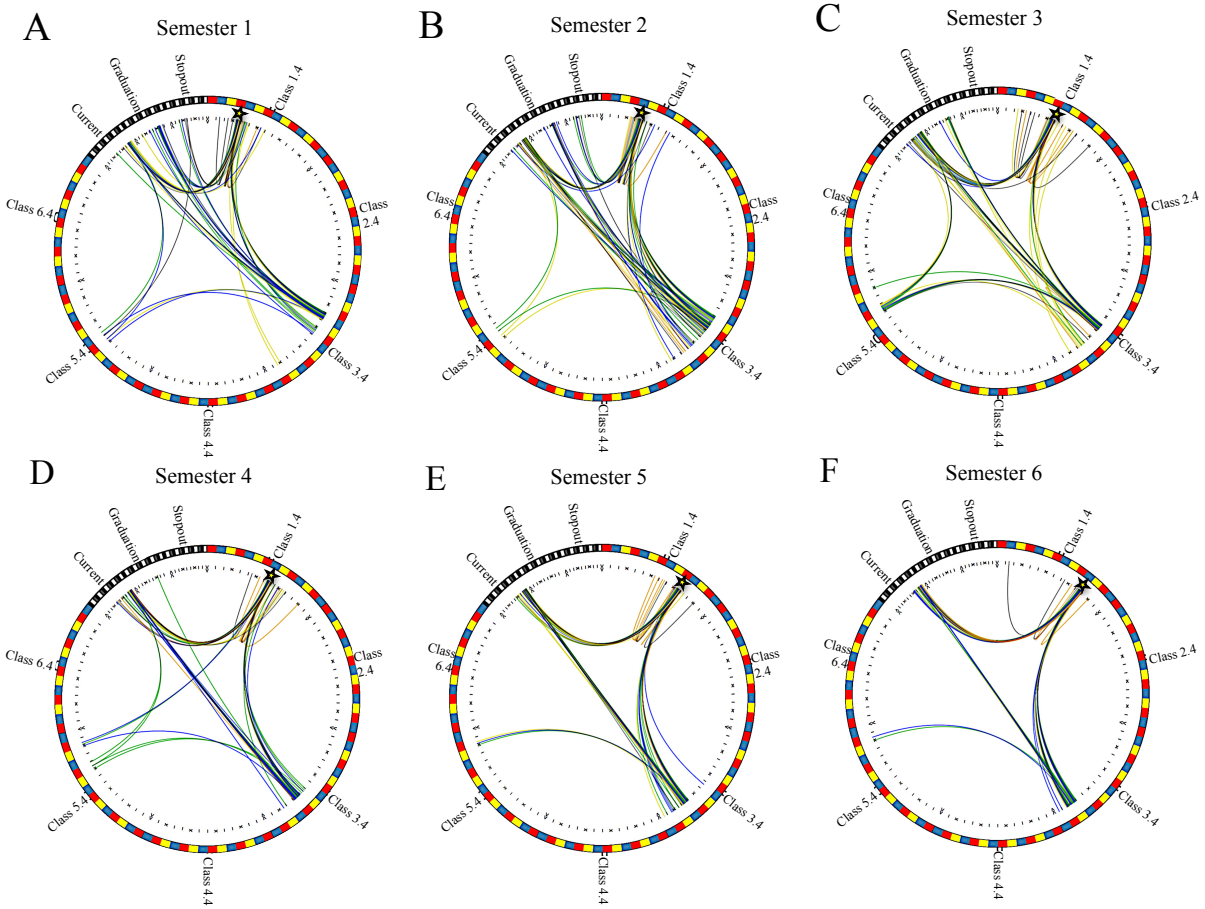


Figure 1.4: Analysis of a change in curriculum. Our institution implemented a change in curriculum for Class 1.4 of this track. Cohorts include students who took Class 1.4 in **A)** Semester 1, N=56 **B)** Semester 2, N=70 **C)** Semester 3, N=63 **D)** Semester 4, N=62 **E)** Semester 5, N=60 **F)** Semester 6, N=85. Without prior knowledge, this tool identified that the curriculum change occurred between semesters 3 and 4, evidenced by the increase of A/B grades coming out of Class 1.4 in addition to a decrease in student delay moving into Class 2.4.

together. During and after Semester 4, the cohorts advance to Class 2.4 with less spread, suggesting they have begun to advance as a cohesive unit and students do not delay over multiple semesters. Students also exhibit better performances in the first two classes, with the majority of

students gaining a higher grade, though not significantly higher. Together these factors correctly indicated Semester 4 was the semester where the curriculum change was implemented. Interestingly, the change in curriculum appears to not address the population of students that only take Class 1.4, evidenced by a consistent migration of students from Class 1.4 to Current. This example demonstrates this tool is capable of identifying trends in student data.

Effects of Grade in a Seed Course

Student Circos plots were used to visualize how student's performance in a seed course (the first in a sequence) affects longitudinal progress through foundational classes of a major. Students who took the seed course in a particular semester were separated by grade into four plots (Figure 1.5). The A and B cohorts tend to maintain a high level of achievement throughout the rest of the track, with few ending in Stopout and many graduating after completing all the courses, as one might expect of students who have early academic success within their major (Figure 1.5A and B). Students that struggle in the first course often continue to have difficulty in subsequent courses (Figure 1.5C). Within the C cohort, multiple students repeat many classes within this major. Seeing the difficulties students have within this cohort highlights an area that could be further analyzed to improve student success. A possible issue at hand is the fixed vs. growth mindset, a common area of discussion in pedagogy.[21] Students starting track 5 with a DWF often terminate at Stopout or Graduation, demonstrating the difficulty of recovering from a failing grade within this major (Figure 1.5D). These two end points are drastically different and more work is required to determine what factors lead the DWF cohort to Graduation compared to stopping out. On the whole, this tool has provided visual evidence that can lead a stakeholder into deeper analysis of the C or DWF cohorts to further elucidate factors that lead to differential success, and ask how the program or support services might be adjusted to improve outocomes for these students.

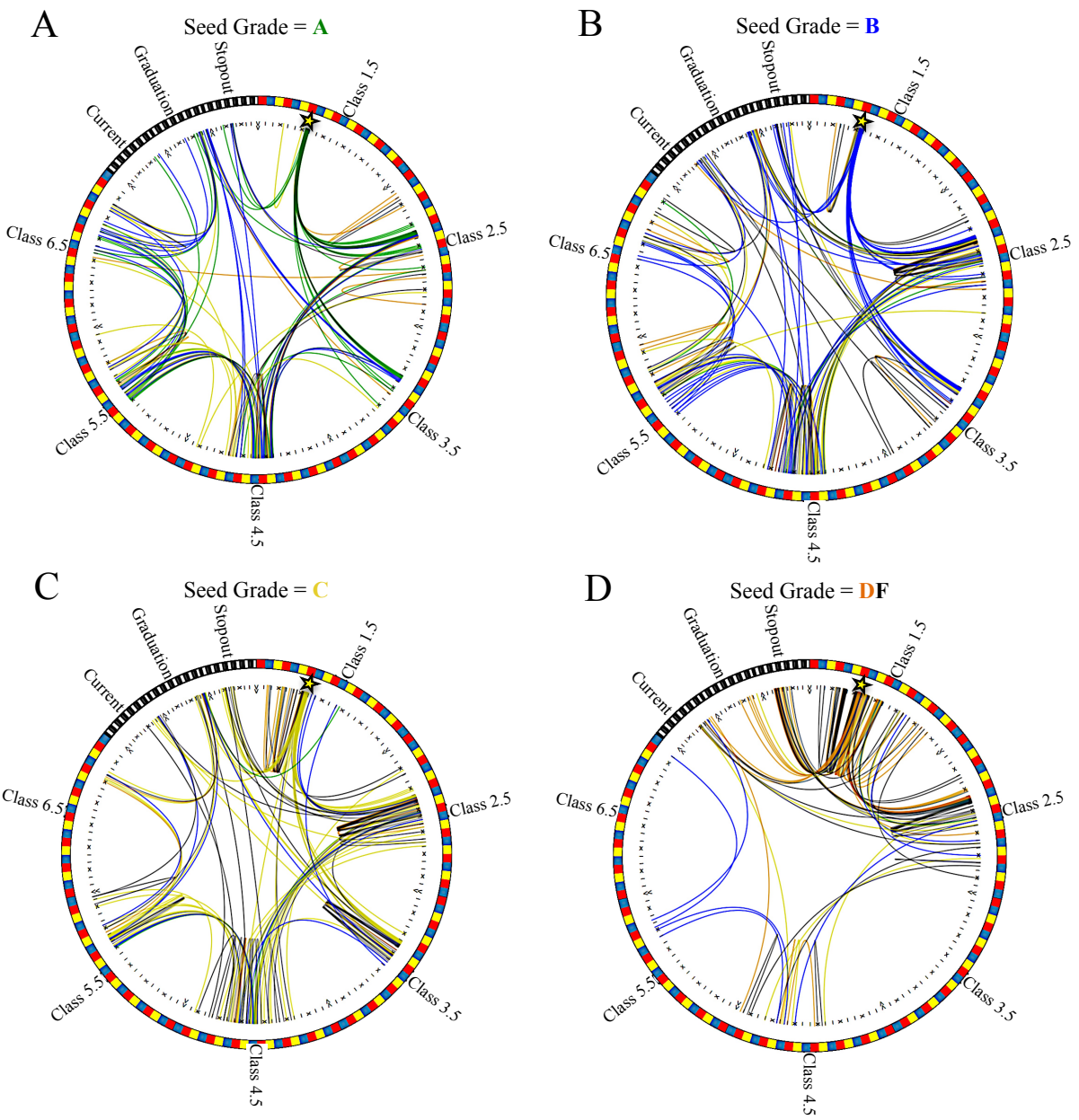


Figure 1.5: Effect of seed course grade on progression through major. Cohort includes students who took Class 1.4 in the semester designated with a star and grade of **A)** A, N=48 **B)** B, N=59 **C)** C, N=54 **D)** DWF, N=68. Many students who didn't pass Class 1.4 (Grade DWF) in the selected semester also failed in prior semesters, with few moving to higher-level courses within this track. Students who do well in their first course (Grade = A/B) tend to maintain success through the entire track, demonstrating that when students start strong, they tend to finish strong. Conversely, when students experience difficulty in the first class (Grade = C/D/F) few students recover to experience high levels of success.

Student Success in a Gateway Course

Gateway courses are foundation-level, with high enrollment, and a high risk of failure and have been an area of intense study within pedagogy for over twenty years.[22-25] This tool was used to compare cohorts that retake the gateway course Class 2.6 (retake gateway course, RGC) and those that progress after one attempt (OA), to analyze if gateway courses hamper student success in this major (Figure 1.6). OA students do significantly better than RGC in all classes, even when comparing the highest attained grade for RGC (Figure 1.7). The RGC cohort takes Class 1.6 multiple times, suggesting that the skills and knowledge gained in Class 1.6 is not appropriately preparing students for Class 2.6. Further analysis of the RGC cohort shows that for the first two classes of this track, the grades attained decrease with each attempt (Figure 1.8). These data suggest that students seldom improve their grade, which likely creates a financial burden on the student. Additionally, the cohorts were analyzed for correlations between class grade and ACT composite score or high school grade point average (HS GPA), and the gender or ethnicity make up but no significant relationships were found (Appendix A Figure 1-4) Regardless of grade or number of attempts many students move from Class 2.6 to Current, Graduation, or Stopout. These plots visually demonstrate a large exodus from this major, supporting the hypothesis that the gateway course truly inhibits students from succeeding and progressing in this major.

Effect of Delaying Within a Major

Every major contains expected pathways, progressions, and milestones that faulty deem successful. For this major, students are expected to progress from the first to fourth class within two years (Figure 1.10). A faculty and staff narrative at our institution involved students' whom delay taking certain courses, suggesting the delay leads to a lower grade in the final course, an affect we wanted to visualize. At first glance, the cohort delaying (delay full, DF) and the cohort

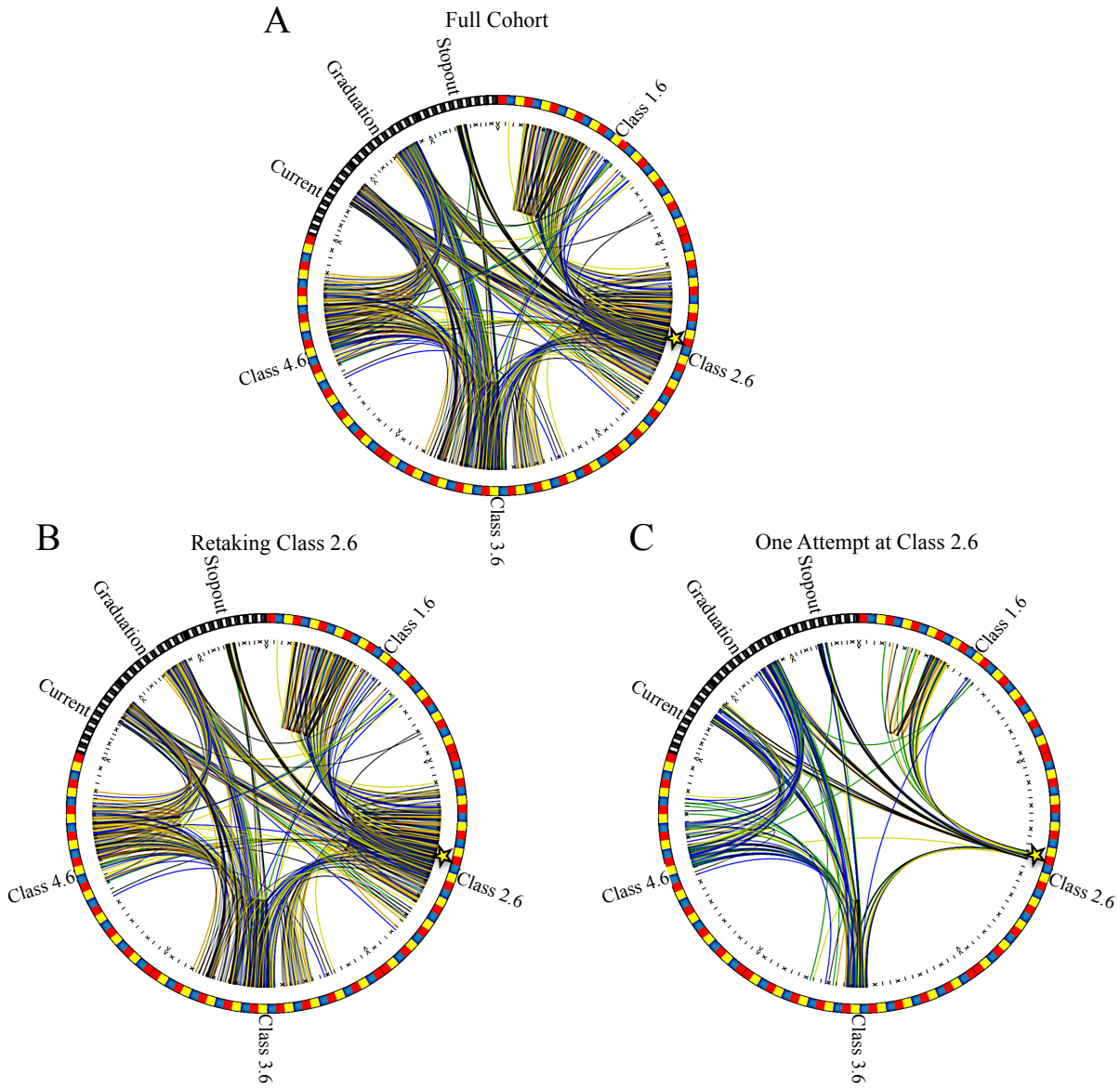


Figure 1.6: Student performance within gateway course. Students who took the gateway course Class 2.5, in the term denoted with a star. **A)** Full cohort, N=407 **B)** Students retaking Class 2.5 (RGC), N=234 **C)** Students who do not retake Class 2.5 (OA), N= 172. The RGC cohort had decidedly poorer performances in Classes 3.5 and 4.5. These students also exhibit a wide fanning within Classes 1.5 and 3.5, demonstrating that many retake each course over a large time period. OA students most often earn a grade of C and seldom stumble within Classes 1.5 or 3.5. Collectively these plots demonstrate that success in Class 2.5 indicates success in Class 3.5, and confirms Class 2.5 behaves as a gateway course.

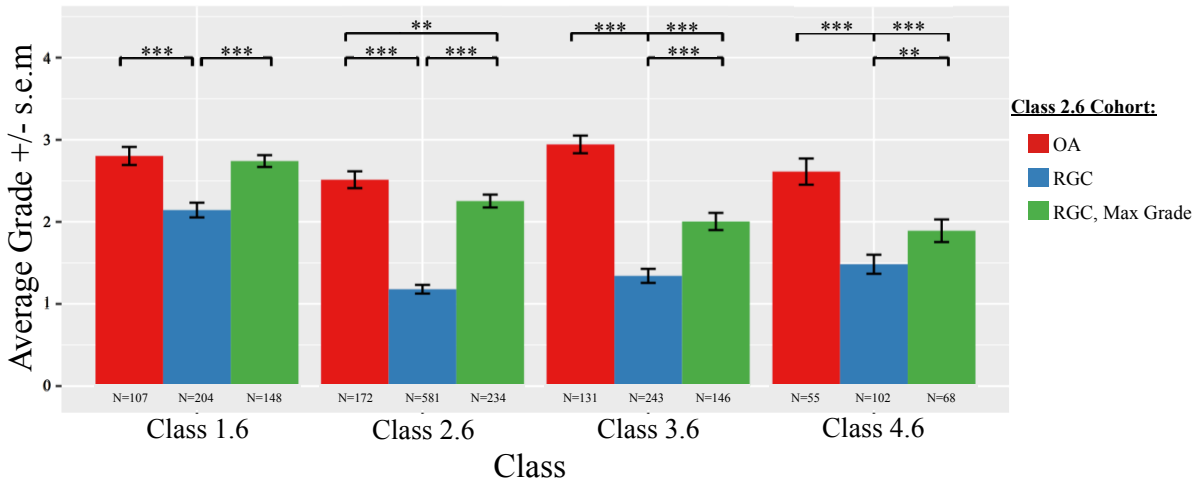


Figure 1.7: Average grade for Class 2.6 cohorts. Students who attempt Class 2.6 once (OA) do significantly better than the cohort that retakes this class (RGC). OA students do better in all but the first class when using the highest grade attained by the RGC cohort (RGC, Max Grade). This supports the conclusion that students who struggle in the gateway course continue to perform poorly throughout the major. *** = p<0.01, ** = p<0.05, student's t-test

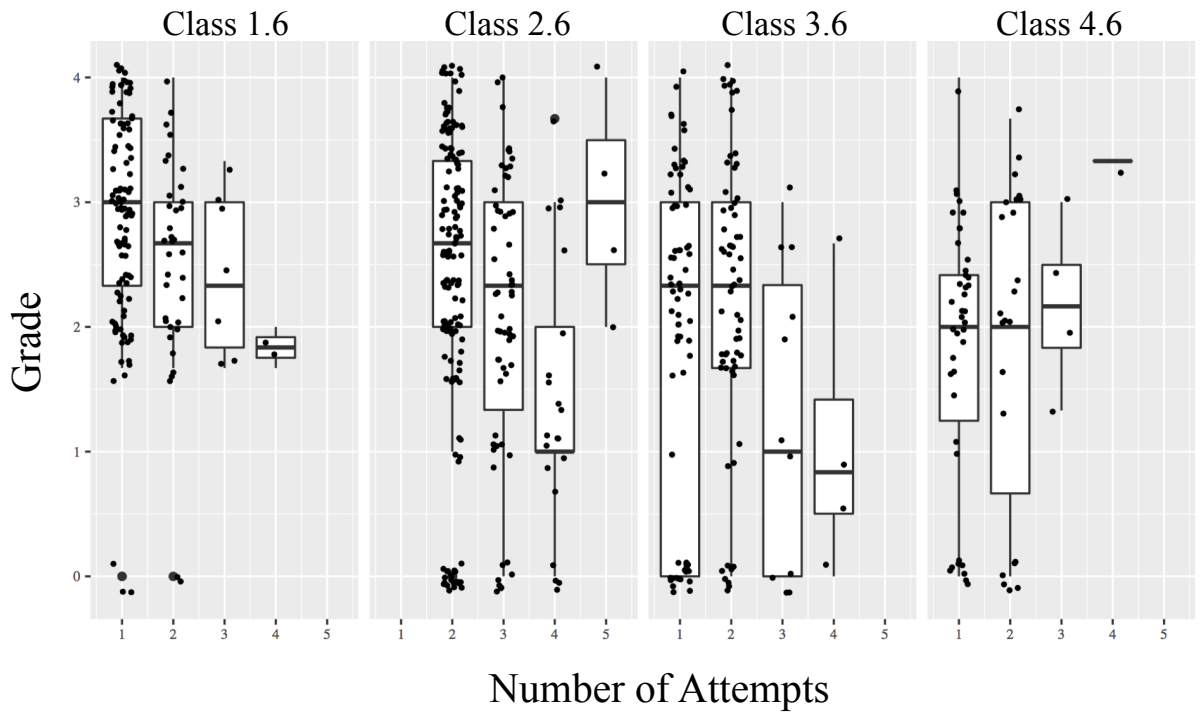


Figure 1.8: Class performance by number of attempts for students retaking Class 2.6. In the early classes of this major, as attempts increase the earned grades decrease. The same trend does not hold for the last two classes. For Class 1.6 and Class 2.6, the data support this institutions policy of requiring special permission to take a class more than three times, as student grades are generally worse at higher attempts.

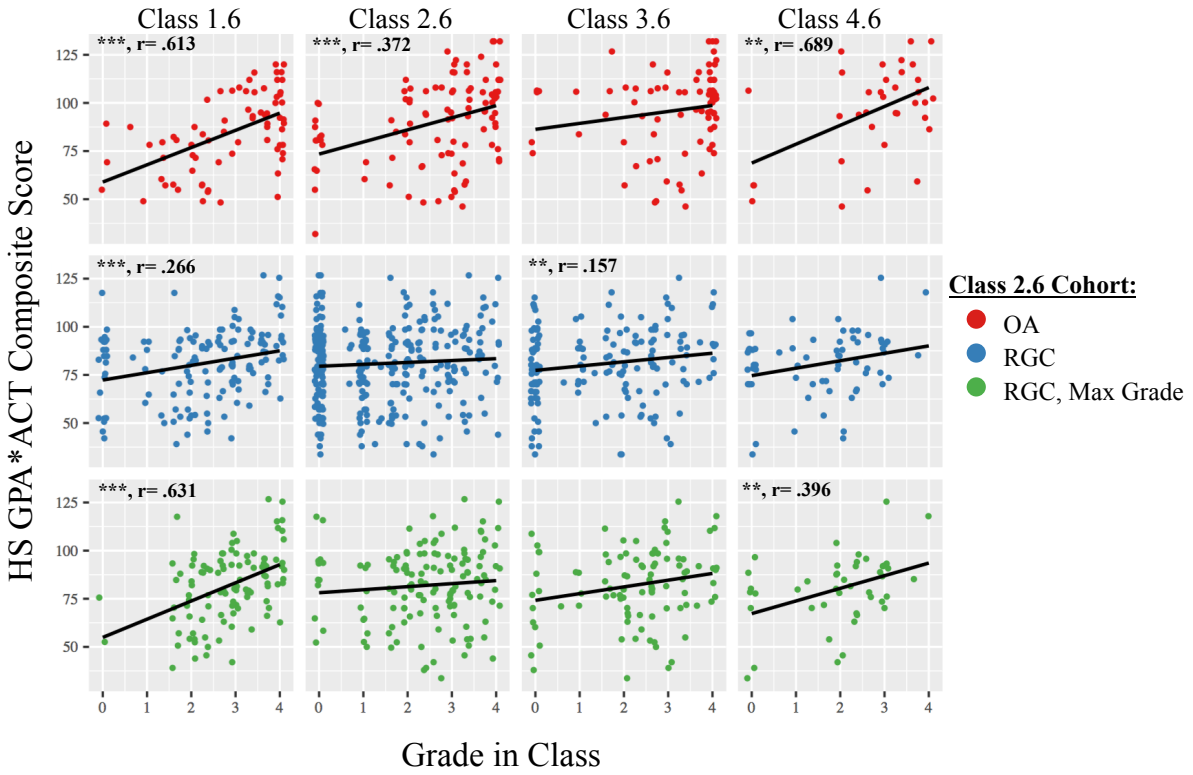


Figure 1.9: Correlation between class grade and HS GPA*ACT Composite score. HS GPA*ACT Composite score is a good predictor of success for students who attempt Class 2.6 once (OA), but is very poor for students retaking the gateway course (RGC). r = Pearson rho, *** = p< 0.01, **= p< 0.05

that stays on time (OT) do not appear to perform differently in the first or last course (Figure 1.10C and D). However, DF can be split into two distinct sub-populations: those who delay because they retake the second class of this progression (delay retake, DR) and those who delay for reasons beyond our knowledge (delay no retake, DNR) (i.e taking courses outside this track, personal factors, etc.). Plotting these populations separately shows the DR cohort performs poorer in the first and last course compared to DNR (Figure 1.11). The grades within Class 4.10 for DNR mirror those of OT. The data unexpectedly show that how long a student takes to progress through this major has minimal impact on success, but the students' path between classes plays a large role. Correlation analyses and gender or ethnicity make-ups or each cohort were conducted as with the gateway course investigation, but no significant results were identified (data not included). Without these plots, identification of DR and DNR cohorts would

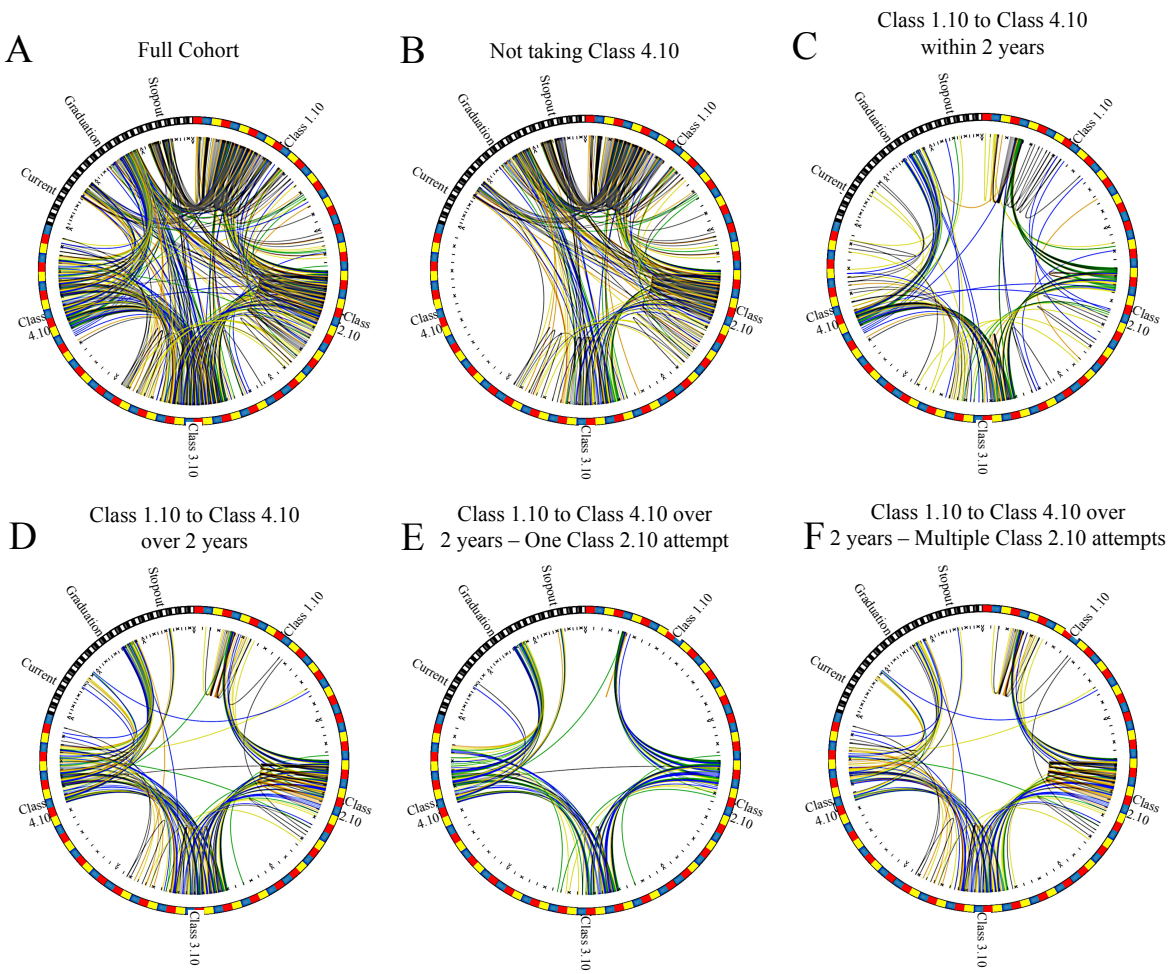


Figure 1.10: Student path from Class 1.9 to Class 4.9 and the effect on Class 4.9 performance. Students who took Class 1.9 in the semester denoted by a star compose this cohort. **A)** Full cohort, N=416 **B)** Students who did not take Class 4.9, N=296 **C)** Students taking Class 4 within two years of taking Class 1.9 (OT), N=41 **D)** Students taking Class 4.2 more than two years after Class 1.9 (DF), N= 79 **E)** Students taking Class 4.9 more than two years after Class 1.9 and did not retake Class 2.9 (DNR), N=43 **F)** Students taking Class 4.9 more than two years after Class 1.9 and retook Class 2.9 (DR), N=36. Students are expected to progress from Class 1.9 to Class 4.9 within two academic years. OT students do very well in Class 4.9, with most earning a B or higher. The DF cohort can be split into two sub-populations: DNR and DR. The DR cohort performs poorly in every class within this track compared to DNR. OT and DF cohorts both show a high success rate after reaching Class 4.9, confirming a trend previously reported at our institution.

have been exceedingly difficult. Additionally, it has previously been reported that when students reach Class 4.10 there is ~ 96% graduation rate. For all of the cohorts that reach Class 4.10, very few students migrate to Stopout confirming the high success rate (Figure 1.10A, C, D). Student Circos plots have verified a trend previously reported at this institution while simultaneously discrediting another.

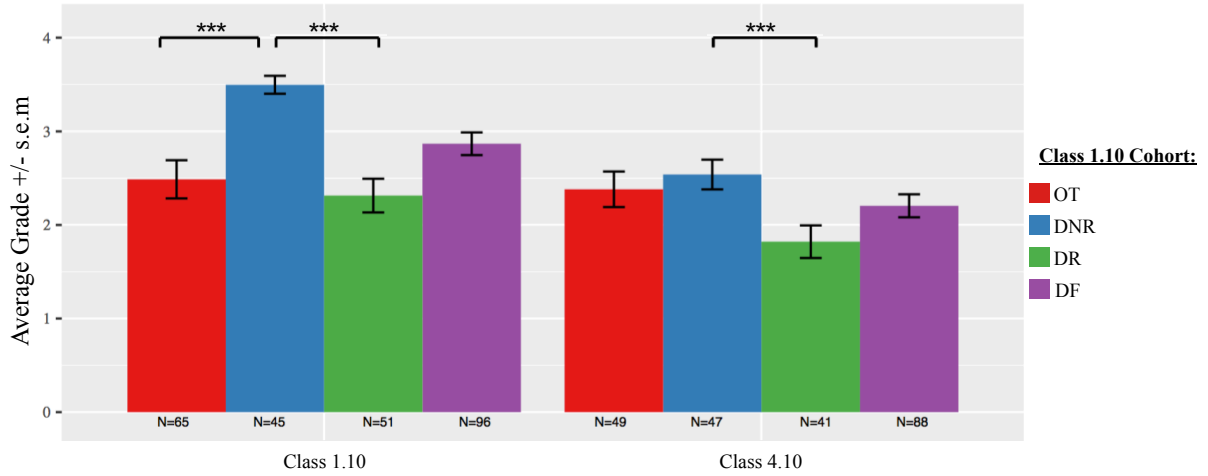


Figure 1.11: Average grade by Class 1.10 cohort. Students delaying within major and not retaking Class 2.10 (DNR) do significantly better in each class than their repeating counterparts (DR). Interestingly, DNR students perform better in the first class than the cohort that progresses through this major on time (OT). DF = full delay cohort ***= $p<0.01$

Discussion

Improving student success in higher education is heavily dependent on data being in the hands of faculty and staff who have the ability to generate positive changes at their institution. Visualization tools depicting student data in a longitudinal manner are required for faculty to draw meaningful conclusions from historic student data. Current predictive algorithms and student migration plots fail to accomplish the task of presenting data for longitudinal analysis of student cohorts and often utilize current data and not historic. The Student Circos tool achieves the goal of depicting student data as longitudinal progression and has exhibited a wide variety of applications. Plotting student data in a circular fashion maximizes the quantity of data presented

in each plot and shows longitudinal student progression through majors. Using this tool has identified questions that require further analysis to generate actionable outcomes for improving student success. For example, the DWF cohort of a seed course contains students that attain success by graduating and the less optimal endpoint of Stopout. An institution can use these data to set the stage for an in depth analysis, determining what factors lead to the two very different outcomes. Students can use these data as a precautionary tale to start their college career on a positive note. Student Circos plots have also generated data visualizations that challenge previous understandings, as is the case with students who delay within a major. Without this tool, the effect of a student's progression through compared to the amount of time delaying within a major would likely never have been identified.

It is important to note that this tool can be altered to many ends. For example, there is a version of this tool that generates single student plots that are useful for advising purposes. Single student plots depict student progression of a major and the major requirements that have been met. Additionally, one can alter plots by adjusting the order in which data is printed, to better represent the question at hand. Currently, term, class, and grade are used to sort course data, in that order, before plotting. By adjusting these parameters one can change the emphasis of a plot or mine for different trends. The level of personalization and control over cohort and class selections make Student Circos plots an invaluable tool for analyzing longitudinal cohort progression. One major limitation of Student Circos plots is the inability to draw meaningful conclusions from large cohorts of students because the plots become very hard to read. However, the benefit of only plotting smaller samples is that an interested faculty member will need a directed question before starting analyses, preventing them from becoming lost in a sea of data.

Student Circos plots allow faculty and staff to identify important trends in longitudinal student progression and generate actionable outcomes for student success in higher education.

Chapter 1 Materials and Methods

Institutions wishing to implement this program will require three items: student data in .csv format (Appendix A Figure 5), R, and a java editor. If the student data is in JSON format, it can be unpacked within R into a data frame and then used for the plots, but none of the included R scripts contain code for JSON unpacking. Within Java the user must install REngine and Rserve, each java script for the UI contains import script for these packages. Three packages available within R must be added to the users library before use: RCircos, Rserve, and plyr. When using the UI, indexing of each package occurs within the Rconnection of the java script, so they only have to be installed in the operating version of R. However, the Rserve socket must be initiated within R before executing the java script.

Java interface

This tool uses a Java UI that implements RConnection to integrate R and Java (Figure 1.1). The code for the Java UI can be found at the Feig Lab website (http://chem.wayne.edu/feiggroup/) The UI affords investigators unfamiliar with R the opportunity to make Student Circos plots in a high throughput manner. The interface contains two sections; the mandatory input information located at the top and the optional demographic selectors located beneath. Possible demographic selectors include: admission semester, graduation semester, ACT score, declared major, gender, high school GPA, class and/or semester and/or grade, ethnicity, and population cohorts. The major and ethnicity designations, classes within each area of study, and population cohorts (i.e stars cohorts) must be customized to other institutions. Areas in the code that require this attention have been appropriately annotated.

Data Manipulation in R

Two user-created R functions achieve the proper data manipulations to generate Student Circos plots and can be found within two individual R scripts located at Feig lab website (http://chem.wayne.edu/feiggroup/). Both of these functions are accessed within Java and executed within R through the Rconnection to generate the desired image(s). The first function generates a data frame that termed the "karyotype", defining the outermost circle of the plots and denotes the semesters, classes of interest, and tick marks. The three possible terminations of study Current, Graduation, and Stopout remain constant within the karyotype. A Stopout is defined as a student who has not registered for a class in over two years. User input of starting semester, ending semester, and class track of choice generates the karyotype (Appendix A Figure 6). The second function takes user defined cohort selectors, either demographic or class/term, to generate a list of students who fit into the desired cohort. This function then selects all classes for the cohort that fit into the desired class track and transforms the data into a format suitable for plotting with RCircos (Appendix A Figure 7). RCircos uses the resulting data frames from each function to create the final plots as .pdf files on the desktop (or the working directory in the users R). These R scripts will work with any institutions data provided the starting course and demographic data are in the same format as ours (Appendix A Figure 8). However, if the starting data were in a different format, a small amount of programming time would have to be dedicated to transforming the data into appropriate Circos format. Data reported here have been masked for the privacy of students and faculty.

Chapter 2
Adapting TcdA into a Cell-Specific Protein Delivery Vehicle

Introduction

The cellular membrane provides a formidable barrier to the intracellular delivery of exogenous proteins. Since proteins have the ability to alter biochemical pathways, it would be desirable to introduce a cloned protein to analyze its effect on cell processes or cellular localization. One strategy for exogenous protein introduction is through expression in situ using systems like the T7/lac or tetracycline responsive elements.[26, 27] Systems like these often lead to protein levels that are not biologically relevant. Recent development of synthetic promoter libraries and predictive mRNA design tools has improved the control of gene expression.[28, 29] However, induction over a large concentration range requires creation of multiple constructs and ample trial and error. These shortcomings make cellular delivery desirable to exogenous expression. Current methodologies for protein delivery rely on appending polycationic tags such as the HIV trans activator of transcription (Tat) or Drosophila antennapedia (Antp) to proteins of interest.[30] These cell-penetrating peptides (CPPs) efficiently achieve cellular delivery, but there are drawbacks to these systems. Due to the highly cationic character of CPPs, protein chimeras often experience undesired delivery to different cellular compartments such as the nucleus.[31] There has been some evidence that the nuclear localization is predominantly an artifact of the fixation process often associated with analyzing delivery of exogenous proteins.[32] CPPs also have a high level of toxicity associated with treatment.[33-35] Finally, due to the general mechanism of CPP uptake, these systems also suffer from a lack of cellular specificity, dramatically limiting

the *in vivo* applications. These major drawbacks create a need for a more efficient, less toxic, and cellular specific delivery vehicle.

Clostridium difficile (*C. difficile*) is a pathogenic spore-forming bacterium most commonly associated with pseudomembranous colitis.[36] *C. difficile* is a common nosocomial infection often colonizing after the normal flora of the gut is disrupted by antibiotic treatment. *C. difficile* produces two major virulence factors, Enterotoxin A (TcdA) and Entertoxin B (TcdB) (Figure 2.1).[37] Each of these toxins contains multiple subunits that behave in a well-orchestrated fashion to intoxicate epithelial cells of the gut with an active cargo, leading to cell death (Figure 2.2). Due to the large level of homology and shared mechanism of intoxication, only TcdA will be discussed further. TcdA contains a C-terminal repeating oligopeptide (CROP) region that elicits endocytosis via cell-surface receptor recognition and aggregation.[38-40] Upon endosomal

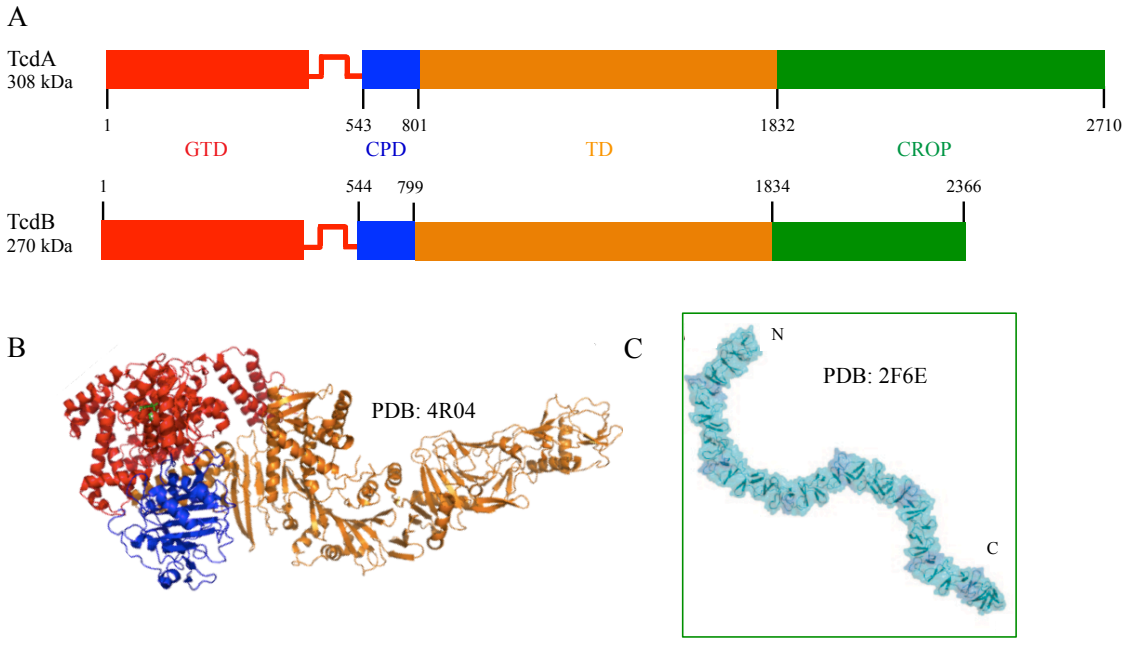


Figure 2.1: Major virulence factors of *C. difficile*. A) Subunits of Toxin A (TcdA) and Toxin B (TcdB) from *C. difficile* from N- to C-terminus: glucosyl transferase domain (red, GTD), cysteine protease domain (blue, CPD), translocation domain (orange, TD), and C-terminal repetitive oligopeptide (green, CROP).[37] B) Crystal structures of TcdA GTD, CPD, and TD.[42.] C) NMR structure of CROP.[38]

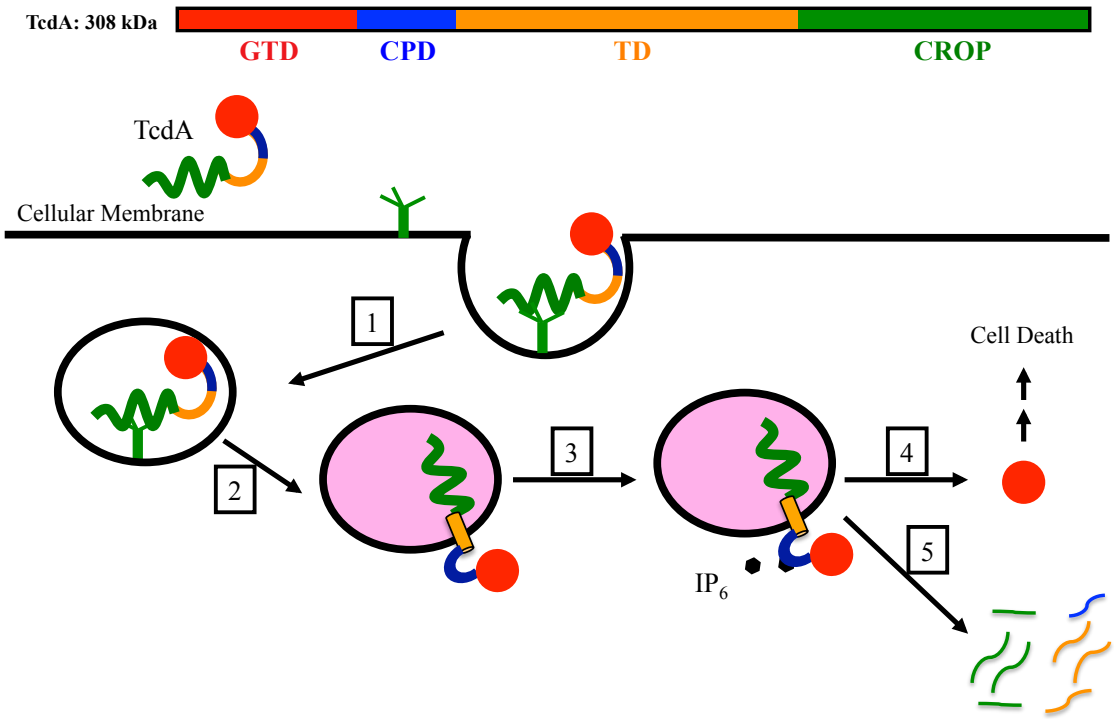


Figure 2.2: Mechanism of TcdA intoxication. 1) CROP binds and aggregates epithelial cell surface receptors inducing endocytosis. 2) Endosomal acidification causes TD to change conformations and form a pore in the endosomal membrane through which CPD and GTD can be fed into the cytosol. 3) In the cytosol CPD binds IP_6 (black hexagon) and 4) cleaves GTD from the rest of the toxin. 5) Remainder of toxin remains in the endosome and degraded.

acidification, the translocation domain (TD) undergoes a conformational change inserting itself into the endosomal membrane creating a pore.[41, 42] Once the pore is formed, the cysteine protease domain (CPD) and glucosyltransferase domain (GTD) are translocated from the endosome into the cytosol. CPD then binds the eukaryotic-specific inositol hexakisphosphate (IP_6) activating the protease moiety, cleaving GT from the rest of the toxin.[43, 44] Once liberated, GTD is able to glycosylate Rho family GTPases inducing actin disregulation, ultimately leading to cell death.[45] TcdA processing within the cell leads to the delivery of an active cargo while the rest of the protein is degraded, making TcdA an attractive possibility as a delivery vehicle for cargos other than GTD. The tagless delivery eliminates many of the negative CPP-chimera effects. Previous reports have shown that by replacing GTD with reporter proteins such as luciferase, the TcdA scaffold is capable of delivering non-natural cargo to endothelial cells.[46] Additionally, the TcdB

has been used to specifically target neurons by replacing CROP with the receptor-binding domain of Botulinum neurotoxin.[47] Combining both of these lines of research, the goal was to create TcdA chimeras that are cell specific protein delivery vehicles capable of altering cellular pathways.

