Data Exploration, Preparation, and Pilot Studies for Building a Knowledge Model of the Cayo Santiago Rhesus Monkeys

Martin Q. Zhao
Department of Computer Science
Mercer University
Macon, GA, USA
zhao_mq@mercer.edu

Mehakpreet Kaur
Department of Computer Science
Mercer University
Macon, GA, USA
kaur_m@mercer.edu

Soumik Kundu
School of Engineering
Mercer University
Macon, GA, USA
soumik.kundu@live.mercer.edu

Qian Wang
Dept of Biomedical Sciences
TAMU School of Dentistry
Dallas, TX, USA
qian.wang@tamu.edu

***Abstract*--The Cayo Santiago rhesus, established and maintained for 85 years, has evolved into a valuable resource for researchers across various disciplines. This research paper outlines an ongoing NSF project aimed at developing a comprehensive database and user-friendly software application, CSViewer, to uncover hidden knowledge. Using a big data approach, the paper focuses on key events in the colony's population dynamics, emphasizing gender-specific analyses. It also explores data exploration and preparation processes, along with the application of the genealogy model in inbreeding analysis and genetic tracing. Future efforts, including the expansion of CSViewer's functions, are also addressed.**

Keywords—Cayo Santiago Rhesus Monkey Colony, CSViewer for analysts, information integration, big data approach, female vs male reproduction patterns

I. INTRODUCTION

The Cayo Santiago rhesus macaque colony has been kept for over eight decades and has become an invaluable asset within the realms of biomedical and anthropological sciences. Its potential for informative contributions is increasingly evident with the application of data science technologies ([8], [10], [13], [14]). Armed with computer science and big data techniques, an ongoing NSF project aimed at building a Knowledge Model that (1) integrates related data on different aspects into a comprehensive database, and (2) develops a software application, CSViewer for Analysts to provide user friendly interfaces and appropriate data analytical tools, and (3) helps uncover hidden knowledge ([10], [11]). This paper adopts a big data approach to provide a retrospective view for selected key events of the CS colony, such as population control, social group evolution, and reproduction. Gender-specific analyses are carried out wherever possible to explain sex-related differences. After introducing various sources of data, essential data exploration and preparation efforts are discussed. The paper then demonstrates how specific user cases involving assessing reproductive patterns may be supported and visualized using CSViewer, as well as how the genealogy model as contained in the Knowledge Model can be used in inbreeding analysis and genetic tracing. Future initiatives are addressed as well, including the possibility of adding pertinent features to the CSViewer app.

II. DATA COLLECTION AND MANAGEMENT NEEDS

A. Data Sources and Collection Purposes

The Cayo Santiago (CS) rhesus macaque (*Macaca mulatta*) colony has been kept by the Caribbean Primate Research Center (CPRC) since 1950. Matrilineal pedigree and specific details about each animal have been meticulously kept. Additionally, CPRC has curated s Skelton for thousands of CS animals. The purpose of this NSF project is to build a non-human primate model for the studies of human conditions such as aging, the impact of genetic and environmental factors on health and diseases, the impact of natural disasters and resiliency, based mainly on skeletal collections. A background for the study based on skeleton collections is first established via a consolidation of census data. This study only uses data from field census records, but it highlights how computer science and big data techniques can be used to creatively examine topics like population dynamics, mating behavior, and inbreeding trends.

B. The CSViewer App for Data Access and Analytical Support

To effectively manage the comprehensive data collection described above, a relational data model, as outlined in [10] was introduced. Subsequently, experimental implementations were developed using SQL Server within both local network and AWS cloud settings. Additional NoSQL databases are also being researched for storing textual data for the CPRC skeletal catalog, deriving images and measuring data.

A Java-based application, CSViewer for Analysts has been developed [8] to provide interactive user interfaces to allow researchers to select and analyze the data and visualize their results. A researcher can create a research "project" that includes several models, each of which focuses on an aspect of the family tree, Measure or Image of bone specimens cataloged at the CPRC museum, and related Analytical tasks, using the functions made possible by its menus (as shown in Fig. 1). A project and its models can be saved as files for later use.

