SCDS Project Proposal (2019-2020)

1000 Plagues in the Genomics Era: Databases and Digital Exhibits

This manuscript (permalink) was automatically generated from ktmeaton/obsidian-public@c2e4bac4 on May 10, 2021.

Authors

- **Katherine Eaton**
  0000-0001-6862-7756 · ktmeaton
  McMaster Ancient DNA Center, McMaster University

Project Overview

Background

When used within a multi-disciplinary approach, genetic evidence provides an intriguing window into the disease experience of past populations. Ancient DNA techniques have been developed to identify bacterial and viral DNA present in skeletal remains, and even affected artifacts [1]. These methods have additionally been used to explore the patterns of spread and geographic origins of several high-profile diseases, including the plague [2]. However, the ability to accurately reconstruct spatial patterns of an ancient disease requires robust comparative datasets. In order to produce such datasets, extensive sampling is required across both time and place, followed by careful curation. Fortunately, recent advances in DNA sequencing technology have facilitated efforts to conduct this extensive sampling on a global scale, and data repositories continue to grow at unprecedented rates [3].

However, this technological progress has also placed a strain on our ability to effectively manage, curate, and present the avalanche of data effectively. Over the course of my residency in the 2017/2018 period, I curated and visualized approximately 300 plague records. In 2019, there are now more than 1000 available records, of which the vast majority are underutilized in published literature [4]. The rapid pace at which this information is generated, and the slow pace at which it is incorporated into ongoing research, suggests an urgent need for new strategies for curation, visualization, and dissemination of genetic disease records.

Despite my previous efforts to curate and visualize a portion of this information, it was in the realm of dissemination that I encountered substantial obstacles. Firstly, I was uncertain about the process of publishing software, as opposed to a typical research paper, as digital tools have a public life before and after publication that must be considered. Secondly, the dynamic, interactive maps I created did not translate well to figure generation for sharing in a static document. Finally, I wasn't easily able to incorporate a multidisciplinary approach, as the software platforms I selected did not allow for integrating heterogeneous data types (ex. evolutionary trees, historical records, and archaeological artifacts). I thus needed to rethink my objectives for completing software development and explore alternative dissemination outlets for my analysis.

Proposal

The project I propose for the duration of the residency program would be conducted in two phases. The first phase (Fall Term) is a direct continuation of my work from the 2017/2018 residency by moving the database tool I created towards release, publication, and continuing support infrastructure. In this aspect of the project, I would be primarily focused on how to "complete" a digital humanities project and navigate the hurdles therein.

The second phase would involve putting the database software to work by analyzing ancient and modern samples of plague DNA. In my 2017/2018 residency, I generated 2 initial visualizations: a global map of disease distribution, and a dynamic time series of historical outbreaks. While overall a productive experience, this output also suffered from issues such as an inflexibility in incorporating disparate data forms and a restriction to only being functionally displayed on my local computer. Thus I am seeking new avenues for web-based visualization that could allow greater flexibility in terms of public dissemination, collaborator participation from other disciplines, and storytelling engagement.

The intended audience for this project would be two-fold. First, for the content provided, the database and exhibits produced would primarily be targeted towards plague geneticists, archaeologists, and historians. The second audience group I would like to focus on are teams pursuing cross-disciplinary work who are interested in exploring creative means of communicating. This project emphasizes the role that digital exhibits play not just as showpieces, but as integral parts of the analytical methodology. As such, I hope to explore how digital humanities tools might present new opportunities for multidisciplinary teams to create and investigate research questions using visual elements as a primary means of exchange.

Influences

This project draws inspiration from diverse sources, including academic articles, software repositories, and online exhibits. My original database software was motivated by projects such as SRAdb [5] and MetaSRA [6]. While these provided an excellent foundation, they also left methodological issues outstanding, which gave me a clear vision of possible future contributions. These works also introduced me to different platforms for project management, regarding how to package, release, update, and document code changes. In a similar vein, engaging with the Ancient DNA project repository EAGER revealed to me many unknown functions of Github that I aspire to learn in terms of user support and issue tracking [7].

The inspiration for the data visualization and exhibiting aspects of this project first come from the NextStrain project [8]. This is an international effort primarily focused on mapping the evolutionary history of bacterial and viral diseases, of which I find the tuberculosis exhibit particular striking and fun to explore (https://nextstrain.org/tb/global). This interface style is suitable for interrogating evolutionary relationships and patterns, as well as conducting a superficial exploration of the contextual metadata. My initial goal will be to recreate this display for the plague data that is publicly available, along with the data I have generated throughout the course of my dissertation's lab work.