Chimeric Toxin Design

Multiple features have been engineered into the chimeric proteins using the TcdA scaffold. Unique digestion sites flanking the cargos and receptor binding domains (RBDs) were engineered for swift interchanging of different cargos and RBDs (Figure 2.3A). The chimeras also include a C-terminus hexahistidine tag (His_6) for affinity purification as well as orthogonal labeling sites. An YbbR12 sequence has been placed at the C-terminus of cargos upstream of the CPD [48, 49]. YbbR12 is a recognition sequence for the *Bacillus subtilis* phosphopantethein transferase Sfp, enabling the labeling of cargos with CoA substrates. The YbbR12 tag is located such that the label remains with the cargo upon CPD cleavage, important for visualizing cellular delivery and localization of cargo. In addition, the RBDs contain a Sortase A (SrtA) recognition

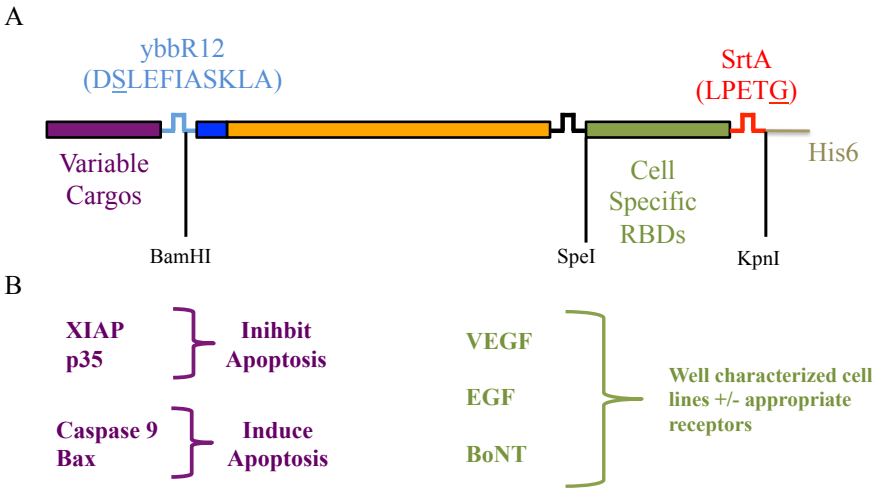


Figure 2.3: Design of chimeric toxins with variable cargos and receptor binding domains (RBD). A) Unique digestion sites have been engineered into the chimeric toxins for efficient subcloning multiple cargos and RBDs. Each cargo has been subcloned to contain a C-terminal ybbR12 recognition sequence and the RBDs have a C-terminal SrtA recognition sequence. Together these sites allow for orthogonal labeling of the chimeric delivery vehicles. B) Variable cargos and RBDs for chimeric toxin production.

motif located at the C-terminus between the RBD and His_6.[50] Sortase A is a transpeptidase that appends a substrate containing three N-terminal glycines in place of the terminal glycine in the recognition sequence (Figure 2.3A). Together these sites allow for orthogonal labeling of the chimeric toxins. A possible application for these sites could be labeling the chimeric toxins with a FRET pair of fluorophores as a qualitative means of determining cytosolic delivery efficiency, or as a method for tracking intracellular localization of the scaffold and cargo independently.

Results

Chimeric Subcloning

Caspase 9 (Casp9), Bcl-2-associated X protein (Bax), X-linked inhibitor of apoptosis (XIAP), and Cyclin-Dependent Kinase 5 activator 1 (p35) were chosen as cargo proteins (Figure 2.3B). These proteins were chosen because of their pronounced effect on cellular signaling related to apoptosis. Cell death is a convenient qualitative and quantitative measure for active cargo delivery. The cell-specific RBDs chosen were epidermal growth factor (EGF), vascular endothelial growth factor (VEGF), and the heavy chain of botulinum neurotoxin (BoNT). Well-characterized cell lines that exhibit positive and negative phenotypes for RBD specific receptors are commercially available allowing cellular specificity to be explored. Given the immense size of the chimera plasmids (>11 kb), the subcloning process required multiple steps (Figure 2.4). First, the RBD (cargo subcloning was carried out by Amit Kumar) were amplified out of commercial vectors to introduce the desired digestion sites and labeling tags, then ligated into pCR2.1®TOPO® TA vector (TOPO). RBD-TOPO vectors were digested and ligated into a modified pWH1520* vector (Figure 2.4A). pWH1520* is a parent vector containing the lab's cloned TcdA constructs and designed for expression within *Bacillus megaterium*. Digesting the RBD-shuttle and ligating with variable Cargo-TcdA constructs created by Amit resulted in

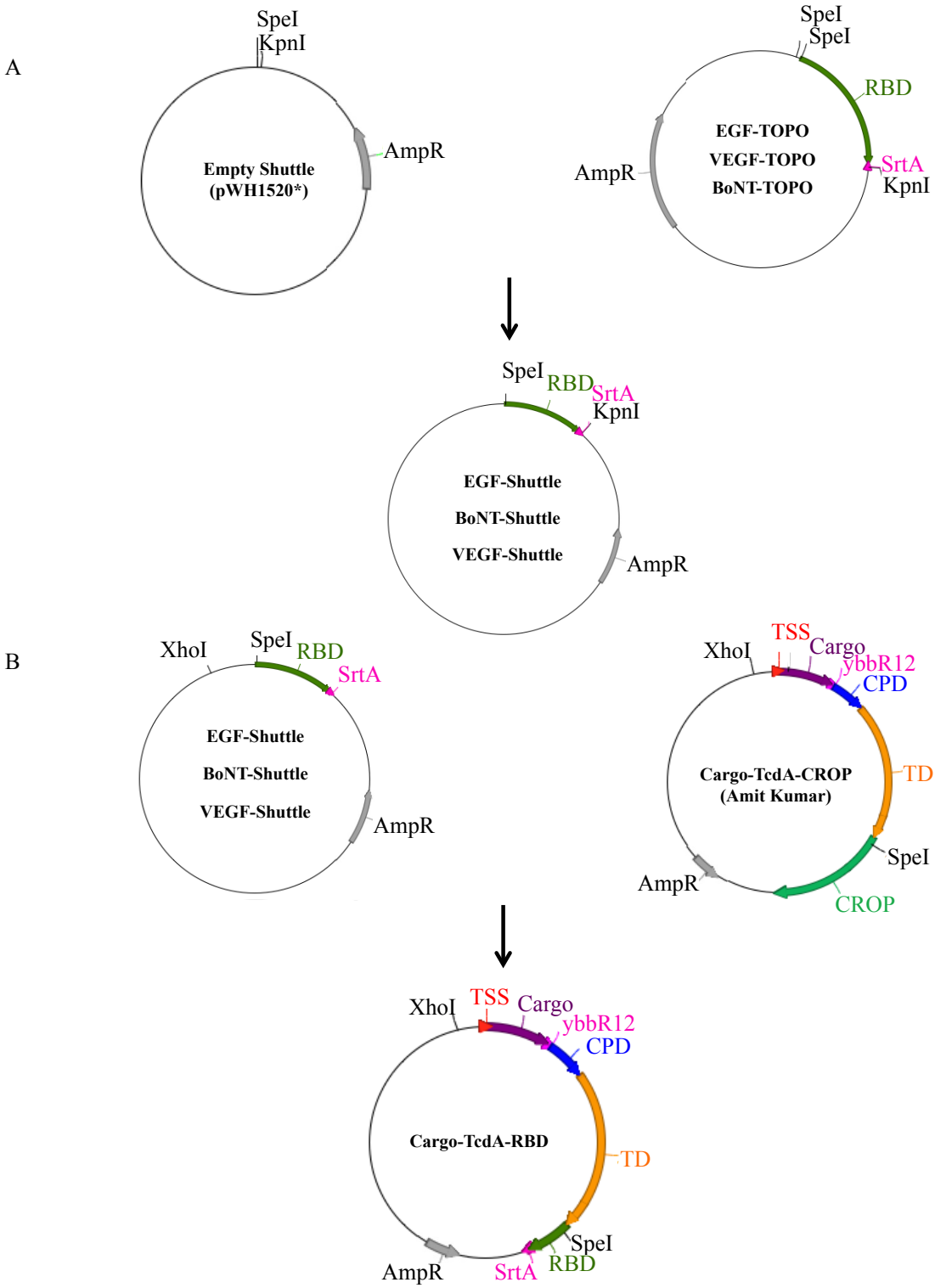


Figure 2.4: Subcloning strategy to create Cargo-TcdA-RBD chimeras. A) RBDs EGF, VEGF, and BoNT were PCR amplified out of commercial vectors and TOPO cloned. The resulting vectors were digested with SpeI and KpnI to insert into a modified MoBiTech expression vector pWH1520*. B) The resulting RBD-shuttle vectors were digested with XhoI and SpeI for insertion into Cargo-TcdA-CROP constructs created by another member in the Feig lab.

Cargo-TcdA-RBDs, the final products ready for expression after DNA sequence verification (Figure 2.4B).

The subcloning scheme allows for interchanging cargo and RBD moieties efficiently. Due to the inherent repetition within the subcloning process each step of Casp9-TcdA-EGF (CAE) chimera generation will be included, but only the final alignments are included for all other constructs. As stated above, the first step of chimera subcloning required PCR amplifying EGF out of the commercial Lambda-EGF116 cDNA vector and ligating the product into TOPO

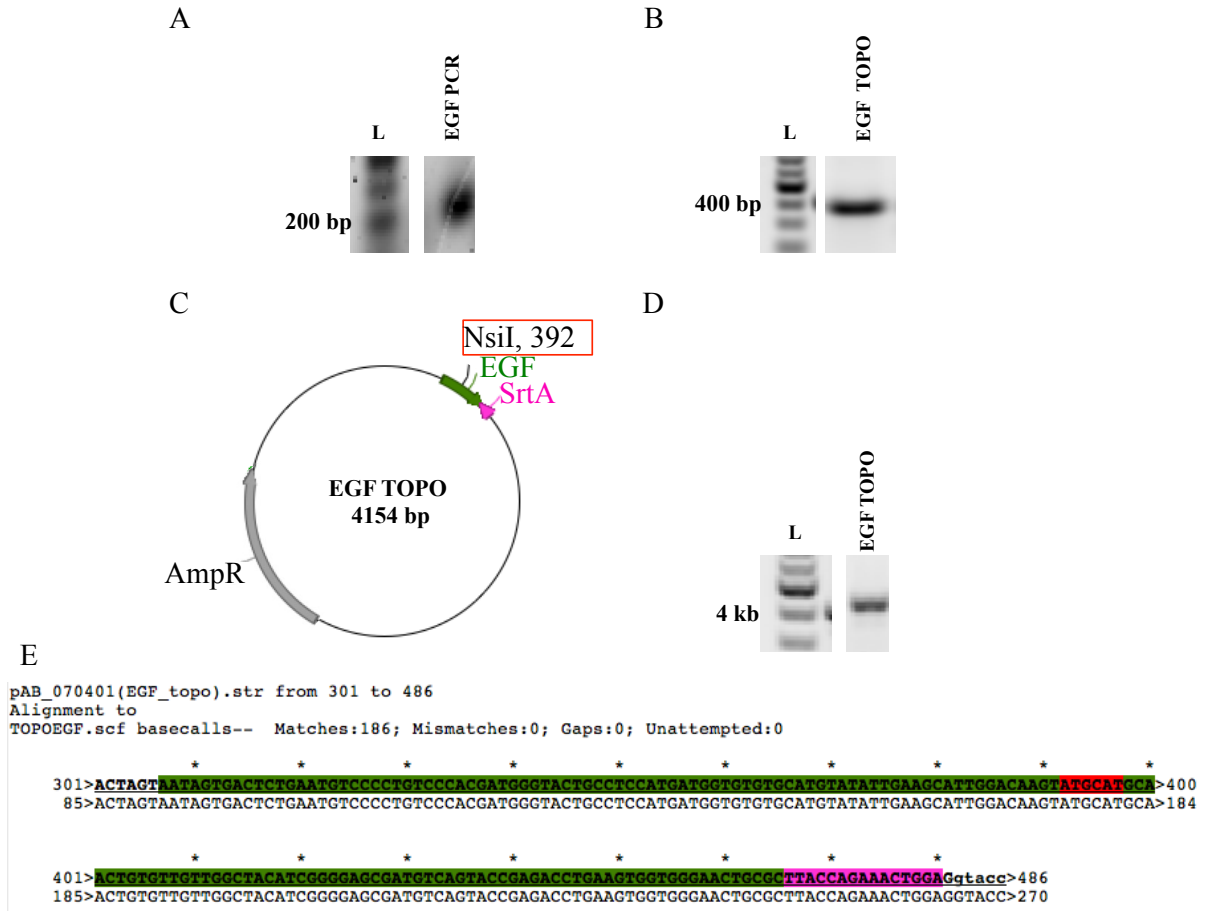


Figure 2.5: Creation and verification of EGF-TOPO vector. Visualization of DNA fragments conducted in 1.2% agarose (containing 0.7 ug/mL EB) gels at 75V for 1-1.5 hours. A) PCR product of EGF amplification out of Lambda-EGF116 (ATCC). Expected product of 198 bp can be seen and was used for TOPO-TA cloning. B) Colony PCR verification for proper EGF insertion into TOPO, 364 bp product expected. C) Plasmid map for EGF-TOPO highlighting lone NsiI digestion site located within EGF used for digest verification of proper ligation. Expect linearization product of about 4.1 kb. E) ApE sequencing alignment for EGF-TOPO verifying insertion and the presence of engineered digestion sites and tag.

(Figure 2.5A). Colony PCR (cPCR) was conducted to identify colonies with EGF-TOPO and plasmids isolated from those colonies were further verified by endonuclease digestion with NsiI, whose only recognition site is located within the EGF gene (Figure 2.5B-D). Direct DNA sequencing definitively verified the EGF sequence in TOPO with the designed KpnI/SpeI digestion sites and SrtA tag (Figure 2.5E).

The large size of TcdA and chimeras along with codon biases requires expression in *B. megaterium* using the MoBiTech expression system, requiring final constructs be held within a modified pWH1520 (pWH1520*) vector.[51, 52] EGF was subcloned into the pWH1520* from TOPO and digest verified (Figure 2.6). EGF Shuttle was not sequenced because of the small

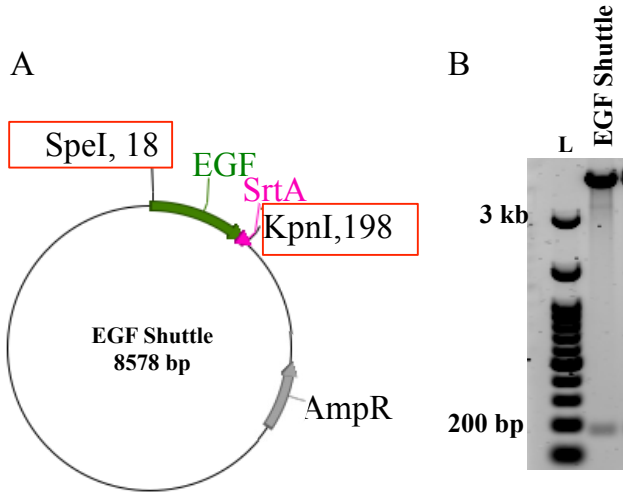


Figure 2.6: Subcloning EGF into Shuttle vector. EGF TOPO was digested with SpeI and KpnI and subsequently ligated into Shuttle. A) EGF Shuttle (8.5 kb) should have single SpeI and KpnI digestion sites. B) Expected 184 bp excision product can be seen verifying EGF Shuttle for ligation into a TcdA Scaffold.

gene size and prior sequence verification, as was true for all the RBD Shuttle constructs. The verified EGF-pWH1520* and Casp9-TcdA were digested with MluI and SpeI and the desired DNA fragment was gel purified. MluI was only used with the Caspase 9 (Casp9) constructs because of an intragenic XhoI site, all other chimeras used XhoI/SpeI. The purified fragments were ligated creating Casp9-TcdA-EGF (CAE). Digesting CAE with XhoI and SpeI confirmed

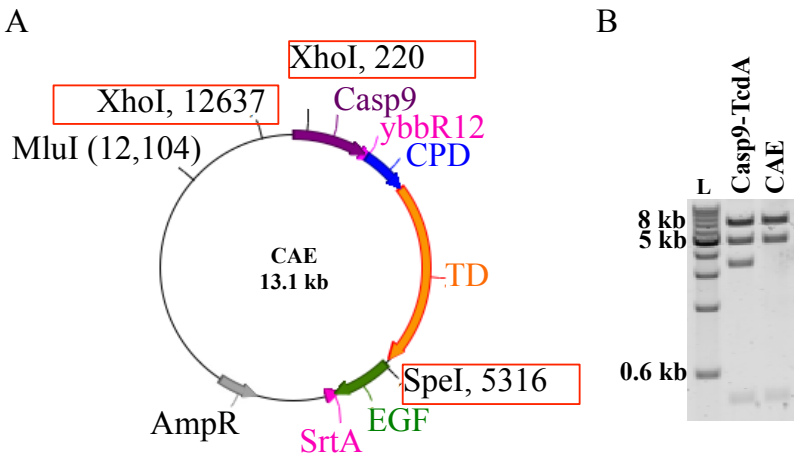


Figure 2.7: Verification of CAE subcloning. A) EGF Shuttle and Casp9-TcdA were digested with MluI and SpeI, and ligated together to form Casp9-TcdA-EGF (CAE). B) CAE test digested with XhoI and SpeI. Expected 7.3, 5.1, and 0.8 kb fragments evidenced. A digestion band that results from an SpeI site within CROP is also absent. The digest verified CAE was then sequence verified.

the proper replacement of CROP with EGF, further verified by DNA sequencing (Figure 2.7, Appendix B Figure 1). The subcloning process was repeated for the generation of the constructs listed in Table 2.1. Each RBD was cloned from a commercial vector using the appropriate primer in Table 2.2 and sequenced with the primers listed in Table 2.3. Sequence alignments can be found in Appendix B Figures 2-8.

Table 2.1: Complete chimeric constructs

Casp9-TcdA-EGF	CAE
Casp9-TcdA-VEGF	CAV
p35-TcdA-EGF	PAE
p35-TcdA-VEGF	PAV
Casp9-TcdA-BoNT	CAV
Bax-TcdA-BoNT	BAB
Luciferase-TcdA-EGF	LAE
Luciferase-TcdA-VEGF	LAV
Luciferase-TcdA-BoNT	LAB

Table 2.2: Chimera subcloning primers

Primer Number	Primer Name	Sequence	Polymerase Used	Anneal Temp (°C)	Elongation Time (min)	Cycles	Restriction Site	Use
1	VEGF-F	GCTTTAACTAGTACGGACAGACA GACAGACACC	Pfu	63.8	1.5	35	Spe I	PCR out of pENTR221
2	VEGF-R	gtgatgggtacCTCCAGTTTCTGGTAA CCGCCTCGGCTTGTCACATTTTTC	Pfu	63.8	1.5	35	Kpn I	PCR out of pENTR221
3	EGF-F	GCTTTAACTAGTAATAGTGACTCT GAATGTCCCCTGTCCC	Taq	68	0.5	40	Spe I	PCR out of LambdaEGF116
4	EGF-R	gtgatgggtacCTCCAGTTTCTGGTAAG CGCAGTTCCCACCACTTCAGG	Taq	68	0.5	40	Kpn I	PCR out of LambdaEGF116
5	BLyS-F	GCTTTAACTAGTATGGATGACTCC ACAGAAAGGGAGCAG	Taq	68	1.5	35	Spe I	PCR out of pMD18-t
6	BLyS-R	gtgatgggtacCTCCAGTTTCTGGTAAC AGCAGTTTCAATGCACCAAAAAA TGTGACATC	Taq	68	1.5	35	Kpn I	PCR out of pMD18-t
7	SpeI_XIAP_F_v2	TGACAAATGGTCCAAACTAGTAG ATCTATG	Pfu	64	4	25	SpeI	Amplify out of XIAP-CROP to ligate into pHis1522
8	SpeI_XIAP_R_v2	AAAGGGATCCCGCCAGTTTACTA GC	Pfu	64	4	25	BamHI	Amplify out of XIAP-CROP to ligate into pHis1523
9	TcdC-R	CCGAGCCTCGAGATTAATTTTCTC TACAGCTATC	Pfu	47	1.5	35	XhoI	Amplify out of TcdC-pET30a and remove residues 1-151
10	TcdC152-F	GGAACCCATATGAAAGACGACGA AAAGAAAGC	Pfu	47	1.5	35	NdeI	Amplify out of TcdC-pET30a/TcdC-pET28a and remove residues 1-152
11	TcdC152N-R	CAGTGCCTCGAGTTAATTAATTTT CTCTACAGCTATC	Pfu	47	1.5	35	XhoI	Amplify out of TcdC-pET28a and remove residues 1-153

Table 2.3: Chimera sequencing primers

Primer Number	Primer Name	Primer Sequnce	EGF TOPO	BLyS TOPO	VEGF TOPO	CAE	CAV	PAE	PAV	PAB	CAB	BAB	LAE	LAV	LAB
1	M13-F	CAGGAAACAGCTATGAC	X	X	X										
2	M13-R	GTAAAACGACGCCAGT	X	X	X										
3	W5	GTTGATGGAT AAACTTGTTC				X	X	X	X	X	X	X	X	X	X
4	W3	CATCCAGCCTCGCGTC								X	X	X		X	X
5	TcdA 2326-R	CGCTTGTGTTGAATTCATC				X	X					X			
6	TcdA 2326-F	GATGAATTCAACACAAGCG					X	X	X	X	X	X	X	X	X
7	TcdA 3065-F	CAAAAGTAATGGTGAGTC				X	X	X	X	X	X	X	X	X	X
8	TcdA 3727-F	CTATTTTAATCATTTGTCTG				X	X	X	X	X	X	X	X	X	X
9	TcdA 4258-F	GCCAACTATAACTACTAAC				X	X								
10	TcdA 4502-F	CTTATTATAGGCAATCAAAC				X	X	X	X	X	X	X	X	X	X
11	TcdA 5034-F	CCGTATACTCATCTTACC				X	X			X					
12	TcdA 5268-F	CATCGTCATCTAAAAGCAC				X	X	X	X	X	X	X	X	X	X
13	TcdA 5720-F	TCATTAGGATATATAATGAG								X					
14	CF	GAACTTCTGCCGTGAGTCC				X	X				X				
15	VF	GAAGGAAGAGGAGAGGG					X							X	X
16	PF	TGTCCCCCACGGACC						X	X						

Expression and Purification of TcdA and Chimeric Constructs

Before attempting purification of the new chimera constructs, TcdA was used as practice purification following the laboratory protocol. TcdA expression was induced in *B. megaterium* and purified with a Ni-NTA column. The lab purification protocol worked as evidenced by a strong TcdA band in the primary elution fraction (E2) (Figure 2.8A). Using the same protocol, CAE and PAB were expressed and purified over a Ni-NTA column, but not evidenced in the primary elution fractions (Figure 2.8A and B). Expression and lysis conditions were altered after

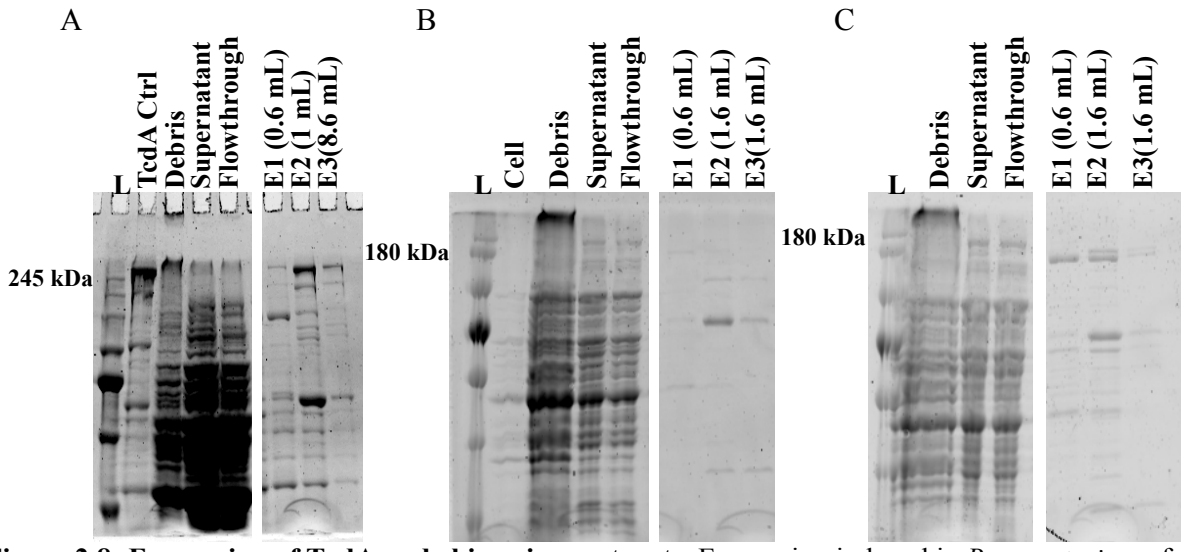


Figure 2.8: Expression of TcdA and chimeric constructs. Expression induced in *B. megaterium* of A) TcdA, 310 kDa B) CAE, 210 kDa C) PAB, 240 kDa and partially purified over Ni-NTA column. The primary elution fraction (E2) of each protein prep shows clear evidence of soluble TcdA, but not CAE or PAB.

multiple attempts to purify chimeras using the protocol for TcdA purification failed to produce soluble protein (Figure 2.9). Using PAB, chimera expression was induced at OD_{600} 0.67 or 1.1 for two hours or induced at OD_{600} 0.4 for 16 or 23 hours. In both instances there was no evidence of soluble chimera in the supernatant, suggesting that altered induction conditions do not improve PAB expression. In some instances, recombinant proteins require small molecule additives to improve expression efficiency [53]. Tween (0.01% v/v) and sucrose (5% w/v) showed no appreciable difference in PAB expression as determined in the supernatant. It is important to

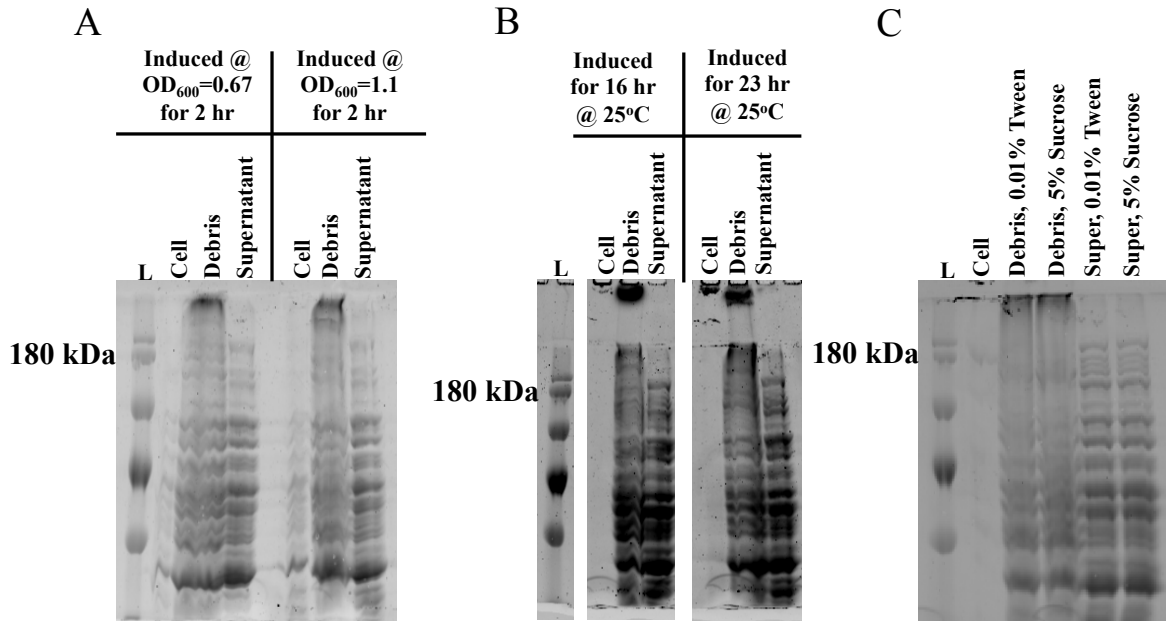


Figure 2.9: Altering expression and lysis conditions attempting to improve PAB isolation. A) PAB expression was induced at increased OD_{600} (0.67,1.1) for the normal duration. Cells were lysed and visualized for chimera expression in the supernatant. No appreciable amount of PAB can be seen. B) PAB induction was carried out at a decreased temperature for long time periods (25°C for 16/23 hours), negative for soluble PAB in the supernatant. C) Addition of Tween or sucrose osmolytes addition during lysis did not improve PAB isolation.

note that this result on its own does not definitively conclude a lack chimera expression. TcdA exhibits a very minimal amount recombinant protein in the supernatant fraction (Figure 2.8A). However, the induction conditions and stabilizing additives study, together with experiments run by Amit Kumar in parallel, strongly suggest that the chimeras are not properly expressing.

Cargos and RBDs within the chimeras are eukaryotic in origin, and the translation of these proteins could rely on tRNA that are not abundant within *B. megaterium*. A codon usage table (Kazusa) identified six codons that account for less than 10% of the total codons used for that particular amino acid (problem codons) within *B. megaterium*. The differential usage of the problem codons within each cargo and RBD are summarized in Table 2.4. The two chimeric constructs purified (CAE and PAB) contain a strong negative bias for the CUG codon in the cargo moiety. The heavy negative bias at the N-terminus could be stalling protein translation, ultimately leading to termination.[54, 55] Amit Kumar attempted to verify this hypothesis by

Table 2.4: Negative codon bias of TcdA and chimera domains

			Differential codon usage										
			Cargo				Native TcdA				RBD		
	B. meg codon usage	Prop. of total codons for amino acid	Casp9	XIAP	p35	Bax (native)	GT	CPD	TD	CROP	EGF	VEGF	BoNT
CUG (L)	7.7	0.0842	-50	14.1	-67.2	-39.2	5.5	3.8	6.7	-48.1	-48.9	-19.9	5.4
UCC (S)	4.8	0.0745	-21.6	6	-50.6	-16	-1.7	-6.8	1.9	2.5	-14.1	-7.5	4.8
UCG (S)	4.9	0.0761	-7.1	0	-14.6	4.9	4.9	4.9	2	1.5	4.9	-1.2	4.9
CCC (P)	3.1	0.0881	-18.5	4	-23	-22.9	3.1	3.1	3.1	2	-15.8	-30.6	3.1
CGG (R)	2.5	0.0622	-23.9	6	-13.8	-13.1	2.5	2.5	2.5	-0.9	2.5	-31.2	2.5
AGG (R)	3	0.0746	-13.8	12.1	-0.3	-2.2	-3.5	-0.9	3	-0.4	3	-3.1	-8.5

B. meg codon usage = frequency of codon /1000 codons; Prop. of toal codons = (*B. meg* usage of individual codon)/Σ(usage for all codons for amino acid); Differential codon usage = (Usage of domain) – (*B. meg* usage); Highlighted values are considered problematic.

expressing chimeras that contain XIAP as the cargo. Despite the lack of negative codon bias within XIAP, there was still no evidence of protein expression. Comparing codon usage of luciferase-TcdA (Luc-TcdA), a construct expressed previously in this lab, to the new chimeras showed the presence of a problem codon near the site of translation initiation (Figure 2.10). Cargo subcloning abolished a 5'-SpeI site in Shuttle and in doing so an AGG problem codon was introduced. The exact effect of codon bias on translation is under debate and the impaired translation could be due to RNA structure and not bias.[56] Regardless of mechanism, the current hypothesis is introduction of a problem codon causes the lack of chimeric expression, and work correcting the codon was started but is incomplete.

Luc-TcdA M V Q T S R S M K
atg gtc caa act agt aga tct atg **aaa**

Cargo-TcdA-RBD M V Q T R R S M T
atg gtc caa act ag**G** aga tct atg **act**

Figure 2.10: Problem codon introduced near translation start site. During cargo subcloning, abolishing a 5'-SpeI site lead to the introduction of the problem AGG codon. Translation start site indicated in red, the added nucleotide introduced in subcloning and creating the problem codon is capital and underlined. Start of the cargo genes are colored purple.

While fixing the problem codon near translation initiation of chimeric proteins, focus shifted to expressing luciferase (Luc) containing chimeras in the hopes of investigating cell-specific delivery and intracellular localization. First, Luc-TcdA-EGF (LAE) was purified using

the lab protocol. After the first nickel column, LAE is evidenced in the primary elution fraction (E2) (Figure 2.11). Before completing purification, whole cell lysates were assayed for luciferase activity to ensure the chimeras were expressed and active. Additionally, whole cell lysates of *B. megaterium* carrying LAV and a TcdA construct that has GT removed (Cargoless) were also assayed. Whole cell lysates were standardized for total protein content by a Bradford assay before conducting the experiment. LAE and LAV both exhibited luciferase activity, while cargoless had no signal (Figure 2.12). These data suggest that LAE is expressed at a higher level than LAV and both constructs are properly expressed and active. Measuring intracellular protein delivery with the luciferase assay requires a stable signal over a long time frame. Optimization of this assay was done with Luc-TcdA at multiple concentrations, with the addition of 50,75, or 100 uL of reaction solution (Figure 2.13). The lowest volume of reaction solution had the most stable signal over the timeframe studied, but for each concentration of Luc-TcdA the signal decay was rapid, losing 50% of the signal within three minutes for 150 nM Luc-TcdA (Figure 2.13A). The

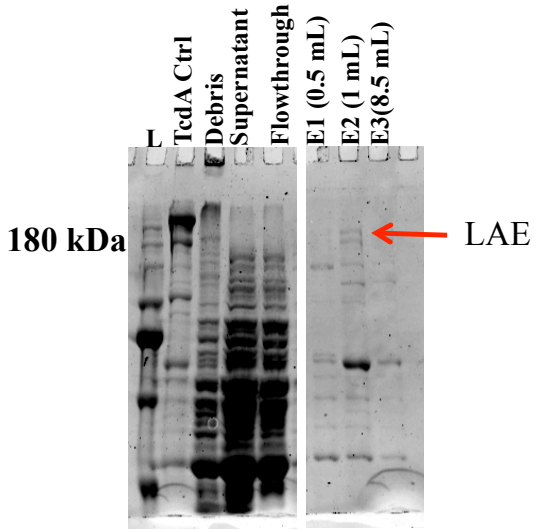


Figure 2.11: Purification of LAE. LAE (177 kDa) expression was induced in *B. megaterium* at OD_{600} 0.4 for 3 hours, lysed via sonication, and partially purified over a nickel column. Soluble protein evidenced in primary elution fraction.

lower concentration of protein lead to more stable measurements over time but contain large

amounts of error in the measurements. Difficulties purifying the luc chimeras and inability to reproducibly and stably conduct the luciferase assay lead to ceasing work on this project.

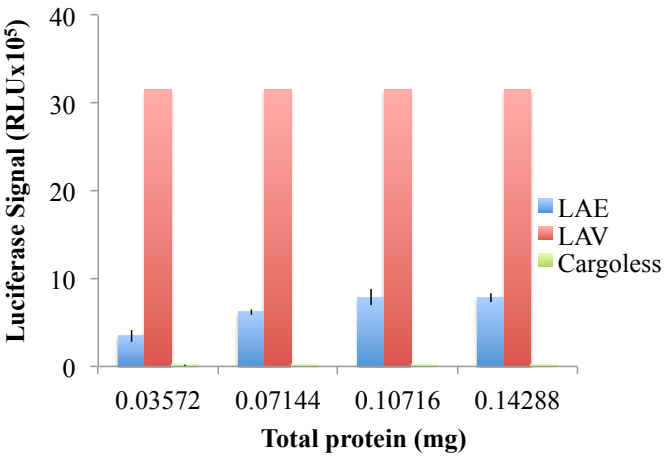


Figure 2.12: Luciferase activity assay of whole cell lysates. *B. megaterium* containing LAE, LAV, and Cargoless were lysed and clarified. The resulting lysates were measured for total protein using a Bradford assay and normalized. LAE and LAV both show luciferase activity, though LAV at a much lower level. The negative control shows no activity validating that LAE and LAV are properly expressed and active. Activity expressed as mean +/- standard deviation

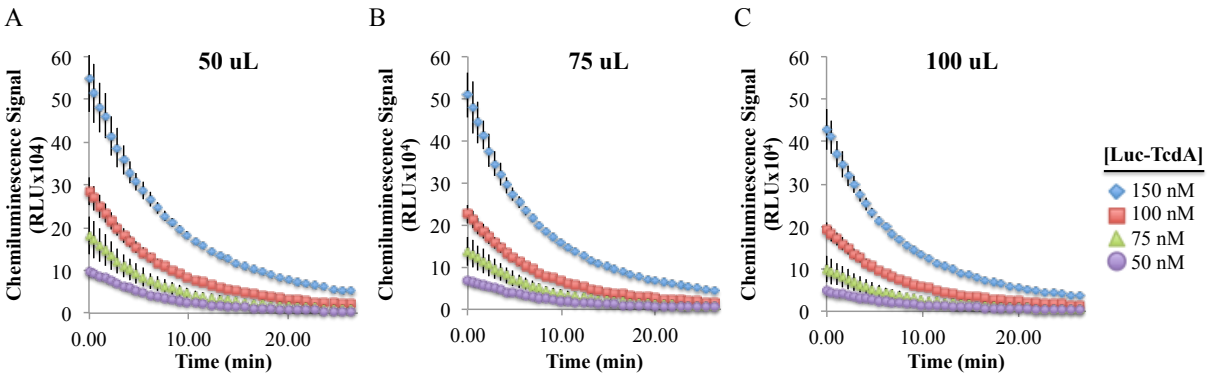


Figure 2.13: Stability of Luc-TcdA luciferase assay. Luc-TcdA (50,75,100,150 nM) was plated in triplicate in a 96 well plate. A) 50 uL B) 75 uL C) 100 uL of reaction solution was added to each well. Chemiluminescent signal was measured 35 times over 25 minutes. 50 nM Luc-TcdA had the most stable signal for each volume of reaction solution, but the overall signal is very low and poor for a cellular assay.

The subcloning strategy to make multiple Cargo-RBD chimeras was intended to allow for swift interchanging and straightforward work process, which was true for creation of the initial constructs. However, introduction of a problematic codon at translation initiation lead to poor or absent expression of chimeric proteins that contained XIAP, Casp9, p35 or Bax as the cargo moiety. Luc chimeras were able to be expressed as active proteins but could not be purified in appreciable quantities. Luc-TcdA was purified completely, but the activity was no reproducible.

Table 2.5: Chimera subcloning clone reference table

Plasmid Code	Plasmid Name	Digest verified (Book,page)	Digest Enzymes	Sequence Verified	Comments
pAB_70401	EGF_TOPO	Y (I,49)	NsiI	Y	Amplify out of commercial vector and pass into shuttle
pAB_70402	BLyS_TOPO	N	n.a	Y	Amplify out of commercial vector and pass into shuttle
pAB_70403	VEGF_TOPO	N	n.a	Y	Amplify out of commercial vector and pass into shuttle
n.a	Cas9_TOPO	N	n.a	N	incomplete
pAB_80401	EGF_Shuttle	Y (I,132)	KpnI/SpeI		50:50 ligation with Cargo-TcdA
pAB_80403	VEGF_Shuttle	Y (I,132)	KpnI/SpeI		50:50 ligation with Cargo-TcdA
pAB_80404	Casp9_TcdA_EGF	Y (II,15)	XhoI/SpeI	Y	complete chimera
pAb_80405	Casp9_TcdA_VEGF	Y (II,15)	XhoI/SpeI	Y	complete chimera
pAB_80406	p35_TcdA_EGF	Y (II,27)	XhoI/SpeI	Y	complete chimera
pAB_80407	p35_TcdA_VEGF	Y (II,27)	XhoI/SpeI	Y	complete chimera
pAB_80408_A	BoNT_Shuttle	Y (II,36)	XhoI/PvuI	N	50:50 ligation with Cargo-TcdA
pAB_80408	p35_TcdA_BoNT	Y (II,40)	BamHI/SpeI	Y	complete chimera
pAB_80410	Casp9_TcdA_BoNT	Y (II,42)	BamHI/SpeI	Y	complete chimera
pAB_80411	Bax_TcdA_BoNT	Y (II,43	BamHI/SpeI	Y	complete chimera
pAB_80412	Luc_TcdA_EGF	Y (II,112)	DraIII	Y	complete chimera
pAB_80413	XIAP-pHIS	n.a	n.a	n.a	New shuttle vetor to alleviate problem codon at TSS
pAB_80414	Luc_TcdA_VEGF	Y (II,143)	SpeI	Y	complete chimera
pAB_80415	Luc_TcdA_BoNT	Y (II,144)	SpeI	Y	complete chimera
n.a	Cas9_shuttle	N	n.a	N	incomplete
n.a	XIAP2_TcdA_EGF	N	n.a	N	incomplete

Chapter 2 Materials and Methods

Cell pelleting and clarification of lysate carried out on Beckman Coulter Avanti J-E centrifuge. SDS-PAGE and agarose gels imaged with Typhoon 9210 variable mode imager. Ligations performed with T4 DNA ligase (NEB) according to protocol. FPLC purifications used Superdex200 size exclusion column operated by Bio-Rad NGC Chromatography system. Sequencing was completed by Applied Genomics Technology Center (http://www.agtc.med.wayne.edu). PCR reactions carried out with either Taq polymerase (Fisher, 1 mM Taq buffer, 2.5 mM $MgCl_2$, 0.8 mM dNTPs, primers 0.2 μm, 5 units Taq) or Pfu polymerase (1x Pfu Buffer, 0.8 mM dNTPs, primers 0.2 μM, 1 unit Pfu), both following published protocols using values. Temperature conditions for PCR include 30s/95^{o}C melt, 25s anneal time, 78^{o}C elongation temp. Anneal temperature and elongation times included in Table 2.2.