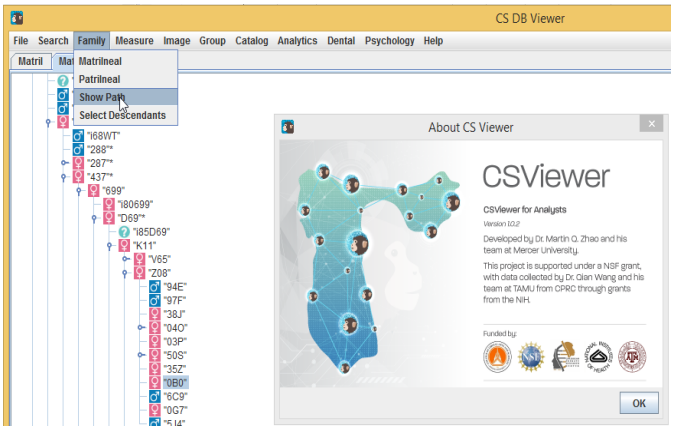

Figure 1-Menu Structure and the Main Panel of the CSViewer App

The main window of the CSViewer app displays family trees or analytical artifacts (like charts or life table) in its main panel. Summary information of an animal selected, as well as bone measures and photos as available will be displayed in several additional panels.

C. *Benefits of using the Knowledge Model*

In this project, we are aiming to build a knowledge model, which consists of three key components as described earlier. The benefits can be summarized in the following aspects:

1) Data/knowledge asset: In addition to a comprehensive database of original data from various sources, it also provides a fundamental genealogy model which can be cross-referenced to other data sources such as measures and photos of skeleton or dental specimens. The genealogical framework can be presented in multiple formats, including matrilineal, patrilineal, or kinship tree structures, facilitating genetic tracing.

2) User Experience: Accessing animal data and related measure data are supported through intuitive menu structures, enhancing user experience. In a single setting, this simplified interface smoothly transitions from operations involving data exploration, selection, and visualization to those involving modeling and analyses.

3) Modeling and genetic tracing: This platform, which emphasizes the necessity of genealogical models and cross-referencing, equips researchers to carry out basic analytical activities easily, along with visualization of results. The system also enables accurate custom analysis with CSViewer or the generation of datasets in standard formats (such as .csv or .json) for export to the user's choice analytical tool.

The following sections will discuss data processing and results in more detail.

III. Data Exploration and Preparation

As introduced in the previous section, the CS colony provides abundant data assets which will be included in the allocated database for the target Knowledge Model. This paper will focus predominantly on conducting analytical investigations related to the population dynamics and reproduction patterns within the CS colony. Therefore, only the work relating to the censor datasets from 2013 and 2020 [20] will be described. Data preparation efforts regarding other datasets will be briefly mentioned in the section V-B.

A. *Merging Censor Datasets to Generate a Complete Animal List*

The genealogical data, sourced from [20], is provided in the form of two distinct spreadsheets. The initial spreadsheet, referred to as the 2013 censor set, enumerates 9611 animals born from 1955 onwards, extending up to 2013. Notably, this dataset lacks any sire information for its entries. Conversely, the 2020 set encompasses 5388 animals born between 1966 and 2020, with most entries containing meticulously tracked sire data.

To manage and consolidate this data effectively, the Microsoft Access application is used to create a roster of distinct animals. Initially, the 2020 dataset is loaded into a table to ensure that the most current data is retained for each entry. During this process, when an attempt is made to insert an animal entry from the 2013 censor set that shares a tattoo number with an existing entry in the table, it is prevented from entry due to a UNIQUE key violation. Consequently, only 10,949 distinct animal entries are identified and integrated into the comprehensive list.

Each CS animal is given a unique tattoo number that is cast onto its body and utilized as its identification when compiling the complete list as previously outlined. The tattoo numbers are nonetheless classified as private information (like the names of the animals) and cannot be used in publications, according to the MOU signed with CPRC. To prevent privacy invasion in reports and documents, a "Unicode" scheme has been developed to replace tattoos.