When going beyond a superficial exploration of contextual evidence, a StoryMap exhibit may be particularly suited to weaving together rich historical, archaeological, and genetic sources. My first StoryMap inspiration was the Black Death plague map where the viewer can move throughout Europe learning about plague history with visual aids [9]. With regards to form, I'm enthralled with the Naonaiyaotit Traditional Knowledge Project Atlas and how it presents a tab-navigable map series of

different regions and narratives [ntkp2019NaonaiyaotitTraditionalKnowledge]. Currently, I envision a similar format for the plague narratives, in which the user selects tabs to navigate between notable regions, events, and evidence in the history of this disease.

Materials and Methods

Data and Access

The data for this project consists of two sources:

1. Publicly acquired plague records and DNA sequences from online genetic repositories [10,11]. The records and accompanying metadata are retrieved using an API provided by the National Centre for Biotechnology Information (NCBI) and my customizable software [12].
2. Ancient plague DNA samples I have sequenced throughout the course of my dissertation. Contextual metadata is provided by archaeologist contacts and surviving historical records.

The primary metadata fields I will be visualizing are geographic location and time period, although supplemental variables (of which there are over 50 available, including environmental data) may also prove informative when incorporating socioecological theories of disease.

Curation and Comparison

In my previous work, I laid out an initial curation and metadata standardization strategy for the plague record database. However, I did not have a way to organize and link to the literature I was using to perform this curation so that others may critique and build on this strategy. For this project, I first plan to create an open bibliography through Zotero in order to organize the primary source documents to improve curation transparency, reproducibility, and facilitate collaboration if possible.

The foundation of the first exhibit ("NextStrain"), involves evolutionary analysis to compare the genetic relatedness of plague strains against each other. This approach shares many similarities with how stylometry methodology operates in digital humanities projects [13]. A document is compared to other texts in order to construct a pedigree of similarities, so as to answer a specific question (perhaps authorship of an unidentified object). But instead of English documents, I have DNA "texts" to compare against each other. This is an intensive computational step, and is conducted on a high performance computing cluster where an account is generously provided by Dr. Brian Golding.

Following comparison and curation, the database contents and the relationships within can be exhibited. The first exhibit, a NextStrain project, provides the visualization engine in the form of a web application that connects to a users' local GitHub repository. The website also hosts community projects on the front page for higher visibility and thus I hope to apply for a display there. The second exhibit is planned to be a shared community map series of StoryMaps using ArcGIS online. The primary localizing points on the map would be genetic plague records, while the textual and visual element of the story would be historical and archaeological evidence that ties the disease experiences together across the globe.

Technical Expertise, Skills, and Requirements

The technical skills needed to accomplish these objectives expand upon my previous work in database design and geovisualization (2017/2018 Residency), as well as web design and pedagogical resources (2018/2019 Residency). The skills that I am looking to develop first include furthering my GitHub

project management and user support skills. I also hope to gain more familiarity in geocoding and georeferencing skills, and will take relevant DMDS workshops to continue that skill progression. In addition to instructional workshops, it would be beneficial to discuss with others about tackling subjectivity with ambiguous data and the politics of place names in a historical context. In addition, I will be learning two new visualization engines, NextStrain and ArcGIS Online, which I have only used in a learning capacity for tutorial completion. I am therefore looking forward to more advanced design options and to improve my short-form communication skills through these media. Finally, I will be developing my ability to work in a multidisciplinary context as I communicate with collaborating archaeologists and historian co-authors to contribute to the StoryMap.

Project Timeline

The two phases of this project are divided between Fall Term 2019 and Winter Term 2020. As "completing" a digital project is a high priority, I will begin with publishing and disseminating my database software (Phase I) before proceeding to learning about exhibiting the products of that project (Phase II).

Phase I: September - December 2019

- **Sept**: Software Packaging and Release
- **Sept-Oct**: Publication Writing
- **Nov-Dec**: Project Page improvements
- **Nov-Dec**: Documentation and Support Infrastructure

This period also includes the foundational analysis for Phase II.

Phase II: January - April 2020

- **Jan-Feb**: NextStrain Phylogeny Exhibit
- **Mar-Apr**: ArcGIS StoryMap Exhibit
- **May 7-10**: Conference Presentation

The final objective will be a podium presentation at a digital exhibit-centric session at the International Congress of Medieval Studies (Kalamazoo, MI, May 2020) entitled "Curating Medieval Plague and Pestilence."

Dissemination

Dissemination outcomes for this project are distributed across four categories. First would be the Github repositories and project pages that host the source code for the software and exhibits. Second are the academic papers, and third are the two digital exhibits themselves, NextStrain and StoryMap. Fourth, and finally, is an international conference presentation to showcase the exhibits created and exchange ideas about digital exhibits with other ancient disease scholars.