Creation of RBD-Shuttles

EGF (ATCC, 59957), VEGF (Life Technologies, IOH81488), and BLyS (Sino Biological, HG10056) were PCR amplified out of commercial vectors with primers 1-6 according to profile above using values specified in Table 2.2. PCR products purified with Microelute cycle-Pure Kit (Omega, D6293-02) and TOPO cloned (Invitrogen, 45-0641) according to manufacturer protocol (to introduce 5'-SpeI digestion site and 3'-SrtA tag/KpnI digestion site). T-10 electrocompetent cells transformed with ligated RBD-TOPO vectors via electroporation using a Bio-Rad micropulser. A synthetic gene for BoNT for nucleotides 2583-3888 was purchased (Gene Script) and to contain the same features within pUC57-Kan. After isolating plasmids EGF-, VEGF-, and BLyS-TOPO with E.Z.N.A Plasmid DNA Mini Kit II (Omega, D6945-02), they were verified via digestion and sequencing. Each RBD was digested

out of TOPO with engineered sites and ligated into pWH1520* and transformed into T-10 cells via electroporation, then verified via endonuclease digestion creating RBD-Shuttle vectors.

Creation of Cargo-TcdA-RBD Chimeras

p35-TcdA and XIAP-TcdA (created and verified by Amit Kumar) were digested with SpeI/XhoI and Casp9-TcdA digested with SpeI/MluI and crystal violet gel purified with E.Z.N.A Gel Extraction Kit (Omega, D2500-01). VEGF-, EGF-, and BoNT-Shuttle digested according to Cargo-TcdA constructs to be combined and gel purified. Purified fragments were ligated and transformed into T-10 cells via electroporation. Constructs were verified by digestion and sequencing.

SrtA Purification

BL21(DE3) cells transformed pET23 containing SrtA (gracious gift from Dr. Woody Guo) were grown to an OD_{600}=0.4 in 1L-LB miller media and induced with 1 μM IPTG for 3.75 hrs. After induction cells were pelleted (4,000 g, 10 min, 4^{o}C, F10.5 rotor) and stored at -80^{o}C. A cell pellet from 0.5L prep thawed with Lysis buffer (50 mM Tris-HCl, 150mM NaCl, pH 8.0) containing ¼ tablet of cOmplete, mini, EDTA free protease inhibitor cocktail tablet (Sigma-Aldrich, 11836170001). Solubilized cells lysed via sonication (Branson Digital Sonifier) for 3 cycles, 25s pulses, 37% power, 60s rest between cycles. Cell lysate clarified by centrifugation (15,000 rpm, 30 min, 4^{o}C, JA-17 rotor) and passed through 0.8μm and 0.2 μm (Pall Corporation Acrodisc Syringe Filter, Supor Membrane, 4618/4612). SrtA purified with a HisTrap HP column (GE Healthcare). The column was washed with 5 column volumes (CV) H_2O, charged with 2 CV 100 mM $NiSO_4$, washed with 5 CV H_2O, and equilibrated with 5 CV Lysis Buffer. The filtered lysate was loaded onto the column, washed with 20 CV of Lysis buffer then with 5 CV wash buffer (50 mM Tris-HCl, 250 mM NaCl, 25 mM imidazole, pH 8.0). SrtA was eluted with

10 CV of elution buffer (50 mM Tris-HCl, 150 mM NaCl, 500 mM imidazole, pH 8.0) collected in 10 1mL aliquots. Purification characterized with SDS-PAGE electrophoresis (5% stacking, 15% running 29:1) at 200V for 40 minutes, visualized with Coomassie stain. Pure fractions were pooled and protein concentration was determined with a Bradford assay using Bio-Rad protein assay according to protocol

Purification of RhlB^{SrtA}

Glycerol stock of RhlB-BL21(DE3) (a gift from Amit Kumar) containing the SrtA recognition sequence used for overnight culture grown in Kanamycin (30 ug/mL) LB from. 500 mL of LB inoculated with overnight culture to an OD_{600} = 0.4 and induced with 1 μM IPTG for 3 hours. Cell pellet sonicated for 3 cycles, 25s pulses, 37% power in lysis buffer (25 mM Tris-HCl, 300 mM KCl, 10 mM imidazole, pH 8.5) and clarified at 15,000 rpm for 30 minutes at 4^{o}C. RhlB was loaded on a HisTrap HP column equilibrated with lysis buffer, washed with 5 CV lysis buffer, 5 CV wash buffer 1 (10mM HEPES, 300 mM KCl, 50 mM imidazole, pH 7.5), 5 CV wash buffer 2 (10mM HEPES, 300 mM KCl, 1M urea, pH 7.5), 5 CV wash buffer 3 (10 mM HEPES, 1M KCl, 10 mM imidazole, pH7.5), and eluted in 8 CV (10 mM HEPES, 300 mM KCl, 300 mM imidazole, pH 7.5). Dialyzed (10 mM HEPES, 300 mM KCl, 100 μM EDTA, 1 mM DTT, pH 7.5) for 12 hours, changing the buffer halfway through.

SrtA Activity Assay

To test SrtA activity, 25uM RhlB and 0,35, or 50 μM GGG-FITC substrate in SrtA reaction buffer (50 mM Tris-HCl, 60 mM $CaCl_2$, 150 mM NaCl) reactions were ran at 40 ^{o}C 0.5-7 hours, after which each was treated with a stop solution (10 mM EDTA). Unreacted GGG-FITC was removed using equilibrated micro bio-spin 6 chromatography columns (BIO-Rad,732-6200) eluting at 1,000 g for 4 minutes. Filtrates were analyzed for A_{495} and A_{280}.

***Bacillus megaterium* Transformation and Protein Purification**

All chimeric toxin work was carried out according to Biosafety Level II requirements.

B. megaterium protoplasts generated as previously described[52]. Sequence verified chimeric toxins were transformed following the MoBitech protocol. Properly transformed cells were grown in 1L LB Miller media to an OD_{600}=0.4-0.8 then induced with 0.5% or 1% D-Xylose for 2.5-20 hrs at 37°C or 25°C. After induction, cells were pelleted (9,000 g, 10 min, 4°C, F10.5 rotor) and stored at -80 °C until purification. 1L of pelleted cells were thawed on ice with lysis buffer (50 mM Sodium Phosphate, 300 mM NaCl, 10 mM Imidazole, pH 8.0 or 50 mM Tris-HCl, 150 mM NaCl, 10 mM imidazole, pH 7.5) with ½ protease tablet alone, protease tablet with 0.01 % Tween 20, or protease tablet with 5% Sucrose. Solubilized cells were sonicated 5 times, 30s cycles, at 37% power, with 90s rest period between cycles. Samples were then clarified by centrifuge (15,000 rpm, 40 min, 4°C, JA-17 rotor). Supernatant was then filtered through 0.8 μm filter, then a 0.2 μm filter. Sonication was then analyzed via SDS-PAGE (4% stacking gel and 10% running gel 29:1) at 165V for 3 hours. Filtered supernatant was then loaded onto HisTrap HP column equilibrated with either lysis or binding buffer (50 mM Sodium Phosphate, 300 mM NaCl, 20 mM imidazole, pH 8.0). Column washed first with 10 CV lysis or binding buffer and then with wash buffer (50 mM Sodium Phosphate, 300 mM NaCl, 50 mM Imidazole, pH 8.0). Protein eluted with 10 CV Elution buffer (50 mM Sodium Phosphate, 300 mM NaCl, 250 mM Imidazole, pH 8.0), with the first ~0.5 mL being collected, the next 1-5 mL were collected and termed primary elution fraction, the final volume was also collected. The column was then stripped with strip buffer (50 mM Sodium Phosphate, 300 mM NaCl, 50 mM EDTA, pH 8.0) and stored in 20% EtOH.

Primary elution fractions were filtered with 0.2 μm filter directly into a 3 mL plastic syringe with the bottom sealed by parafilm. The plunger was carefully placed back in the syringe and parafilm removed to inject sample into FPLC super loop. All loaded samples were diluted with lysis buffer in a second injection into super loop to wash protein from the injection line and dilute the sample to prevent overloading the column. Elution fractions from FPLC that contained protein were then concentrated over HistTrap HP column and eluted in < 2.5 mL of elution or strip buffer.

His6 Dot Blot Assay

Immunoblot PVDF membrane (BIO-RAD, 162-0175) permeabilized with 100 % methanol and equilibrated in 1x transfer buffer (25 mM Tris-HCl, 192 mM glycine, 0.1% SDS, pH 8.3). Protein samples transferred to the equilibrated membrane using a Hybri-slot vacuum manifold. Staining for His6 proteins carried out with Thermo Scientific His Probe- HRP according to protocol (ThermoScientific, HisProbe: 15165, SuperSignal West Pico Chemiluminescent Substrate: 34087).

Luc Activity Assay

Luc activity measured with Tecan Genios Plus Plate reader under luminescence setting. Assays were carried out according the supplied protocol with *Gaussia* Luciferase Cellular Assays (XACTAGEN, 31001). In short, 20 uL protein samples of varying concentrations were added to white 96 well plates. Glum.1 Assay Solution was administered to each well using a multichannel pipette a row at a time. Measurements made in kinetic mode for 35 cycles. For whole cell lysates total protein was quantified using a Bradford assay normalized with BSA.

***B. megaterium* Codon Bias**

Codon usage determined using Kazusa Codon Usage Database (http://www.kazusa.or.jp/codon/). These data were then used to determine the usage of each codon compared to all possible for each amino acid. Codons that account for less than 10% of all codons for an amino acid are termed problem codon. Gene sequences for each cargo and RBD as well as TcdA analyzed for codon usage. Differential codon usage = (*B. megaterium* codon usage) – (codon usage).

Chapter 3
Mechanism of TcdC control of *C. difficile* Toxin Expression

Introduction

In 2011, nearly half a million instances of *C. difficile* infections (CDI) resulted in almost 29,000 deaths in the United States.[57] With the morbidity and mortality of CDI rising over the last two decades, understanding the pathogenesis and mechanism of control for *C. difficile* virulence factors, TcdA and TcdB, is of paramount importance. The genes that produce these two toxins are located in an ~18.5 kb region of the genome referred to as the pathogenicity locus (PaLoc, Figure 3.1).[58] Additionally the PaLoc contains proteins involved in controlling toxin production

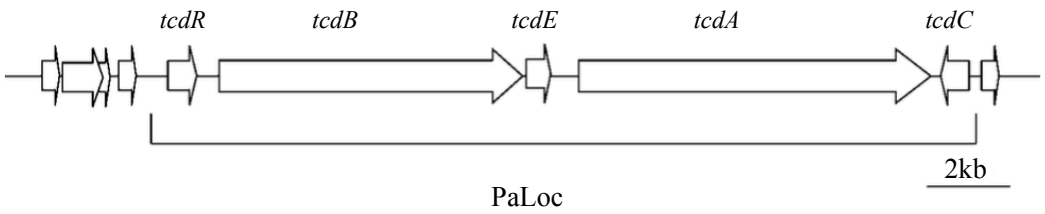


Figure 3.1: Pathogenicity locus of *C. difficile*. The 18.5 kb pathogenicity locus (PaLoc) of *C. difficile* contains the genes encoding virulence factors TcdA and TcdB. The PaLoc also contains the genes for a putative holin-like protein for toxin secretion, TcdE, an alternative sigma factor responsible for inducing toxin expression, TcdR, and the putative anti-sigma factor that inhibits toxin expression (TcdC).[58]

(*tcdR, tcdC)* and secretion (*tcdE).* There is consensus that the alternative sigma factor, TcdR, increases TcdA/B expression, and higher levels of energy sources such as glucose, biotin, and amino acids decrease toxin production.[59-62] However, the exact role TcdC plays in controlling expression of TcdA/B is unknown. Multiple reports claim that point or frame shift mutations lead to hyper-virulence in *C. difficile* strains.[58, 63, 64] Others have found that mutations in TcdC do

not predict increased TcdA/B expression or hyper-virulence.[65, 66] The field has yet to reach a consensus on how or if TcdC affects toxin synthesis.

While work continues on elucidating the importance of TcdC in CDI, van Leeuwen *et. al.* have recently reported that TcdC binds DNA fragments adopting the G-quadruplex (G4) structure.[67] G4s are a class of secondary structure with four guanines hydrogen bonded through Hoogsteen base-pairing (Figure 3.2A).[68] When these motifs occur in multiples, they stack into a highly stable structure (Figure 3.2B). G4s have been linked to many biological processes such as transcription, telomere maintenance, prokaryotic evasion of eukaryotic immune system, and neurological diseases.[69, 70] At first glance, the concept of TcdC binding G4s does not seem biologically relevant, given that the *C. difficile* is an AT-rich genome with 29% GC content, compared to 52% GC content of *Escherichia coli*.[71, 72] My goal was to validate the biological importance of TcdC's affinity for G4 nucleic acids and elucidate the mechanism by which TcdC controls toxin expression.

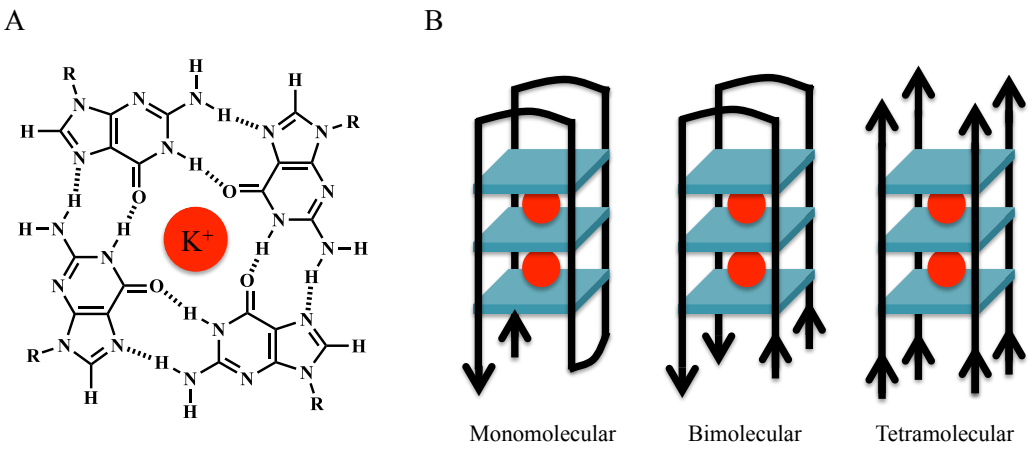


Figure 3.2: Structure of G-quadruplex. A) Four guanines hydrogen bond through Hoogsteen base-pairing forming a G-quadruplex (G4). The resulting central cavity is stabilized by a metal cation, most often K^+, shown here, or Na^+. B) Successive G4s stack in a planar fashion and can be composed of a single strand of nucleic acid, two strands of nucleic acid, or four stands of nucleic acid [68].

Results

TcdC's affinity for G4 DNA lead to the hypothesis that sequences within the *C. difficile* genome leads to transcribed oligonucleotides that contain G4 structures. These sequences, if present, are hypothesized to behave as a protein sink, where G4 transcripts bind with TcdC inhibiting it from fulfilling the biological role of controlling toxin expression.

Mining of *C. difficile* Genome for G4 Plausibility

Genomes of three difference *C. difficile* strains were mined for the G4 motif $G_4N_{1\text{-}7}G_4$, where N is any nucleotide within R using the gregexpr function.[73] The resulting 61 predicted G4 forming sequences between strains 630, CD196, and R20291 are summarized in Table 3.1. The identified sequences map to many genes throughout the *C. difficile* genome of each strain. The sequences highlighted are of high interest because they belong to transposon (Tn) 6103.

The genome of *C. difficile* has acquired many transposable genetic elements, often associated with antibiotic resistance genes.[74-77] If G4 sequences exist in *C. difficile* that bind TcdC, it seems more likely that the G4 would be located within an acquired piece of DNA rather than the natural AT-rich genome. Using published Tn found within *C. difficile*, twelve G4 sequences were identified within predicted transcripts (Table 3.2). From this table, five sequences were selected for binding studies (Table 3.3). (+) Ctrl corresponds to the oligonucleotide identified previously to bind with TcdC and (-) Ctrl contains no G4 motif.

G4 Oligonucleotide binding study

The oligonucleotides (oligos) in Table 3.3 were validated for the formation of G4 using the dye ETC that specifically binds nucleic acids adopting the G4 structure (Figure 3.3A).[78, 79] All of the predicted G4 oligos were stained with ETC and (-) Ctrl failed to stain confirming the

***Table 3.1 Predicted G4 sequences indentified in multiple strains of* C. difficile**

Number	Sequence	Strain	Genome Start	Genome Stop	Hypervirulent?
1	ggggccgggggt	630	31780	31791	no
2	ggggatttggggt	630	55736	55748	no
3	ggggccgggggt	630	130408	130419	no
4	ggggccgggggt	630	130706	130717	no
5	ggggaggtgggggа	630	671858	671871	no
6	ggggttgcagggggа	630	775131	775145	no
7	ggggttgctatgggga	630	1016437	1016452	no
8	gggggggttggggt	630	1028455	1028467	no
9	ggggtaaggggc	630	1303284	1303295	no
10	gggggagatttaggggc	630	1605949	1605965	no
11	ggggaaataatgggga	630	1676593	1676608	no
12	gggggggggc	630	2125011	2125020	no
13	ggggaagaaaagggggc	630	2350140	2350156	no
14	ggggagaaagtgggga	630	3414810	3414825	no
15	ggggaaaaactgggga	630	3414843	3414858	no
16	ggggtattattggggt	630	3680768	3680783	no
17	ggggccgggggt	CD196	31663	31674	no
18	ggggatttggggt	CD196	55619	55631	no
19	ggggccgggggt	CD196	134033	134044	no
20	ggggccgggggt	CD196	134331	134342	no
21	gggggggagcctgtgggga	CD196	474616	474634	no
22	ggggaggtgggggа	CD196	591183	591196	no
23	ggggttgcagggggа	CD196	697627	697641	no
24	ggggttgctatgggga	CD196	944160	944175	no
25	gggggggttggggt	CD196	956187	956199	no
26	ggggtttttgggggа	CD196	975274	975288	no
27	gggggagatttaggggc	CD196	1464338	1464354	no
28	ggggaaataatgggga	CD196	1537913	1537928	no
29	gggggattcttggggt	CD196	2840174	2840189	no
30	ggggtattattggggt	CD196	3491333	3491348	no
31	ggggaaaggggt	CD196	4027844	4027856	no
46	ggggccgggggt	R20291	31565	31576	yes
47	ggggatttggggt	R20291	55521	55533	yes
48	gggggggagcctgtgggga	R20291	471948	471966	yes
49	ggggaggtgggggа	R20291	588569	588582	yes
50	ggggttgcagggggа	R20291	695148	695162	yes
51	ggggttgctatgggga	R20291	941776	941791	yes
52	gggggggttggggt	R20291	953803	953815	yes
53	ggggtttttgggggа	R20291	972890	972904	yes
54	gggggagatttaggggc	R20291	1461959	1461975	yes
55	ggggaaataatgggga	R20291	1535534	1535549	yes
56	ggggtgcaggggc	R20291	2062669	2062681	yes
57	ggggtcaaggggc	R20291	2062696	2062708	yes
58	ggggacaggggc	R20291	2107351	2107362	yes
59	gggggattcttggggt	R20291	2920731	2920746	yes
60	ggggtattattggggt	R20291	3572064	3572079	yes
61	ggggaaaggggt	R20291	4108630	4108642	yes

Highlighted sequences map to Tn6103

***Table 3.2: Predicted G4 sequences identified in transposons found in* C. difficile**

Number	G4 Sequence	Transposon Name	Transposon Accesion Number	Start	Stop	Predicted QGRS
1	ggggtgcaggggc	Tn6103	BK008007.1	22270	22282	5
2	ggggtcaaggggc	Tn6103	BK008007.1	22297	22309	5
3	ggggacaggggc	Tn6103	BK008007.1	66957	66968	2
4	ggggccgggga	Tn6110	BK008009.1	51233	51243	6
5	cccctgtcccct	Tn6110	BK008009.1	17997	18008	2
6	ccccttgacccca	Tn6110	BK008009.1	51388	51400	5
7	cccctgcaccccg	Tn6110	BK008009.1	51415	51427	5
8	ggggatatgctggggc	Tn6073	BK008006.1	12681	12696	8
9	ggggatatgctggggc	Tn6194Like	HG475346.1	12421	12436	7
10	ggggacgggga	Tn6215	KC166248.1	4462	4472	1
11	ggggctgccccggggc	Tn6215	KC166248.1	6020	6035	3

QGRS (Quadruplex Forming G-Rich Sequences) determined for the entire transcript that contains predicted G4

Table 3.3: Oligonucleotides for TcdC

Name	Sequence
(+) Ctrl	cgttcgatagggatagggag
(-) Ctrl	ggcgatgtcaaacagaatcgt
Tn6103	gccggggtgcaggggcggca
Tn6103-2	gctggggtcaaggggcaacg
Tn6103-3	ggatttaaggggacaggggcag
Tn6110	ggggccggggaaggggcggcg
Tn6073	cctgctggggatatgctggggctt

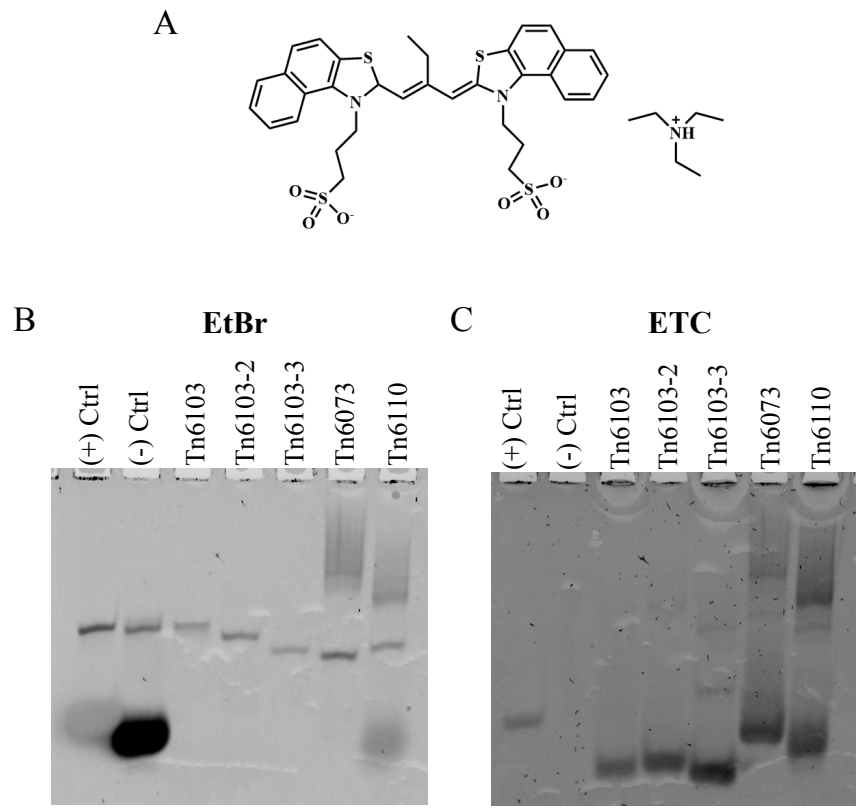


Figure 3.3: Verification of G4 structure for oligonucleotides intended for TcdC binding studies. A) Structure of the G4 specific dye ETC. Native PAGE gel stained with B) EtBr or C) ETC. Positively stained (-) Ctrl with EtBr and not with ETC confirms ETC specificity for G4 structure, but does not label G4 structures. Each oligo excluding the (-) Ctrl, stained positive for G4. However G4 adopting sequences are not positively stained with EtBr. The stability of G4 and presence of internal sodium ion likely preventing EtBr to inter-chelate and positively stain.

identified G4 oligos form G4 structures and the (-) Ctrl does not (Figure 3.3B). Presence of the (-) Ctrl oligo was verified using ethidium bromide (EtBr) staining after ETC had been washed out of the gel (Figure 3.3C). These data confirm that the oligos selected form G4 and can be used in binding assays with TcdC.

TcdC containing a C-terminal His_6 tag ($TcdC^C$) was purified and used for the initial binding studies of (+) Ctrl oligo (Figure 3.4). ^{32}P labeled (+) Ctrl was incubated with $TcdC^C$ at 0^oC or room temperature, both resulted in no evidence of binding. Van Leeuwen *et. al.* performed their binding studies with a truncated TcdC by removing the hydrophobic domain (Figure 3.5). To more closely replicate their binding experiments, subcloning of a truncated form

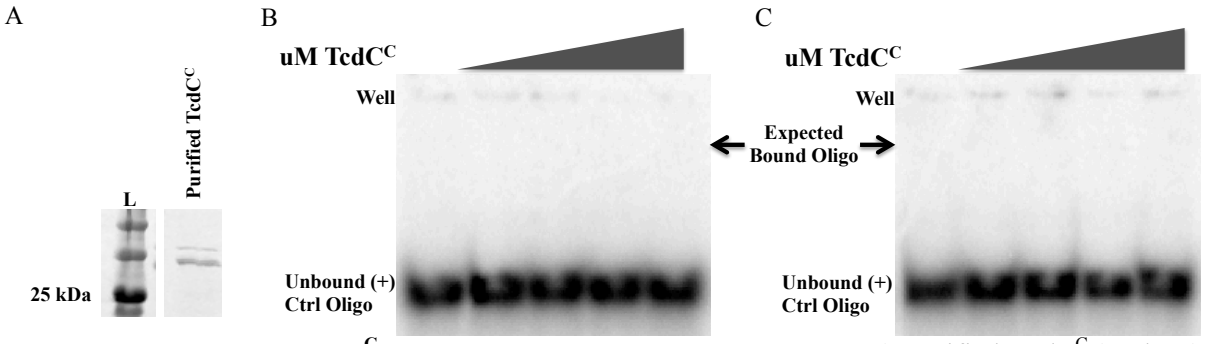


Figure 3.4: Full length TcdC^{C} binding reaction with (+) Ctrl oligo. A) Purified TcdC^{C} (24 kDa) migrates higher on the gel than expected due to the highly positive nature of the protein. Small amount of impurity is present. Binding reactions conducted with 0,1,5,10, or 20 uM TcdC incubated at B) 0^{o}C or C) room temperature for one hour with 20 kcpm/lane of (+) Ctrl, and reactions were visualized using PI exposure cassette. In both cases there is no evidence of TcdC^{C} binding.

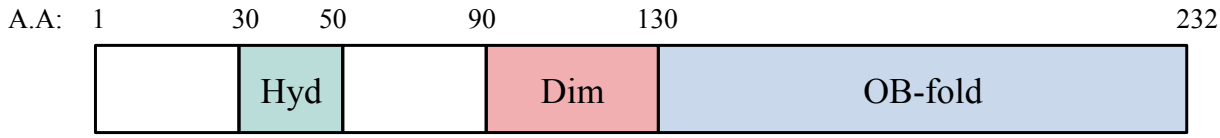


Figure 3.5: Domains of TcdC. Three predicted domains of TcdC, A.A = amino acid, Hyd = hydrophobic domain, Dim = dimerization domain, OB-fold = predicted oligonucleotide binding fold.[67]

of TcdC containing a C-terminal His_6 (TcdC^{152C}) was completed (Alignment Appendix C Figure 1). Binding reactions described above were replicated using TcdC^{152C} and also failed to bind (+) Ctrl oligo (data not shown). A potential explanation for the inability to replicate TcdC binding of (+) Ctrl could be too low an oligo concentration. To test this hypothesis binding reactions were run in the presence of non-radiolabeled (cold) oligo (Figure 3.6). Increasing the concentration of oligo did not lead to binding by TcdC^{152C}. The lack of (+) Ctrl binding by full-length or truncated TcdC lead to the hypothesis that the C-terminus His_6 could be interfering with the predicted OB-fold domain. Subcloning was then completed to create TcdC^{152N} that contains the His_6 at the N-terminus (Alignment Appendix C Figure 2). Due to the timescale of subcloning and the difficulties in running the assays, (+) Ctrl was labeled with Cy3 rather than ^{32}P for the remainder of binding assays. Binding assays were next conducted with the addition of heparin and tRNA, limiting nonspecific interactions to promote TcdC^{152N} binding of (+) Ctrl (Figure 3.7). However there was still no evidence of TcdC^{152N} binding the oligo. No evidence of TcdC^{C}, TcdC^{152C}, or

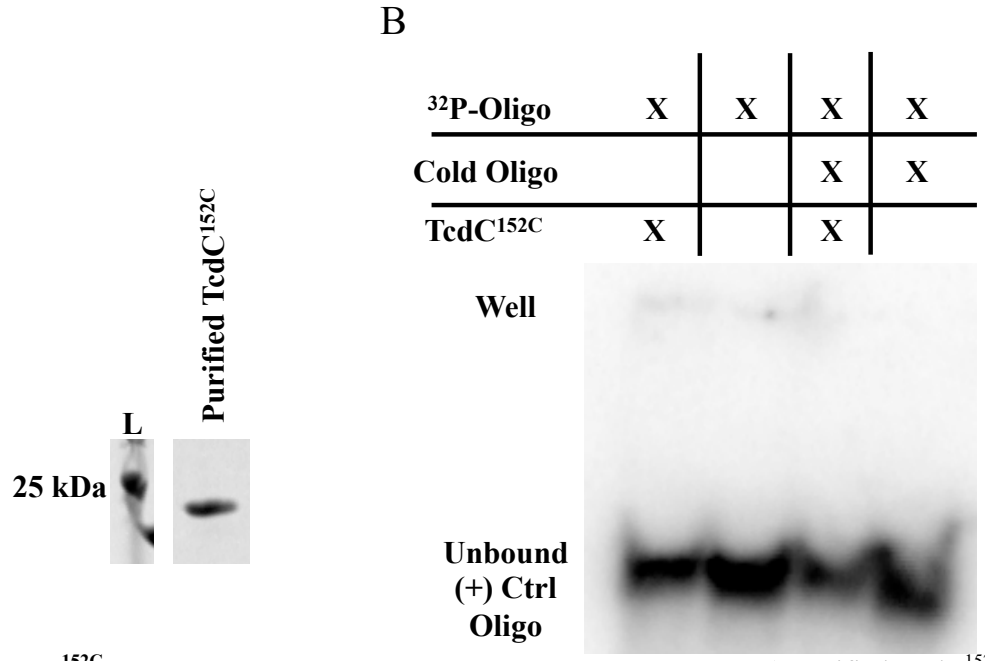


Figure 3.6: **TcdC^{152C} binding assay doped with non-radiolabeled oligo.** A) Purified TcdC^{152C}. B) Binding reaction (520 nM TcdC^{152C}, 20 kcpm/lane ^{32}P-Oligo) doped with non-radiolabeled (cold) oligo (24 nM) increasing the concentration within the reaction without altering the kcpm/lane, which is constant. With added oligo there is still no evidence of TcdC^{152C} binding (+) Ctrl oligo.

TcdC^{152N} binding the (+) Ctrl oligo despite altering reaction conditions, concentrations, and addition tRNA/heparin lead to attaining the TcdC construct (TcdC^{152N*}) used previously, which initially indicated TcdC's specificity for (+) Ctrl [67]. Using TcdC^{152N*} under the same reaction conditions as TcdC^{152N} still did not show any positive signs of binding (+) Ctrl^{Cy3} (Figure 3.8). To date, this work was unable to replicate the work previously done to demonstrate TcdC affinity for G4 containing oligo. This is not to say that the results found previously by Van Leeuwen *et. al.* are not reproducible, only that I couldn't do it effectively.

Conclusion

The exact biological role of TcdC is still under intense debate. Some reports suggest truncation mutations in *tcdC* leads to an increase in toxin production within *C. difficile*, while others report evidence that TcdC plays no role in toxin production. A report that TcdC binds G4-forming oligonucleotides opened a new possible mechanism through which TcdC affects on toxin production could be altered. Through the work presented here, it is still unclear whether

TcdC's affinity for G4 is biologically relevant. Further work will need to be done to validate TcdC binding of G4 oligonucleotides, and whether transcripts in *C. difficile* can adopt this structure and alter TcdC function *in vivo*.

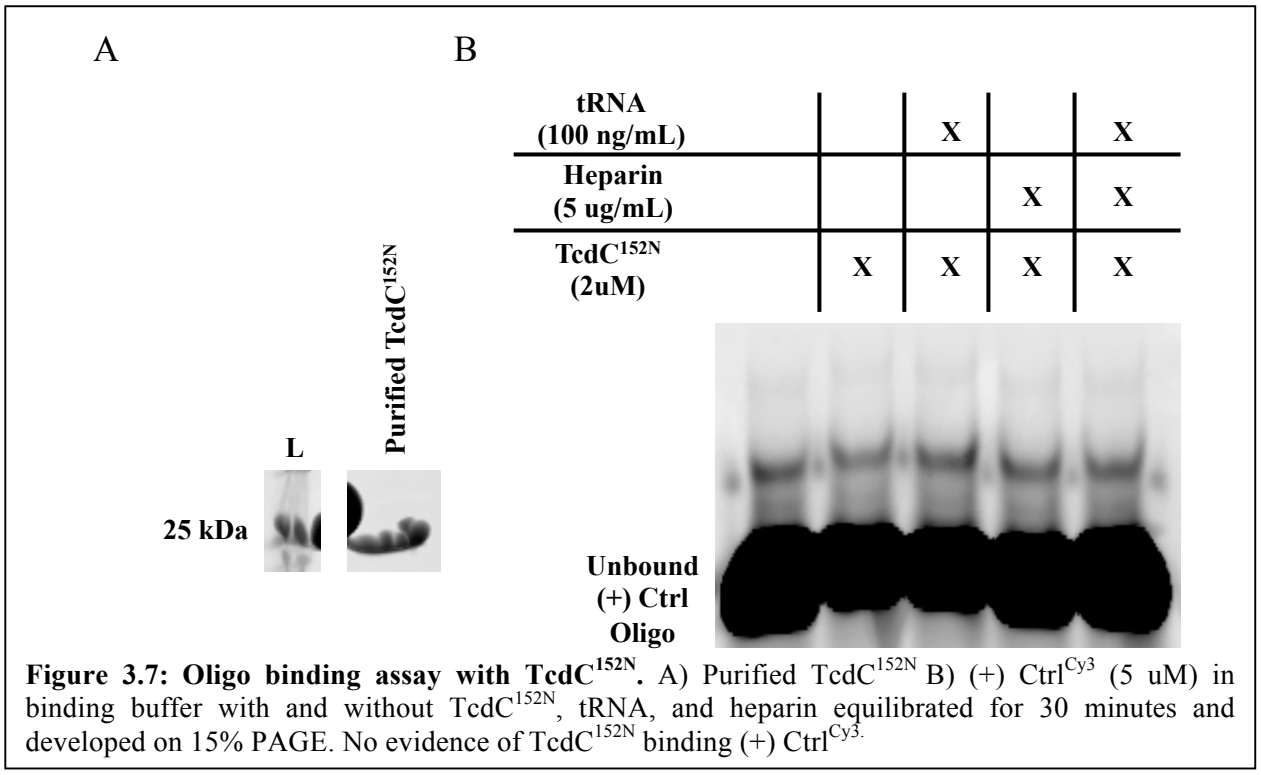


Figure 3.7: Oligo binding assay with TcdC^{152N}. A) Purified TcdC^{152N} B) (+) Ctrl^{Cy3} (5 uM) in binding buffer with and without TcdC^{152N}, tRNA, and heparin equilibrated for 30 minutes and developed on 15% PAGE. No evidence of TcdC^{152N} binding (+) Ctrl$^{Cy3.}$

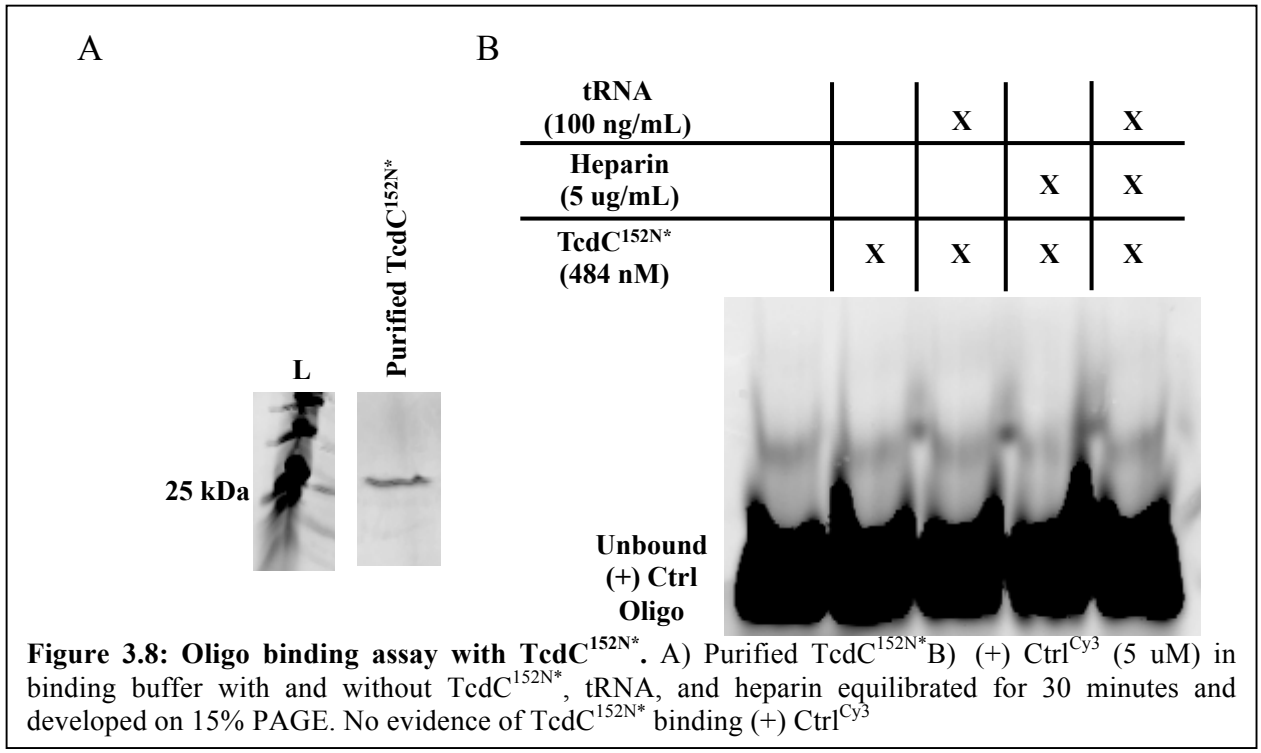


Figure 3.8: Oligo binding assay with TcdC^{152N*}. A) Purified TcdC^{152N*}B) (+) Ctrl^{Cy3} (5 uM) in binding buffer with and without TcdC^{152N*}, tRNA, and heparin equilibrated for 30 minutes and developed on 15% PAGE. No evidence of TcdC^{152N*} binding (+) Ctrl^{Cy3}

Chapter 3 Materials and Methods

Mining *C. difficile* genomes for G4 motif

C. difficile strains 630, CD196, and R20291 we scanned for the G4 motif (G_4-N_{1-7}-G_4) in R. The identified sequences were analyzed with BLAST to find the genomic locus and proposed gene that contains the G4 sequence. Published transposons that are relevant to *C. difficile* (Tn6107, Tn6073, Tn6103, Tn6110) we analyzed for G4 motif as the genomes above. Transposons with identified G4s were analyzed for Quadruplex Forming G-Rich sequences (bioinformatics.ramapo.edu/QGRS/analyze.php).

Analysis of G4 Oligos

G4 oligos were purchased from Sigma-Aldrich and received as a dried pellet. Oligos solubilized in 1x PBS to make 100 uM stock Solutions. G4 formation with and without heating and with 0,50,75, or 100 mM KCl developed on 20 % native page (37.5:1) with 1x TBE at 400 V for 2 hours. Gels either stained with EtBr (0.5 ug/mL, 15 min with 2 15 min H_2O washes) or ETC (20 uM in PBS 1 hour with 1X PBS rinse) and then imaged. The same process was applied to each of the oligos without added heat or salt.

TcdC Expression and Purification

Rosetta(DE3) cells with a plasmid carrying full length $TcdC^C$ or $TcdC^N$ were grown in 10 mL LB containing chloramphenicol (34 ug/mL) and kanamycin (30 ug/mL) for 18 hours at 37 ^oC, 250 rpm. Inoculate fresh 10 mL LB with 100 uL of overnight and grow for 6 hours at 37 ^oC, 250 rpm. Add full 10 mL into 1L LB and grow for 18 hours at 25^oC. Expression induced with 1 uM IPTG for 3-5 hours. Cells harvested at 5,000 g for 15 minutes and pellets stored at -80^oC.

Frozen pellets thawed in lysis buffer (20 mM HEPES, 300 mM NaCl, 10 % [w/v] glycerol, 250 uM TCEP, pH 8.0) and sonicated 8 cycles, at 40% power, 30s pulse, with 90s rest

between pulses. Triton X-100 added to lysate for a final concentration of 0.01% (v/v) and rotated end-over-end for ~15 minutes then clarified at 15,000 rpm for 30 min. Supernatant filtered with 0.8 um then 0.2 um filter and loaded on a Ni-NTA column equilibrated with load buffer (20 mM HEPES, 300 mM NaCl, 0.1% [v/v] Triton X-100, 10 % [w/v] glycerol, 10 mM imidazole, 250 uM TCEP), washed with 30 CV (20 mM HEPES, 300 mM NaCl, 0.1% [v/v] Triton X-100, 10 % [w/v] glycerol, 120 mM imidazole, 250 uM TCEP), and eluted (20 mM HEPES, 300 mM NaCl, 0.1% [v/v] Triton X-100, 10 % [w/v] glycerol, 250 mM imidazole, 250 uM TCEP). Elution fractions were pooled and concentrated using Amicon Ultra 15 Centrifugal Filter Device 3,000 MWCO according to manufacture protocol to a final volume of ~2.5 mL. Protein quantified using Bradford assay.

$TcdC^{152}$ subcloning and purification

$TcdC^{152C}$ amplified out of full length $TcdC^{C}$-pET30a using primers 26 and 27. PCR product digested with XhoI/NdeI and ligated into pET30a. Transformed into T10, verified by digest and sequencing then transformed into Rosetta(DE3).

$TcdC^{152N}$ amplified out of full length $TcdC^{C}$-pET30a primers 27 and 28. PCR product digested with NdeI/XhoI and ligated into pET28a. Transformed into T10, verified by digest and sequencing then transformed into Rosetta(DE3)

$TcdC^{152*}$ was a generous gift from Dr. Jeroen Corver from Leids Universitair Medisch Centrum.

$TcdC^{152C}$, $TcdC^{152N}$, and $TcdC^{152*}$ were expressed and purified according to the protocol above for TcdC with the following buffer changes: Lysis buffer (50 mM $NaPO_4$, 250 uM TCEP, 300 mM NaCl, pH 8.0, 0.1% NP-40 post sonication), load buffer (50 mM $NaPO_4$, 250 uM TCEP, 300 mM NaCl, 20 mM imidazole, 5 % [w/v] glycerol, 0.1% NP-40, pH 7.0), wash buffer

(50 mM $NaPO_4$, 250 uM TCEP, 300 mM NaCl, 120 mM imidazole, 5 % [w/v] glycerol, pH 7.0), elution buffer (50 mM $NaPO_4$, 250 uM TCEP, 300 mM NaCl, 250 mM imidazole, 5 % [w/v] glycerol, pH 7.0), and dialysis buffer (50 mM $NaPO_4$, 250 uM TCEP, 150 mM NaCl, 5 % [w/v] glycerol, pH 7.0). Some elution or wash fractions were further purified using FPLC as described above for chimeric protein purification when needed.