A Unicode is formatted as "L-###", with a capital letter starting from "A" and three digits starting from 000. It follows a similar format to the 77 founders' code, which is "Q####": "Q" followed by four numerals. Between "A-000" and "P-999," which can be continued starting with "R-000," this method offers 16000 different, unique codes. Considering this, it offers enough space for all CS animals to be added in the future.

The CSViewer app will provide cross-referencing between Unicode and tattoo numbers as well as other data source identifiers (such as "catalog" numbers for the skeleton specimens kept in the CPRC museum's collection). Using Unicode in publications (as in [3]) met the requirements for referencing to specific animals under discussion while still maintaining animal privacy.

B. *Headcount & Gender Ratio*

We have performed standard exploration tasks using the compiled comprehensive animal list. In preparation for the discussions on gender-specific group interactions and reproductive studies in this paper, we present here a selection of pertinent statistics to set the stage.

It is important to note that the animal genders are labeled as "F", "M", and "U" for female, male, and unknown (often undetermined due to stillbirth/miscarriage or dead/removed shortly after birth). One animal's label, which uses the Unicode

Supported by NSF grants to M.Q.Z. and Q.W. (#1926402, #1926601).

character L-393, was however left empty; then changed to label "U".

TABLE I. OVERALL HEADCOUNT BY GENDER

Gender	*Female*	*Male*	*Unknown*	Total
Count	5271	5524	153	10948*

With a slight male preponderance, the total female-to-male ratio is 48.8 to 51.2 (based on 10795 respondents with known gender). Only emigrants or removals of CS animals were made for scientific purposes. The gender ratio fluctuates as a result of natural births, deaths, and removals [16][17] for population control. The gender ratio has been kept balanced (between 40% and 55%), as can be seen in Fig. 2, which is based on animals born between 1955 and 2018 with a known gender label ("F" or "M", totaling 10795).

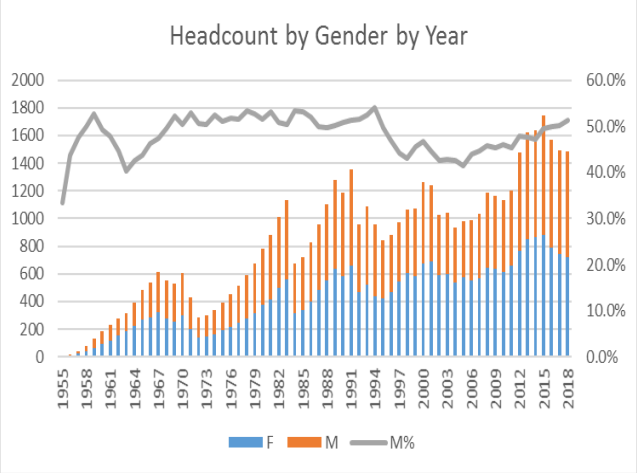

Fig. 2. CS Population and Gender Ratio By Gender

C. *Gender-Specific Life Expectancy*

To create gender-specific life tables, 10795 animals with gender labels were used. For male and female animals, respectively, E.x_m and E.x_f represent estimated life span at age x. Ages at the time of death or removal are taken from the censor datasets.

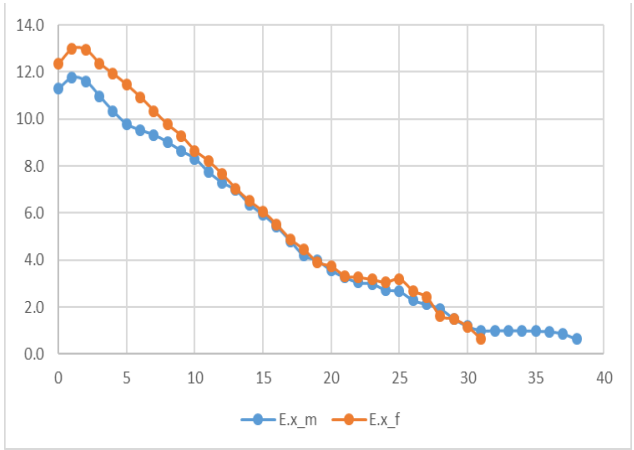

Fig. 3. Gender-specific Life Expetancy for CS Population

Females are anticipated to live longer from birth to age 30, as seen in Fig. 3. The longest age at date of death (Age at DOD) for females was 31, compared to one male individual who lived to the age of 38. As a result, the life expectancy curve was longer for males.