While these dissemination products all build on each other in a logical progression, I learned from my 2018/2019 residency that my goals tend to be overly ambitious and full completion is not a realistic expectation. Therefore, I will approach Phase II being reasonably satisfied with reporting works in progress and try to embrace the productive failures emerge throughout the digital humanities research process.

References

1. **17th Century Variola Virus Reveals the Recent History of Smallpox**
   Ana T. Duggan, Maria F. Perdomo, Dario Piombino-Mascali, Stephanie Marciniak, Debi Poinar, Matthew V. Emery, Jan P. Buchmann, Sebastian Duchêne, Rimantas Jankauskas, Margaret Humphreys, … Hendrik N. Poinar
   Current Biology (2016-12) https://linkinghub.elsevier.com/retrieve/pii/S0960982216313240
   DOI: 10.1016/j.cub.2016.10.061

2. **A draft genome of *Yersinia pestis* from victims of the Black Death**
   Kirsten I. Bos, Verena J. Schuenemann, G. Brian Golding, Hernán A. Burbano, Nicholas Waglechner, Brian K. Coombes, Joseph B. McPhee, Sharon N. DeWitte, Matthias Meyer, Sarah Schmedes, … Johannes Krause
   Nature (2011-10) http://www.nature.com/articles/nature10549
   DOI: 10.1038/nature10549

3. **Big Data: Astronomical or Genomical?**
   Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, Gene E. Robinson
   PLOS Biology (2015-07-07) https://dx.plos.org/10.1371/journal.pbio.1002195
   DOI: 10.1371/journal.pbio.1002195

4. **Ancient *Yersinia pestis* genomes from across Western Europe reveal early diversification during the First Pandemic (541–750)**
   Marcel Keller, Maria A. Spyrou, Christiana L. Scheib, Gunnar U. Neumann, Andreas Kröpelin, Brigitte Haas-Gebhard, Bernd Päffgen, Jochen Haberstroh, Albert Ribera i Lacomba, Claude Raynaud, … Johannes Krause
   Proceedings of the National Academy of Sciences (2019-06-18) https://www.pnas.org/content/116/25/12363
   DOI: 10.1073/pnas.1820447116 · PMID: 31164419

5. **SRAdb: query and use public next-generation sequencing data from within R**
   Yuelin Zhu, Robert M Stephens, Paul S Meltzer, Sean R Davis
   BMC Bioinformatics (2013) http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-19
   DOI: 10.1186/1471-2105-14-19

6. **MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive**
   Matthew N Bernstein, AnHai Doan, Colin N Dewey
   Bioinformatics (2017-09-15) https://academic.oup.com/bioinformatics/article/33/18/2914/3848915
   DOI: 10.1093/bioinformatics/btx334

7. **EAGER**
   Alexander Peltzer
   (2019) https://github.com/nf-core/eager

8. **Nextstrain: real-time tracking of pathogen evolution**
   James Hadfield, Colin Megill, Sidney M. Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, Richard A. Neher

Bioinformatics (2018-12-01) https://academic.oup.com/bioinformatics/article/34/23/4121/5001388
DOI: 10.1093/bioinformatics/bty407

9. **The Black Death**
   The Black Death
   The Black Death (2019) https://uploads.knightlab.com/storymapjs/428a8c7a4899449a5949f9fe5bd79d06/the-spread-of-the-black-death/index.html

10. **Assembly: a resource for assembled genomes at NCBI**
    Paul A Kitts, Deanna M Church, Françoise Thibaud-Nissen, Jinna Choi, Vichet Hem, Victor Sapojnikov, Robert G Smith, Tatiana Tatusova, Charlie Xiang, Andrey Zherikov, ... Avi Kimchi
    Nucleic acids research (2016-01-04) https://www.ncbi.nlm.nih.gov/pubmed/26578580
    DOI: 10.1093/nar/gkv1226

11. **The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity**
    Zhemin Zhou, Nabil-Fareed Alikhan, Khaled Mohamed, Yulei Fan, Mark Achtman
    Genome Research (2020-01) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6961584/
    DOI: 10.1101/gr.251678.119 · PMID: 31809257 · PMCID: PMC6961584

12. **NCBImeta: efficient and comprehensive metadata retrieval from NCBI databases**
    Katherine Eaton
    Journal of Open Source Software (2020) https://doi.org/10.21105/joss.01990
    DOI: 10.21105/joss.01990

13. **Stylometry with R: A Package for Computational Text Analysis**
    Maciej Eder, Jan Rybicki, Mike Kestemont
    The R Journal (2016) https://journal.r-project.org/archive/2016/RJ-2016-007/index.html
    DOI: 10.32614/rj-2016-007