Radiolabeling G4-Oligos and TcdC Binding Assay

Each G4-Oligo was labeled with Perkin Elmer Adenosine 5'-triphosphate[γ-^{32}P] (NEG002A250UC) using PNK from NEB following published protocol and purified by EtOH precipitation. Radioactivity of labeled/purified oligos measured with Beckman Coulter LS 6500 Multipurpose Scintilation Counter. Filter paper was blotted with 1 uL of oligo solution in a scintillation vial. 10 mL of Fischer Scientific ScintiSafe Econo 1 Cocktail and samples were measured for kcpm/uL.

All binding reactions carried out with 20 kcpm Oligo per lane or 703 ng of (+)Ctrl^{FITC} in oligo binding buffer (20 mM HEPES, 50 mM NaCl, 40 mM KCl, 7% [w/v] glycerol, 1 mM EDTA, 0.1 mM DTT at varying concentrations of TcdC in a final volume of 15 or 20 uL. After a 0.5-1 hour equilibration period reactions developed on pre-ran 20% PAGE (37.5:1) unless otherwise stated, at 200V 3-4 hours. For radiolabel experiments gels were dried using BIO-RAD Model 583 Gel Dryer at 50^{o}C 4-8 hours then exposed to Amersham Biosciences PI Screen for at least 2 hours and then imaged. For (+)Ctrl^{FITC} experiments, binding reactions contained tRNA mixture (100 ng/mL) and/heparin (5 ug/mL) and the gel was directly imaged after development.

Appendix A: Secondary student data analyses

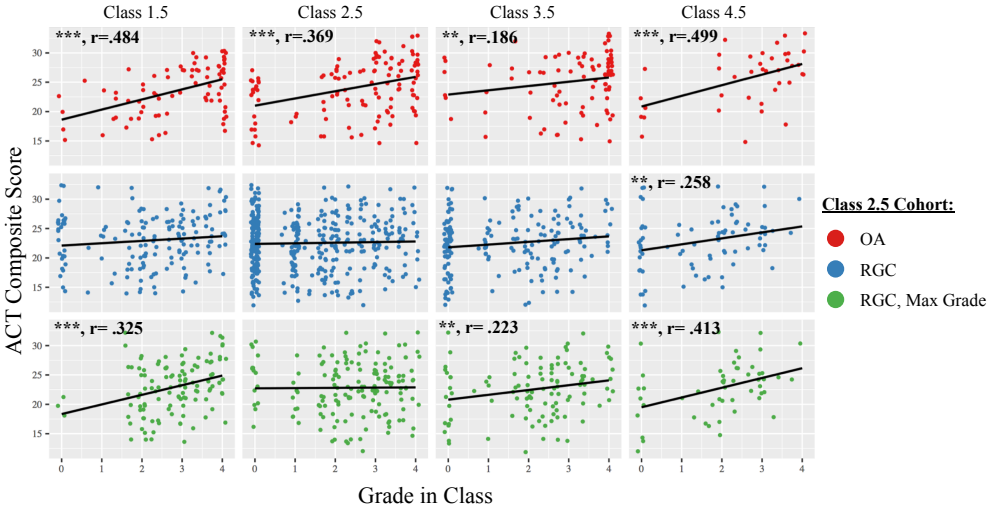


Figure 1-A: Correlation between ACT Composite score and grade for Class 2.5 cohorts. Grades for students attempting Class 2.5 once (OA), retaking Class 2.5 (RGC), and the max grade attained by students retaking Class 2.5 (RGC, Max Grade) were plotted against their ACT composite score. A linear regression line was then cast on the scatter plots and significant correlations were determined by using the Pearson, Kendall, and Spearman tests of correlation. Correlations that were statistically significant at the levels stated for all three tests are reported here. *** = p-value < 0.01, **= p-value < 0.05. r = Pearson rho.

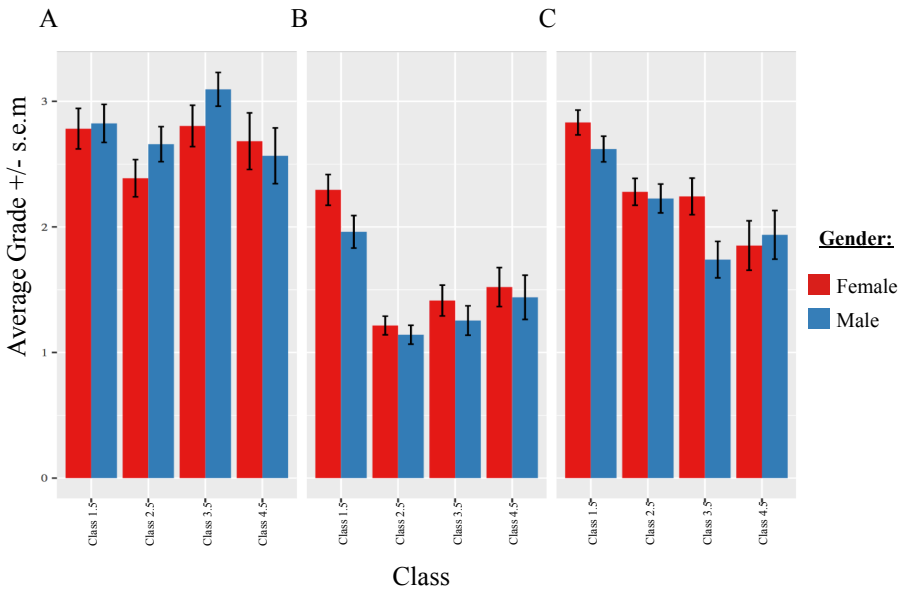


Figure 2-A: Grade by gender for Class 2.5 cohorts. A) Students attempting Class 2.5 once B) Student taking Class 2.5 multiple times C) Highest grade attained by students taking Class 2.5 multiple times. There is no clear trend in grade as it relates to gender between all three cohorts.

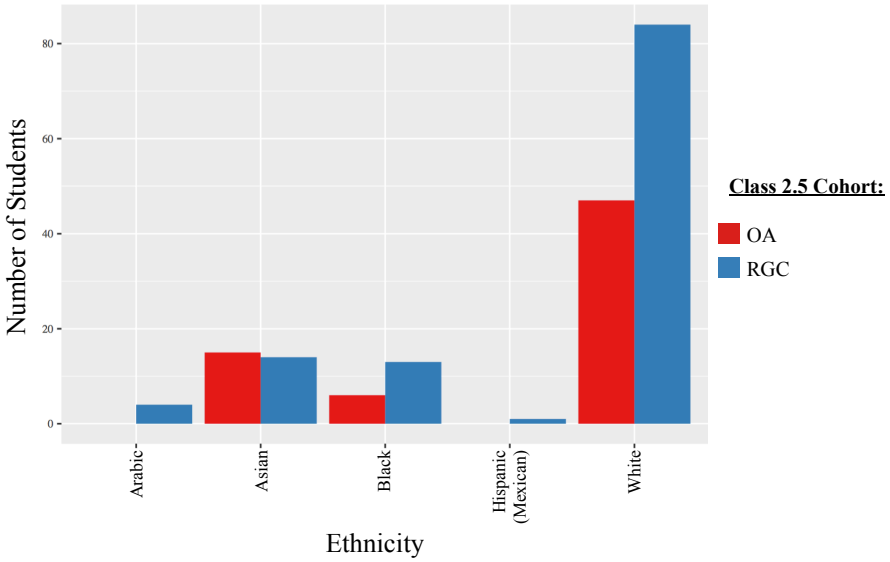


Figure 3-A: Ethnicity summary for Class 2.5 Cohorts. There is no indication that ethnicity is a predictor for success in the gateway course Class 2.5. There are a large proportion of students that contain missing ethnicity information (OA = 104, RG = 118), limiting the ability to draw conclusions from this data.

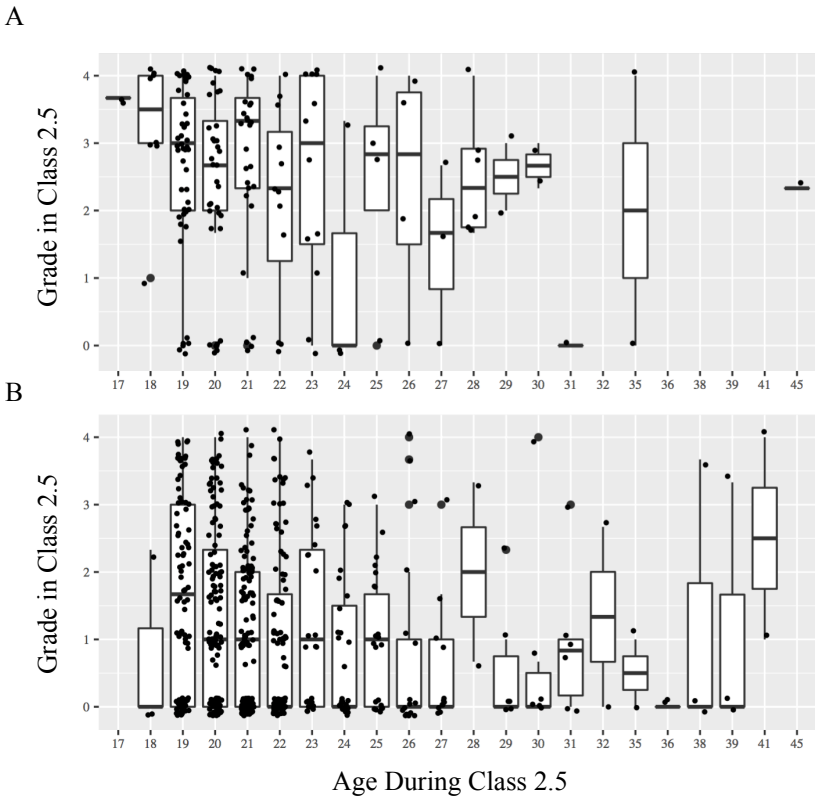


Figure 4-A: Age during Class 2.5 by cohort. A) Students attempting Class 2.5 once B) Student taking Class 2.5 multiple times. There is no clear indication that age is a predictor of success in Class 2.5.

	ID	termCourse	grade
1	9BD20B2B2C73B0F6597C336FC05969C8	201501_ENG3050	A
2	9BD20B2B2C73B0F6597C336FC05969C8	201601_ANT3150	
3	9BD20B2B2C73B0F6597C336FC05969C8	201409_MAT1800	B
4	9BD20B2B2C73B0F6597C336FC05969C8	201501_PS1010	A
5	9BD20B2B2C73B0F6597C336FC05969C8	201509_ENG3060	
6	9BD20B2B2C73B0F6597C336FC05969C8	201501_MAT2010	B
7	9BD20B2B2C73B0F6597C336FC05969C8	201601_CSC2200	
8	7C9D5020B1A90D92A24989DF39433008	201409_COM1010	A
9	7C9D5020B1A90D92A24989DF39433008	201409_HON1000	B+
10	7C9D5020B1A90D92A24989DF39433008	201509_CHM1220	

Figure 5-A: Student data structure. Students are referenced by a unidentifiable code (ID) and each course a student enrolls in is logged along with the semester (termCourse). When the student completes a course the grade is recorded (grade) and if the student withdraws before any grade is assigned this value is left blank

	Chromosome	ChromStart	ChromEnd	Band	Stain
1	chrBIO1050	1	6000	F2007	acen
2	chrBIO1050	6001	12000	W2008	stalk
3	chrBIO1050	12001	18000	SS2008	yellow
4	chrBIO1050	18001	24000	F2008	acen
5	chrBIO1050	24001	30000	W2009	stalk
6	chrBIO1050	30001	36000	SS2009	yellow
7	chrBIO1050	36001	42000	F2009	acen
8	chrBIO1050	42001	48000	W2010	stalk
9	chrBIO1050	48001	54000	SS2010	yellow
10	chrBIO1050	54001	60000	F2010	acen
11	chrBIO1050	60001	66000	W2011	stalk
12	chrBIO1050	66001	72000	SS2011	yellow
13	chrBIO1050	72001	78000	F2011	acen
14	chrBIO1050	78001	84000	W2012	stalk
15	chrBIO1050	84001	90000	SS2012	yellow
16	chrBIO1050	90001	96000	F2012	acen
17	chrBIO1050	96001	102000	W2013	stalk
18	chrBIO1050	102001	108000	SS2013	yellow
19	chrBIO1050	108001	114000	F2013	acen
20	chrBIO1050	114001	120000	W2014	stalk
21	chrBIO1050	120001	126000	SS2014	yellow
22	chrBIO1050	126001	132000	F2014	acen
23	chrBIO1050	132001	138000	W2015	stalk
24	chrBIO1050	138001	144000	SS2015	yellow
25	chrBIO1050	144001	150000	F2015	acen

Figure 6-A: Data structure of karyotype. Each class (Chromosome) contains a start and stop position for every semester (Band) selected. The color of each semester box is designated by the Stain column

Chromosome	chromStart	chromEnd	Chromosome.1	chromStart.1	chromEnd.1	PlotColor
PHY2170	110501	110551	PHY2170	126001	126051	orange3
PHY2170	116501	116551	PHY2170	126834	126884	orange3
PHY2170	118001	118051	PHY2170	127667	127717	orange3
PHY2170	119501	119551	PHY2170	128501	128551	gray30
PHY2170	128501	128551	PHY2180	132001	132051	green4
PHY2170	128631	128681	PHY2180	132313	132363	green4
PHY2170	128761	128811	PHY2180	144001	144051	green4
PHY2170	128892	128942	PHY2180	138001	138051	green4
PHY2170	129022	129072	Current	48001	48051	green4
PHY2170	129153	129203	PHY2180	132626	132676	green4
PHY2170	129283	129333	PHY2180	132938	132988	green4
PHY2170	129414	129464	Current	48167	48217	green4
PHY2170	129544	129594	PHY2180	140501	140551	green4
PHY2170	129674	129724	PHY2180	133251	133301	blue
PHY2170	129805	129855	PHY2180	145251	145301	blue
PHY2170	129935	129985	PHY2180	133563	133613	blue
PHY2170	130066	130116	Current	46001	46051	blue
PHY2170	130196	130246	PHY2180	133876	133926	blue
PHY2170	130327	130377	Current	48334	48384	blue
PHY2170	130457	130507	PHY2180	146501	146551	blue
PHY2170	130587	130637	PHY2180	134188	134238	blue
PHY2170	130718	130768	Current	48501	48551	yellow3
PHY2170	130848	130898	Current	48667	48717	yellow3
PHY2170	130979	131029	Current	48834	48884	yellow3
PHY2170	131109	131159	PHY2180	134501	134551	yellow3
PHY2170	131240	131290	Current	49001	49051	yellow3
PHY2170	131370	131420	Current	49167	49217	yellow3
PHY2170	131501	131551	PHY2170	144001	144051	gray30
PHY2180	134501	134551	Current	49334	49384	blue
PHY2180	136001	136051	Current	44001	44051	yellow3
PHY2180	137501	137551	Current	49501	49551	orange3

Figure 7-A: Data structure of student data for plotting. For every student grade entry the class is specified (Chromosome), the position within the class is determine by chromStart and chromEnd, and the second class (Chromosome.1) position by chromStart.1 and chromEnd.1. Student grade in class coded by PlotColor.

	ID	degreeAward	ethnicity	gender	admitTerm	courseOrigin	major
1	8C996F9D761181BE6A58D395BE7C4336	0	UN	F	201309	NA	GEN
2	1AB39C6F28F9003EBDAD3AF8FAB318DE	0	UN	F	201309	NA	PNUR
3	1B68E6AE0441E168D40644D8904279AC	0	UN	M	201509	NA	BCCB
4	26A5C270CA3045EB506A6A099DD4075D	0	UN	F	201409	NA	ELHS
5	D48FBF2BB2E6EFD09B84F0D6561B9F27	0	UN	F	201509	NA	P S

actComposite	actEnglish	actMath	actReading	hsGpa	starsCohort
15	15	15	15	3.15	RH-TT,LC-DCE,LC-ALL-AY,GRS,LC-EME0993,FTC,_LC-ALL
21	19	17	25	3.21	LC-WSUCUS,LC-ALL-AY,GRS,FTC,_LC-ALL
33	31	35	32	4.00	HON,GRS
27	32	26	30	4.00	GRS_BAK,WSCH_GOLD,HON,GRS,FTC
24	25	25	23	3.52	GRS

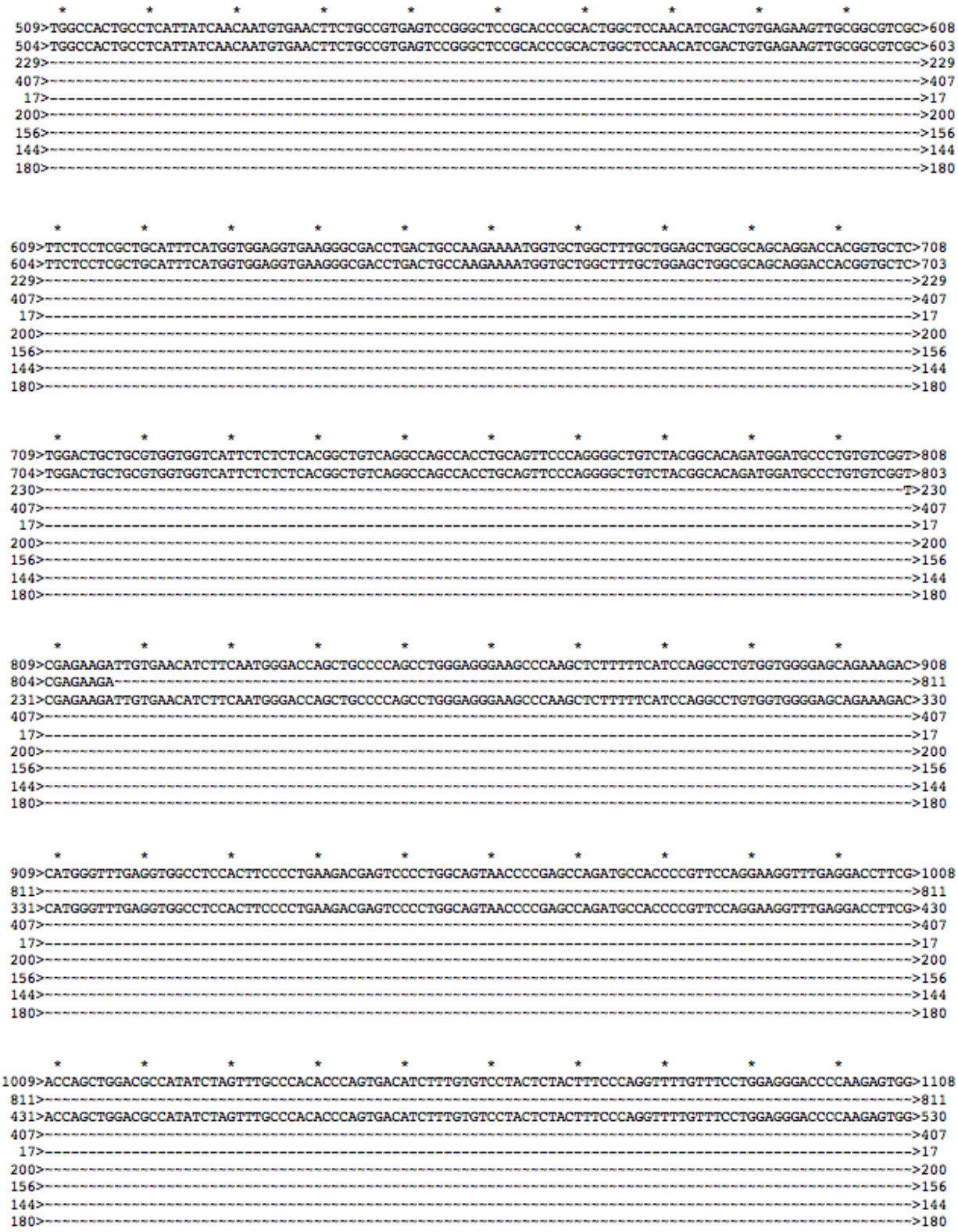
Figure 8-A: Student demographic data structure.

Appendix B: Chimera sequence alignments

Full sequence alignment is included for CAE only. Each chimera is built on the same TcdA backbone, so sequence alignments are only included for the variable reagions of each chimeric construct.

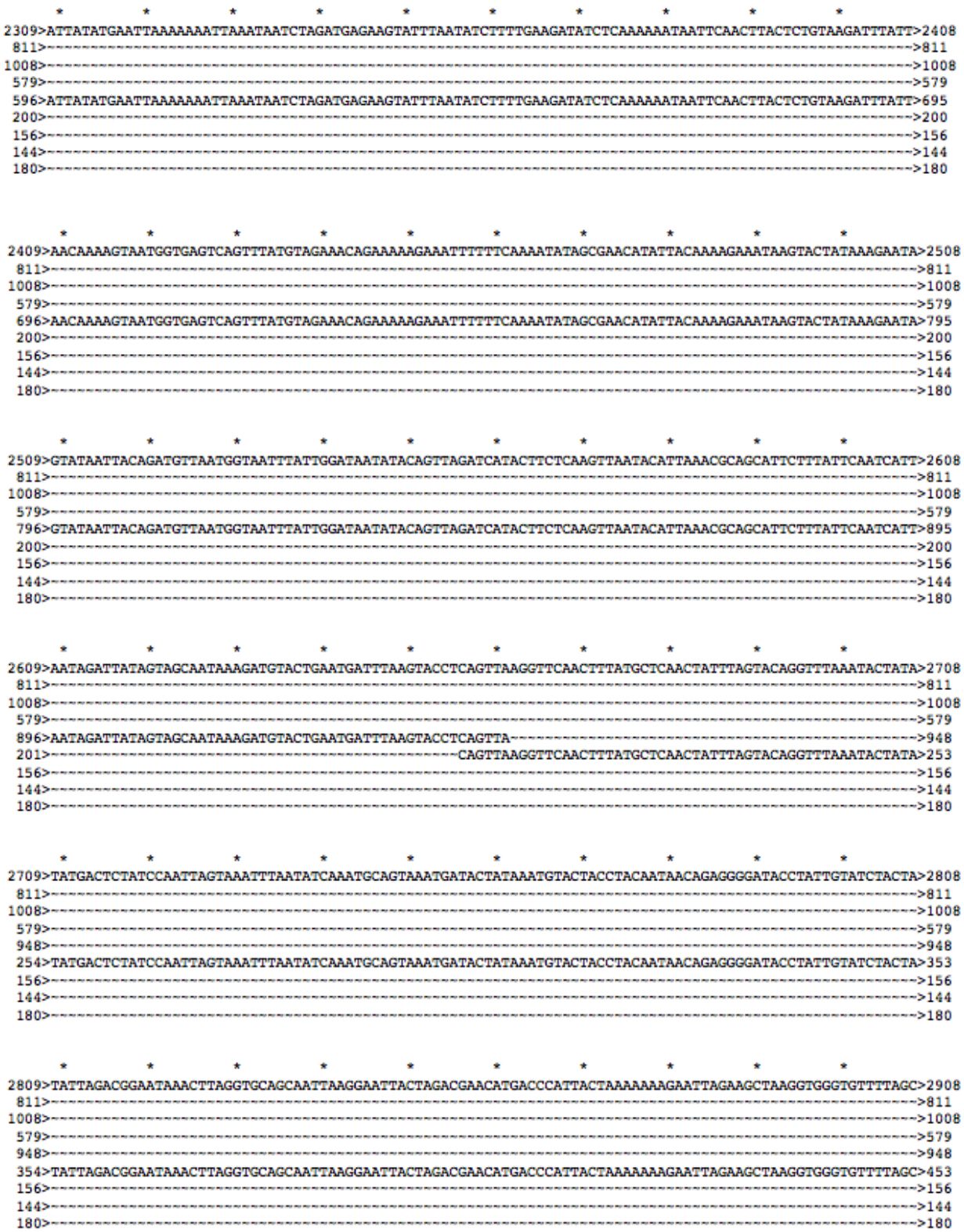
Figure 1-B: CAE alignment

```
Fri Oct 17, 2014 12:43 EDT
pAB_80404(Casp9_TcdA_EGF).ape from 9 to 5522
Alignment to
CAE__W5.seq--  Matches:806; Mismatches:1; Gaps:4707; Unattempted:0
CAE__CF.seq--  Matches:779; Mismatches:0; Gaps:4735; Unattempted:0
3502_A01_CAE__CF_008.seq--  Matches:172; Mismatches:0; Gaps:5342; Unattempted:0
CAE__T2326F.seq--  Matches:932; Mismatches:0; Gaps:4582; Unattempted:0
CAE__T3065F.seq--  Matches:628; Mismatches:0; Gaps:4886; Unattempted:0
CAE__T3727F.seq--  Matches:767; Mismatches:0; Gaps:4747; Unattempted:0
CAE__T4502F.seq--  Matches:808; Mismatches:0; Gaps:4706; Unattempted:0
CAE_T5268F.seq--  Matches:687; Mismatches:0; Gaps:4827; Unattempted:0

        *         *         *         *         *         *         *         *         *         *
  9>atggtccaaaCTAGgAGATCTATGGACGAAGCGGATCGGCGGCTCCTGCGGCGGTGCCGGCTGCGGCTGGTGGAAGAGCTGCAGGTGGACCAGCTCTGGG>108
  5>NTGGTCC-AACTAGGAGATCTATGGACGAAGCGGATCGGCGGCTCCTGCGGCGGTGCCGGCTGCGGCTGGTGGAAGAGCTGCAGGTGGACCAGCTCTGGG>103
229>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>229
407>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>407
 17>A~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>17
200>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>200
156>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>156
144>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>144
180>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>180


    *         *         *         *         *         *         *         *         *         *
109>ACGCCCTGCTGAGCCGCGAGCTGTTCAGGCCCCATATGATCGAGGACATCCAGCGGGCAGGCTCTGGATCTCGGCGGGATCAGGCCAGGCAGCTGATCAT>208
104>ACGCCCTGCTGAGCCGCGAGCTGTTCAGGCCCCATATGATCGAGGACATCCAGCGGGCAGGCTCTGGATCTCGGCGGGATCAGGCCAGGCAGCTGATCAT>203
229>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>229
407>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>407
 17>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>17
200>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>200
156>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>156
144>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>144
180>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>180


    *         *         *         *         *         *         *         *         *         *
209>AGATCTGGAGACTCGAGGGAGTCAGGCTCTTCCTTTGTTCATCTCCTGCTTAGAGGACACAGGCCAGGACATGCTGGCTTCGTTTCTGCGAACTAACAGG>308
204>AGATCTGGAGACTCGAGGGAGTCAGGCTCTTCCTTTGTTCATCTCCTGCTTAGAGGACACAGGCCAGGACATGCTGGCTTCGTTTCTGCGAACTAACAGG>303
229>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>229
407>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>407
 17>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>17
200>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>200
156>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>156
144>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>144
180>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>180


    *         *         *         *         *         *         *         *         *         *
309>CAAGCAGCAAAGTTGTCGAAGCCAACCCTAGAAAACCTTACCCCAGTGGTGCTCAGACCAGAGATTCGCAAACCAGAGGTTCTCAGACCGGAAACACCCA>408
304>CAAGCAGCAAAGTTGTCGAAGCCAACCCTAGAAAACCTTACCCCAGTGGTGCTCAGACCAGAGATTCGCAAACCAGAGGTTCTCAGACCGGAAACACCCA>403
229>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>229
407>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>407
 17>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>17
200>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>200
156>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>156
144>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>144
180>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>180


    *         *         *         *         *         *         *         *         *         *
409>GACCAGTGGACATTGGTTCTGGCGGATTTGGTGATGTCGGTGCTCTTGAGAGTTTGAGGGGAAATGCAGATTTGGCTTACATCCTGAGCATGGAGCCCTG>508
404>GACCAGTGGACATTGGTTCTGGCGGATTTGGTGATGTCGGTGCTCTTGAGAGTTTGAGGGGAAATGCAGATTTGGCTTACATCCTGAGCATGGAGCCCTG>503
229>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>229
407>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>407
 17>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>17
200>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>200
156>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>156
144>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>144
180>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>180
```


509>TGGCCACTGCCTCATTATCAACAATGTGAACTTCTGCCGTGAGTCCGGGCTCCGCACCCGCACTGGCTCCAACATCGACTGTGAGAAGTTGCGGCGTCGC>608
504>TGGCCACTGCCTCATTATCAACAATGTGAACTTCTGCCGTGAGTCCGGGCTCCGCACCCGCACTGGCTCCAACATCGACTGTGAGAAGTTGCGGCGTCGC>603
229>>229
407>>407
17>>17
200>>200
156>>156
144>>144
180>>180
609>TTCTCCTCGCTGCATTTCATGGTGGAGGTGAAGGGCGACCTGACTGCCAAGAAAATGGTGCTGGCTTTGCTGGAGCTGGCGCAGCAGGACCACGGTGCTC>708
604>TTCTCCTCGCTGCATTTCATGGTGGAGGTGAAGGGCGACCTGACTGCCAAGAAAATGGTGCTGGCTTTGCTGGAGCTGGCGCAGCAGGACCACGGTGCTC>703
229>>229
407>>407
17>>17
200>>200
156>>156
144>>144
180>>180
709>TGGACTGCTGCGTGGTGGTCATTCTCTCTCACGGCTGTCAGGCCAGCCACCTGCAGTTCCCAGGGGCTGTCTACGGCACAGATGGATGCCCTGTGTCGGT>808
704>TGGACTGCTGCGTGGTGGTCATTCTCTCTCACGGCTGTCAGGCCAGCCACCTGCAGTTCCCAGGGGCTGTCTACGGCACAGATGGATGCCCTGTGTCGGT>803
230>T>230
407>>407
17>>17
200>>200
156>>156
144>>144
180>>180
809>CGAGAAGATTGTGAACATCTTCAATGGGACCAGCTGCCCCAGCCTGGGAGGGAAGCCCAAGCTCTTTTTCATCCAGGCCTGTGGTGGGGAGCAGAAAGAC>908
804>CGAGAAGA>811
231>CGAGAAGATTGTGAACATCTTCAATGGGACCAGCTGCCCCAGCCTGGGAGGGAAGCCCAAGCTCTTTTTCATCCAGGCCTGTGGTGGGGAGCAGAAAGAC>330
407>>407
17>>17
200>>200
156>>156
144>>144
180>>180
909>CATGGGTTTGAGGTGGCCTCCACTTCCCCTGAAGACGAGTCCCCTGGCAGTAACCCCGAGCCAGATGCCACCCCGTTCCAGGAAGGTTTGAGGACCTTCG>1008
811>>811
331>CATGGGTTTGAGGTGGCCTCCACTTCCCCTGAAGACGAGTCCCCTGGCAGTAACCCCGAGCCAGATGCCACCCCGTTCCAGGAAGGTTTGAGGACCTTCG>430
407>>407
17>>17
200>>200
156>>156
144>>144
180>>180
1009>ACCAGCTGGACGCCATATCTAGTTTGCCCACACCCAGTGACATCTTTGTGTCCTACTCTACTTTCCCAGGTTTTGTTTCCTGGAGGGACCCCAAGAGTGG>1108
811>>811
431>ACCAGCTGGACGCCATATCTAGTTTGCCCACACCCAGTGACATCTTTGTGTCCTACTCTACTTTCCCAGGTTTTGTTTCCTGGAGGGACCCCAAGAGTGG>530
407>>407
17>>17
200>>200
156>>156
144>>144
180>>180

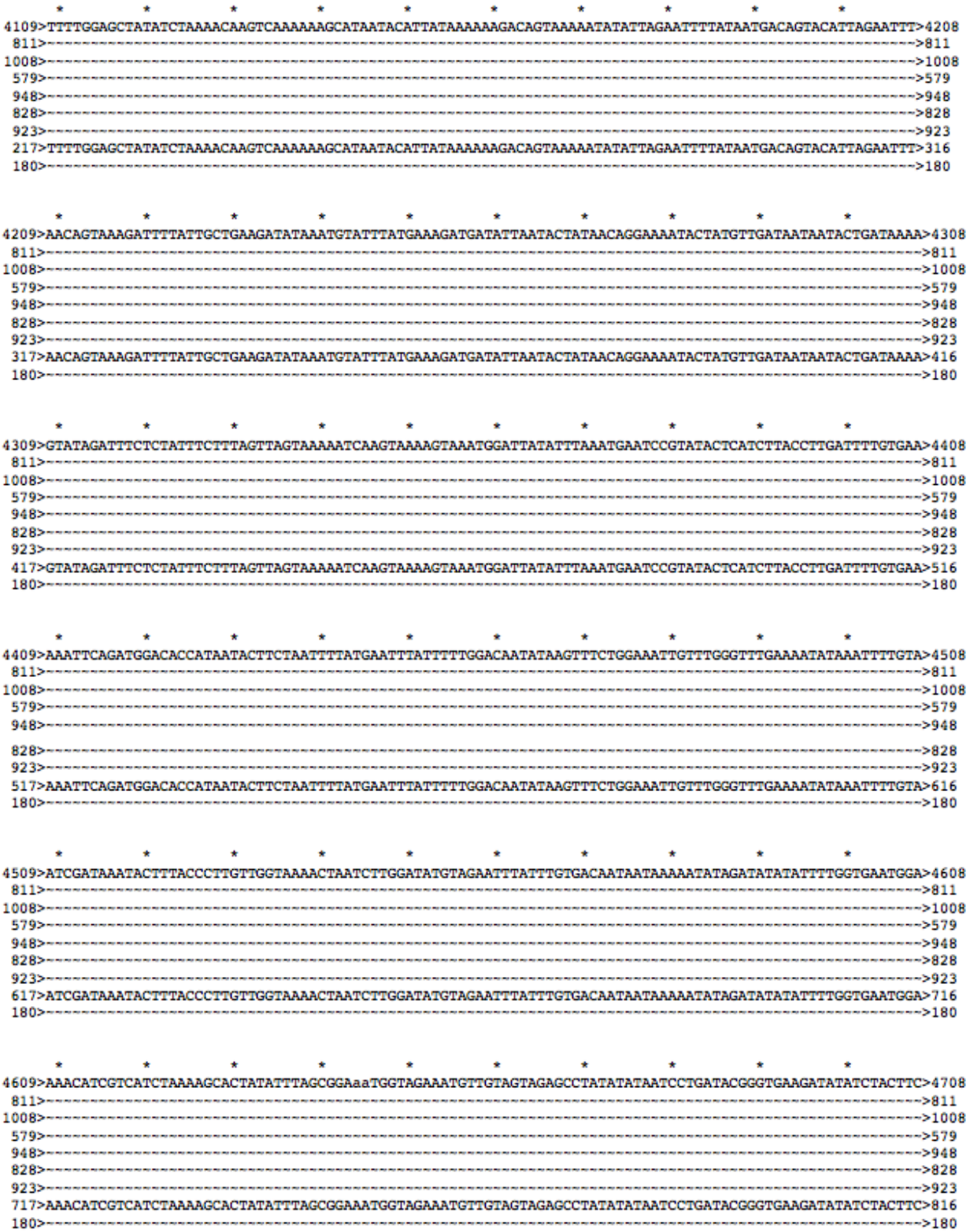
```
          *         *         *         *         *         *         *         *         *         *
1109>CTCCTGGTACGTTGAGACCCTGGACGACATCTTTGAGCAGTGGGCTCACTCTGAAGACCTGCAGTCCCTCCTGCTTAGGGTCGCTAATGCTGTTTCGGTG>1208
 811>---------------------------------------------------------------------------------------------------->811
 531>CTCCTGGTACGTTGAGACCCTGGACGACATCTTTGAGCAGTGGGCTCACTCTGAAGACCTGCAGTCCCTCCTGCTTAGGGTCGCTAATGCTGTTTCGGTG>630
 407>---------------------------------------------------------------------------------------------------->407
  17>---------------------------------------------------------------------------------------------------->17
 200>---------------------------------------------------------------------------------------------------->200
 156>---------------------------------------------------------------------------------------------------->156
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180

          *         *         *         *         *         *         *         *         *         *
1209>AAAGGGATTTATAAACAGATGCCTGGTTGCTTTAATTTCCTCCGGAAAAAACTTTTCTTTAAAACATCAGTGTTGGATTCTCTTGAATTTATTGCTAGTA>1308
 811>---------------------------------------------------------------------------------------------------->811
 631>AAAGGGATTTATAAACAGATGCCTGGTTGCTTTAATTTCCTCCGGAAAAAACTTTTCTTTAAAACATCAGTGTTGGATTCTCTTGAATTTATTGCTAGTA>730
 407>---------------------------------------------------------------------------------------------------->407
  17>---------------------------------------------------------------------------------------------------->17
 200>---------------------------------------------------------------------------------------------------->200
 156>---------------------------------------------------------------------------------------------------->156
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180

          *         *         *         *         *         *         *         *         *         *
1309>AACTGGCGgGATCcCTTTCTGAAGACAATGGGGTAGACTTTAATAAAAATACTGCCCTCGACAAAAACTATTTATTAAATAATAAAATTCCATCAAACAA>1408
 811>---------------------------------------------------------------------------------------------------->811
 731>AACTGGCGGGATCCCTTTCTGAAGACAATGGGGTAGACTTTAATAAAAATACTGCCCTCGACAAAAACTATTTATTAAATAATAAAATTCCATCAAACAA>830
 407>---------------------------------------------------------------------------------------------------->407
  17>---------------------------------------------------------------------------------------------------->17
 200>---------------------------------------------------------------------------------------------------->200
 156>---------------------------------------------------------------------------------------------------->156
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180

          *         *         *         *         *         *         *         *         *         *
1409>TGTAGAAGAAGCTGGAAGTAAAAATTATGTTCATTATATCATACAGTTACAAGGAGATGATATAAGTTATGAAGCAACATGCAATTTATTTTCTAAAAAT>1508
 811>---------------------------------------------------------------------------------------------------->811
 831>TGTAGAAGAAGCTGGAAGTAAAAATTATGTTCATTATATCATACAGTTACAAGGAGATGATATAAGTTATGAAGCAACATGCAATTTATTTTCTAAAAAT>930
 407>---------------------------------------------------------------------------------------------------->407
  17>---------------------------------------------------------------------------------------------------->17
 200>---------------------------------------------------------------------------------------------------->200
 156>---------------------------------------------------------------------------------------------------->156
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180

          *         *         *         *         *         *         *         *         *         *
1509>CCTAAAAATAGTATTATTATACAACGAAATATGAATGAAAGTGCAAAAAGCTACTTTTTAAGTGATGATGGAGAATCTATTTTAGAATTAAATAAATATA>1608
 811>---------------------------------------------------------------------------------------------------->811
 931>CCTAAAAATAGTATTATTATACAACGAAATATGAATGAAAGTGCAAAAAGCTACTTTTTAAGTGATGATGGAGAATCT---------------------->1008
 408>----------------------------------------------------------TAAGTGATGATGGAGAATCTATTTTAGAATTAAATAAATATA>449
  17>---------------------------------------------------------------------------------------------------->17
 200>---------------------------------------------------------------------------------------------------->200
 156>---------------------------------------------------------------------------------------------------->156
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180

          *         *         *         *         *         *         *         *         *         *
1609>GGATACCTGAAAGATTAAAAAATAAGGAAAAAGTAAAAGTAACCTTTATTGGACATGGTAAAGATGAATTCAACACAAGCGAATTTGCTAGATTAAGTGT>1708
 811>---------------------------------------------------------------------------------------------------->811
1008>---------------------------------------------------------------------------------------------------->1008
 450>GGATACCTGAAAGATTAAAAAATAAGGAAAAAGTAAAAGTAACCTTTATTGGACATGGTAAAGATGAATTCAACACAAGCGAATTTGCTAGATTAAGTGT>549
  17>---------------------------------------------------------------------------------------------------->17
 200>---------------------------------------------------------------------------------------------------->200
 156>---------------------------------------------------------------------------------------------------->156
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180
```