The three-part equation below can be used to predict the instantaneous mortality rate [18], which is another way to compare life expectancy:

$$m_t = a_1 e^{-b_1 t} + a_2 + a_3 e^{b_3 t} \quad (1)$$

Table 2 displays gender-specific coefficients, while Figure 4 displays death rates and fitting curves. The constant a_2 for females is marginally greater than that for males (0.0736 vs. 0.0660), indicating that a female has a somewhat higher chance of dying in middle age than her male counterpart does. However, because the mortality curve (in Fig. 4) rises more slowly for females, they often live longer. (The outlier male subject lived to 38 years was excluded in this curve fitting case.)

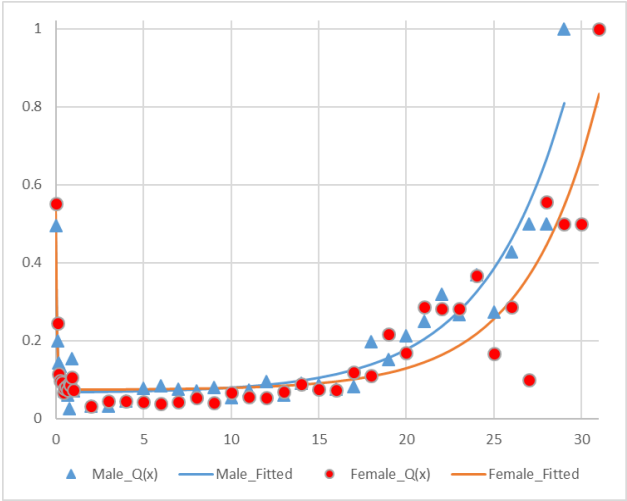

Fig. 4. Mortality Rate Fitting by Gender

D. *Matrilineal and Patrilineal Family Trees*

Matrilineal lineage from the 77 founders is traced from the censor datasets, through the "behavior mom" recorded for each animal. Since the adoption of DNA testing in the mid-1970s, "Dam Genetics" can be used to trace the "biological mother" for each subject. Fig. 5 shows the trend of DNA testing in the CS colony.

TABLE II. OVERALL HEADCOUNT BY GENDER

Coefficients	*Male*	*Female*
a_1	0.4176	0.4781
b_1	11.1388	13.12966
a_2	0.0660	0.0736
a_3	0.0017	0.0005
b_3	0.2105	0.2333

It is important to note that there are only 58 animals, as proven by, whose "behavior mom" is different. In cases where

the dam's DNA is known, the biological mother is used to create the matrilineal family tree.

Matrilineal and patrilineal family trees are built using Java programs, as a part of the CSViewer for Analysts app [8]. To initially arrange animal entries based on founder lineage and pedigree information, siblings for each mother are sequenced using data structures using the Java Collections Framework API. Sire DNA is then used to build patrilineal trees starting from the earliest known fathers.

Kinship trees that combine ancestors (including both sire and dam information when available) and descendants can be built when genetic studies (such as in [3]) or inbreeding analysis (to be discussed in section IV.C) are needed.

The interactive tree structure in CSViewer for Analysts, which is developed with the Java JTree API standard, is a highly helpful tool that allows users to visualize matrilineal and patrilineal trees by expanding or contracting tree nodes. Kinship trees can be generated and edited using a separate dialog box (popped up from the CSViewer main window), which is developed using a simple graph framework [2] used in similar applications. [9]