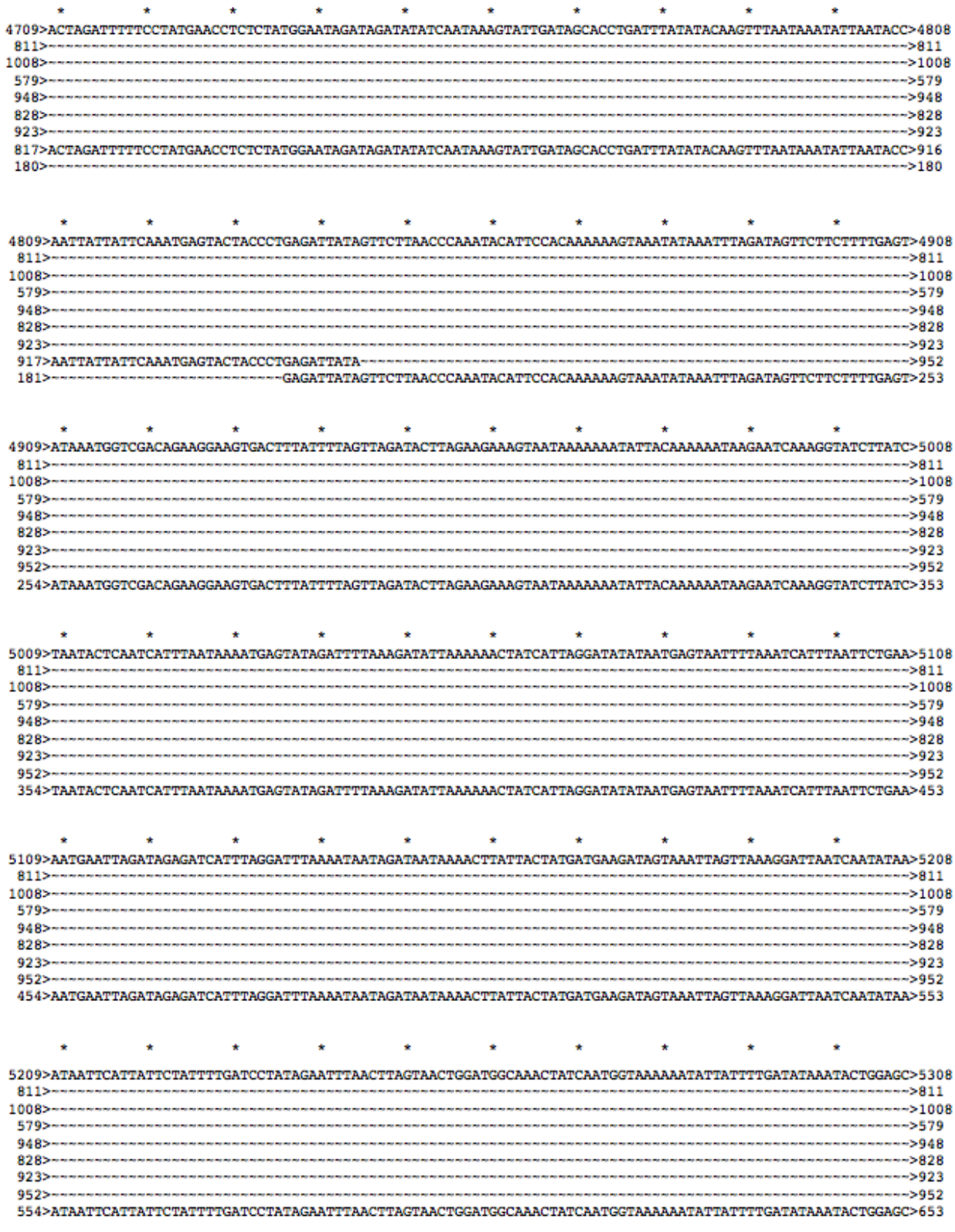
```
     *         *         *         *         *         *         *         *         *         *
1709>AGATTCACTTTCCAATGAGATAAGTTCATTTTTAGATACCATAAAATTAGATATATCACCTAAAAATGTAGAAGTAAACTTACTTGGATGTAATATGTTT>1808
 811>-------------------------------------------------------------------------------------------------->811
1008>-------------------------------------------------------------------------------------------------->1008
 550>AGATTCACTTTCCAATGAGATAAGTTCATT-------------------------------------------------------------------->579
  18>----------------------AGTTCATTTTTAGATACCATAAAATTAGATATATCACCTAAAAATGTAGAAGTAAACTTACTTGGATGTAATATGTTT>95
 200>-------------------------------------------------------------------------------------------------->200
 156>-------------------------------------------------------------------------------------------------->156
 144>-------------------------------------------------------------------------------------------------->144
 180>-------------------------------------------------------------------------------------------------->180


     *         *         *         *         *         *         *         *         *         *
1809>AGTTATGATTTTAATGTTGAAGAAACTTATCCTGGGAAGTTGCTATTAAGTATTATGGACAAAATTACTTCCACTTTACCTGATGTAAATAAAAATTCTA>1908
 811>-------------------------------------------------------------------------------------------------->811
1008>-------------------------------------------------------------------------------------------------->1008
 579>-------------------------------------------------------------------------------------------------->579
  96>AGTTATGATTTTAATGTTGAAGAAACTTATCCTGGGAAGTTGCTATTAAGTATTATGGACAAAATTACTTCCACTTTACCTGATGTAAATAAAAATTCTA>195
 200>-------------------------------------------------------------------------------------------------->200
 156>-------------------------------------------------------------------------------------------------->156
 144>-------------------------------------------------------------------------------------------------->144
 180>-------------------------------------------------------------------------------------------------->180


     *         *         *         *         *         *         *         *         *         *
1909>TTACTATAGGAGCAAATCAATATGAAGTAAGAATTAATAGTGAGGGAAGAAAAGAACTTCTGGCTCACTCAGGTAAATGGATAAATAAAGAAGAAGCTAT>2008
 811>-------------------------------------------------------------------------------------------------->811
1008>-------------------------------------------------------------------------------------------------->1008
 579>-------------------------------------------------------------------------------------------------->579
 196>TTACTATAGGAGCAAATCAATATGAAGTAAGAATTAATAGTGAGGGAAGAAAAGAACTTCTGGCTCACTCAGGTAAATGGATAAATAAAGAAGAAGCTAT>295
 200>-------------------------------------------------------------------------------------------------->200
 156>-------------------------------------------------------------------------------------------------->156
 144>-------------------------------------------------------------------------------------------------->144
 180>-------------------------------------------------------------------------------------------------->180


     *         *         *         *         *         *         *         *         *         *
2009>TATGAGCGATTTATCTAGTAAAGAATACATTTTTTTTGATTCTATAGATAATAAGCTAAAAGCAAAGTCCAAGAATATTCCAGGATTAGCATCAATATCA>2108
 811>-------------------------------------------------------------------------------------------------->811
1008>-------------------------------------------------------------------------------------------------->1008
 579>-------------------------------------------------------------------------------------------------->579
 296>TATGAGCGATTTATCTAGTAAAGAATACATTTTTTTTGATTCTATAGATAATAAGCTAAAAGCAAAGTCCAAGAATATTCCAGGATTAGCATCAATATCA>395
 200>-------------------------------------------------------------------------------------------------->200
 156>-------------------------------------------------------------------------------------------------->156
 144>-------------------------------------------------------------------------------------------------->144
 180>-------------------------------------------------------------------------------------------------->180


     *         *         *         *         *         *         *         *         *         *
2109>GAAGATATAAAAACATTATTACTTGAcGCAAGTGTTAGTCCTGATACAAAATTTATTTTAAATAATCTTAAGCTTAATATTGAATCTTCTATTGGTGATT>2208
 811>-------------------------------------------------------------------------------------------------->811
1008>-------------------------------------------------------------------------------------------------->1008
 579>-------------------------------------------------------------------------------------------------->579
 396>GAAGATATAAAAACATTATTACTTGACGCAAGTGTTAGTCCTGATACAAAATTTATTTTAAATAATCTTAAGCTTAATATTGAATCTTCTATTGGTGATT>495
 200>-------------------------------------------------------------------------------------------------->200
 156>-------------------------------------------------------------------------------------------------->156
 144>-------------------------------------------------------------------------------------------------->144
 180>-------------------------------------------------------------------------------------------------->180


     *         *         *         *         *         *         *         *         *         *
2209>ACATTTATTATGAAAAATTAGAGCCTGTTAAAAATATAATTCACAATTCTATAGATGATTTAATAGATGAGTTCAATCTACTTGAAAATGTATCTGATGA>2308
 811>-------------------------------------------------------------------------------------------------->811
1008>-------------------------------------------------------------------------------------------------->1008
 579>-------------------------------------------------------------------------------------------------->579
 496>ACATTTATTATGAAAAATTAGAGCCTGTTAAAAATATAATTCACAATTCTATAGATGATTTAATAGATGAGTTCAATCTACTTGAAAATGTATCTGATGA>595
 200>-------------------------------------------------------------------------------------------------->200
 156>-------------------------------------------------------------------------------------------------->156
 144>-------------------------------------------------------------------------------------------------->144
 180>-------------------------------------------------------------------------------------------------->180
```


2309>ATTATATGAATTAAAAAAATTAAATAATCTAGATGAGAAGTATTTAATATCTTTTGAAGATATCTCAAAAAATAATTCAACTTACTCTGTAAGATTTATT>2408
811>811
1008>1008
579>579
596>ATTATATGAATTAAAAAAATTAAATAATCTAGATGAGAAGTATTTAATATCTTTTGAAGATATCTCAAAAAATAATTCAACTTACTCTGTAAGATTTATT>695
200>200
156>156
144>144
180>180
2409>AACAAAAGTAATGGTGAGTCAGTTTATGTAGAAACAGAAAAAGAAATTTTTTCAAAATATAGCGAACATATTACAAAAGAAATAAGTACTATAAAGAATA>2508
811>811
1008>1008
579>579
696>AACAAAAGTAATGGTGAGTCAGTTTATGTAGAAACAGAAAAAGAAATTTTTTCAAAATATAGCGAACATATTACAAAAGAAATAAGTACTATAAAGAATA>795
200>200
156>156
144>144
180>180
2509>GTATAATTACAGATGTTAATGGTAATTTATTGGATAATATACAGTTAGATCATACTTCTCAAGTTAATACATTAAACGCAGCATTCTTTATTCAATCATT>2608
811>811
1008>1008
579>579
796>GTATAATTACAGATGTTAATGGTAATTTATTGGATAATATACAGTTAGATCATACTTCTCAAGTTAATACATTAAACGCAGCATTCTTTATTCAATCATT>895
200>200
156>156
144>144
180>180
2609>AATAGATTATAGTAGCAATAAAGATGTACTGAATGATTTAAGTACCTCAGTTAAGGTTCAACTTTATGCTCAACTATTTAGTACAGGTTTAAATACTATA>2708
811>811
1008>1008
579>579
896>AATAGATTATAGTAGCAATAAAGATGTACTGAATGATTTAAGTACCTCAGTTA>948
201>CAGTTAAGGTTCAACTTTATGCTCAACTATTTAGTACAGGTTTAAATACTATA>253
156>156
144>144
180>180
2709>TATGACTCTATCCAATTAGTAAATTTAATATCAAATGCAGTAAATGATACTATAAATGTACTACCTACAATAACAGAGGGGATACCTATTGTATCTACTA>2808
811>811
1008>1008
579>579
948>948
254>TATGACTCTATCCAATTAGTAAATTTAATATCAAATGCAGTAAATGATACTATAAATGTACTACCTACAATAACAGAGGGGATACCTATTGTATCTACTA>353
156>156
144>144
180>180
2809>TATTAGACGGAATAAACTTAGGTGCAGCAATTAAGGAATTACTAGACGAACATGACCCATTACTAAAAAAAGAATTAGAAGCTAAGGTGGGTGTTTTAGC>2908
811>811
1008>1008
579>579
948>948
354>TATTAGACGGAATAAACTTAGGTGCAGCAATTAAGGAATTACTAGACGAACATGACCCATTACTAAAAAAAGAATTAGAAGCTAAGGTGGGTGTTTTAGC>453
156>156
144>144
180>180

```
      *         *         *         *         *         *         *         *         *         *
2909>AATAAATATGTCATTATCTATAGCTGCAACTGTAGCTTCAATTGTTGGAATAGGTGCTGAAGTTACTATTTCTTATTACCTATAGCTGGTATATCTGCA>3008
 811>---------------------------------------------------------------------------------------------------->811
1008>---------------------------------------------------------------------------------------------------->1008
 579>---------------------------------------------------------------------------------------------------->579
 948>---------------------------------------------------------------------------------------------------->948
 454>AATAAATATGTCATTATCTATAGCTGCAACTGTAGCTTCAATTGTTGGAATAGGTGCTGAAGTTACTATTTCTTATTACCTATAGCTGGTATATCTGCA>553
 156>---------------------------------------------------------------------------------------------------->156
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180

      *         *         *         *         *         *         *         *         *         *
3009>GGAATACCTTCATTAGTTAATAATGAATTAATATTGCATGATAAGGCAACTTCAGTGGTAAACTATTTTAATCATTTGTCTGAATCTAAAAAATATGGCC>3108
 811>---------------------------------------------------------------------------------------------------->811
1008>---------------------------------------------------------------------------------------------------->1008
 579>---------------------------------------------------------------------------------------------------->579
 948>---------------------------------------------------------------------------------------------------->948
 554>GGAATACCTTCATTAGTTAATAATGAATTAATATTGCATGATAAGGCAACTTCAGTGGTAAACTATTTTAATCATTTGTCTGAATCTAAAAAATATGGCC>653
 156>---------------------------------------------------------------------------------------------------->156
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180

      *         *         *         *         *         *         *         *         *         *
3109>CTCTTAAGACAGAAGATGATAAAATTTTAGTTCCTATTGATGATTTAGTAATATCAGAAATAGATTTTAATAATAATTCGATAAAACTAGGAACATGTAA>3208
 811>---------------------------------------------------------------------------------------------------->811
1008>---------------------------------------------------------------------------------------------------->1008
 579>---------------------------------------------------------------------------------------------------->579
 948>---------------------------------------------------------------------------------------------------->948
 654>CTCTTAAGACAGAAGATGATAAAATTTTAGTTCCTATTGATGATTTAGTAATATCAGAAATAGATTTTAATAATAATTCGATAAAACTAGGAACATGTAA>753
 156>---------------------------------------------------------------------------------------------------->156
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180

      *         *         *         *         *         *         *         *         *         *
3209>TATATTAGCAATGGAGGGGGGATCAGGACACACAGTGACTGGTAATATAGATCACTTTTTCTCATCTCCATCTATAAGTTCTCATATTCCTTCATTATCA>3308
 811>---------------------------------------------------------------------------------------------------->811
1008>---------------------------------------------------------------------------------------------------->1008
 579>---------------------------------------------------------------------------------------------------->579
 948>---------------------------------------------------------------------------------------------------->948
 754>TATATTAGCAATGGAGGGGGGATCAGGACACACAGTGACTGGTAATATAGATCACTTTTTCTCATCTCCATCTAT------------------------->828
 157>--------------------------------------------------------------------CATCTATAAGTTCTCATATTCCTTCATTATCA>188
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180

      *         *         *         *         *         *         *         *         *         *
3309>ATTTATTCTGCAATAGGTATAGAAACAGAAAATCTAGATTTTTCAAAAAAAATAATGATGTTACCTAATGCTCCTTCAAGAATGTTTTGGTGGGAAACTG>3408
 811>---------------------------------------------------------------------------------------------------->811
1008>---------------------------------------------------------------------------------------------------->1008
 579>---------------------------------------------------------------------------------------------------->579
 948>---------------------------------------------------------------------------------------------------->948
 828>---------------------------------------------------------------------------------------------------->828
 189>ATTTATTCTGCAATAGGTATAGAAACAGAAAATCTAGATTTTTCAAAAAAAATAATGATGTTACCTAATGCTCCTTCAAGAATGTTTTGGTGGGAAACTG>288
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180

      *         *         *         *         *         *         *         *         *         *
3409>GAGCAGTTCCAGGTTTAAGATCATTGGAAAATGACGGAACTAGATTACTTGATTCAATAAGAGATTTATACCCAGGTAAATTTTACTGGAGATTCTATGC>3508
 811>---------------------------------------------------------------------------------------------------->811
1008>---------------------------------------------------------------------------------------------------->1008
 579>---------------------------------------------------------------------------------------------------->579
 948>---------------------------------------------------------------------------------------------------->948
 828>---------------------------------------------------------------------------------------------------->828
 289>GAGCAGTTCCAGGTTTAAGATCATTGGAAAATGACGGAACTAGATTACTTGATTCAATAAGAGATTTATACCCAGGTAAATTTTACTGGAGATTCTATGC>388
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180
```

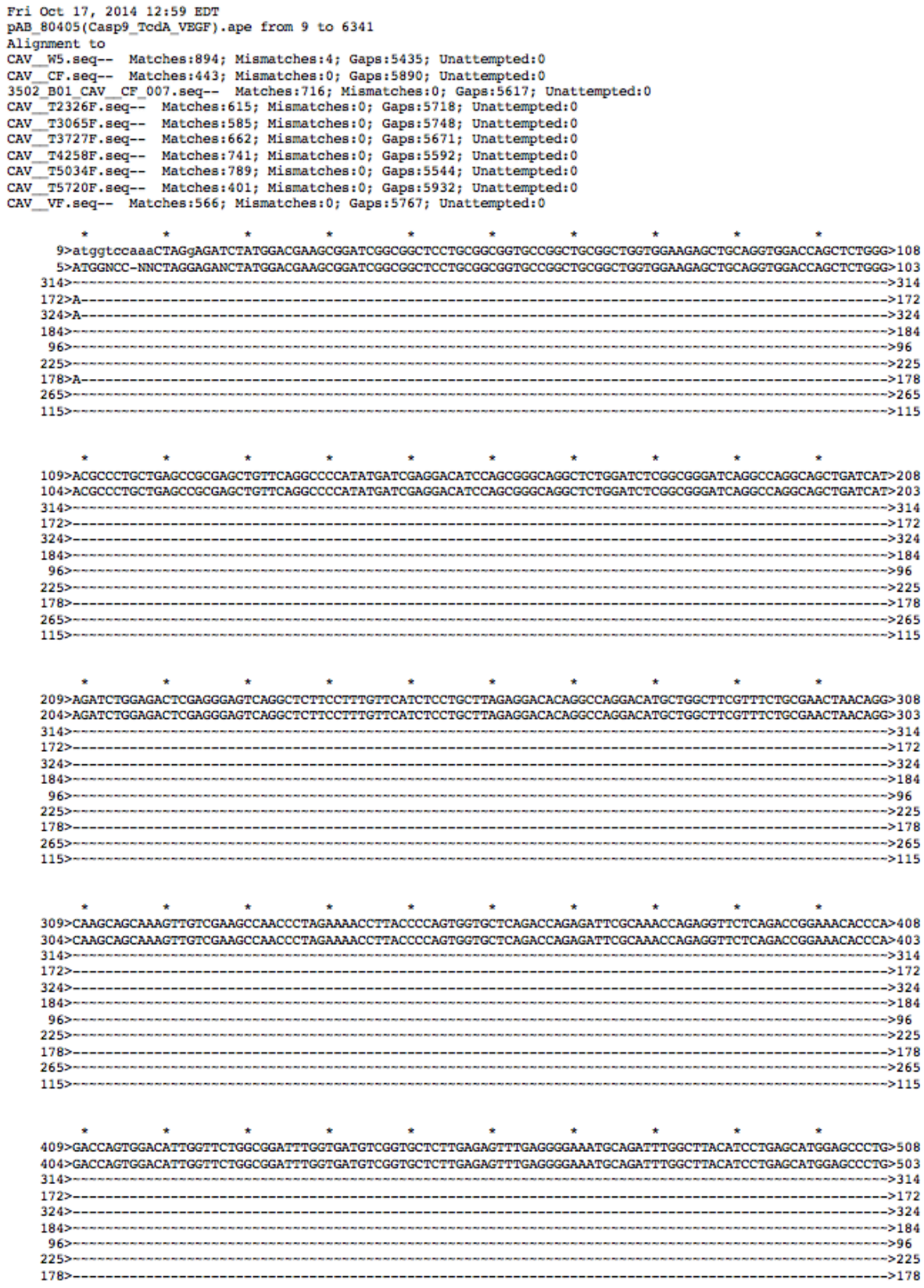
```
      *         *         *         *         *         *         *         *         *         *
3509>TTTTTCGATTATGCAATAACTACATTAAAACCAGTTTATGAAGACACTAATATTAAAATTAAACTAGATAAAGATACTAGAAACTTCATAATGCCAACT>3608
 811>---------------------------------------------------------------------------------------------------->811
1008>---------------------------------------------------------------------------------------------------->1008
 579>---------------------------------------------------------------------------------------------------->579
 948>---------------------------------------------------------------------------------------------------->948
 828>---------------------------------------------------------------------------------------------------->828
 389>TTTTTCGATTATGCAATAACTACATTAAAACCAGTTTATGAAGACACTAATATTAAAATTAAACTAGATAAAGATACTAGAAACTTCATAATGCCAACT>488
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180

      *         *         *         *         *         *         *         *         *         *
3609>ATAACTACTAACGAAATTAGAAACAAATTATCTTATTCATTTGATGGAGCAGGAGGAACTTACTCTTTATTATTATCTTCATATCCAATATCAACGAATA>3708
 811>---------------------------------------------------------------------------------------------------->811
1008>---------------------------------------------------------------------------------------------------->1008
 579>---------------------------------------------------------------------------------------------------->579
 948>---------------------------------------------------------------------------------------------------->948
 828>---------------------------------------------------------------------------------------------------->828
 489>ATAACTACTAACGAAATTAGAAACAAATTATCTTATTCATTTGATGGAGCAGGAGGAACTTACTCTTTATTATTATCTTCATATCCAATATCAACGAATA>588
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180

      *         *         *         *         *         *         *         *         *         *
3709>TAAATTTATCTAAAGATGATTTATGGATATTTAATATTGATAATGAAGTAAGAGAAATATCTATAGAAAATGGTACTATTAAAAAAGGAAAGTTAATAAA>3808
 811>---------------------------------------------------------------------------------------------------->811
1008>---------------------------------------------------------------------------------------------------->1008
 579>---------------------------------------------------------------------------------------------------->579
 948>---------------------------------------------------------------------------------------------------->948
 828>---------------------------------------------------------------------------------------------------->828
 589>TAAATTTATCTAAAGATGATTTATGGATATTTAATATTGATAATGAAGTAAGAGAAATATCTATAGAAAATGGTACTATTAAAAAAGGAAAGTTAATAAA>688
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180

      *         *         *         *         *         *         *         *         *         *
3809>AGATGTTTTAAGTAAAATTGATATAAATAAAAATAAACTTATTATAGGCAATCAAACAATAGATTTTTCAGGCGATATAGATAATAAAGATAGATATATA>3908
 811>---------------------------------------------------------------------------------------------------->811
1008>---------------------------------------------------------------------------------------------------->1008
 579>---------------------------------------------------------------------------------------------------->579
 948>---------------------------------------------------------------------------------------------------->948
 828>---------------------------------------------------------------------------------------------------->828
 689>AGATGTTTTAAGTAAAATTGATATAAATAAAAATAAACTTATTATAGGCAATCAAACAATAGATTTTTCAGGCGATATAGATAATAAAGATAGATATATA>788
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180

      *         *         *         *         *         *         *         *         *         *
3909>TTCTTGACTTGTGAGTTAGATGATAAAATTAGTTTAATAATAGAAATAAATCTTGTTGCAAAATCTTATAGTTTGTTATTGTCTGGGGATAAAAATTATT>4008
 811>---------------------------------------------------------------------------------------------------->811
1008>---------------------------------------------------------------------------------------------------->1008
 579>---------------------------------------------------------------------------------------------------->579
 948>---------------------------------------------------------------------------------------------------->948
 828>---------------------------------------------------------------------------------------------------->828
 789>TTCTTGACTTGTGAGTTAGATGATAAAATTAGTTTAATAATAGAAATAAATCTTGTTGCAAAATCTTATAGTTTGTTATTGTCTGGGGATAAAAATTATT>888
 144>---------------------------------------------------------------------------------------------------->144
 180>---------------------------------------------------------------------------------------------------->180

      *         *         *         *         *         *         *         *         *         *
4009>TGATATCCAATTTATCTAATACTATTGAGAAAATCAATACTTTAGGCCTAGATAGTAAAAATATAGCGTACAATTACACTGATGAATCTAATAATAAATA>4108
 811>---------------------------------------------------------------------------------------------------->811
1008>---------------------------------------------------------------------------------------------------->1008
 579>---------------------------------------------------------------------------------------------------->579
 948>---------------------------------------------------------------------------------------------------->948
 828>---------------------------------------------------------------------------------------------------->828
 889>TGATATCCAATTTATCTAATACTATTGAGAAAATC----------------------------------------------------------------->923
 145>----------------------------GAAAATCAATACTTTAGGCCTAGATAGTAAAAATATAGCGTACAATTACACTGATGAATCTAATAATAAATA>216
 180>---------------------------------------------------------------------------------------------------->180
```

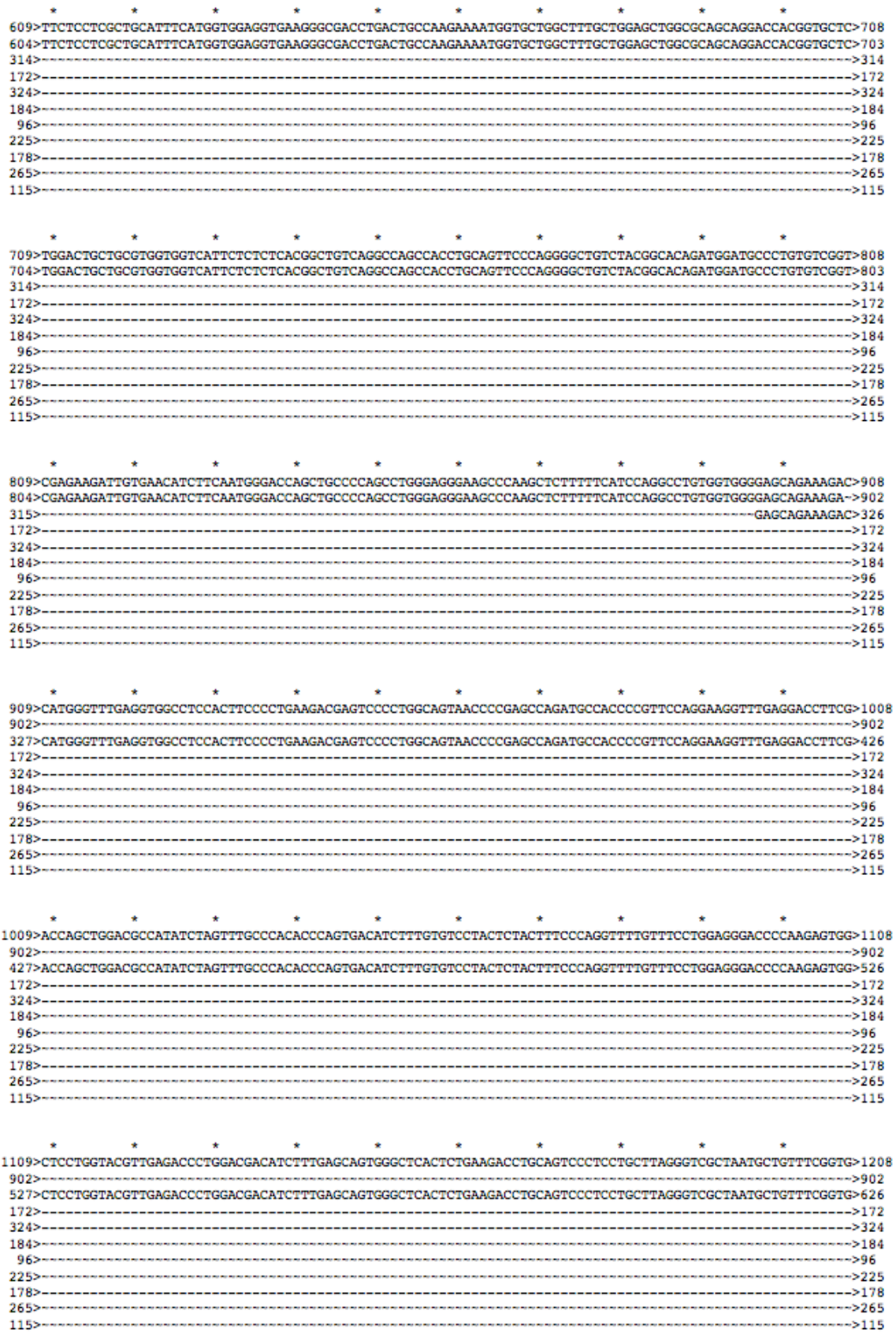

4109>TTTTGGAGCTATATCTAAAACAAGTCAAAAAAGCATAATACATTATAAAAAAGACAGTAAAAATATATTAGAATTTTATAATGACAGTACATTAGAATTT>4208
811>>811
1008>>1008
579>>579
948>>948
828>>828
923>>923
217>TTTTGGAGCTATATCTAAAACAAGTCAAAAAAGCATAATACATTATAAAAAAGACAGTAAAAATATATTAGAATTTTATAATGACAGTACATTAGAATTT>316
180>>180
4209>AACAGTAAAGATTTTATTGCTGAAGATATAAATGTATTTATGAAAGATGATATTAATACTATAACAGGAAAATACTATGTTGATAATAATACTGATAAAA>4308
811>>811
1008>>1008
579>>579
948>>948
828>>828
923>>923
317>AACAGTAAAGATTTTATTGCTGAAGATATAAATGTATTTATGAAAGATGATATTAATACTATAACAGGAAAATACTATGTTGATAATAATACTGATAAAA>416
180>>180
4309>GTATAGATTTCTCTATTTCTTTAGTTAGTAAAAATCAAGTAAAAGTAAATGGATTATATTTAAATGAATCCGTATACTCATCTTACCTTGATTTTGTGAA>4408
811>>811
1008>>1008
579>>579
948>>948
828>>828
923>>923
417>GTATAGATTTCTCTATTTCTTTAGTTAGTAAAAATCAAGTAAAAGTAAATGGATTATATTTAAATGAATCCGTATACTCATCTTACCTTGATTTTGTGAA>516
180>>180
4409>AAATTCAGATGGACACCATAATACTTCTAATTTTATGAATTTATTTTTGGACAATATAAGTTTCTGGAAATTGTTTGGGTTTGAAAATATAAATTTTGTA>4508
811>>811
1008>>1008
579>>579
948>>948
828>>828
923>>923
517>AAATTCAGATGGACACCATAATACTTCTAATTTTATGAATTTATTTTTGGACAATATAAGTTTCTGGAAATTGTTTGGGTTTGAAAATATAAATTTTGTA>616
180>>180
4509>ATCGATAAATACTTTACCCTTGTTGGTAAAACTAATCTTGGATATGTAGAATTTATTTGTGACAATAATAAAAATATAGATATATATTTTGGTGAATGGA>4608
811>>811
1008>>1008
579>>579
948>>948
828>>828
923>>923
617>ATCGATAAATACTTTACCCTTGTTGGTAAAACTAATCTTGGATATGTAGAATTTATTTGTGACAATAATAAAAATATAGATATATATTTTGGTGAATGGA>716
180>>180
4609>AAACATCGTCATCTAAAAGCACTATATTTAGCGGAaaTGGTAGAAATGTTGTAGTAGAGCCTATATATAATCCTGATACGGGTGAAGATATATCTACTTC>4708
811>>811
1008>>1008
579>>579
948>>948
828>>828
923>>923
717>AAACATCGTCATCTAAAAGCACTATATTTAGCGGAAATGGTAGAAATGTTGTAGTAGAGCCTATATATAATCCTGATACGGGTGAAGATATATCTACTTC>816
180>>180

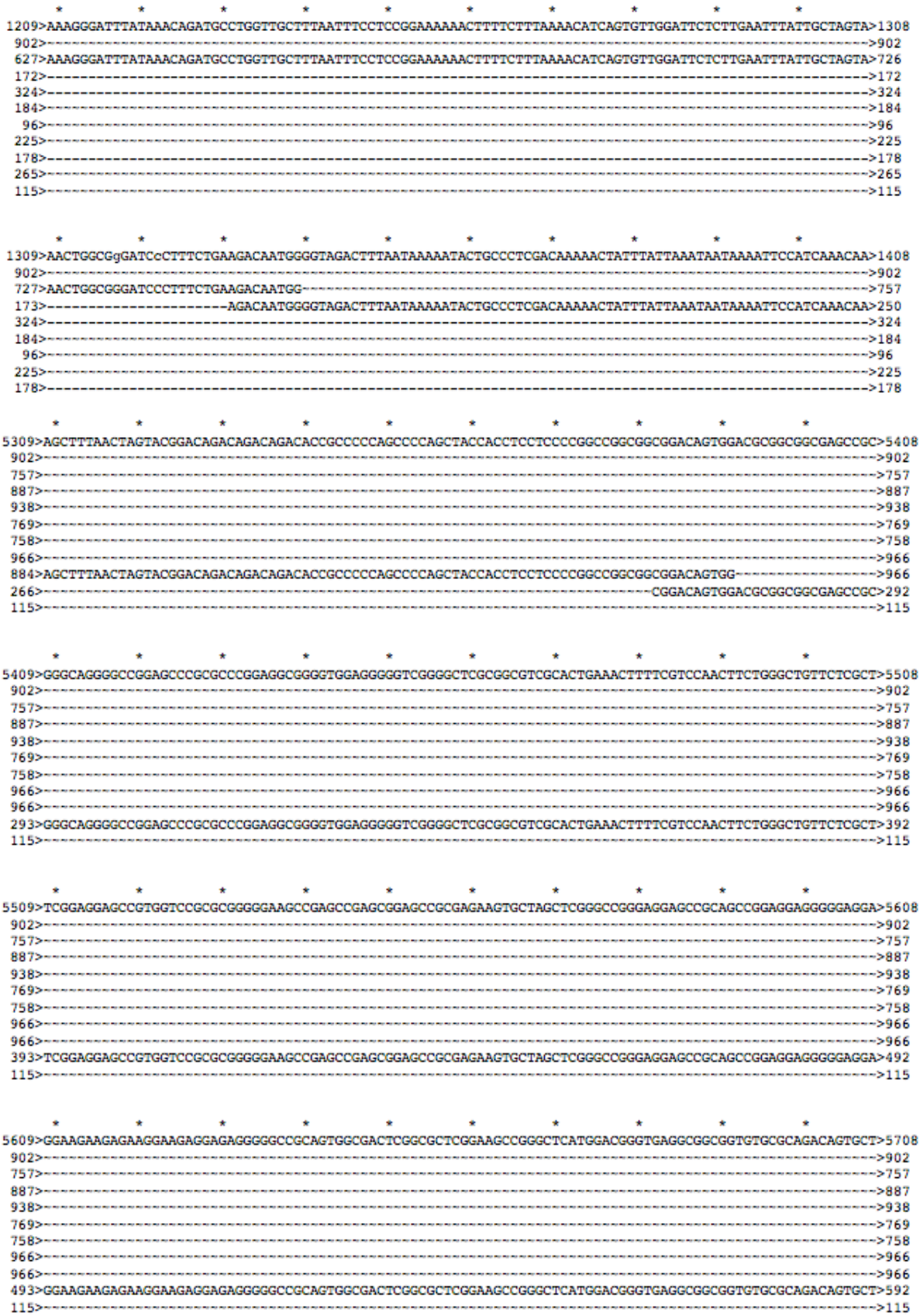

4709>ACTAGATTTTTCCTATGAACCTCTCTATGGAATAGATAGATATATCAATAAAGTATTGATAGCACCTGATTTATATACAAGTTTAATAAATATTAATACC>4808
811>>811
1008>>1008
579>>579
948>>948
828>>828
923>>923
817>ACTAGATTTTTCCTATGAACCTCTCTATGGAATAGATAGATATATCAATAAAGTATTGATAGCACCTGATTTATATACAAGTTTAATAAATATTAATACC>916
180>>180
4809>AATTATTATTCAAATGAGTACTACCCTGAGATTATAGTTCTTAACCCAAATACATTCCACAAAAAAGTAAATATAAATTTAGATAGTTCTTCTTTTGAGT>4908
811>>811
1008>>1008
579>>579
948>>948
828>>828
923>>923
917>AATTATTATTCAAATGAGTACTACCCTGAGATTATA>952
181>GAGATTATAGTTCTTAACCCAAATACATTCCACAAAAAAGTAAATATAAATTTAGATAGTTCTTCTTTTGAGT>253
4909>ATAAATGGTCGACAGAAGGAAGTGACTTTATTTTAGTTAGATACTTAGAAGAAAGTAATAAAAAAATATTACAAAAAATAAGAATCAAAGGTATCTTATC>5008
811>>811
1008>>1008
579>>579
948>>948
828>>828
923>>923
952>>952
254>ATAAATGGTCGACAGAAGGAAGTGACTTTATTTTAGTTAGATACTTAGAAGAAAGTAATAAAAAAATATTACAAAAAATAAGAATCAAAGGTATCTTATC>353
5009>TAATACTCAATCATTTAATAAAATGAGTATAGATTTTAAAGATATTAAAAAACTATCATTAGGATATATAATGAGTAATTTTAAATCATTTAATTCTGAA>5108
811>>811
1008>>1008
579>>579
948>>948
828>>828
923>>923
952>>952
354>TAATACTCAATCATTTAATAAAATGAGTATAGATTTTAAAGATATTAAAAAACTATCATTAGGATATATAATGAGTAATTTTAAATCATTTAATTCTGAA>453
5109>AATGAATTAGATAGAGATCATTTAGGATTTAAAATAATAGATAATAAAACTTATTACTATGATGAAGATAGTAAATTAGTTAAAGGATTAATCAATATAA>5208
811>>811
1008>>1008
579>>579
948>>948
828>>828
923>>923
952>>952
454>AATGAATTAGATAGAGATCATTTAGGATTTAAAATAATAGATAATAAAACTTATTACTATGATGAAGATAGTAAATTAGTTAAAGGATTAATCAATATAA>553
5209>ATAATTCATTATTCTATTTTGATCCTATAGAATTTAACTTAGTAACTGGATGGCAAACTATCAATGGTAAAAAATATTATTTTGATATAAATACTGGAGC>5308
811>>811
1008>>1008
579>>579
948>>948
828>>828
923>>923
952>>952
554>ATAATTCATTATTCTATTTTGATCCTATAGAATTTAACTTAGTAACTGGATGGCAAACTATCAATGGTAAAAAATATTATTTTGATATAAATACTGGAGC>653

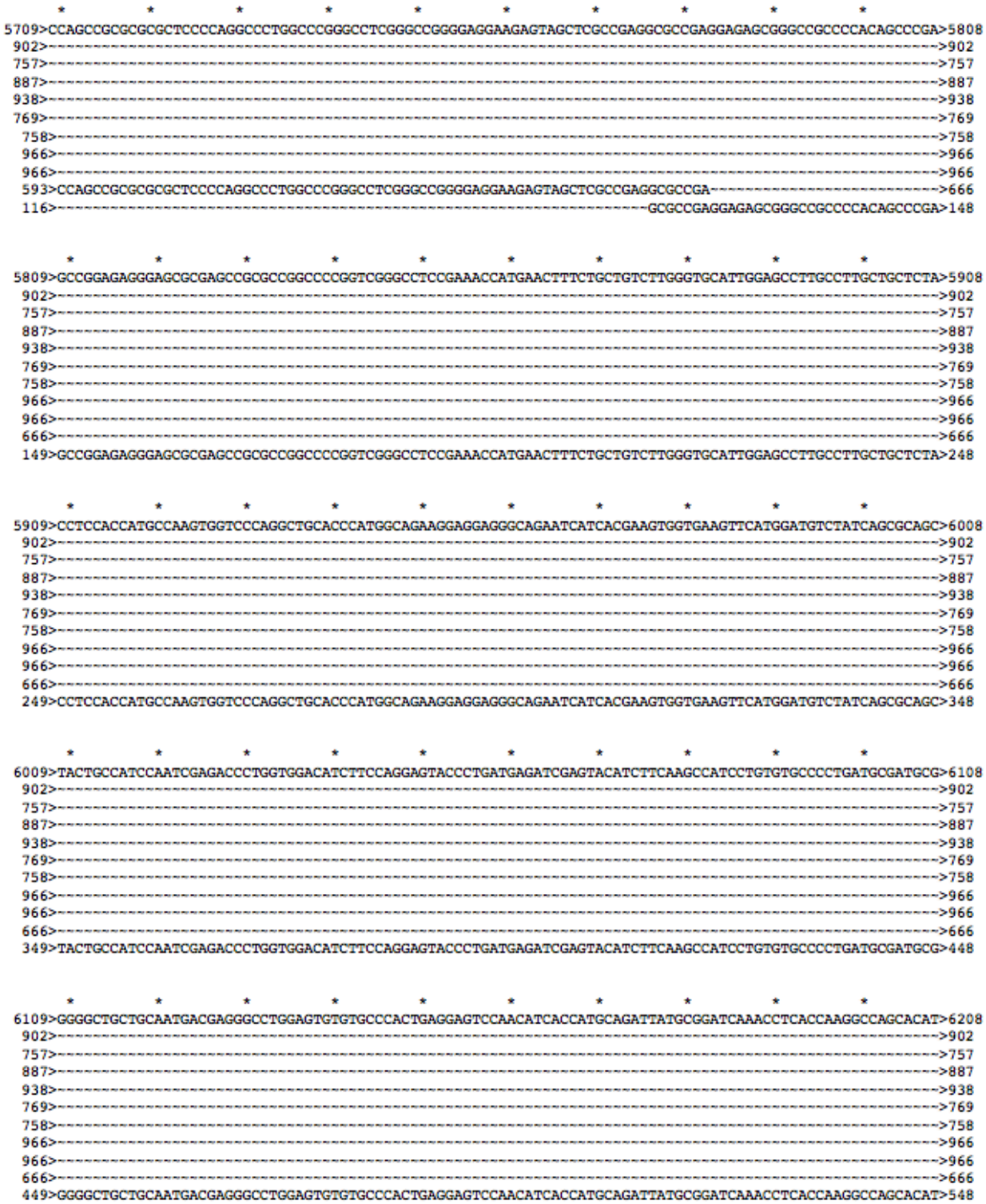
```
     *         *         *         *         *         *         *         *         *         *
5309>AGCTTTAACTAGTAATAGTGACTCTGAATGTCCCCTGTCCCACGATGGGTACTGCCTCCATGATGGTGTGTGCATGTATATTGAAGCATTGGACAAGTAT>5408
 811>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>811
1008>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>1008
 579>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>579
 948>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>948
 828>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>828
 923>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>923
 952>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>952
 654>AGCTTTAACTAGTAATAGTGACTCTGAATGTCCCCTGTCCCACGATGGGTACTGCCTCCATGATGGTGTGTGCATGTATATTGAAGCATTGGACAAGTAT>753


     *         *         *         *         *         *         *         *         *         *
5409>GCATGCAACTGTGTTGTTGGCTACATCGGGGAGCGATGTCAGTACCGAGACCTGAAGTGGTGGGAACTGCGCTTACCAGAAACTGGAGgtacccatcacc>5508
 811>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>811
1008>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>1008
 579>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>579
 948>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>948
 828>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>828
 923>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>923
 952>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>952
 754>GCATGCAACTGTGTTGTTGGCTACATCGGGGAGCGATGTCAGTACCGAGACCTGAAGTGGTGGGAACTGCGCTTACCAGAAACTGGAGGTACCCATCACC>853


     *         *
5509>atcatcaccactaa>5522
 811>~~~~~~~~~~~~~>811
1008>~~~~~~~~~~~~~>1008
 579>~~~~~~~~~~~~~>579
 948>~~~~~~~~~~~~~>948
 828>~~~~~~~~~~~~~>828
 923>~~~~~~~~~~~~~>923
 952>~~~~~~~~~~~~~>952
 854>ATCATCACCACTAA>867
```

Figure 2-B: CAV alignment




609>TTCTCCTCGCTGCATTTCATGGTGGAGGTGAAGGGCGACCTGACTGCCAAGAAAATGGTGCTGGCTTTGCTGGAGCTGGCGCAGCAGGACCACGGTGCTC>708
604>TTCTCCTCGCTGCATTTCATGGTGGAGGTGAAGGGCGACCTGACTGCCAAGAAAATGGTGCTGGCTTTGCTGGAGCTGGCGCAGCAGGACCACGGTGCTC>703
314>>314
172>>172
324>>324
184>>184
96>>96
225>>225
178>>178
265>>265
115>>115
709>TGGACTGCTGCGTGGTGGTCATTCTCTCTCACGGCTGTCAGGCCAGCCACCTGCAGTTCCCAGGGGCTGTCTACGGCACAGATGGATGCCCTGTGTCGGT>808
704>TGGACTGCTGCGTGGTGGTCATTCTCTCTCACGGCTGTCAGGCCAGCCACCTGCAGTTCCCAGGGGCTGTCTACGGCACAGATGGATGCCCTGTGTCGGT>803
314>>314
172>>172
324>>324
184>>184
96>>96
225>>225
178>>178
265>>265
115>>115
809>CGAGAAGATTGTGAACATCTTCAATGGGACCAGCTGCCCCAGCCTGGGAGGGAAGCCCAAGCTCTTTTTCATCCAGGCCTGTGGTGGGGAGCAGAAAGAC>908
804>CGAGAAGATTGTGAACATCTTCAATGGGACCAGCTGCCCCAGCCTGGGAGGGAAGCCCAAGCTCTTTTTCATCCAGGCCTGTGGTGGGGAGCAGAAAGA->902
315>GAGCAGAAAGAC>326
172>>172
324>>324
184>>184
96>>96
225>>225
178>>178
265>>265
115>>115
909>CATGGGTTTGAGGTGGCCTCCACTTCCCCTGAAGACGAGTCCCCTGGCAGTAACCCCGAGCCAGATGCCACCCCGTTCCAGGAAGGTTTGAGGACCTTCG>1008
902>>902
327>CATGGGTTTGAGGTGGCCTCCACTTCCCCTGAAGACGAGTCCCCTGGCAGTAACCCCGAGCCAGATGCCACCCCGTTCCAGGAAGGTTTGAGGACCTTCG>426
172>>172
324>>324
184>>184
96>>96
225>>225
178>>178
265>>265
115>>115
1009>ACCAGCTGGACGCCATATCTAGTTTGCCCACACCCAGTGACATCTTTGTGTCCTACTCTACTTTCCCAGGTTTTGTTTCCTGGAGGGACCCCAAGAGTGG>1108
902>>902
427>ACCAGCTGGACGCCATATCTAGTTTGCCCACACCCAGTGACATCTTTGTGTCCTACTCTACTTTCCCAGGTTTTGTTTCCTGGAGGGACCCCAAGAGTGG>526
172>>172
324>>324
184>>184
96>>96
225>>225
178>>178
265>>265
115>>115
1109>CTCCTGGTACGTTGAGACCCTGGACGACATCTTTGAGCAGTGGGCTCACTCTGAAGACCTGCAGTCCCTCCTGCTTAGGGTCGCTAATGCTGTTTCGGTG>1208
902>>902
527>CTCCTGGTACGTTGAGACCCTGGACGACATCTTTGAGCAGTGGGCTCACTCTGAAGACCTGCAGTCCCTCCTGCTTAGGGTCGCTAATGCTGTTTCGGTG>626
172>>172
324>>324
184>>184
96>>96
225>>225
178>>178
265>>265
115>>115


1209>AAAGGGATTTATAAACAGATGCCTGGTTGCTTTAATTTCCTCCGGAAAAAACTTTTCTTTAAAACATCAGTGTTGGATTCTCTTGAATTTATTGCTAGTA>1308
902>>902
627>AAAGGGATTTATAAACAGATGCCTGGTTGCTTTAATTTCCTCCGGAAAAAACTTTTCTTTAAAACATCAGTGTTGGATTCTCTTGAATTTATTGCTAGTA>726
172>>172
324>>324
184>>184
96>>96
225>>225
178>>178
265>>265
115>>115
1309>AACTGGCGgGATCcCTTTCTGAAGACAATGGGGTAGACTTTAATAAAAATACTGCCCTCGACAAAAACTATTTATTAAATAATAAAATTCCATCAAACAA>1408
902>>902
727>AACTGGCGGGATCCCTTTCTGAAGACAATGG>757
173>AGACAATGGGGTAGACTTTAATAAAAATACTGCCCTCGACAAAAACTATTTATTAAATAATAAAATTCCATCAAACAA>250
324>>324
184>>184
96>>96
225>>225
178>>178
5309>AGCTTTAACTAGTACGGACAGACAGACAGACACCGCCCCCAGCCCCAGCTACCACCTCCTCCCCGGCCGGCGGCGGACAGTGGACGCGGCGGCGAGCCGC>5408
902>>902
757>>757
887>>887
938>>938
769>>769
758>>758
966>>966
884>AGCTTTAACTAGTACGGACAGACAGACAGACACCGCCCCCAGCCCCAGCTACCACCTCCTCCCCGGCCGGCGGCGGACAGTGG>966
266>CGGACAGTGGACGCGGCGGCGAGCCGC>292
115>>115
5409>GGGCAGGGGCCGGAGCCCGCGCCCGGAGGCGGGGTGGAGGGGGTCGGGGCTCGCGGCGTCGCACTGAAACTTTTCGTCCAACTTCTGGGCTGTTCTCGCT>5508
902>>902
757>>757
887>>887
938>>938
769>>769
758>>758
966>>966
966>>966
293>GGGCAGGGGCCGGAGCCCGCGCCCGGAGGCGGGGTGGAGGGGGTCGGGGCTCGCGGCGTCGCACTGAAACTTTTCGTCCAACTTCTGGGCTGTTCTCGCT>392
115>>115
5509>TCGGAGGAGCCGTGGTCCGCGCGGGGGAAGCCGAGCCGAGCGGAGCCGCGAGAAGTGCTAGCTCGGGCCGGGAGGAGCCGCAGCCGGAGGAGGGGGAGGA>5608
902>>902
757>>757
887>>887
938>>938
769>>769
758>>758
966>>966
966>>966
393>TCGGAGGAGCCGTGGTCCGCGCGGGGGAAGCCGAGCCGAGCGGAGCCGCGAGAAGTGCTAGCTCGGGCCGGGAGGAGCCGCAGCCGGAGGAGGGGGAGGA>492
115>>115
5609>GGAAGAAGAGAAGGAAGAGGAGAGGGGGCCGCAGTGGCGACTCGGCGCTCGGAAGCCGGGCTCATGGACGGGTGAGGCGGCGGTGTGCGCAGACAGTGCT>5708
902>>902
757>>757
887>>887
938>>938
769>>769
758>>758
966>>966
966>>966
493>GGAAGAAGAGAAGGAAGAGGAGAGGGGGCCGCAGTGGCGACTCGGCGCTCGGAAGCCGGGCTCATGGACGGGTGAGGCGGCGGTGTGCGCAGACAGTGCT>592
115>>115


5709>CCAGCCGCGCGCGCTCCCCAGGCCCTGGCCCGGGCCTCGGGCCGGGGAGGAAGAGTAGCTCGCCGAGGCGCCGAGGAGAGCGGGCCGCCCCACAGCCCGA>5808
902>>902
757>>757
887>>887
938>>938
769>>769
758>>758
966>>966
966>>966
593>CCAGCCGCGCGCGCTCCCCAGGCCCTGGCCCGGGCCTCGGGCCGGGGAGGAAGAGTAGCTCGCCGAGGCGCCGA>666
116>GCGCCGAGGAGAGCGGGCCGCCCCACAGCCCGA>148
5809>GCCGGAGAGGGAGCGCGAGCCGCGCCGGCCCCGGTCGGGCCTCCGAAACCATGAACTTTCTGCTGTCTTGGGTGCATTGGAGCCTTGCCTTGCTGCTCTA>5908
902>>902
757>>757
887>>887
938>>938
769>>769
758>>758
966>>966
966>>966
666>>666
149>GCCGGAGAGGGAGCGCGAGCCGCGCCGGCCCCGGTCGGGCCTCCGAAACCATGAACTTTCTGCTGTCTTGGGTGCATTGGAGCCTTGCCTTGCTGCTCTA>248
5909>CCTCCACCATGCCAAGTGGTCCCAGGCTGCACCCATGGCAGAAGGAGGAGGGCAGAATCATCACGAAGTGGTGAAGTTCATGGATGTCTATCAGCGCAGC>6008
902>>902
757>>757
887>>887
938>>938
769>>769
758>>758
966>>966
966>>966
666>>666
249>CCTCCACCATGCCAAGTGGTCCCAGGCTGCACCCATGGCAGAAGGAGGAGGGCAGAATCATCACGAAGTGGTGAAGTTCATGGATGTCTATCAGCGCAGC>348
6009>TACTGCCATCCAATCGAGACCCTGGTGGACATCTTCCAGGAGTACCCTGATGAGATCGAGTACATCTTCAAGCCATCCTGTGTGCCCCTGATGCGATGCG>6108
902>>902
757>>757
887>>887
938>>938
769>>769
758>>758
966>>966
966>>966
666>>666
349>TACTGCCATCCAATCGAGACCCTGGTGGACATCTTCCAGGAGTACCCTGATGAGATCGAGTACATCTTCAAGCCATCCTGTGTGCCCCTGATGCGATGCG>448
6109>GGGGCTGCTGCAATGACGAGGGCCTGGAGTGTGTGCCCACTGAGGAGTCCAACATCACCATGCAGATTATGCGGATCAAACCTCACCAAGGCCAGCACAT>6208
902>>902
757>>757
887>>887
938>>938
769>>769
758>>758
966>>966
966>>966
666>>666
449>GGGGCTGCTGCAATGACGAGGGCCTGGAGTGTGTGCCCACTGAGGAGTCCAACATCACCATGCAGATTATGCGGATCAAACCTCACCAAGGCCAGCACAT>548

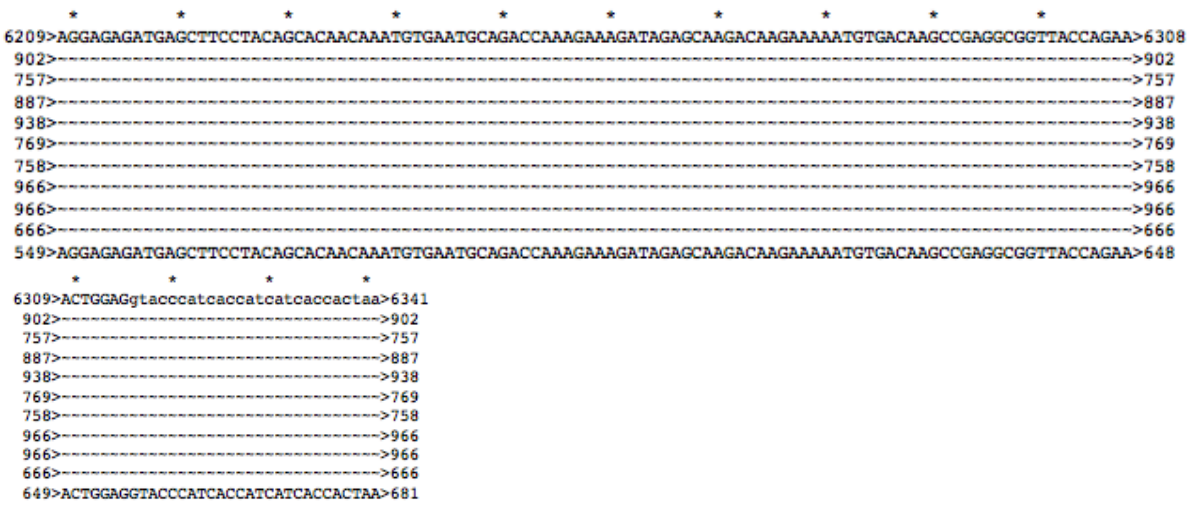
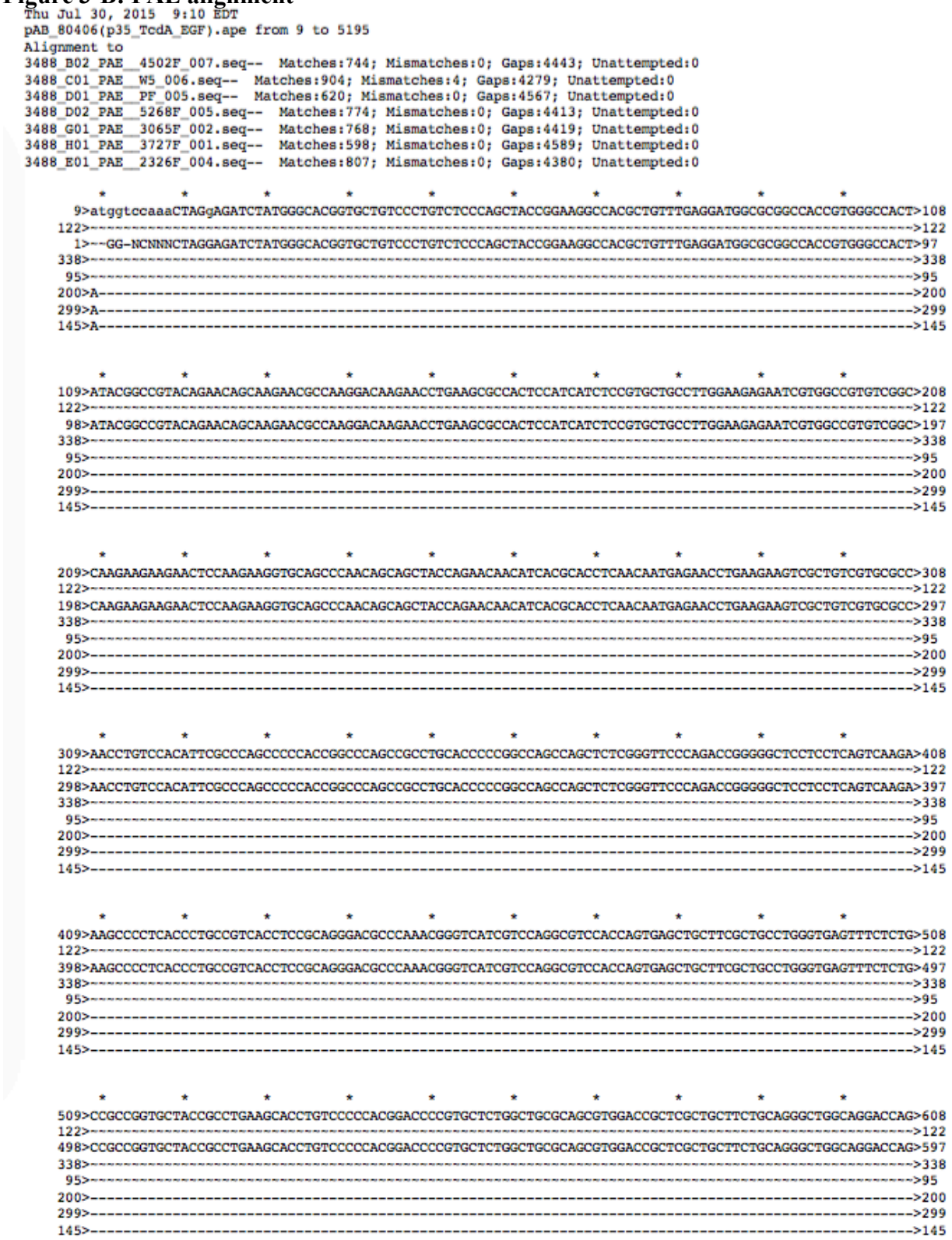

6209>AGGAGAGATGAGCTTCCTACAGCACAACAAATGTGAATGCAGACCAAAGAAAGATAGAGCAAGACAAGAAAAATGTGACAAGCCGAGGCGGTTACCAGAA>6308
902>>902
757>>757
887>>887
938>>938
769>>769
758>>758
966>>966
966>>966
666>>666
549>AGGAGAGATGAGCTTCCTACAGCACAACAAATGTGAATGCAGACCAAAGAAAGATAGAGCAAGACAAGAAAAATGTGACAAGCCGAGGCGGTTACCAGAA>648
6309>ACTGGAGgtacccatcaccatcatcaccactaa>6341
902>>902
757>>757
887>>887
938>>938
769>>769
758>>758
966>>966
966>>966
666>>666
649>ACTGGAGGTACCCATCACCATCATCACCACTAA>681

Figure 3-B: PAE alignment



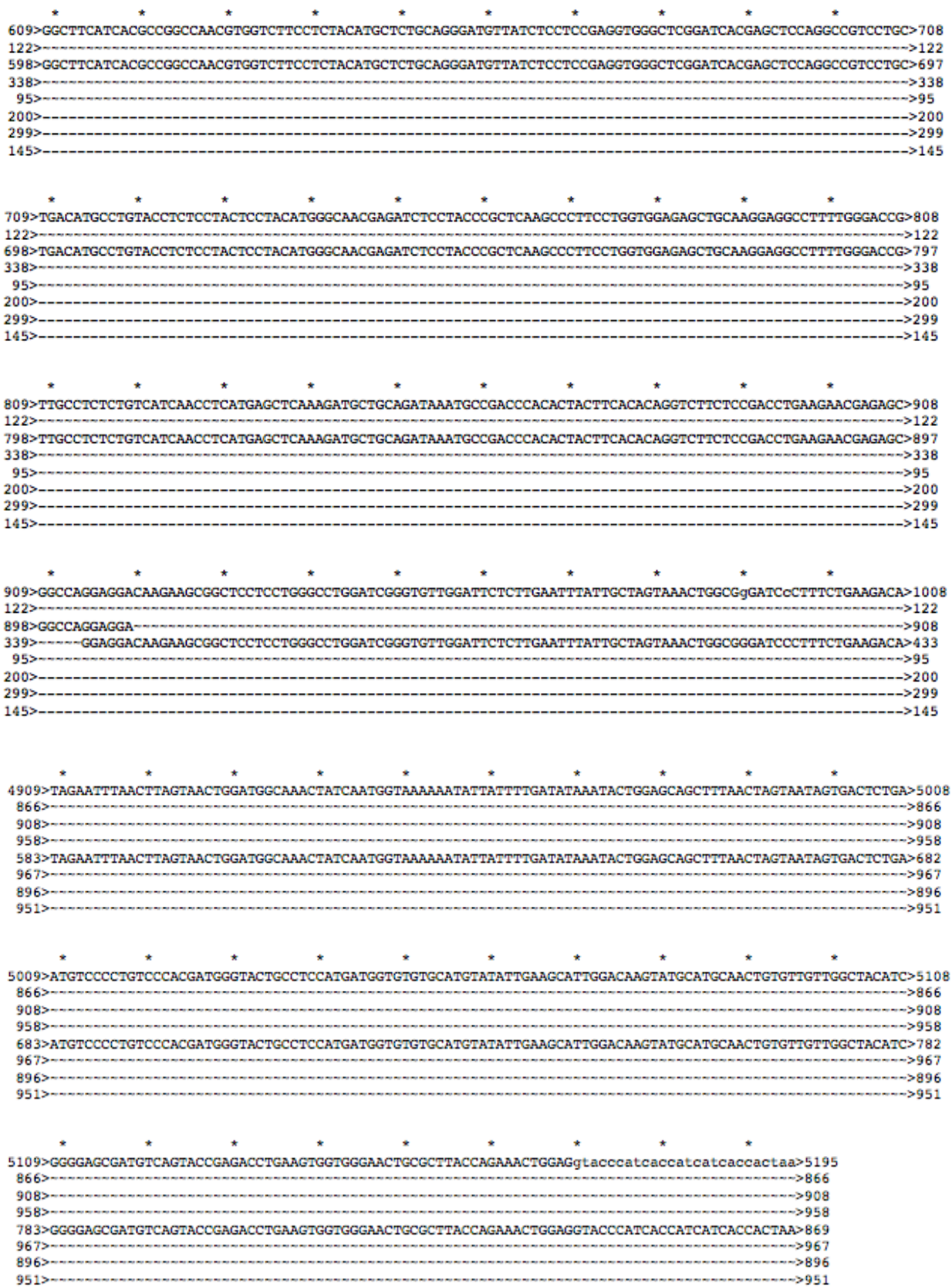
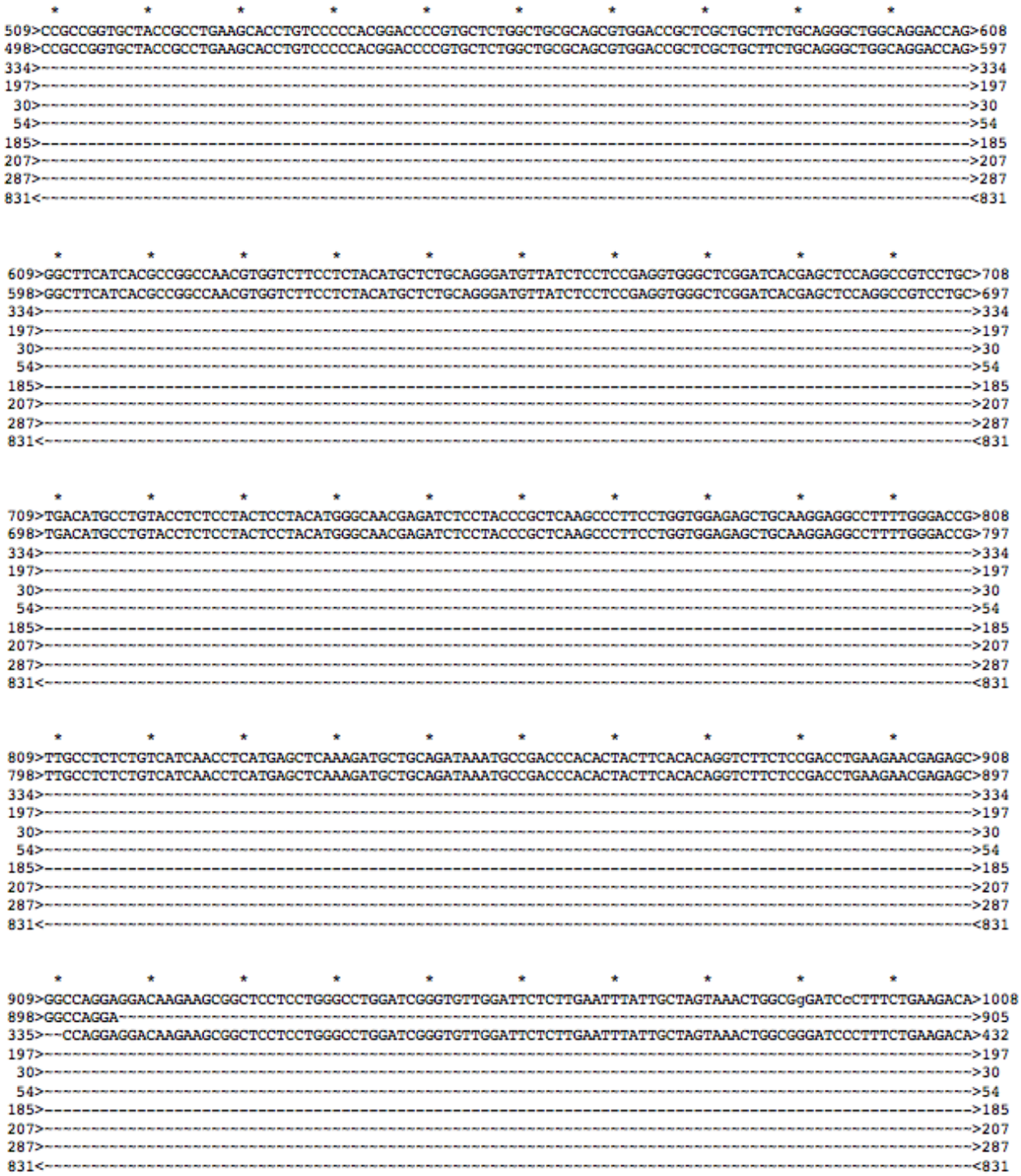
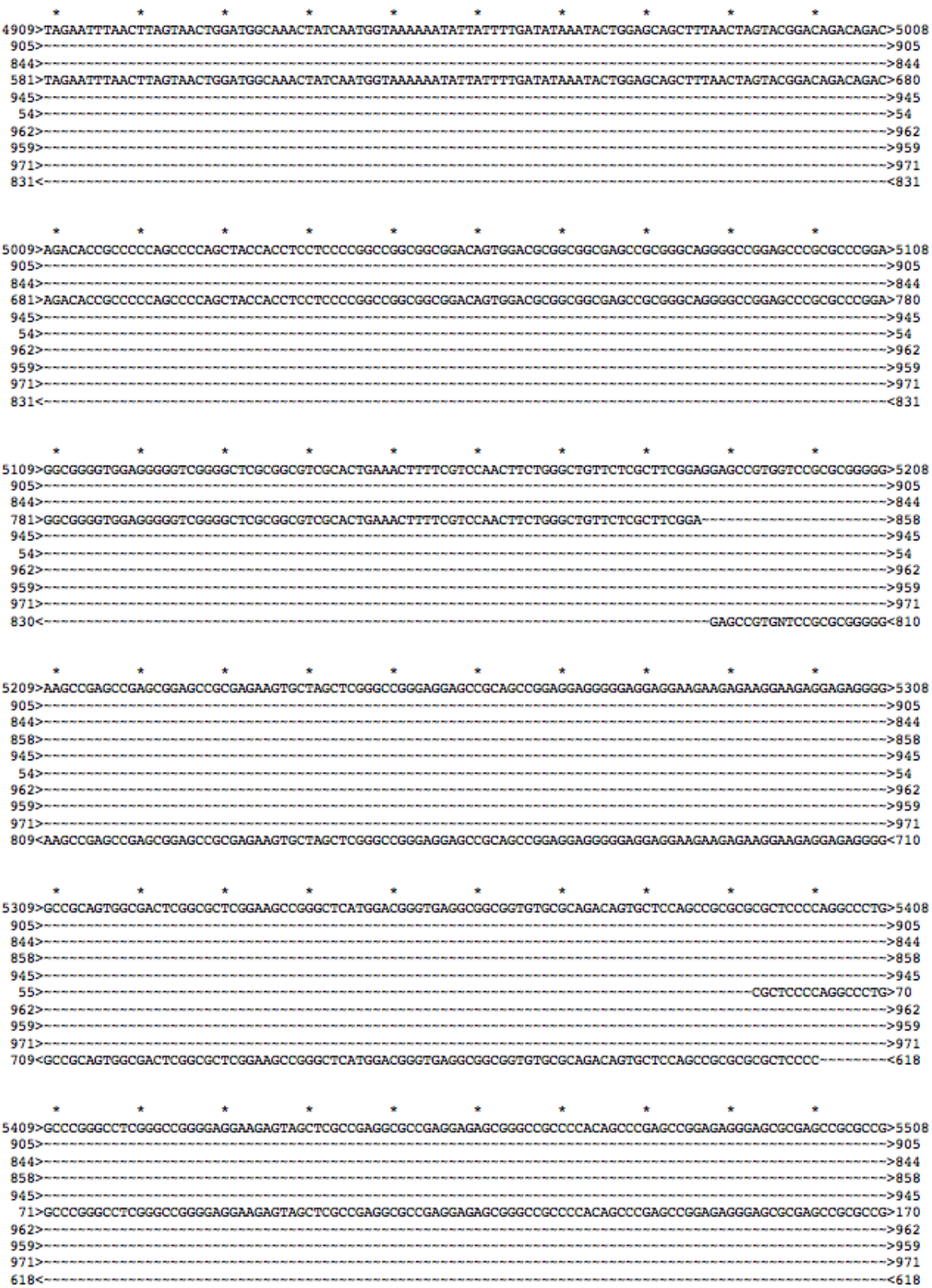

609>GGCTTCATCACGCCGGCCAACGTGGTCTTCCTCTACATGCTCTGCAGGGATGTTATCTCCTCCGAGGTGGGCTCGGATCACGAGCTCCAGGCCGTCCTGC>708
122>>122
598>GGCTTCATCACGCCGGCCAACGTGGTCTTCCTCTACATGCTCTGCAGGGATGTTATCTCCTCCGAGGTGGGCTCGGATCACGAGCTCCAGGCCGTCCTGC>697
338>>338
95>>95
200>>200
299>>299
145>>145
709>TGACATGCCTGTACCTCTCCTACTCCTACATGGGCAACGAGATCTCCTACCCGCTCAAGCCCTTCCTGGTGGAGAGCTGCAAGGAGGCCTTTTGGGACCG>808
122>>122
698>TGACATGCCTGTACCTCTCCTACTCCTACATGGGCAACGAGATCTCCTACCCGCTCAAGCCCTTCCTGGTGGAGAGCTGCAAGGAGGCCTTTTGGGACCG>797
338>>338
95>>95
200>>200
299>>299
145>>145
809>TTGCCTCTCTGTCATCAACCTCATGAGCTCAAAGATGCTGCAGATAAATGCCGACCCACACTACTTCACACAGGTCTTCTCCGACCTGAAGAACGAGAGC>908
122>>122
798>TTGCCTCTCTGTCATCAACCTCATGAGCTCAAAGATGCTGCAGATAAATGCCGACCCACACTACTTCACACAGGTCTTCTCCGACCTGAAGAACGAGAGC>897
338>>338
95>>95
200>>200
299>>299
145>>145
909>GGCCAGGAGGACAAGAAGCGGCTCCTCCTGGGCCTGGATCGGGTGTTGGATTCTCTTGAATTTATTGCTAGTAAACTGGCGgGATCcCTTTCTGAAGACA>1008
122>>122
898>GGCCAGGAGGA>908
339>-----GGAGGACAAGAAGCGGCTCCTCCTGGGCCTGGATCGGGTGTTGGATTCTCTTGAATTTATTGCTAGTAAACTGGCGGGATCCCTTTCTGAAGACA>433
95>>95
200>>200
299>>299
145>>145
4909>TAGAATTTAACTTAGTAACTGGATGGCAAACTATCAATGGTAAAAAATATTATTTTGATATAAATACTGGAGCAGCTTTAACTAGTAATAGTGACTCTGA>5008
866>>866
908>>908
958>>958
583>TAGAATTTAACTTAGTAACTGGATGGCAAACTATCAATGGTAAAAAATATTATTTTGATATAAATACTGGAGCAGCTTTAACTAGTAATAGTGACTCTGA>682
967>>967
896>>896
951>>951
5009>ATGTCCCCTGTCCCACGATGGGTACTGCCTCCATGATGGTGTGTGCATGTATATTGAAGCATTGGACAAGTATGCATGCAACTGTGTTGTTGGCTACATC>5108
866>>866
908>>908
958>>958
683>ATGTCCCCTGTCCCACGATGGGTACTGCCTCCATGATGGTGTGTGCATGTATATTGAAGCATTGGACAAGTATGCATGCAACTGTGTTGTTGGCTACATC>782
967>>967
896>>896
951>>951
5109>GGGGAGCGATGTCAGTACCGAGACCTGAAGTGGTGGGAACTGCGCTTACCAGAAACTGGAGgtacccatcaccatcatcaccactaa>5195
866>>866
908>>908
958>>958
783>GGGGAGCGATGTCAGTACCGAGACCTGAAGTGGTGGGAACTGCGCTTACCAGAAACTGGAGGTACCCATCACCATCATCACCACTAA>869
967>>967
896>>896
951>>951

Figure 4-B: PAV alignment

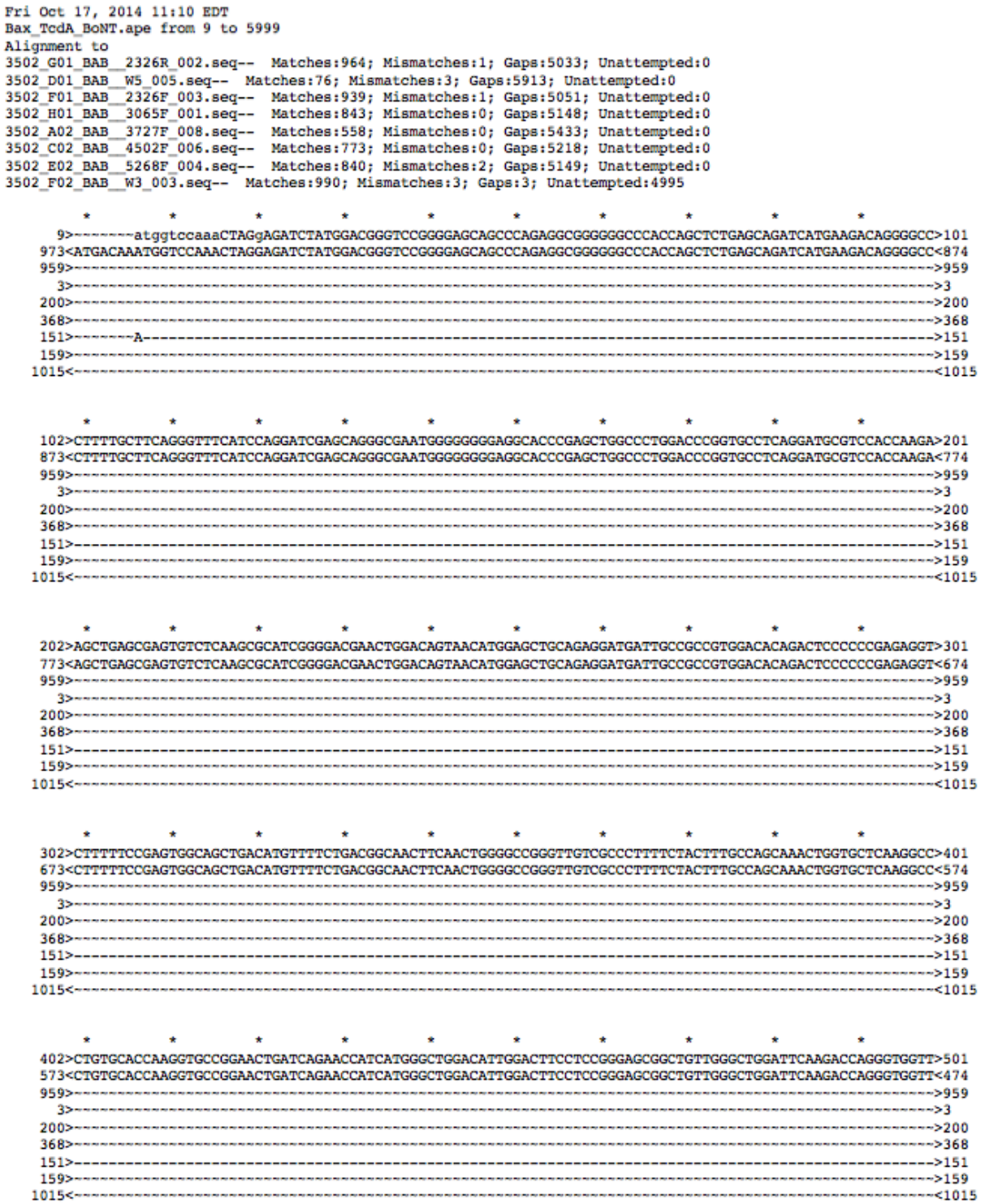
```
Thu Jul 30, 2015  9:31 EDT
pAB_80407(p35_TcdA_VEGF).ape from 9 to 6014
Alignment to
3488_A04_PAV__W5_016.seq--  Matches:901; Mismatches:4; Gaps:5101; Unattempted:0
3488_B04_PAV__PF_015.seq--  Matches:510; Mismatches:0; Gaps:5496; Unattempted:0
3488_B05_PAV__5268F_023.seq--  Matches:661; Mismatches:0; Gaps:5345; Unattempted:0
3488_C04_PAV__2326F_014.seq--  Matches:914; Mismatches:1; Gaps:5091; Unattempted:0
3488_C05_PAV__VF_022.seq--  Matches:622; Mismatches:0; Gaps:5384; Unattempted:0
3488_D06_PAV__4502F_021.seq--  Matches:778; Mismatches:0; Gaps:5228; Unattempted:0
3488_E04_PAV__3065F_012.seq--  Matches:752; Mismatches:0; Gaps:5254; Unattempted:0
3488_F04_PAV__3727F_011.seq--  Matches:683; Mismatches:1; Gaps:5322; Unattempted:0
3488_H06_PAV__W3_017.seq--  Matches:212; Mismatches:1; Gaps:5793; Unattempted:0

     *        *        *        *        *        *        *        *        *        *
  9>atggtccaaaCTAGgAGATCTATGGGCACGGTGCTGTCCCTGTCTCCCAGCTACCGGAAGGCCACGCTGTTTGAGGATGGCGCGGCCACCGTGGGCCACT>108
  1>--GG-NCNNNCTAGGAGATCTATGGGCACGGTGCTGTCCCTGTCTCCCAGCTACCGGAAGGCCACGCTGTTTGAGGATGGCGCGGCCACCGTGGGCCACT>97
334>----------------------------------------------------------------------------------------------------->334
197>----------------------------------------------------------------------------------------------------->197
 30>----------------------------------------------------------------------------------------------------->30
 54>----------------------------------------------------------------------------------------------------->54
185>A---------------------------------------------------------------------------------------------------->185
207>----------------------------------------------------------------------------------------------------->207
287>----------------------------------------------------------------------------------------------------->287
831<-----------------------------------------------------------------------------------------------------<831

     *        *        *        *        *        *        *        *        *        *
109>ATACGGCCGTACAGAACAGCAAGAACGCCAAGGACAAGAACCTGAAGCGCCACTCCATCATCTCCGTGCTGCCTTGGAAGAGAATCGTGGCCGTGTCGGC>208
 98>ATACGGCCGTACAGAACAGCAAGAACGCCAAGGACAAGAACCTGAAGCGCCACTCCATCATCTCCGTGCTGCCTTGGAAGAGAATCGTGGCCGTGTCGGC>197
334>----------------------------------------------------------------------------------------------------->334
197>----------------------------------------------------------------------------------------------------->197
 30>----------------------------------------------------------------------------------------------------->30
 54>----------------------------------------------------------------------------------------------------->54
185>----------------------------------------------------------------------------------------------------->185
207>----------------------------------------------------------------------------------------------------->207
287>----------------------------------------------------------------------------------------------------->287
831<-----------------------------------------------------------------------------------------------------<831

     *        *        *        *        *        *        *        *        *        *
209>CAAGAAGAAGAACTCCAAGAAGGTGCAGCCCAACAGCAGCTACCAGAACAACATCACGCACCTCAACAATGAGAACCTGAAGAAGTCGCTGTCGTGCGCC>308
198>CAAGAAGAAGAACTCCAAGAAGGTGCAGCCCAACAGCAGCTACCAGAACAACATCACGCACCTCAACAATGAGAACCTGAAGAAGTCGCTGTCGTGCGCC>297
334>----------------------------------------------------------------------------------------------------->334
197>----------------------------------------------------------------------------------------------------->197
 30>----------------------------------------------------------------------------------------------------->30
 54>----------------------------------------------------------------------------------------------------->54
185>----------------------------------------------------------------------------------------------------->185
207>----------------------------------------------------------------------------------------------------->207
287>----------------------------------------------------------------------------------------------------->287
831<-----------------------------------------------------------------------------------------------------<831

     *        *        *        *        *        *        *        *        *        *
309>AACCTGTCCACATTCGCCCAGCCCCCACCGGCCCAGCCGCCTGCACCCCCGGCCAGCCAGCTCTCGGGTTCCCAGACCGGGGGCTCCTCCTCAGTCAAGA>408
298>AACCTGTCCACATTCGCCCAGCCCCCACCGGCCCAGCCGCCTGCACCCCCGGCCAGCCAGCTCTCGGGTTCCCAGACCGGGGGCTCCTCCTCAGTCAAGA>397
334>----------------------------------------------------------------------------------------------------->334
197>----------------------------------------------------------------------------------------------------->197
 30>----------------------------------------------------------------------------------------------------->30
 54>----------------------------------------------------------------------------------------------------->54
185>----------------------------------------------------------------------------------------------------->185
207>----------------------------------------------------------------------------------------------------->207
287>----------------------------------------------------------------------------------------------------->287
831<-----------------------------------------------------------------------------------------------------<831

     *        *        *        *        *        *        *        *        *        *
409>AAGCCCCTCACCCTGCCGTCACCTCCGCAGGGACGCCCAAACGGGTCATCGTCCAGGCGTCCACCAGTGAGCTGCTTCGCTGCCTGGGTGAGTTTCTCTG>508
398>AAGCCCCTCACCCTGCCGTCACCTCCGCAGGGACGCCCAAACGGGTCATCGTCCAGGCGTCCACCAGTGAGCTGCTTCGCTGCCTGGGTGAGTTTCTCTG>497
334>----------------------------------------------------------------------------------------------------->334
197>----------------------------------------------------------------------------------------------------->197
 30>----------------------------------------------------------------------------------------------------->30
 54>----------------------------------------------------------------------------------------------------->54
185>----------------------------------------------------------------------------------------------------->185
207>----------------------------------------------------------------------------------------------------->207
287>----------------------------------------------------------------------------------------------------->287
831<-----------------------------------------------------------------------------------------------------<831
```

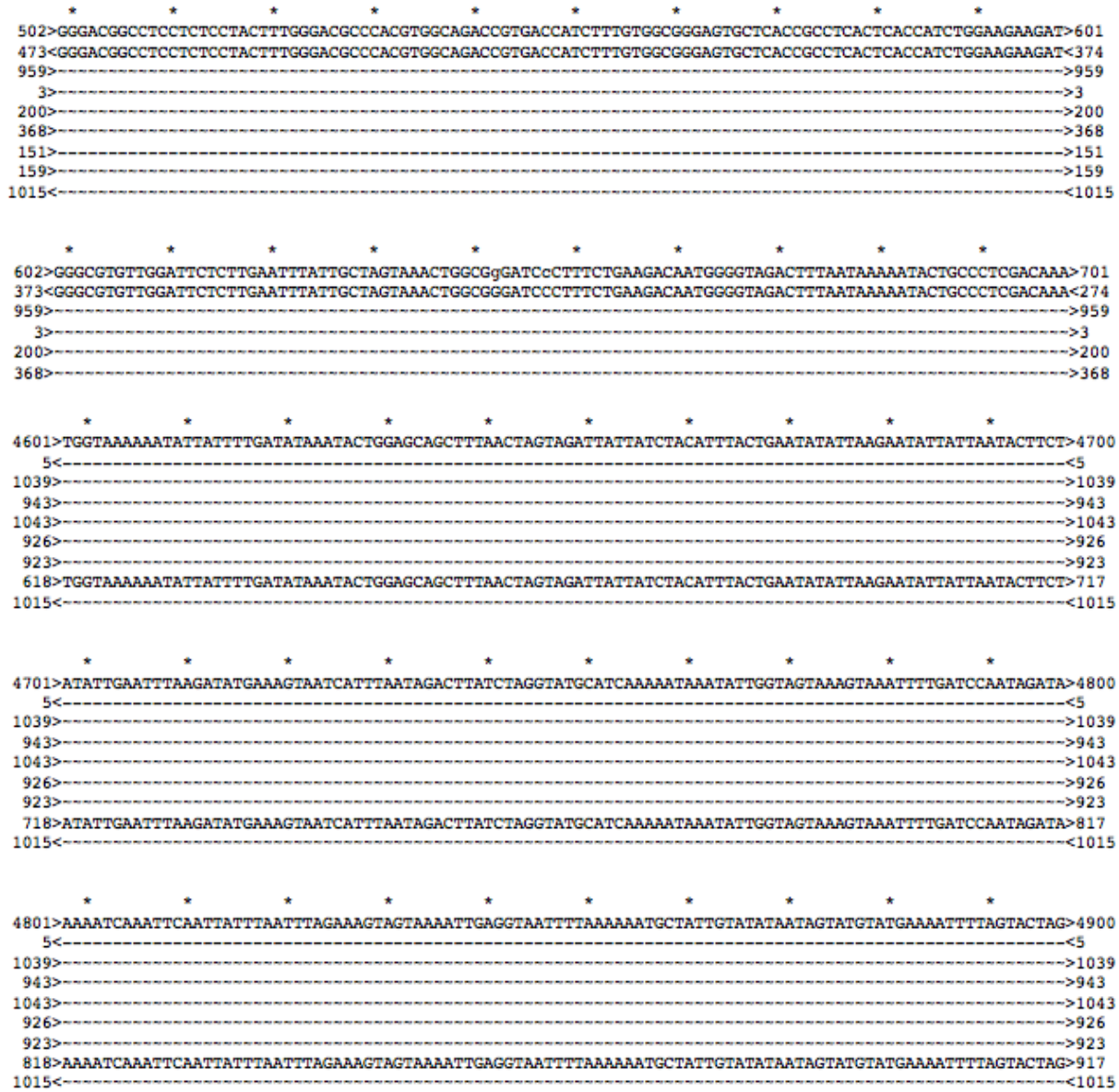

509>CCGCCGGTGCTACCGCCTGAAGCACCTGTCCCCCACGGACCCCGTGCTCTGGCTGCGCAGCGTGGACCGCTCGCTGCTTCTGCAGGGCTGGCAGGACCAG>608
498>CCGCCGGTGCTACCGCCTGAAGCACCTGTCCCCCACGGACCCCGTGCTCTGGCTGCGCAGCGTGGACCGCTCGCTGCTTCTGCAGGGCTGGCAGGACCAG>597
334>>334
197>>197
30>>30
54>>54
185>>185
207>>207
287>>287
831<<831
609>GGCTTCATCACGCCGGCCAACGTGGTCTTCCTCTACATGCTCTGCAGGGATGTTATCTCCTCCGAGGTGGGCTCGGATCACGAGCTCCAGGCCGTCCTGC>708
598>GGCTTCATCACGCCGGCCAACGTGGTCTTCCTCTACATGCTCTGCAGGGATGTTATCTCCTCCGAGGTGGGCTCGGATCACGAGCTCCAGGCCGTCCTGC>697
334>>334
197>>197
30>>30
54>>54
185>>185
207>>207
287>>287
831<<831
709>TGACATGCCTGTACCTCTCCTACTCCTACATGGGCAACGAGATCTCCTACCCGCTCAAGCCCTTCCTGGTGGAGAGCTGCAAGGAGGCCTTTTGGGACCG>808
698>TGACATGCCTGTACCTCTCCTACTCCTACATGGGCAACGAGATCTCCTACCCGCTCAAGCCCTTCCTGGTGGAGAGCTGCAAGGAGGCCTTTTGGGACCG>797
334>>334
197>>197
30>>30
54>>54
185>>185
207>>207
287>>287
831<<831
809>TTGCCTCTCTGTCATCAACCTCATGAGCTCAAAGATGCTGCAGATAAATGCCGACCCACACTACTTCACACAGGTCTTCTCCGACCTGAAGAACGAGAGC>908
798>TTGCCTCTCTGTCATCAACCTCATGAGCTCAAAGATGCTGCAGATAAATGCCGACCCACACTACTTCACACAGGTCTTCTCCGACCTGAAGAACGAGAGC>897
334>>334
197>>197
30>>30
54>>54
185>>185
207>>207
287>>287
831<<831
909>GGCCAGGAGGACAAGAAGCGGCTCCTCCTGGGCCTGGATCGGGTGTTGGATTCTCTTGAATTTATTGCTAGTAAACTGGCGgGATCcCTTTCTGAAGACA>1008
898>GGCCAGGA>905
335>--CCAGGAGGACAAGAAGCGGCTCCTCCTGGGCCTGGATCGGGTGTTGGATTCTCTTGAATTTATTGCTAGTAAACTGGCGGGATCCCTTTCTGAAGACA>432
197>>197
30>>30
54>>54
185>>185
207>>207
287>>287
831<<831