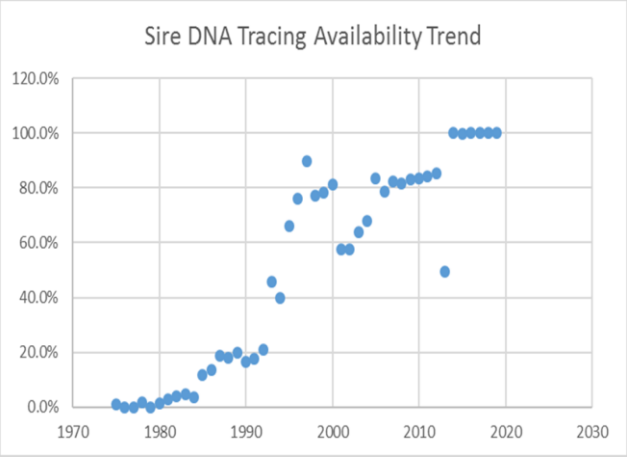

Fig. 5. Sire/Dam DNA Tracing Trend

In the next section, we will discuss how gender-specific reproduction analytics can be facilitated with the comprehensive database. Metrics that are conventionally calculated using solely matrilineal lineage will be expanded to include their patrilineal counterparts owing to the database's sire genetics. The emphasis is on the different kinds of analysis that may be performed with the Knowledge Model that is being built and the ways that the CSViewer app can help analytics.

IV. Reproduction Specific Analytics

Reproduction patterns and metrics have been studied as a part of data exploration effort. Although results related to the CS colony can be found in literature, it is considered beneficial to include them here in the Knowledge Model for easy access by its users. The CSViewer app will provide meaningful ways to allow users to select their dataset of interest and get project-specific results. The focus in this section is to show some initial trials and the way they can be supported in the CSViewer app. We will seek feedback from the user communities to find directions for prioritizing our efforts of including useful features into the system.

Conventionally, reproduction metrics are calculated for matrilineal lineage since that had been the only data available. We are extending the linked concepts to patrilineal counterparts using the subset of 4807 animals with both Dam and Sire Genetics accessible, with 949 different mothers and 610 different fathers involved. DNA tracing has been done on the CS population since 1975.

A. Male and Female Reproduction Curves

It can be seen in Fig 6, females reach sexual maturity at age 3 (with one subject at age 2), whereas males typically start at 5. The peak for female reproduction is between age 4 and 7, where the percentage is above 10%; and the equivalent range for males is between 6 and 10 years old. Both female and male stop reproducing around 25 years of age.

Since there was only one twin birth in the subset, a female has a maximum of 17 offspring, compared to a male's maximum of 53.

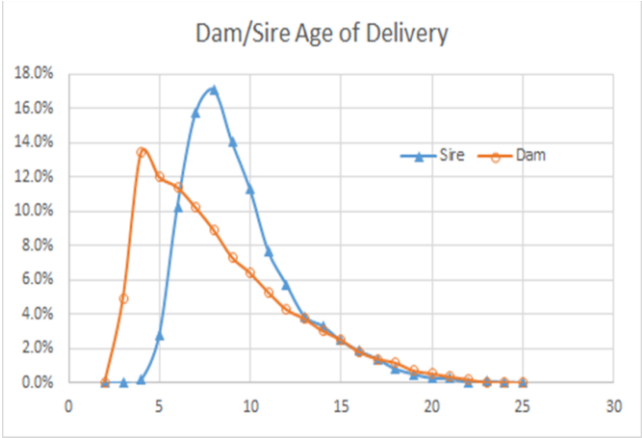

Fig. 6. Male and Female Reproduction Curves

B. Key Reproduction Metrics

In reproductive studies [1][7], the parameters of interest often focus on females. In this context, the reproductive coefficient R0 represents the average number of daughters a female produces throughout her life span, while the generation gap G is the average time interval in years between the mother and her daughters. However, when we extend our analysis to incorporate patrilineal genetics, R0 and G transition to signify the number of sons a male produces and the average age difference between a father and his sons, respectively. The outcomes for the specified CS subset are presented in Table III for reference.

TABLE III. Reproduction Metrics by Gender

Parent Gender	Dam		Sire	
Child Gender	Daughter	Son	Daughter	Son
R0	4.685	5.909	4.768	4.229
G	10.192	-	-	10.949

The greater G value based on sire-son data may be attributed to the fact that male rhesus matures later than female rhesus.

Figure 7 illustrates how both the dam and sire are more likely to have a son than a daughter whether they are either extremely young or very old. Throughout her lifetime, a mother typically produces 1.2 more sons than daughters, but a father typically has 0.5 fewer sons than daughters.