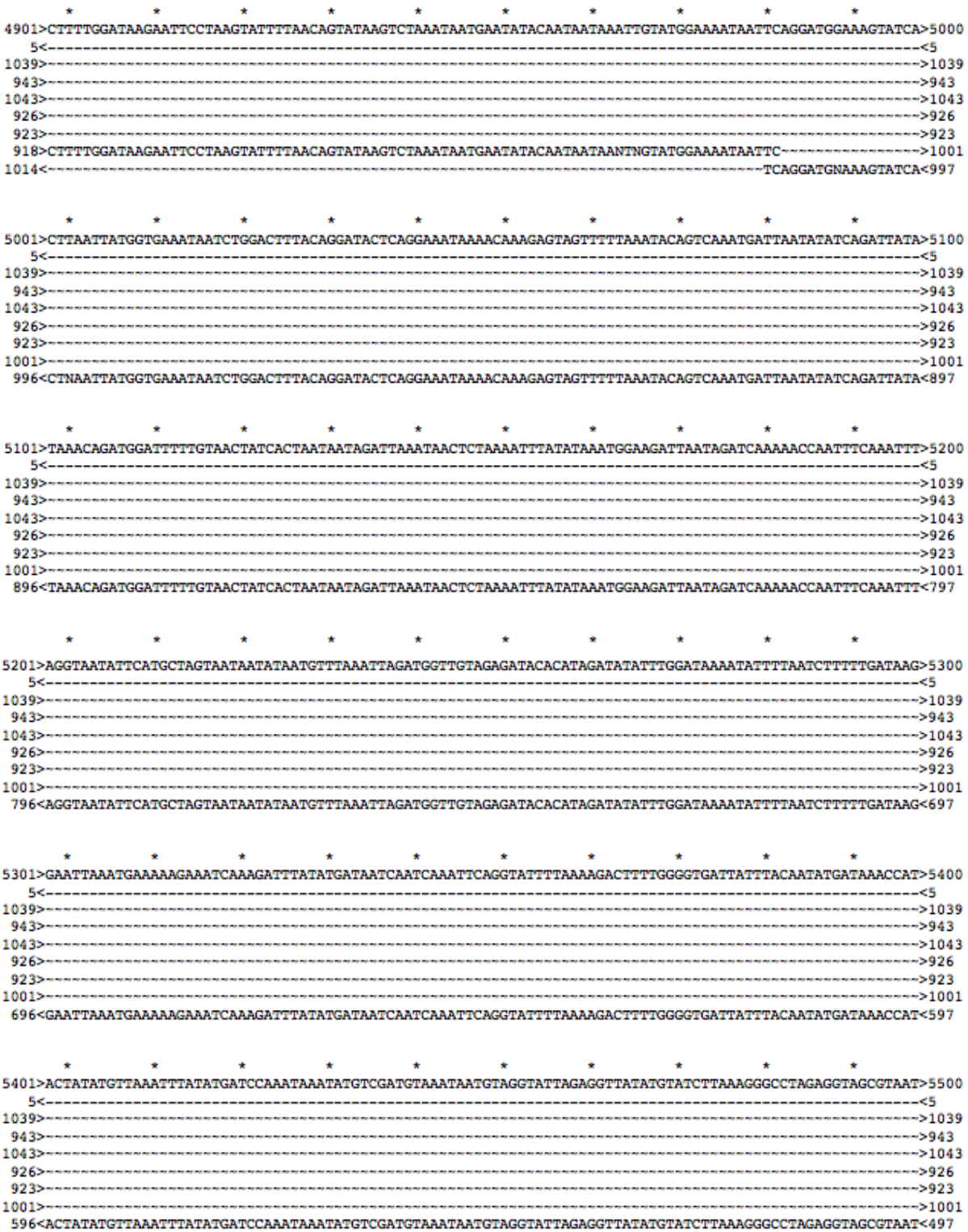

4909>TAGAATTTAACTTAGTAACTGGATGGCAAACTATCAATGGTAAAAAATATTATTTTGATATAAATACTGGAGCAGCTTTAACTAGTACGGACAGACAGAC>5008
905>>905
844>>844
581>TAGAATTTAACTTAGTAACTGGATGGCAAACTATCAATGGTAAAAAATATTATTTTGATATAAATACTGGAGCAGCTTTAACTAGTACGGACAGACAGAC>680
945>>945
54>>54
962>>962
959>>959
971>>971
831<<831
5009>AGACACCGCCCCCAGCCCCAGCTACCACCTCCTCCCCGGCCGGCGGCGGACAGTGGACGCGGCGGCGAGCCGCGGGCAGGGGCCGGAGCCCGCGCCCGGA>5108
905>>905
844>>844
681>AGACACCGCCCCCAGCCCCAGCTACCACCTCCTCCCCGGCCGGCGGCGGACAGTGGACGCGGCGGCGAGCCGCGGGCAGGGGCCGGAGCCCGCGCCCGGA>780
945>>945
54>>54
962>>962
959>>959
971>>971
831<<831
5109>GGCGGGGTGGAGGGGGTCGGGGCTCGCGGCGTCGCACTGAAACTTTTCGTCCAACTTCTGGGCTGTTCTCGCTTCGGAGGAGCCGTGGTCCGCGCGGGGG>5208
905>>905
844>>844
781>GGCGGGGTGGAGGGGGTCGGGGCTCGCGGCGTCGCACTGAAACTTTTCGTCCAACTTCTGGGCTGTTCTCGCTTCGGA----------->858
945>>945
54>>54
962>>962
959>>959
971>>971
830<<----------GAGCCGTGNTCCGCGCGGGGG<810
5209>AAGCCGAGCCGAGCGGAGCCGCGAGAAGTGCTAGCTCGGGCCGGGAGGAGCCGCAGCCGGAGGAGGGGGAGGAGGAAGAAGAGAAGGAAGAGGAGAGGGG>5308
905>>905
844>>844
858>>858
945>>945
54>>54
962>>962
959>>959
971>>971
809<AAGCCGAGCCGAGCGGAGCCGCGAGAAGTGCTAGCTCGGGCCGGGAGGAGCCGCAGCCGGAGGAGGGGGAGGAGGAAGAAGAGAAGGAAGAGGAGAGGGG<710
5309>GCCGCAGTGGCGACTCGGCGCTCGGAAGCCGGGCTCATGGACGGGTGAGGCGGCGGTGTGCGCAGACAGTGCTCCAGCCGCGCGCGCTCCCCAGGCCCTG>5408
905>>905
844>>844
858>>858
945>>945
55>>----------CGCTCCCCAGGCCCTG>70
962>>962
959>>959
971>>971
709<GCCGCAGTGGCGACTCGGCGCTCGGAAGCCGGGCTCATGGACGGGTGAGGCGGCGGTGTGCGCAGACAGTGCTCCAGCCGCGCGCGCTCCCC---------<618
5409>GCCCGGGCCTCGGGCCGGGGAGGAAGAGTAGCTCGCCGAGGCGCCGAGGAGAGCGGGCCGCCCCACAGCCCGAGCCGGAGAGGGAGCGCGAGCCGCGCCG>5508
905>>905
844>>844
858>>858
945>>945
71>GCCCGGGCCTCGGGCCGGGGAGGAAGAGTAGCTCGCCGAGGCGCCGAGGAGAGCGGGCCGCCCCACAGCCCGAGCCGGAGAGGGAGCGCGAGCCGCGCCG>170
962>>962
959>>959
971>>971
618<<618

```
     *         *         *         *         *         *         *         *         *         *
5509>GCCCCGGTCGGGCCTCCGAAACCATGAACTTTCTGCTGTCTTGGGTGCATTGGAGCCTTGCCTTGCTGCTCTACCTCCACCATGCCAAGTGGTCCCAGGC>5608
 905>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>905
 844>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>844
 858>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>858
 945>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>945
 171>GCCCCGGTCGGGCCTCCGAAACCATGAACTTTCTGCTGTCTTGGGTGCATTGGAGCCTTGCCTTGCTGCTCTACCTCCACCATGCCAAGTGGTCCCAGGC>270
 962>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>962
 959>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>959
 971>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>971
 618<~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~<618

     *         *         *         *         *         *         *         *         *         *
5609>TGCACCCATGGCAGAAGGAGGAGGGCAGAATCATCACGAAGTGGTGAAGTTCATGGATGTCTATCAGCGCAGCTACTGCCATCCAATCGAGACCCTGGTG>5708
 905>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>905
 844>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>844
 858>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>858
 945>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>945
 271>TGCACCCATGGCAGAAGGAGGAGGGCAGAATCATCACGAAGTGGTGAAGTTCATGGATGTCTATCAGCGCAGCTACTGCCATCCAATCGAGACCCTGGTG>370
 962>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>962
 959>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>959
 971>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>971
 618<~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~<618

     *         *         *         *         *         *         *         *         *         *
5709>GACATCTTCCAGGAGTACCCTGATGAGATCGAGTACATCTTCAAGCCATCCTGTGTGCCCCTGATGCGATGCGGGGGCTGCTGCAATGACGAGGGCCTGG>5808
 905>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>905
 844>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>844
 858>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>858
 945>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>945
 371>GACATCTTCCAGGAGTACCCTGATGAGATCGAGTACATCTTCAAGCCATCCTGTGTGCCCCTGATGCGATGCGGGGGCTGCTGCAATGACGAGGGCCTGG>470
 962>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>962
 959>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>959
 971>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>971
 618<~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~<618

     *         *         *         *         *         *         *         *         *         *
5809>AGTGTGTGCCCACTGAGGAGTCCAACATCACCATGCAGATTATGCGGATCAAACCTCACCAAGGCCAGCACATAGGAGAGATGAGCTTCCTACAGCACAA>5908
 905>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>905
 844>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>844
 858>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>858
 945>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>945
 471>AGTGTGTGCCCACTGAGGAGTCCAACATCACCATGCAGATTATGCGGATCAAACCTCACCAAGGCCAGCACATAGGAGAGATGAGCTTCCTACAGCACAA>570
 962>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>962
 959>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>959
 971>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>971
 618<~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~<618

     *         *         *         *         *         *         *         *         *         *
5909>CAAATGTGAATGCAGACCAAAGAAAGATAGAGCAAGACAAGAAAAATGTGACAAGCCGAGGCGGTTACCAGAAACTGGAGgtacccatcaccatcatcac>6008
 905>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>905
 844>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>844
 858>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>858
 945>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>945
 571>CAAATGTGAATGCAGACCAAAGAAAGATAGAGCAAGACAAGAAAAATGTGACAAGCCGAGGCGGTTACCAGAAACTGGAGGTACCCATCACCATCATCAC>670
 962>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>962
 959>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>959
 971>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>971
 618<~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~<618

     *
6009>cactaa>6014
 905>~~~~~~>905
 844>~~~~~~>844
 858>~~~~~~>858
 945>~~~~~~>945
 671>CACTAA>676
 962>~~~~~~>962
 959>~~~~~~>959
 971>~~~~~~>971
 618<~~~~~~<618
```

Figure 5-B: BAB alignment




502>GGGACGGCCTCCTCTCCTACTTTGGGACGCCCACGTGGCAGACCGTGACCATCTTTGTGGCGGGAGTGCTCACCGCCTCACTCACCATCTGGAAGAAGAT>601
473<GGGACGGCCTCCTCTCCTACTTTGGGACGCCCACGTGGCAGACCGTGACCATCTTTGTGGCGGGAGTGCTCACCGCCTCACTCACCATCTGGAAGAAGAT<374
959>959
3>3
200>200
368>368
151>151
159>159
1015<1015
602>GGGCGTGTTGGATTCTCTTGAATTTATTGCTAGTAAACTGGCGgGATCcCTTTCTGAAGACAATGGGGTAGACTTTAATAAAAATACTGCCCTCGACAAA>701
373<GGGCGTGTTGGATTCTCTTGAATTTATTGCTAGTAAACTGGCGGGATCCCTTTCTGAAGACAATGGGGTAGACTTTAATAAAAATACTGCCCTCGACAAA<274
959>959
3>3
200>200
368>368
4601>TGGTAAAAAATATTATTTTGATATAAATACTGGAGCAGCTTTAACTAGTAGATTATTATCTACATTTACTGAATATATTAAGAATATTATTAATACTTCT>4700
5<5
1039>1039
943>943
1043>1043
926>926
923>923
618>TGGTAAAAAATATTATTTTGATATAAATACTGGAGCAGCTTTAACTAGTAGATTATTATCTACATTTACTGAATATATTAAGAATATTATTAATACTTCT>717
1015<1015
4701>ATATTGAATTTAAGATATGAAAGTAATCATTTAATAGACTTATCTAGGTATGCATCAAAAATAAATATTGGTAGTAAAGTAAATTTTGATCCAATAGATA>4800
5<5
1039>1039
943>943
1043>1043
926>926
923>923
718>ATATTGAATTTAAGATATGAAAGTAATCATTTAATAGACTTATCTAGGTATGCATCAAAAATAAATATTGGTAGTAAAGTAAATTTTGATCCAATAGATA>817
1015<1015
4801>AAAATCAAATTCAATTATTTAATTTAGAAAGTAGTAAAATTGAGGTAATTTTAAAAAATGCTATTGTATATAATAGTATGTATGAAAATTTTAGTACTAG>4900
5<5
1039>1039
943>943
1043>1043
926>926
923>923
818>AAAATCAAATTCAATTATTTAATTTAGAAAGTAGTAAAATTGAGGTAATTTTAAAAAATGCTATTGTATATAATAGTATGTATGAAAATTTTAGTACTAG>917
1015<1015


4901>CTTTTGGATAAGAATTCCTAAGTATTTTAACAGTATAAGTCTAAATAATGAATATACAATAATAAATTGTATGGAAAATAATTCAGGATGGAAAGTATCA>5000
1039>>1039
943>>943
1043>>1043
926>>926
923>>923
918>CTTTTGGATAAGAATTCCTAAGTATTTTAACAGTATAAGTCTAAATAATGAATATACAATAATAANTNGTATGGAAAATAATTC>1001
1014<TCAGGATGNAAAGTATCA<997
5001>CTTAATTATGGTGAAATAATCTGGACTTTACAGGATACTCAGGAAATAAAACAAAGAGTAGTTTTTAAATACAGTCAAATGATTAATATATCAGATTATA>5100
1039>>1039
943>>943
1043>>1043
926>>926
923>>923
1001>>1001
996<CTNAATTATGGTGAAATAATCTGGACTTTACAGGATACTCAGGAAATAAAACAAAGAGTAGTTTTTAAATACAGTCAAATGATTAATATATCAGATTATA<897
5101>TAAACAGATGGATTTTTGTAACTATCACTAATAATAGATTAAATAACTCTAAAATTTATATAAATGGAAGATTAATAGATCAAAAACCAATTTCAAATTT>5200
1039>>1039
943>>943
1043>>1043
926>>926
923>>923
1001>>1001
896<TAAACAGATGGATTTTTGTAACTATCACTAATAATAGATTAAATAACTCTAAAATTTATATAAATGGAAGATTAATAGATCAAAAACCAATTTCAAATTT<797
5201>AGGTAATATTCATGCTAGTAATAATATAATGTTTAAATTAGATGGTTGTAGAGATACACATAGATATATTTGGATAAAATATTTTAATCTTTTTGATAAG>5300
1039>>1039
943>>943
1043>>1043
926>>926
923>>923
1001>>1001
796<AGGTAATATTCATGCTAGTAATAATATAATGTTTAAATTAGATGGTTGTAGAGATACACATAGATATATTTGGATAAAATATTTTAATCTTTTTGATAAG<697
5301>GAATTAAATGAAAAAGAAATCAAAGATTTATATGATAATCAATCAAATTCAGGTATTTTAAAAGACTTTTGGGGTGATTATTTACAATATGATAAACCAT>5400
1039>>1039
943>>943
1043>>1043
926>>926
923>>923
1001>>1001
696<GAATTAAATGAAAAAGAAATCAAAGATTTATATGATAATCAATCAAATTCAGGTATTTTAAAAGACTTTTGGGGTGATTATTTACAATATGATAAACCAT<597
5401>ACTATATGTTAAATTTATATGATCCAAATAAATATGTCGATGTAAATAATGTAGGTATTAGAGGTTATATGTATCTTAAAGGGCCTAGAGGTAGCGTAAT>5500
1039>>1039
943>>943
1043>>1043
926>>926
923>>923
1001>>1001
596<ACTATATGTTAAATTTATATGATCCAAATAAATATGTCGATGTAAATAATGTAGGTATTAGAGGTTATATGTATCTTAAAGGGCCTAGAGGTAGCGTAAT<497

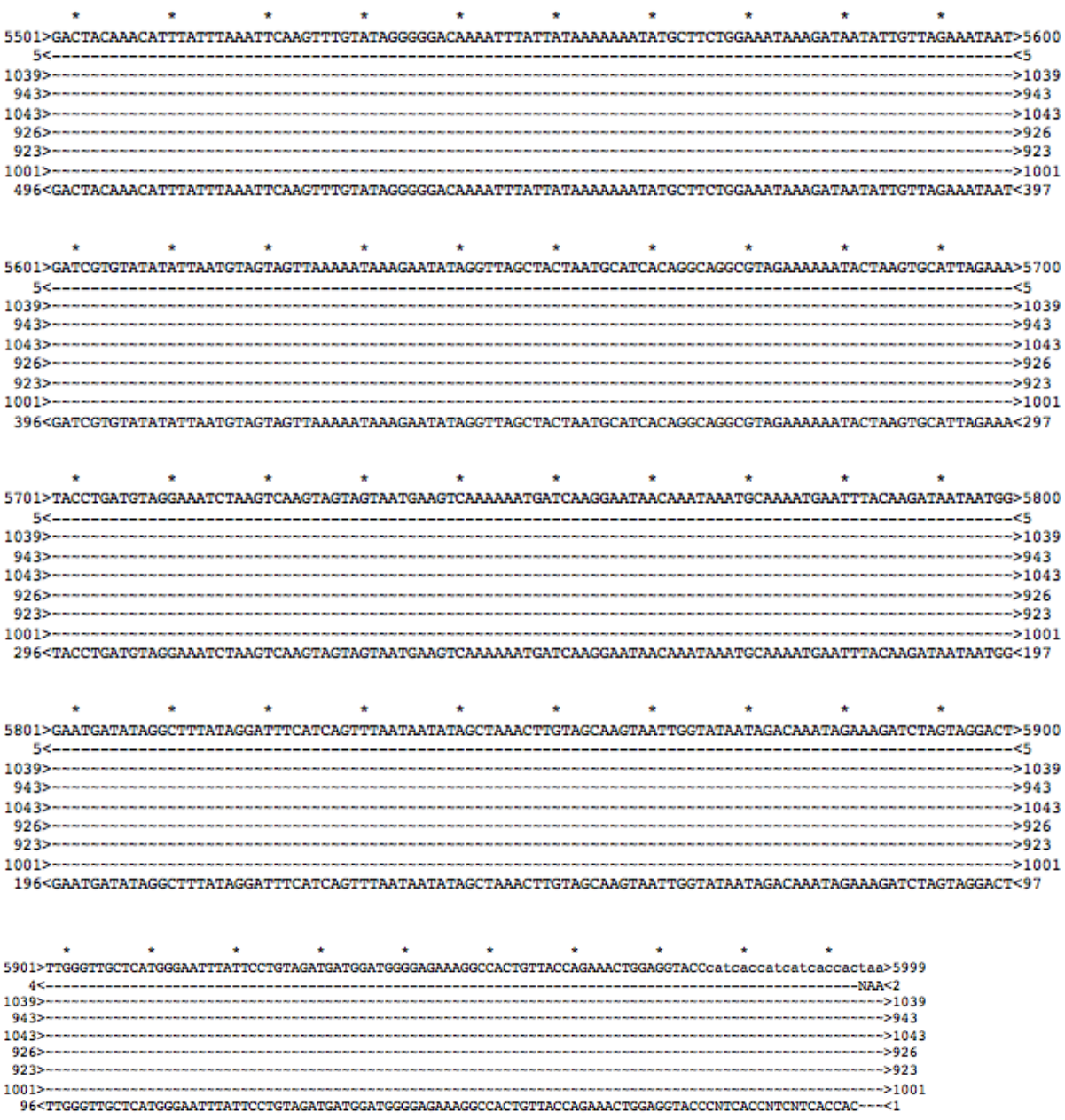
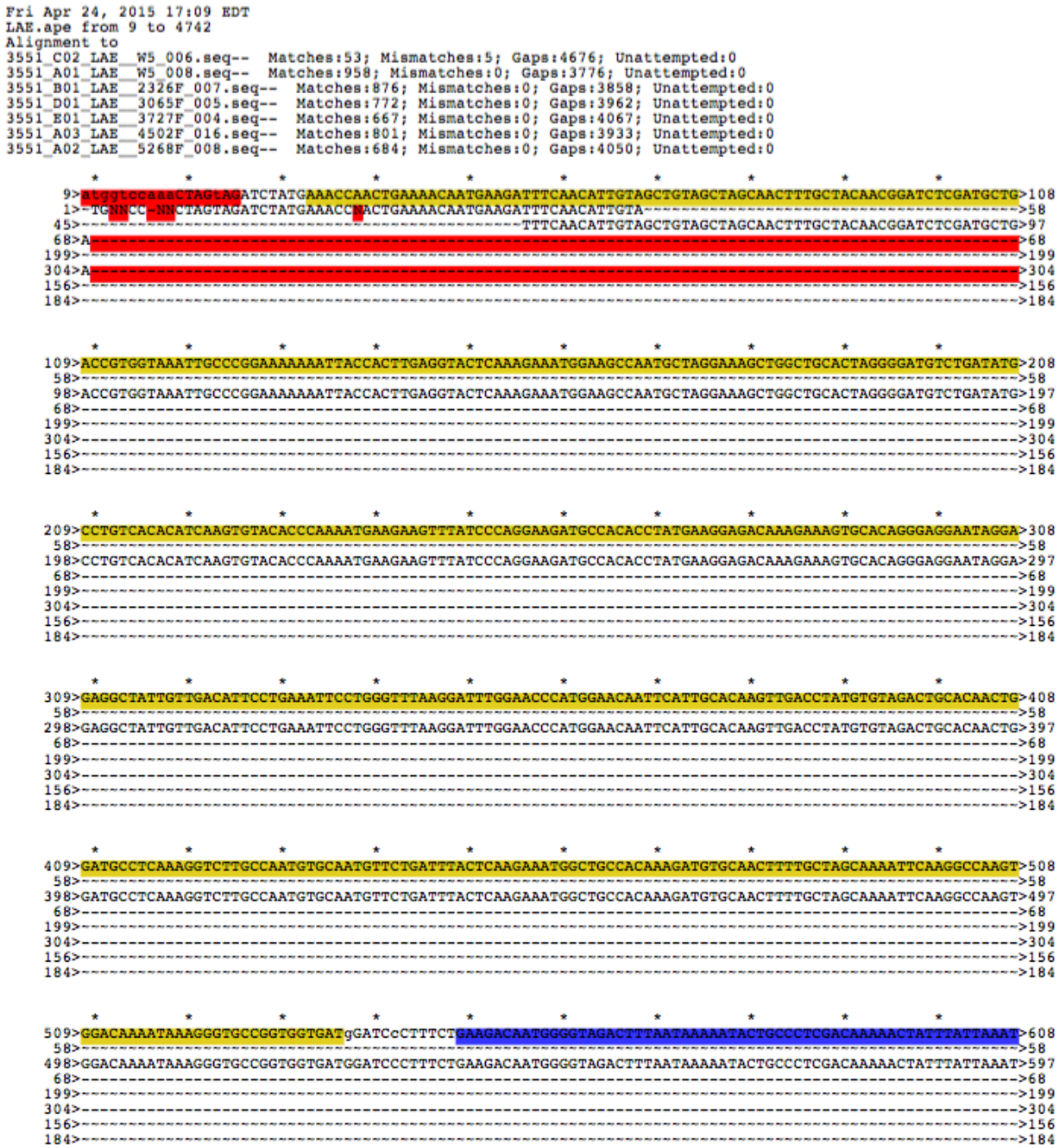

5501>GACTACAAACATTTATTTAAATTCAAGTTTGTATAGGGGGACAAAATTTATTATAAAAAAATATGCTTCTGGAAATAAAGATAATATTGTTAGAAATAAT>5600
5601>GATCGTGTATATATTAATGTAGTAGTTAAAAATAAAGAATATAGGTTAGCTACTAATGCATCACAGGCAGGCGTAGAAAAAATACTAAGTGCATTAGAAA>5700
5701>TACCTGATGTAGGAAATCTAAGTCAAGTAGTAGTAATGAAGTCAAAAAATGATCAAGGAATAACAAATAAATGCAAAATGAATTTACAAGATAATAATGG>5800
5801>GAATGATATAGGCTTTATAGGATTTCATCAGTTTAATAATATAGCTAAACTTGTAGCAAGTAATTGGTATAATAGACAAATAGAAAGATCTAGTAGGACT>5900
5901>TTGGGTTGCTCATGGGAATTTATTCCTGTAGATGATGGATGGGGAGAAAGGCCACTGTTACCAGAAACTGGAGGTACCcatcaccatcatcaccactaa>5999
NAA<2
96<TTGGGTTGCTCATGGGAATTTATTCCTGTAGATGATGGATGGGGAGAAAGGCCACTGTTACCAGAAACTGGAGGTACCCNTCACCNTCNTCACCAC~~~<1

Figure 6-B: LAE alignment



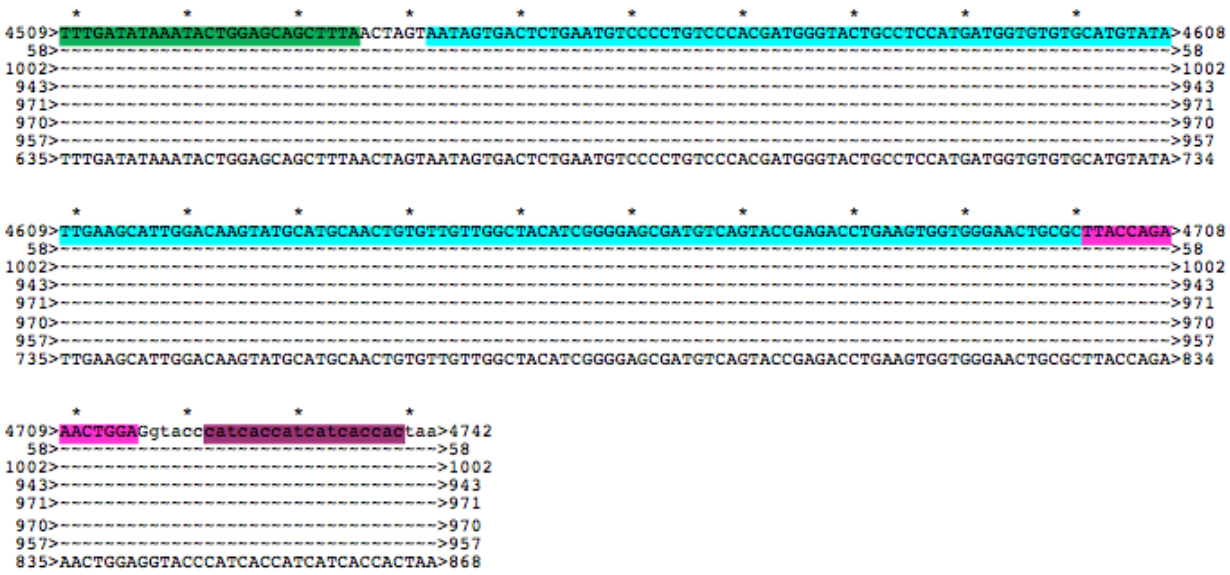
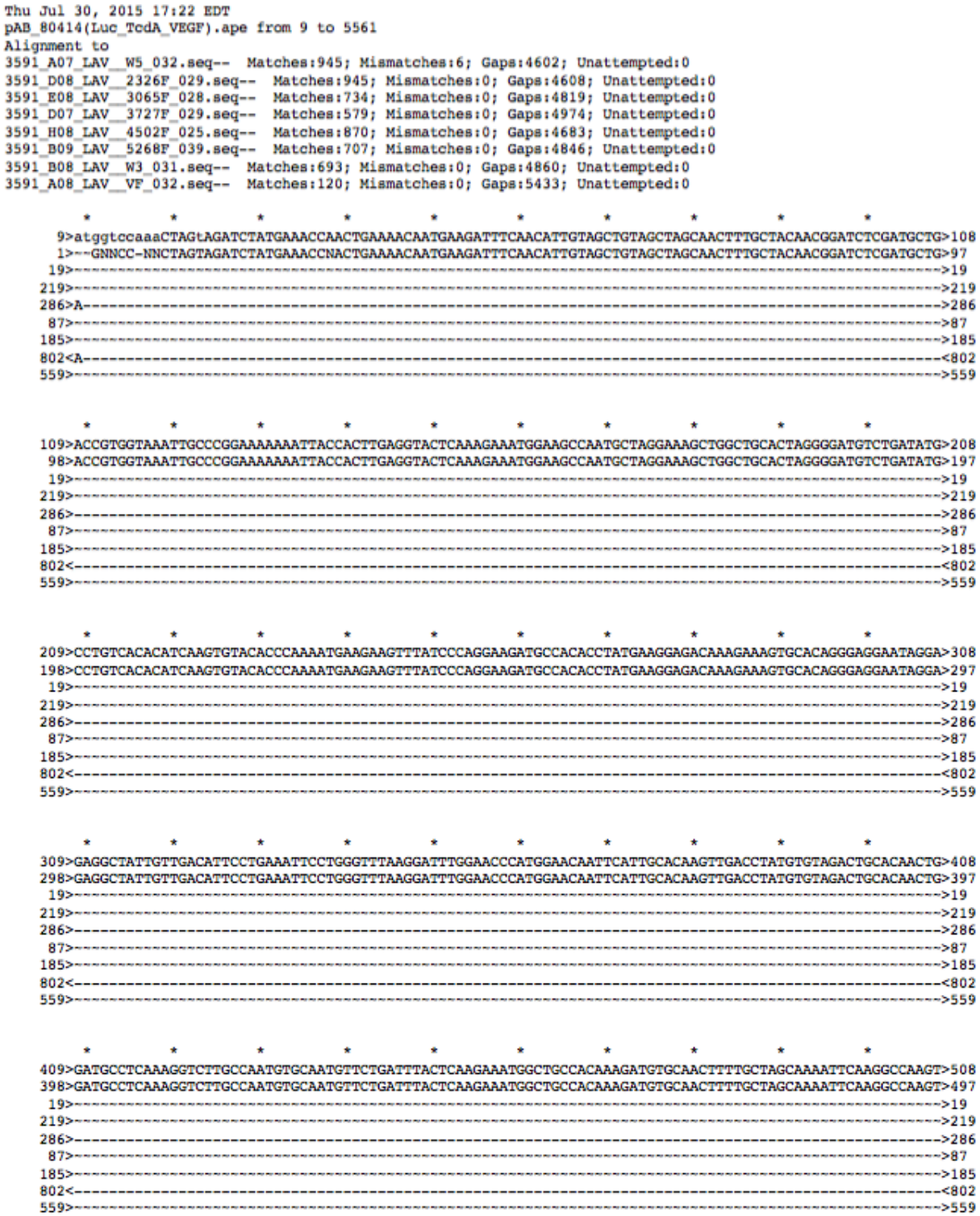
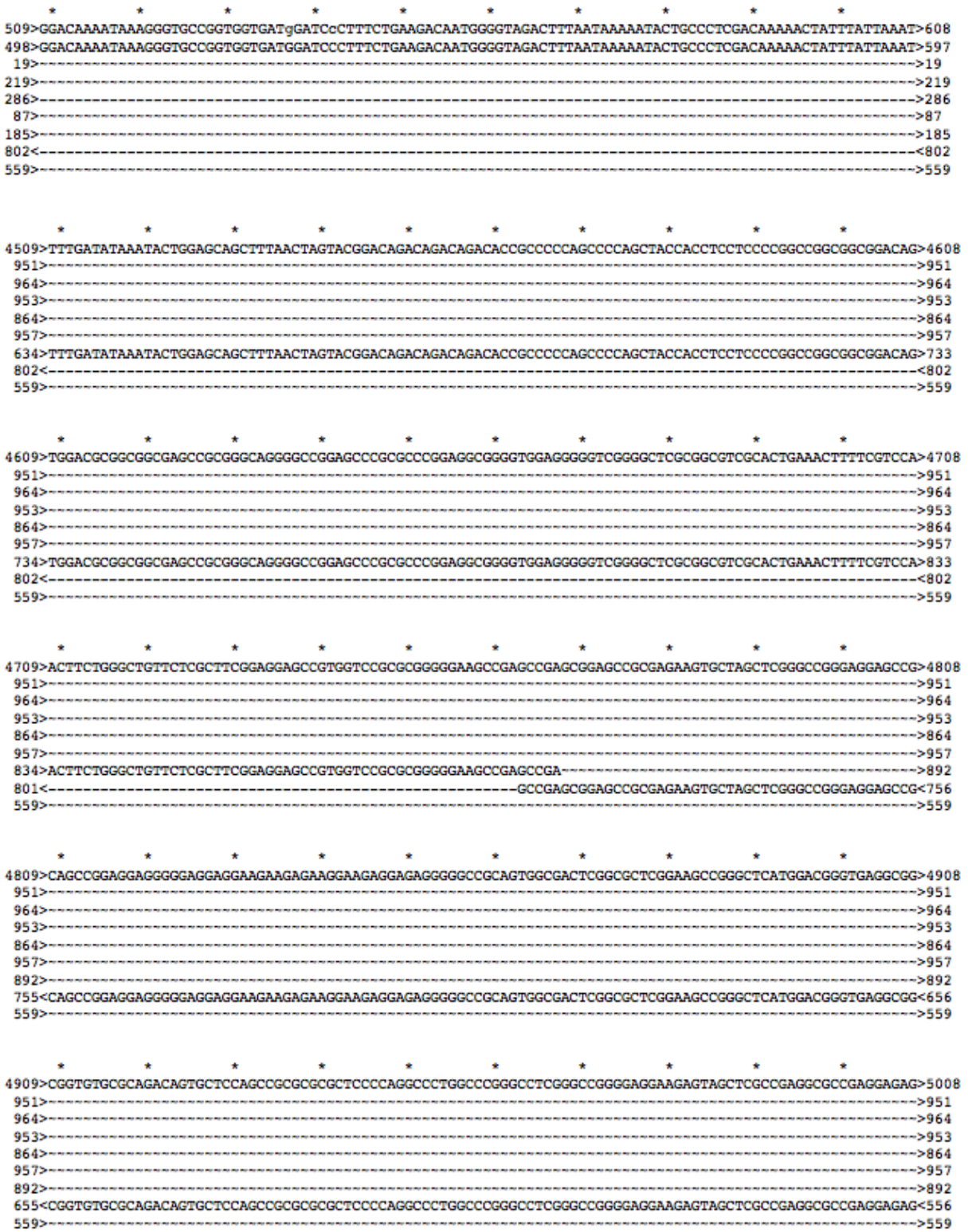

4509>TTTGATATAAATACTGGAGCAGCTTTAACTAGTAATAGTGACTCTGAATGTCCCCTGTCCCACGATGGGTACTGCCTCCATGATGGTGTGTGCATGTATA>4608
58>>58
1002>>1002
943>>943
971>>971
970>>970
957>>957
635>TTTGATATAAATACTGGAGCAGCTTTAACTAGTAATAGTGACTCTGAATGTCCCCTGTCCCACGATGGGTACTGCCTCCATGATGGTGTGTGCATGTATA>734
4609>TTGAAGCATTGGACAAGTATGCATGCAACTGTGTTGTTGGCTACATCGGGGAGCGATGTCAGTACCGAGACCTGAAGTGGTGGGAACTGCGCTTACCAGA>4708
58>>58
1002>>1002
943>>943
971>>971
970>>970
957>>957
735>TTGAAGCATTGGACAAGTATGCATGCAACTGTGTTGTTGGCTACATCGGGGAGCGATGTCAGTACCGAGACCTGAAGTGGTGGGAACTGCGCTTACCAGA>834
4709>AACTGGAGgtaccatcaccatcatcaccactaa>4742
58>>58
1002>>1002
943>>943
971>>971
970>>970
957>>957
835>AACTGGAGGTACCCATCACCATCATCACCACTAA>868

Figure 7-B: LAV alignment




509>GGACAAAATAAAGGGTGCCGGTGGTGATgGATCcCTTTCTGAAGACAATGGGGTAGACTTTAATAAAAATACTGCCCTCGACAAAAACTATTTATTAAAT>608
498>GGACAAAATAAAGGGTGCCGGTGGTGATGGATCCCTTTCTGAAGACAATGGGGTAGACTTTAATAAAAATACTGCCCTCGACAAAAACTATTTATTAAAT>597
4509>TTTGATATAAATACTGGAGCAGCTTTAACTAGTACGGACAGACAGACAGACACCGCCCCCAGCCCCAGCTACCACCTCCTCCCCGGCCGGCGGCGGACAG>4608
4609>TGGACGCGGCGGCGAGCCGCGGGCAGGGGCCGGAGCCCGCGCCCGGAGGCGGGGTGGAGGGGGTCGGGGCTCGCGGCGTCGCACTGAAACTTTTCGTCCA>4708
4709>ACTTCTGGGCTGTTCTCGCTTCGGAGGAGCCGTGGTCCGCGCGGGGGAAGCCGAGCCGAGCGGAGCCGCGAGAAGTGCTAGCTCGGGCCGGGAGGAGCCG>4808
4809>CAGCCGGAGGAGGGGGAGGAGGAAGAAGAGAAGGAAGAGGAGAGGGGGCCGCAGTGGCGACTCGGCGCTCGGAAGCCGGGCTCATGGACGGGTGAGGCGG>4908
4909>CGGTGTGCGCAGACAGTGCTCCAGCCGCGCGCGCTCCCCAGGCCCTGGCCCGGGCCTCGGGCCGGGGAGGAAGAGTAGCTCGCCGAGGCGCCGAGGAGAG>5008

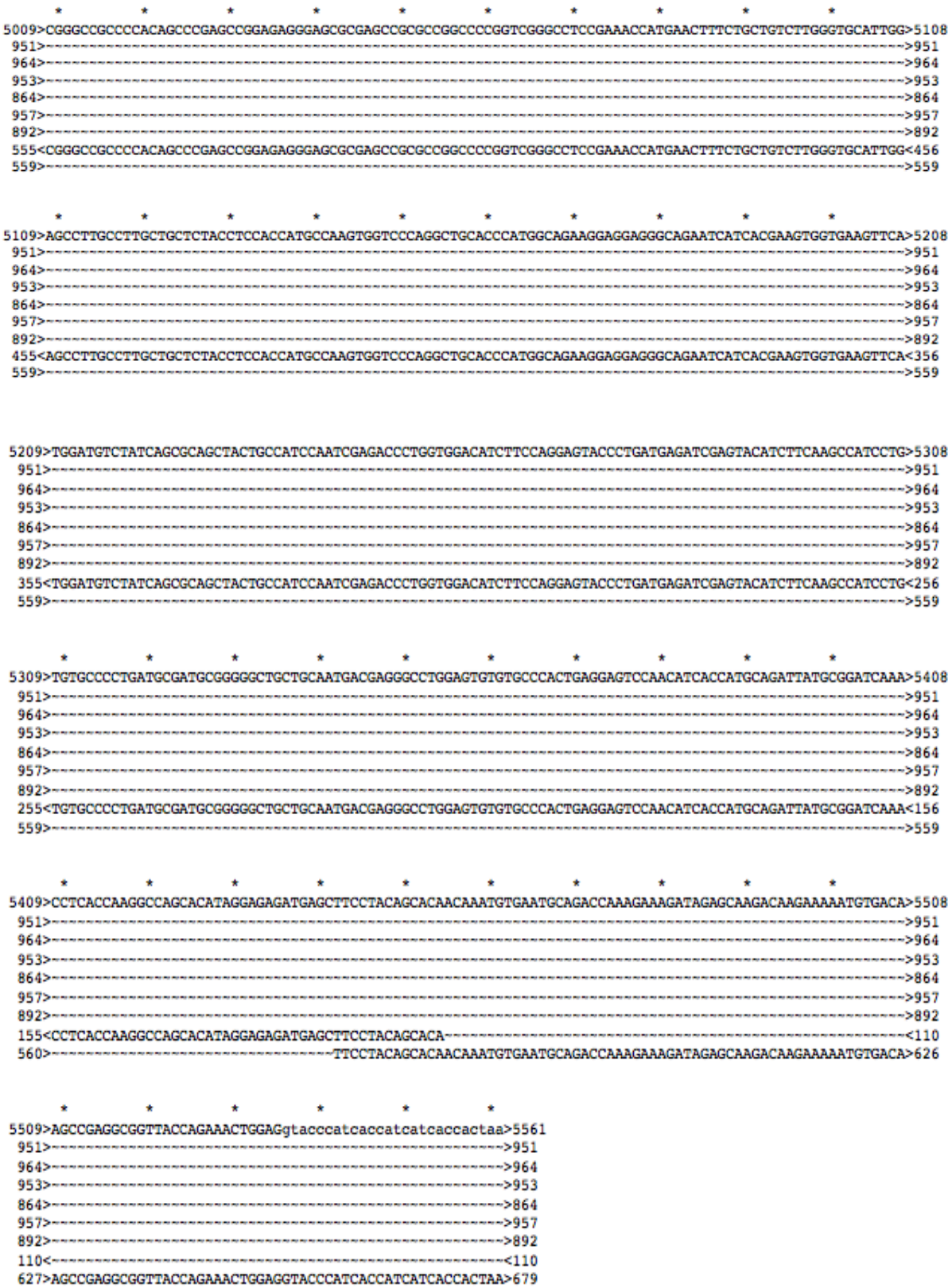
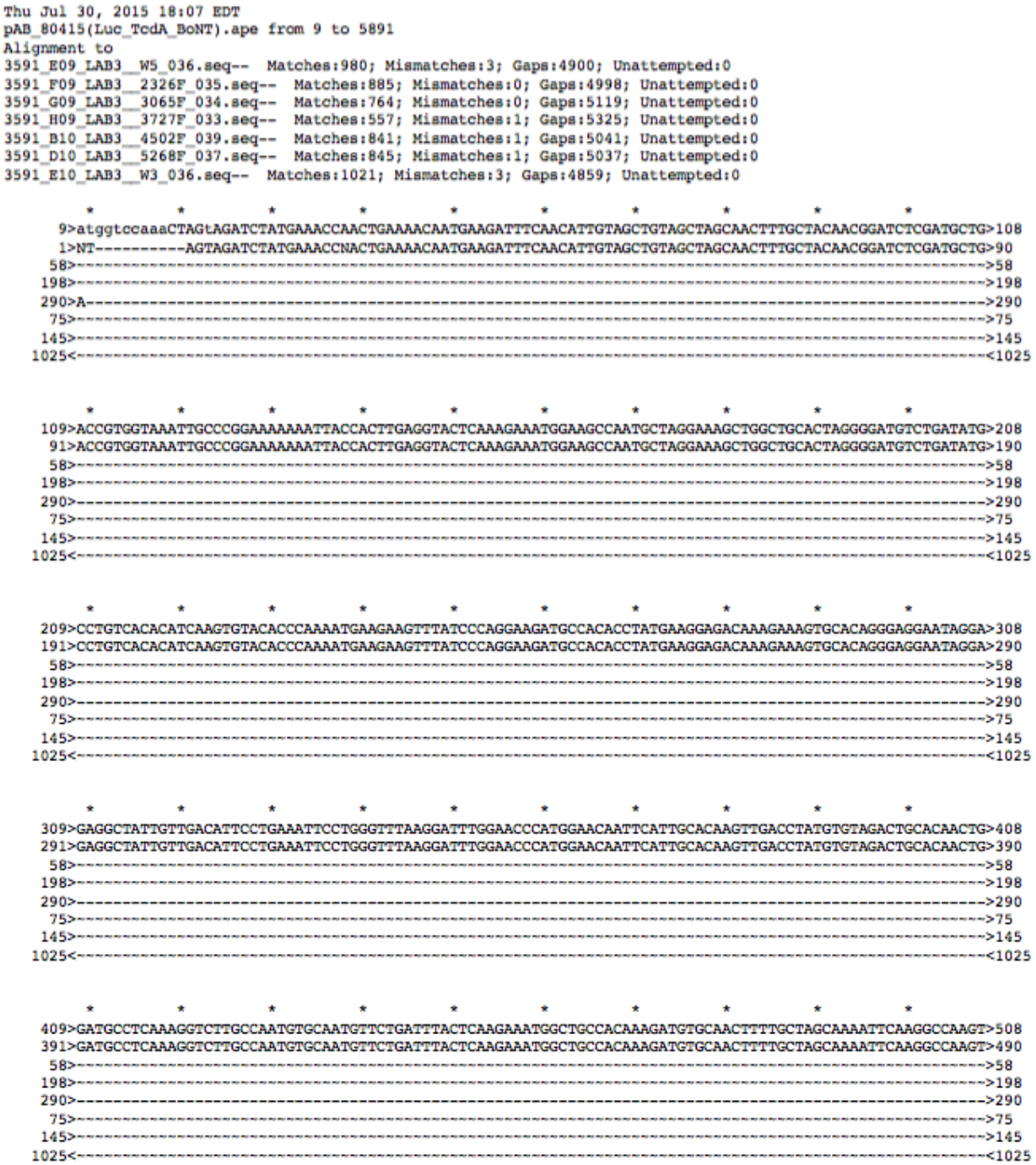
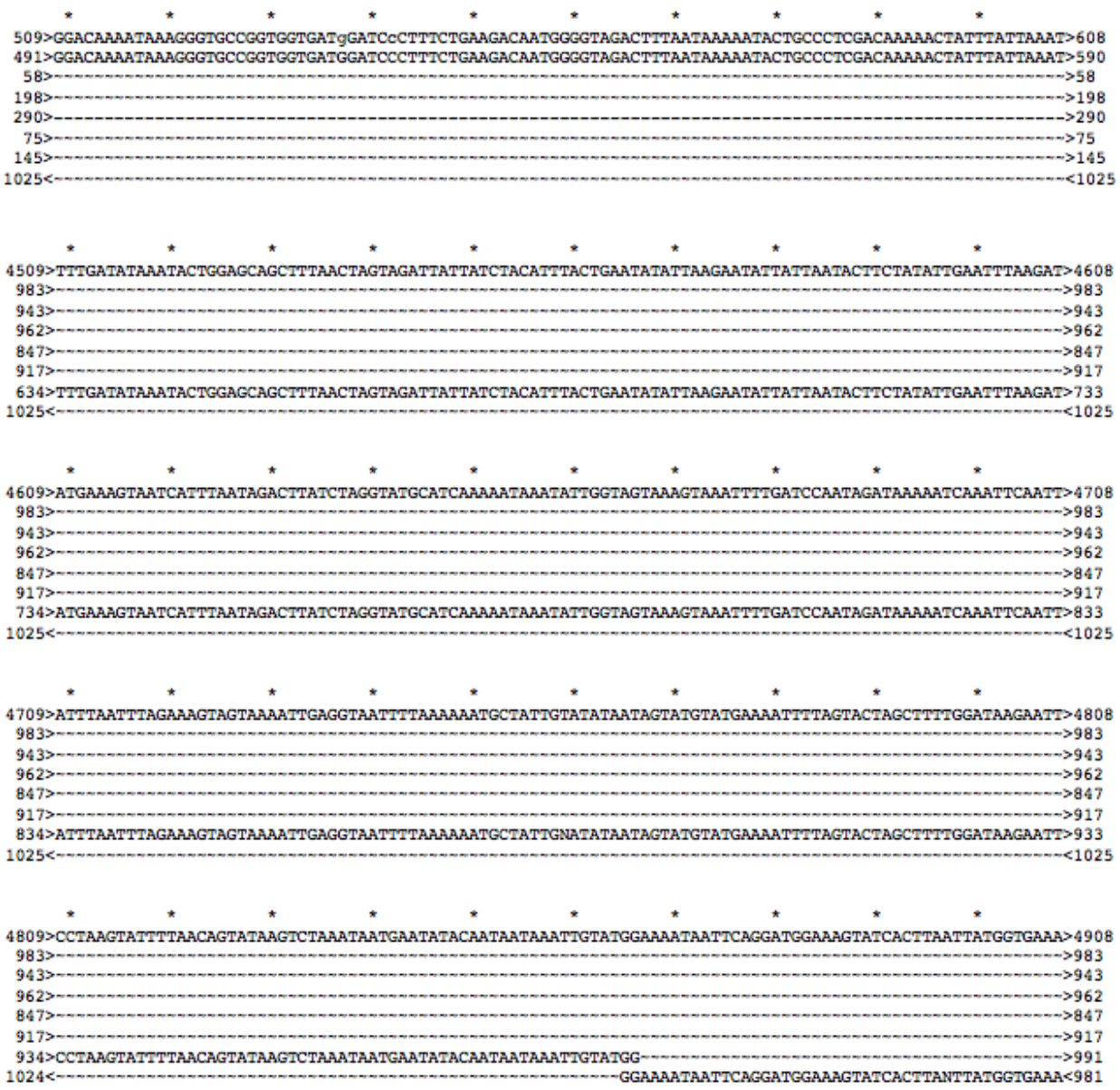
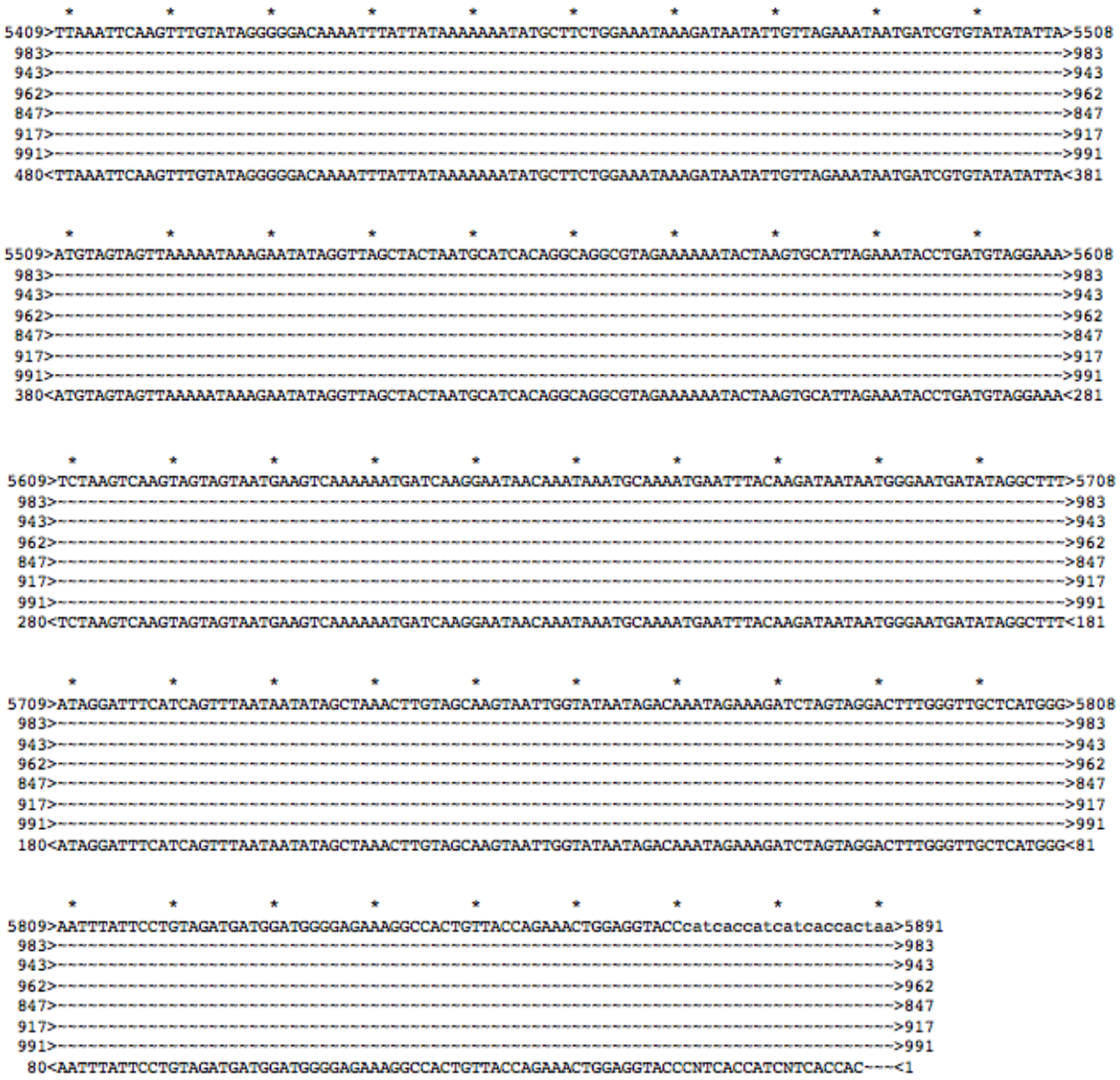

5009>CGGGCCGCCCCACAGCCCGAGCCGGAGAGGGAGCGCGAGCCGCGCCGGCCCCGGTCGGGCCTCCGAAACCATGAACTTTCTGCTGTCTTGGGTGCATTGG>5108
951>>951
964>>964
953>>953
864>>864
957>>957
892>>892
555<CGGGCCGCCCCACAGCCCGAGCCGGAGAGGGAGCGCGAGCCGCGCCGGCCCCGGTCGGGCCTCCGAAACCATGAACTTTCTGCTGTCTTGGGTGCATTGG<456
559>>559
5109>AGCCTTGCCTTGCTGCTCTACCTCCACCATGCCAAGTGGTCCCAGGCTGCACCCATGGCAGAAGGAGGAGGGCAGAATCATCACGAAGTGGTGAAGTTCA>5208
951>>951
964>>964
953>>953
864>>864
957>>957
892>>892
455<AGCCTTGCCTTGCTGCTCTACCTCCACCATGCCAAGTGGTCCCAGGCTGCACCCATGGCAGAAGGAGGAGGGCAGAATCATCACGAAGTGGTGAAGTTCA<356
559>>559
5209>TGGATGTCTATCAGCGCAGCTACTGCCATCCAATCGAGACCCTGGTGGACATCTTCCAGGAGTACCCTGATGAGATCGAGTACATCTTCAAGCCATCCTG>5308
951>>951
964>>964
953>>953
864>>864
957>>957
892>>892
355<TGGATGTCTATCAGCGCAGCTACTGCCATCCAATCGAGACCCTGGTGGACATCTTCCAGGAGTACCCTGATGAGATCGAGTACATCTTCAAGCCATCCTG<256
559>>559
5309>TGTGCCCCTGATGCGATGCGGGGGCTGCTGCAATGACGAGGGCCTGGAGTGTGTGCCCACTGAGGAGTCCAACATCACCATGCAGATTATGCGGATCAAA>5408
951>>951
964>>964
953>>953
864>>864
957>>957
892>>892
255<TGTGCCCCTGATGCGATGCGGGGGCTGCTGCAATGACGAGGGCCTGGAGTGTGTGCCCACTGAGGAGTCCAACATCACCATGCAGATTATGCGGATCAAA<156
559>>559
5409>CCTCACCAAGGCCAGCACATAGGAGAGATGAGCTTCCTACAGCACAACAAATGTGAATGCAGACCAAAGAAAGATAGAGCAAGACAAGAAAAATGTGACA>5508
951>>951
964>>964
953>>953
864>>864
957>>957
892>>892
155<CCTCACCAAGGCCAGCACATAGGAGAGATGAGCTTCCTACAGCACA<110
560>>TTCCTACAGCACAACAAATGTGAATGCAGACCAAAGAAAGATAGAGCAAGACAAGAAAAATGTGACA>626
5509>AGCCGAGGCGGTTACCAGAAACTGGAGgtacccatcaccatcatcaccactaa>5561
951>>951
964>>964
953>>953
864>>864
957>>957
892>>892
110<<110
627>AGCCGAGGCGGTTACCAGAAACTGGAGGTACCCATCACCATCATCACCACTAA>679

Figure 8-B: LAB alignment




509>GGACAAAATAAAGGGTGCCGGTGGTGATgGATCcCTTTCTGAAGACAATGGGGTAGACTTTAATAAAAATACTGCCCTCGACAAAAACTATTTATTAAAT>608
491>GGACAAAATAAAGGGTGCCGGTGGTGATGGATCCCTTTCTGAAGACAATGGGGTAGACTTTAATAAAAATACTGCCCTCGACAAAAACTATTTATTAAAT>590
58>>58
198>>198
290>>290
75>>75
145>>145
1025<<1025
4509>TTTGATATAAATACTGGAGCAGCTTTAACTAGTAGATTATTATCTACATTTACTGAATATATTAAGAATATTATTAATACTTCTATATTGAATTTAAGAT>4608
983>>983
943>>943
962>>962
847>>847
917>>917
634>TTTGATATAAATACTGGAGCAGCTTTAACTAGTAGATTATTATCTACATTTACTGAATATATTAAGAATATTATTAATACTTCTATATTGAATTTAAGAT>733
1025<<1025
4609>ATGAAAGTAATCATTTAATAGACTTATCTAGGTATGCATCAAAAATAAATATTGGTAGTAAAGTAAATTTTGATCCAATAGATAAAAATCAAATTCAATT>4708
983>>983
943>>943
962>>962
847>>847
917>>917
734>ATGAAAGTAATCATTTAATAGACTTATCTAGGTATGCATCAAAAATAAATATTGGTAGTAAAGTAAATTTTGATCCAATAGATAAAAATCAAATTCAATT>833
1025<<1025
4709>ATTTAATTTAGAAAGTAGTAAAATTGAGGTAATTTTAAAAAATGCTATTGTATATAATAGTATGTATGAAAATTTTAGTACTAGCTTTTGGATAAGAATT>4808
983>>983
943>>943
962>>962
847>>847
917>>917
834>ATTTAATTTAGAAAGTAGTAAAATTGAGGTAATTTTAAAAAATGCTATTGNATATAATAGTATGTATGAAAATTTTAGTACTAGCTTTTGGATAAGAATT>933
1025<<1025
4809>CCTAAGTATTTTAACAGTATAAGTCTAAATAATGAATATACAATAATAAATTGTATGGAAAATAATTCAGGATGGAAAGTATCACTTAATTATGGTGAAA>4908
983>>983
943>>943
962>>962
847>>847
917>>917
934>CCTAAGTATTTTAACAGTATAAGTCTAAATAATGAATATACAATAATAAATTGTATGG>991
1024<GGAAAATAATTCAGGATGGAAAGTATCACTTANTTATGGTGAAA<981

```
     *         *         *         *         *         *         *         *         *         *
4909>TAATCTGGACTTTACAGGATACTCAGGAAATAAAACAAAGAGTAGTTTTTAAATACAGTCAAATGATTAATATATCAGATTATATAAACAGATGGATTTT>5008
 983>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>983
 943>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>943
 962>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>962
 847>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>847
 917>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>917
 991>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>991
 980<TAATCTGGACTTTACAGGATACTCAGGAAATAAAACAAAGAGTAGTTTTTAAATACAGTCAAATGATTAATATATCAGATTATATAAACAGATGGATTTT<881

     *         *         *         *         *         *         *         *         *         *
5009>TGTAACTATCACTAATAATAGATTAAATAACTCTAAAATTTATATAAATGGAAGATTAATAGATCAAAAACCAATTTCAAATTTAGGTAATATTCATGCT>5108
 983>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>983
 943>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>943
 962>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>962
 847>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>847
 917>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>917
 991>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>991
 880<TGTAACTATCACTAATAATAGATTAAATAACTCTAAAATTTATATAAATGGAAGATTAATAGATCAAAAACCAATTTCAAATTTAGGTAATATTCATGCT<781

     *         *         *         *         *         *         *         *         *         *
5109>AGTAATAATATAATGTTTAAATTAGATGGTTGTAGAGATACACATAGATATATTTGGATAAAATATTTTAATCTTTTTGATAAGGAATTAAATGAAAAAG>5208
 983>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>983
 943>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>943
 962>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>962
 847>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>847
 917>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>917
 991>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>991
 780<AGTAATAATATAATGTTTAAATTAGATGGTTGTAGAGATACACATAGATATATTTGGATAAAATATTTTAATCTTTTTGATAAGGAATTAAATGAAAAAG<681

     *         *         *         *         *         *         *         *         *         *
5209>AAATCAAAGATTTATATGATAATCAATCAAATTCAGGTATTTTAAAAGACTTTTGGGGTGATTATTTACAATATGATAAACCATACTATATGTTAAATTT>5308
 983>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>983
 943>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>943
 962>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>962
 847>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>847
 917>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>917
 991>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>991
 680<AAATCAAAGATTTATATGATAATCAATCAAATTCAGGTATTTTAAAAGACTTTTGGGGTGATTATTTACAATATGATAAACCATACTATATGTTAAATTT<581

     *         *         *         *         *         *         *         *         *         *
5309>ATATGATCCAAATAAATATGTCGATGTAAATAATGTAGGTATTAGAGGTTATATGTATCTTAAAGGGCCTAGAGGTAGCGTAATGACTACAAACATTTAT>5408
 983>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>983
 943>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>943
 962>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>962
 847>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>847
 917>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>917
 991>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>991
 580<ATATGATCCAAATAAATATGTCGATGTAAATAATGTAGGTATTAGAGGTTATATGTATCTTAAAGGGCCTAGAGGTAGCGTAATGACTACAAACATTTAT<481
```


5409>TTAAATTCAAGTTTGTATAGGGGGACAAAATTTATTATAAAAAAATATGCTTCTGGAAATAAAGATAATATTGTTAGAAATAATGATCGTGTATATATTA>5508
983>>983
943>>943
962>>962
847>>847
917>>917
991>>991
480<TTAAATTCAAGTTTGTATAGGGGGACAAAATTTATTATAAAAAAATATGCTTCTGGAAATAAAGATAATATTGTTAGAAATAATGATCGTGTATATATTA<381
5509>ATGTAGTAGTTAAAAATAAAGAATATAGGTTAGCTACTAATGCATCACAGGCAGGCGTAGAAAAAATACTAAGTGCATTAGAAATACCTGATGTAGGAAA>5608
983>>983
943>>943
962>>962
847>>847
917>>917
991>>991
380<ATGTAGTAGTTAAAAATAAAGAATATAGGTTAGCTACTAATGCATCACAGGCAGGCGTAGAAAAAATACTAAGTGCATTAGAAATACCTGATGTAGGAAA<281
5609>TCTAAGTCAAGTAGTAGTAATGAAGTCAAAAAATGATCAAGGAATAACAAATAAATGCAAAATGAATTTACAAGATAATAATGGGAATGATATAGGCTTT>5708
983>>983
943>>943
962>>962
847>>847
917>>917
991>>991
280<TCTAAGTCAAGTAGTAGTAATGAAGTCAAAAAATGATCAAGGAATAACAAATAAATGCAAAATGAATTTACAAGATAATAATGGGAATGATATAGGCTTT<181
5709>ATAGGATTTCATCAGTTTAATAATATAGCTAAACTTGTAGCAAGTAATTGGTATAATAGACAAATAGAAAGATCTAGTAGGACTTTGGGTTGCTCATGGG>5808
983>>983
943>>943
962>>962
847>>847
917>>917
991>>991
180<ATAGGATTTCATCAGTTTAATAATATAGCTAAACTTGTAGCAAGTAATTGGTATAATAGACAAATAGAAAGATCTAGTAGGACTTTGGGTTGCTCATGGG<81
5809>AATTTATTCCTGTAGATGATGGATGGGGAGAAAGGCCACTGTTACCAGAAACTGGAGGTACCcatcaccatcatcaccactaa>5891
983>>983
943>>943
962>>962
847>>847
917>>917
991>>991
80<AATTTATTCCTGTAGATGATGGATGGGGAGAAAGGCCACTGTTACCAGAAACTGGAGGTACCCNTCACCATCNTCACCAC~~~<1

Appendix C: TcdC sequence alignments

Figure 1-C: TcdC^{152C} alignment

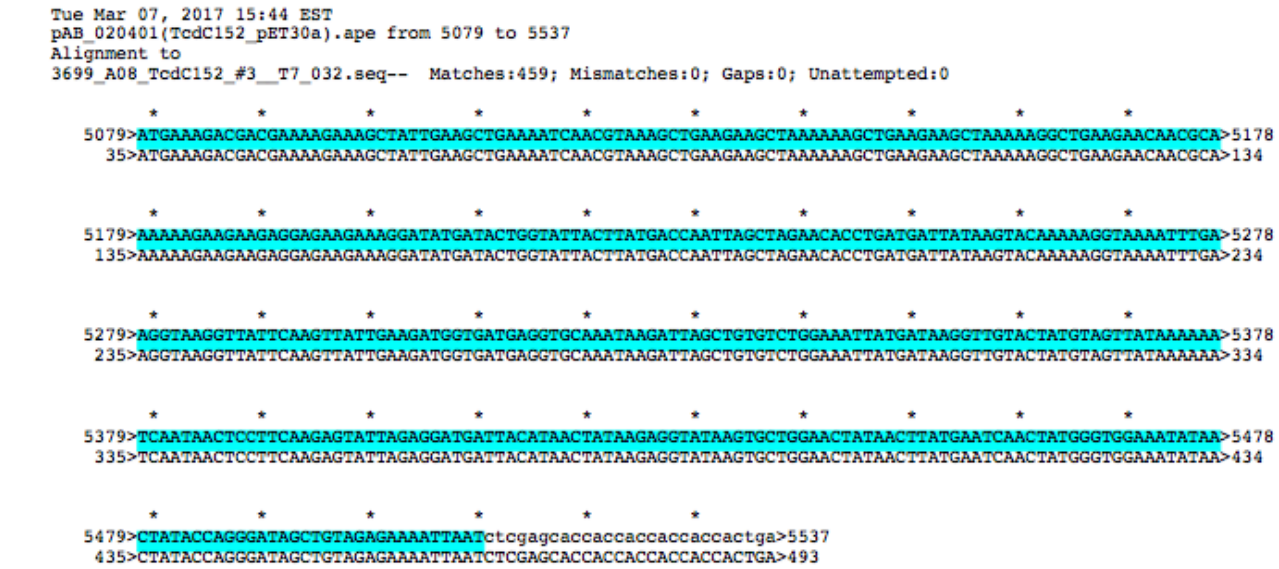
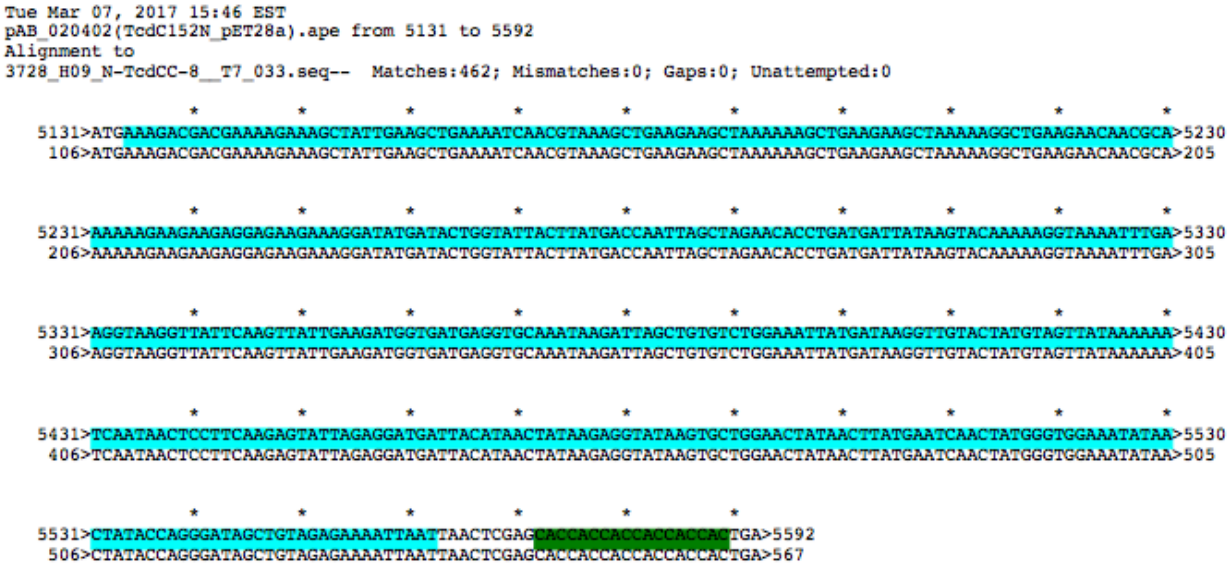


Figure 2-C: TcdC^{152N} alignment



Works Cited

1. C. Romero and S. Ventura (2010). "Educational Data Mining: A Review of the State of the Art." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40(6): 601-618.
2. R. Voorhees and J. Lee (2009). Basics of longitudinal cohort analysis, Lumina Foundation for Education.
3. D. I. Hanauer and G. Hatfull (2015). "Measuring Networking as an Outcome Variable in Undergraduate Research Experiences." CBE Life Sci Educ 14(4): ar38.
4. P. Wilcox, S. Winn and M. Fyvie‐Gauld (2005). "'It was nothing to do with the university, it was just the people': the role of social support in the first‐year experience of higher education." Studies in Higher Education 30(6): 707-722.
5. E. J. Krumrei-Mancuso, F. B. Newton, E. Kim and D. Wilcox (2013). "Psychosocial Factors Predicting First-Year College Student Success." Journal of College Student Development 54(3): 247-266.
6. J. L. Gazley, R. Remich, M. E. Naffziger-Hirsch, J. Keller, P. B. Campbell and R. McGee (2014). "Beyond Preparation: Identity, Cultural Capital, and Readiness for Graduate School in the Biomedical Sciences." J Res Sci Teach 51(8): 1021-1048.
7. B. Dietz-Uhler and J. E. Hurn (2013). "Using learning analytics to predict (and improve) student success: A faculty perspective." J Interactive Online Learning 12(1): 17-26.
8. D. Gašević, S. Dawson, T. Rogers and D. Gasevic (2016). "Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success." The Internet and Higher Education 28: 68-84.

9. L. L. Baer and D. M. Norris (2016). "A call to action for student success analytics." Planning Higher Education Journal: 1-11.
10. S. Slade and P. Prinsloo (2013). "Learning analytics: Ethical issues and dilemmas." American Behavioral Scientist 57(10): 1510-1529.
11. R. Alamuddin, J. Brown and M. Kurzweil (2016). Student data in the digital era: an overview of current practices, Ithaka S+R.
12. A. M. Shahiri, W. Husain and N. a. A. Rashid (2015). "A Review on Predicting Student's Performance Using Data Mining Techniques." Procedia Computer Science 72: 414-422.
13. J. Greer, M. Molinaro, X. Ochoa and T. McKay (2016). Proceedings of the 1st learning analytics for curriculum and program quality improvment workshop, Edinburgh, UK.
14. M. Lovett, O. Meyer and C. Thille (2008). "The open learning initiative: measuring the effectiveness of the OLI statistics course in Accelerating Student Learning." Journ of Interactive Media in Education 13: 1-16.
15. D. Gillborn (2010). "The colour of numbers: surveys, statistics and deficit‐thinking about race and class." Journal of Education Policy 25(2): 253-276.
16. S. B. Garcia and P. L. Guerra (2004). "Deconstructing deficit thinking." Education and Urban Society 36(2): 150-168.
17. J. W. Schofield (2010). "International evidence on ability grouping with curriculum differentiation and the achievement gap in secondary schools." Teachers Collge Record 112(5): 1492-1528.

18. S. E. Bradforth, E. R. Miller, W. R. Dichtel, A. K. Leibovich, A. L. Feig, J. D. Martin, K. S. Bjorkman, Z. D. Schultz and T. L. Smith (2015). "University Learning: Improve undergraduate science education." Nature 523: 282-284.

19. . "Tools for evidence-based action." Retrieved 11/8/16, from http://t4eba.com.

20. M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones and M. A. Marra (2009). "Circos: an information aesthetic for comparative genomics." Genome Res 19(9): 1639-1645.

21. S. Claro, D. Paunesku and C. S. Dweck (2016). "Growth mindset tempers the effects of poverty on academic achievement." Proc Natl Acad Sci U S A 113(31): 8664-8668.

22. C. Alexander, E. Chen and K. Grumback (2009). "How leaky is the health career pipeline? Minority student achievement in college gateway courses." Academic Medicine 84(9): 797-802.

23. L. Gregg-Jolly, J. Swartz, E. Iverson, J. Stern, N. Brown and D. Lopatto (2016). "Situating second-year success: understanding second-year STEM experiences at a liberal arts college." CBE Life Sci Educ 15: 1-10.

24. C. C. Archer and M. K. Miller (2011). "Prioritizing Active Learning: An Exploration of Gateway Courses in Political Science." PS: Political Science & Politics 44(02): 429-434.

25. J. Gainen (1995). "Barriers to success in quantitative gatekeeper courses." New Directions Teaching Learning 41: 5-14.

26. J. W. Dubendorff and F. W. Studier (1991). "Controlling basal expression in an inducible T7 expression system by blocking the target T7 promoter with *lac* repressor." J Mol Biol 219: 45-59.

27. M. Gossen and H. Bujard (1992). "Tight Control of gene expression in mammallian cells by tetracycline-responsive promoters." Proc Natl Acad Sci U S A 89: 5547-5551.

28. K. Hammer, I. Mijakovic and P. R. Jensen (2006). "Synthetic promoter libraries--tuning of gene expression." Trends Biotechnol 24(2): 53-55.

29. S. W. Seo, J. S. Yang, I. Kim, J. Yang, B. E. Min, S. Kim and G. Y. Jung (2013). "Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency." Metab Eng 15: 67-74.

30. H. Brooks, B. Lebleu and E. Vives (2005). "Tat peptide-mediated cellular delivery: back to basics." Adv Drug Deliv Rev 57(4): 559-577.

31. J. L. Zaro, J. E. Vekich, T. Tran and W. C. Shen (2009). "Nuclear localization of cell-penetrating peptides is dependent on endocytosis rather than cytosolic delivery in CHO cells." Mol Pharm 6(2): 337-344.

32. J. P. Richard, K. Melikov, E. Vives, C. Ramos, B. Verbeure, M. J. Gait, L. V. Chernomordik and B. Lebleu (2003). "Cell-penetrating peptides. A reevaluation of the mechanism of cellular uptake." J Biol Chem 278(1): 585-590.

33. S. W. Jones, R. Christison, K. Bundell, C. J. Voyce, S. M. Brockbank, P. Newham and M. A. Lindsay (2005). "Characterisation of cell-penetrating peptide-mediated peptide delivery." Br J Pharmacol 145(8): 1093-1102.

34. S. El-Andaloussi, P. Jarver, H. J. Johansson and U. Langel (2007). "Cargo-dependent cytotoxicity and delivery efficacy of cell-penetrating peptides: a comparative study." Biochem J 407(2): 285-292.

35. A. K. Cardozo, V. Buchillier, M. Mathieu, J. Chen, F. Ortis, L. Ladriere, N. Allaman-Pillet, O. Poirot, S. Kellenberger, J. S. Beckmann, D. L. Eizirik, C. Bonny and F. Maurer

(2007). "Cell-permeable peptides induce dose- and length-dependent cytotoxic effects." Biochim Biophys Acta 1768(9): 2222-2234.

36. D. E. Voth and J. D. Ballard (2005). "Clostridium difficile toxins: mechanism of action and role in disease." Clin Microbiol Rev 18(2): 247-263.

37. D. J. Reinert, T. Jank, K. Aktories and G. E. Schulz (2005). "Structural basis for the function of Clostridium difficile toxin B." J Mol Biol 351(5): 973-981.

38. P. Papatheodorou, C. Zamboglou, S. Genisyuerek, G. Guttenberg and K. Aktories (2010). "Clostridial glucosylating toxins enter cells via clathrin-mediated endocytosis." PLoS ONE 5(5): e10673.

39. A. Greco, J. G. Ho, S. J. Lin, M. M. Palcic, M. Rupnik and K. K. Ng (2006). "Carbohydrate recognition by Clostridium difficile toxin A." Nat Struct Mol Biol 13(5): 460-461.

40. J. G. Ho, A. Greco, M. Rupnik and K. K. Ng (2005). "Crystal structure of receptor-binding C-terminal repeats from Clostridium difficile toxin A." Proc Natl Acad Sci U S A 102(51): 18373-18378.

41. Z. Zhang, M. Park, J. Tam, A. Auger, G. L. Beilhartz, D. B. Lacy and R. A. Melnyk (2014). "Translocation domain mutations affecting cellular toxicity identify the Clostridium difficile toxin B pore." Proc Natl Acad Sci U S A 111(10): 3721-3726.

42. N. M. Chumbler, S. A. Rutherford, Z. Zhang, M. A. Farrow, J. P. Lisher, E. Farquhar, D. P. Giedroc, B. W. Spiller, R. A. Melnyk and D. B. Lacy (2016). "Crystal structure of *Clostridium difficile* toxin A." Nature Microbiology 1: 1-6.

43. A. Shen, P. J. Lupardus, M. M. Gersch, A. W. Puri, V. E. Albrow, K. C. Garcia and M. Bogyo (2011). "Defining an allosteric circuit in the cysteine protease domain of Clostridium difficile toxins." Nat Struct Mol Biol 18(3): 364-371.

44. I. Kreimeyer, F. Euler, A. Marckscheffel, H. Tatge, A. Pich, A. Olling, J. Schwarz, I. Just and R. Gerhard (2011). "Autoproteolytic cleavage mediates cytotoxicity of Clostridium difficile toxin A." Naunyn Schmiedebergs Arch Pharmacol 383(3): 253-262.

45. S. Nottrott, J. Schoentaube, H. Genth, I. Just and R. Gerhard (2007). "Clostridium difficile toxin A-induced apoptosis is p53-independent but depends on glucosylation of Rho GTPases." Apoptosis 12(8): 1443-1453.

46. S. M. Kern and A. L. Feig (2011). "Adaptation of Clostridium difficile toxin A for use as a protein translocation system." Biochem Biophys Res Commun 405(4): 570-574.

47. G. Krautz-Peterson, Y. Zhang, K. Chen, G. A. Oyler, H. Feng and C. B. Shoemaker (2012). "Retargeting Clostridium difficile Toxin B to Neuronal Cells as a Potential Vehicle for Cytosolic Delivery of Therapeutic Biomolecules to Treat Botulism." J Toxicol 2012: 760142.

48. J. Yin, A. J. Lin, D. E. Golan and C. T. Walsh (2006). "Site-specific protein labeling by Sfp phosphopantetheinyl transferase." Nat Protoc 1(1): 280-285.

49. J. Yin, P. D. Straight, S. M. McLoughlin, Z. Zhou, A. J. Lin, D. E. Golan, N. L. Kelleher, R. Kolter and C. T. Walsh (2005). "Genetically encoded short peptide tag for versatile protein labeling by Sfp phosphopantetheinyl transferase." Proc Natl Acad Sci U S A 102(44): 15815-15820.

50. D. A. Levary, R. Parthasarathy, E. T. Boder and M. E. Ackerman (2011). "Protein-protein fusion catalyzed by sortase A." PLoS ONE 6: 1-6.

51. B. Bunk, A. Schultz, S. Stammen, R. Muench, M. J. Warren, M. Rohde, D. Jahn and R. Biedendieck (2010). "A short story about a big magic bug." Bioeng Bugs 1: 85-91.

52. G. Hadlaczky, K. Fodor and L. Alfoldi (1976). "Morphological study of the revrsion to bacillary form of *Bacillus megaterium* protoplasts." Journal of Bacteriology 125: 1172-1179.

53. D. J. Leibly, T. N. Nguyen, L. T. Kao, S. N. Hewitt, L. K. Barrett and W. C. Van Voorhis (2012). "Stabilizing additives added during cell lysis aid in the solubilization of recombinant proteins." PLoS ONE 7(12): e52482.

54. T. Doerks, R. R. Copley, J. Schultz, C. P. Ponting and P. Bork (2002). "Systematic identification of novel protein domain families associated with nuclear functions." Genome Res 12(1): 47-56.

55. C. Gustafsson, S. Govindarajan and J. Minshull (2004). "Codon bias and heterologous protein expression." Trends Biotechnol 22(7): 346-353.

56. D. B. Goodman, G. M. Church and S. Kosuri (2013). "Causes and effects of N-terminal codon bias in bacterial genes." Science 342(6157): 475-479.

57. F. C. Lessa, Y. Mu, W. M. Bamberg, Z. G. Beldavs, G. K. Dumyati, J. R. Dunn, M. M. Farley, S. M. Holzbauer, J. I. Meek, E. C. Phipps, L. E. Wilson, L. G. Winston, J. A. Cohen, B. M. Limabago, S. K. Fridkin, D. N. Gerding and L. C. McDonald (2015). "Burden of *Clostridium difficile* infection in the United States." N Engl J Med 372: 825-834.

58. B. Dupuy, R. Govind, A. Antunes and S. Matamouros (2008). "Clostridium difficile toxin synthesis is negatively regulated by TcdC." J Med Microbiol 57(Pt 6): 685-689.

59. N. Mani, D. Lyras, L. Barroso, P. Howarth, T. Wilkins, J. I. Rood, A. L. Sonenshein and B. Dupuy (2002). "Environmental Response and Autoregulation of Clostridium difficile TxeR, a Sigma Factor for Toxin Gene Expression." Journal of Bacteriology 184(21): 5971-5978.

60. S. S. Dineen, S. M. McBride and A. L. Sonenshein (2010). "Integration of metabolism and virulence by Clostridium difficile CodY." J Bacteriol 192(20): 5350-5362.

61. T. Dubois, M. Dancer-Thibonnier, M. Monot, A. Hamiot, L. Bouillaut, O. Soutourina, I. Martin-Verstraete and B. Dupuy (2016). "Control of Clostridium difficile Physiopathology in Response to Cysteine Availability." Infect Immun 84(8): 2389-2405.

62. N. Mani and B. Dupuy (2001). "Regulation of toxin synthesis in Clostridium difficile by an alternative RNA polymerase sigma factor." Proc Natl Acad Sci U S A 98(10): 5844-5849.

63. S. Matamouros, P. England and B. Dupuy (2007). "Clostridium difficile toxin expression is inhibited by the novel regulator TcdC." Mol Microbiol 64(5): 1274-1288.

64. G. P. Carter, G. R. Douce, R. Govind, P. M. Howarth, K. E. Mackin, J. Spencer, A. M. Buckley, A. Antunes, D. Kotsanas, G. A. Jenkin, B. Dupuy, J. I. Rood and D. Lyras (2011). "The anti-sigma factor TcdC modulates hypervirulence in an epidemic BI/NAP1/027 clinical isolate of Clostridium difficile." PLoS Pathog 7(10): e1002317.

65. R. Murray, D. Boyd, P. N. Levett, M. R. Mulvey and M. J. Alfa (2009). "Truncation in the tcdC region of the Clostridium difficile PathLoc of clinical isolates does not predict increased biological activity of Toxin B or Toxin A." BMC Infect Dis 9: 103.

66. D. Bakker, W. K. Smits, E. J. Kuijper and J. Corver (2012). "TcdC does not significantly repress toxin expression in Clostridium difficile 630DeltaErm." PLoS ONE 7(8): e43247.

67. H. C. van Leeuwen, D. Bakker, P. Steindel, E. J. Kuijper and J. Corver (2013). "Clostridium difficile TcdC protein binds four-stranded G-quadruplex structures." Nucleic Acids Res 41(4): 2382-2393.

68. V. Brazda, L. Haronikova, J. C. Liao and M. Fojta (2014). "DNA and RNA quadruplex-binding proteins." Int J Mol Sci 15(10): 17493-17517.

69. N. Maizels and L. T. Gray (2013). "The G4 genome." PLoS Genet 9(4): e1003468.

70. A. Cammas and S. Millevoi (2016). "RNA G-quadruplexes: emerging mechanisms in disease." Nucleic Acids Res 45(4): 1584-1595.

71. D. R. Knight, B. Elliott, B. J. Chang, T. T. Perkins and T. V. Riley (2015). "Diversity and Evolution in the Genome of Clostridium difficile." Clin Microbiol Rev 28(3): 721-741.

72. F. Lassalle, S. Perian, T. Bataillon, X. Nesme, L. Duret and V. Daubin (2015). "GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands." PLoS Genet 11(2): e1004941.

73. R. A. Becker, J. M. Chambers and A. R. Wilks (1988). The new S language, Wadsworth & Brooks/Cole.

74. M. S. Brouwer, P. J. Warburton, A. P. Roberts, P. Mullany and E. Allan (2011). "Genetic organisation, mobility and predicted functions of genes on integrated, mobile genetic elements in sequenced strains of Clostridium difficile." PLoS ONE 6(8): e23014.

75. M. Sebaihia, B. W. Wren, P. Mullany, N. F. Fairweather, N. Minton, R. Stabler, N. R. Thomson, A. P. Roberts, A. M. Cerdeno-Tarraga, H. Wang, M. T. Holden, A. Wright, C. Churcher, M. A. Quail, S. Baker, N. Bason, K. Brooks, T. Chillingworth, A. Cronin, P. Davis, L. Dowd, A. Fraser, T. Feltwell, Z. Hance, S. Holroyd, K. Jagels, S. Moule, K. Mungall, C. Price, E. Rabbinowitsch, S. Sharp, M. Simmonds, K. Stevens, L. Unwin, S.

Whithead, B. Dupuy, G. Dougan, B. Barrell and J. Parkhill (2006). "The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome." Nat Genet 38(7): 779-786.

76. J. Corver, D. Bakker, M. S. Brouwer, C. Harmanus, M. P. Hensgens, A. P. Roberts, L. Lipman, E. Kuijper and H. C. van Leeuwen (2012). "Analysis of a *Clostridium difficile* PCR ribotype 078 100 kilobase island reveals the presence of a novel transposon, Tn*6164*." BMC Microbiol 12(130): 1-13.
77. J. Amy, P. Johanesen and D. Lyras (2015). "Extrachromosomal and integrated genetic elements in Clostridium difficile." Plasmid 80: 97-110.
78. Q. Yang, J. Xiang, S. Yang, Q. Zhou, Q. Li, Y. Tang and G. Xu (2009). "Verification of specific G-quadruplex structure by using a novel cyanine dye supramolecular assembly: I. recognizing mixed G-quadruplex in human telomeres." Chem Commun (Camb)(9): 1103-1105.
79. Q. Yang, J. Xiang, S. Yang, Q. Li, Q. Zhou, A. Guan, X. Zhang, H. Zhang, Y. Tang and G. Xu (2010). "Verification of specific G-quadruplex structure by using a novel cyanine dye supramolecular assembly: II. The binding characterization with specific intramolecular G-quadruplex and the recognizing mechanism." Nucleic Acids Res 38(3): 1022-1033.

Abstract

DEVELOPMENT OF STUDENT DATA VISUALIZATION TOOL, ADAPTION OF *CLOSTRIDIUM DIFFICILE* TOXIN A INTO PROTEIN DELIVERY VEHICLE, AND ELUCIDATION OF TCDC MECHANISM OF TOXIN CONTROL

ADAM BOYDEN

August 2017

Dr. Andrew Feig, Chemistry (Biochemistry), Masters

Advancing student success in higher education is of paramount importance, and is in need for a tool that visualizes student data in a longitudinal manner. Student Circos plots achieve this by plotting student data in circular plots, depicting the timeline and grades for students selected by demographic or performance information. Cellular delivery of exogenous proteins is a bountiful area of research. However, most current systems have limited *in vivo* applications and most lack cellular specificity. By adapting Toxin A from *Clostridium difficile,* the goal was to create a cell specific protein delivery vehicle that would be robust *in vivo*. However, the chimeric constructs produced were unable to be isolated for study. Control of Toxin A and B in *C. difficile* has been linked to the protein TcdC. However , no clear mechanism has been developed and there is debate on whether TcdC truly plays a role in toxin production. The goal of this project was to identify DNA or RNA molecules within *C. difficile* that could behave as a protein sink, binding TcdC and preventing the native behavior. Recreation of TcdC binding DNA molecules that adopt a G-quartet structure was not successful and further analyses was not carried out.