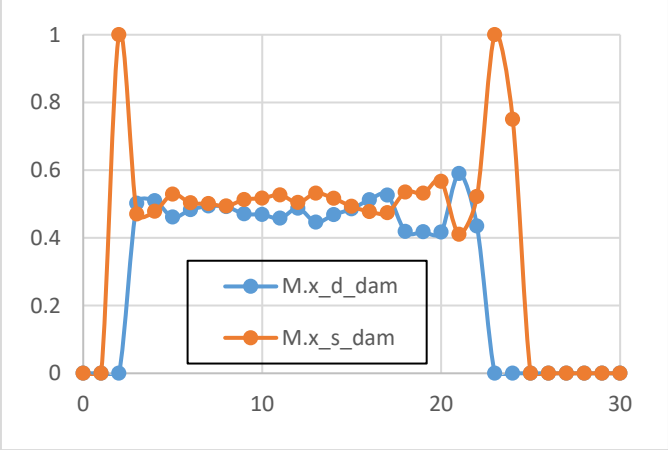

Fig. 7a

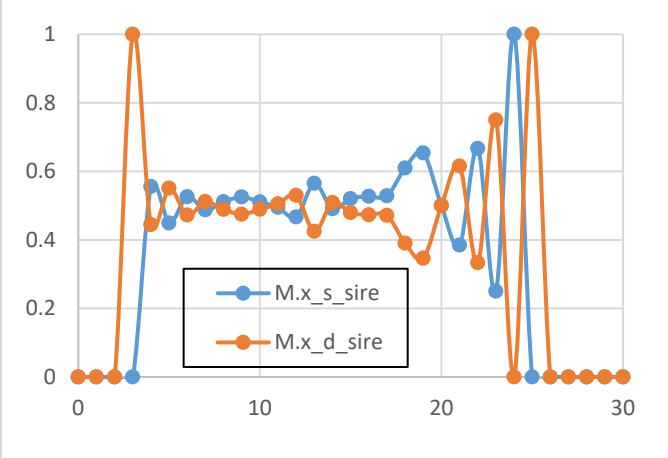

Fig. 7b

Fig. 7. Dam and Sire Tendency for Having Fe/Male Children by Age

C. *Inbreeding Analysis*

Social groups play a crucial role in regulating inbreeding within the CS colony [15]. In adherence to this strategy, male rhesus monkeys are regularly migrated to a group different from their birth group once they reach adulthood. Studies (such as [19]) concluded that offspring's born between closely related parents were rare and uncommon.

Using Gephi [5], we performed a clustering analysis and discovered intriguing group and dyad interaction patterns. The results will be presented in additional papers.

Results of qualitative studies using standard inbreeding coefficient are summarized in this paper. Since patrilineal lineage has only been available since the mid-1970s, potential inbreeds with parents from the same founder line were chosen using SQL queries, and "lower-bound" inbreeding coefficients (avoiding parent inbreeding coefficients) were calculated using Java program modules to be integrated into the CSViewer system. The inbreeding coefficients are often relatively low, with the exception of a few outliers, as can be seen in the box-and-whisker graph in Fig. 8a. The highest (0.125 or 1/8) is associated with a sibling-inbred animal.

To support genetic tracing needed in this inbreeding analysis, a kinship can be generated using a module to be included in CSViewer, as shown in Fig. 8b.

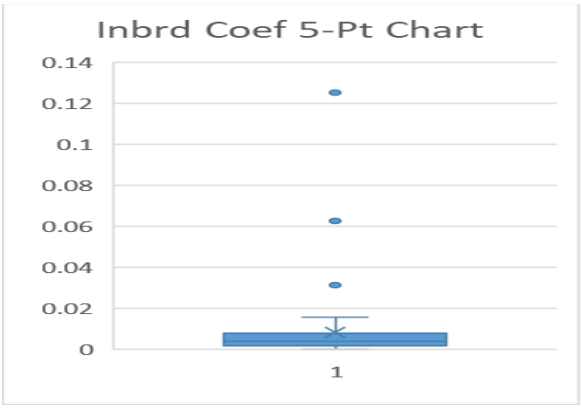

Fig. 8a

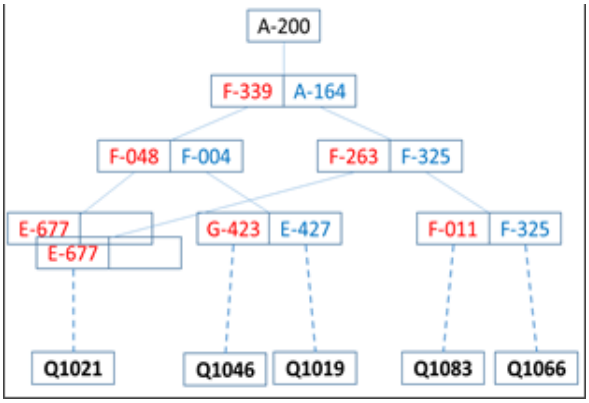

Fig. 8b

Fig. 8. Male and Female Reproduction Curves

A kinship tree can be either extended in a top-down format (i.e., descendants above ancestors, as shown in Fig. 8b) or a bottom up format. Dyads are represented with double boxes, color-coded to show dam-sire identifiers. Solid links are used to show parent-child relationships, while dashed links trace to founders (with intermediate pedigrees omitted).

V. Conclusions and Future Work

The Cayo Santiago rhesus colony provides a well-established genealogical record for about 11000 monkeys over the last 85 years. Efforts supported through a NSF collaborative grant have resulted in a software application, *CSViewer for Analysts* that can support the users to access the comprehensive database to select and analyze data cross-referenced to family tree structure.

A. *Conclusions:*

Data quality analysis and exploration have been conducted with existing and newly collected data from various sources. Family tree structures have been constructed, which can be

manipulated through the CSViewer interface for researchers to use in searching for related measurement and imagery data. This environment provides a framework for easy cross-referencing between biological, pathological, and genetic data sets, and can facilitate a growing set of modeling and visualization functions to be added for finding new knowledge.

Typical analytics regarding the population dynamics and reproduction have been conducted using a big data approach to provide a retrospective view for one of the most useful primate sources in biomedical and anthropological research. Some significant ways of incorporating these analytical tasks in the CSViewer system have been discussed. It can help demonstrate the capabilities of the application and attract the research communities to give specific directions for further development.

B. Future work:

Continued efforts will be made to include newly collected measurements and imagery data into the comprehensive database. In addition to the proposed relational database, NoSQL databases, such as graph database (for family tree) and document database (for catalog records) will be considered. Modules developed for cross-referencing, fine-tuned selection, and various analytical functions will be included into the CSViewer app in an incremental process. We will continue to encourage undergraduate students to work on this project [4].

A workshop is planned in the spring of 2024 to promote the forthcoming version of CSViewer and a working product is projected to be released to the scientific community in late 2024.

ACKNOWLEDGMENTS

The CPRC Skeletal Collection has been supported by National Institutes of Health NIH contracts NIH 5P40OD012217. This project is supported by NSF grants to M.Q.Z. and Q.W. (#1926402 and #1926601). Thanks go to Melween I. Martinez Rodriguez (Current CPRC Director), Bonn V. Liong Aure, Terry B. Kensler, Elizabeth Maldonado, Giselle Caraballo Cruz, other CPRC staff members, Debbie Guatelli-Steinburg, Luci Kohn, and George Francis for their support and help.

We acknowledge work by students at Mercer University's Computer Science Department (especially Kerry Oedel) for their input through various course works. Special thanks to our colleagues, Robert Allen, Jesse Sowell, Cristina Petruso, and Rui Gong for providing their support and contributions.

REFERENCES

[1] A. M. Bronikowski et al, Female and male life tables for seven wild primate species, doi: 10.1038/sdata.2016.6

[2] C. Horstmann, "Object-Oriented Design and Patterns," 2nd Edition, Wiley, 2005.

[3] D. Guatelli-Steinberg et al, Talon Cusp Expression in Cayo Santiago Rhesus Macaques, a brief communication submitted to AJBA, July, 2023

[4] E. R. Widener, S. Kundu, R. Chapagain, P. Sapkota, and M. Q. Zhao, Engaging Students in Undergraduate Research: Teaching Through Design, Development, and Collaboration, Proceedings of FECS'23 - The 19th Int'l Conf on Frontiers in Education: Computer Science and Computer Engineering, Las Vegas, NV, Jul 24-27, 2023.

[5] Gephi, https://gephi.org/, (Retrieved: Sept 15, 2023)

[6] JFreeChart, https://www.jfree.org/jfreechart/, (Retrieved: Sept 15, 2023)

[7] M. J. Kessler et al, Long-Term Effects of Tetanus Toxoid Inoculation on the Demography and Life Expectancy of the Cayo Santiago Rhesus Macaques, American Journal of Primatology 77:211–221 (2015)

[8] M. Q. Zhao, E. R. Widener, G. Francis, Q. Wang. Building a Knowledge Model of Cayo Santiago Rhesus Macaques: Engaging Undergraduate Students in Developing Graphical User Interfaces for an NSF Funded Research Project. In: Daimi, K., Al Sadoon, A. (eds) Proceedings of the Second International Conference on Innovations in Computing Research (ICR'23). ICR 2023. Lecture Notes in Networks and Systems, vol 721. Springer, Cham. https://doi.org/10.1007/978-3-031-35308-6 29

[9] M. Q. Zhao, Knowledge Models for SA Applications and User Interface Development for the SITA System", Technical Report to AFRL/RI, Rome, NY, May 12, 2011 (Revised July 14, 2011).

[10] M. Q. Zhao, Maldonado M, Kensler TB, Kohn LAP, Guatelli-Steinburg D, Wang Q. 2021. Conceptual Design and Prototyping for a Primate Health History Model. In: Arabnia HR, Deligiannidis L, Tinetti FG, Tran Q-N (Editors). Advances in Computer Vision and Computational Biology. New York: Springer. p.511-522.

[11] M. Q. Zhao, T. B. Kensler, D. Guatelli-Steinberg, L. A. P. Kohn, G. Francis, Q. Wang. 2023. Reproduction of CS Cayo Santiago Rhesus Colony based on the Patrilineal Family Trees: The Missing Patterns. American Journal of Biological Anthropology 180(S75):201.

[12] MS SQL Server Online Docs, https://learn.microsoft.com/en-us/sql/sql-server/?view=sql-server-ver16, (Retrieved: Sept 15, 2023)

[13] Q. Wang. 2023. Coming to the Caribbean - Eighty-five years of Rhesus macaques (Macaca mulatta) at Cayo Santiago: A long-term model for the studies of adaptation, diseases, natural disasters and resilience. American Journal of Biological Anthropology 180(S75):189.

[14] Q. Wang. 2012. Bones, Genetics, and Behavior of Rhesus macaques: Macaca mulatta of Cayo Santiago and Beyond. New York: Springer.

[15] R.G., Rawlins, Kessler, M.J. and Turnquist, J.E. (1984), Reproductive Performance, Population Dynamics and Anthropometrics of the Free-Ranging Cayo Santiago Rhesus Macaques. Journal of Medical Primatology, 13: 247-259.

[16] R. Hernandez-Pacheco et al, Managing the Cayo Santiago rhesus macaque population: The role of density, AJP, 2015

[17] D. S. Sade, K. Cushing, P. Cushing, J. Dunaif, A. Figueroa, J. R. Kaplan, C. Lauer, D. Rhodes, J. Schneider, Population dynamics in relation to social structure on Cayo Santiago,Ybk phys. Anthropol, 20 (1977), pp. 253-262

[18] T. B. Gage and B. Dyke, Model Life Tables for the Larger Old World Monkeys, American Journal of Primatology 16:305-320 (1988)

[19] Widdig et al, Genetic Studies on the Cayo Santiago Rhesus Macaques: A Review of 40 Years Of Research, AJP 78:44-62, 2016

[20] Cayo Santiago rhesus colony censor datasets from 2013 and 2020, provided by the Caribbean Primate Research Center (CPRC), unpublished.

[21] Cayo Santiago rhesus colony founder list, provided by the Caribbean Primate Research Center (CPRC), unpublished.