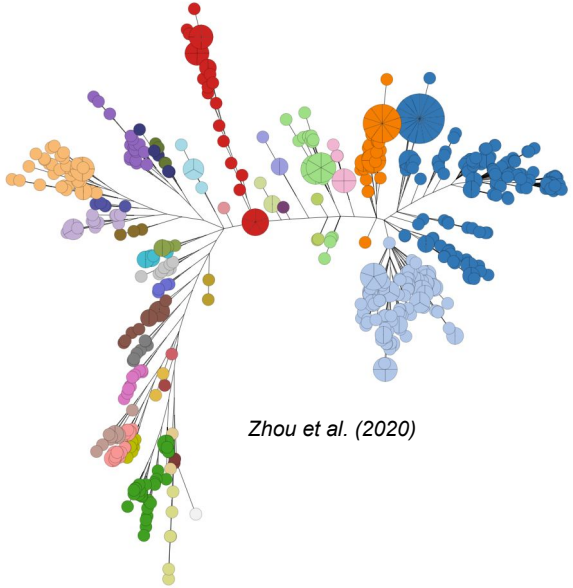
Plagues, Pipes, and Genotypes

Phylogenetics of “resurrecting” disease foci.

Katherine Eaton

30 January, 2020



Zhou et al. (2020)

Li et al. (2008)

Presentation Roadmap

Background

- Why Study Plague?
- Problems
- Previous Work
- Questions

Methods

- Data Collection
- Processing
- Pipes

Results

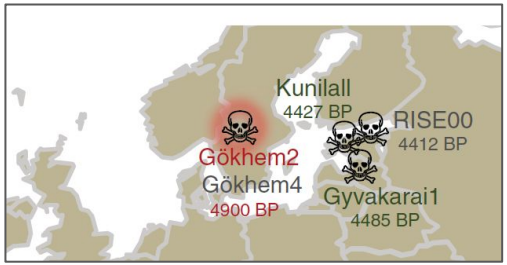
- 2018
- Short Term Goals
- Long Term Goals

Background: Plague

The What, Why, and How of plague research.

Why Do We Study “The Plague”?

Neolithic (5000 BP)



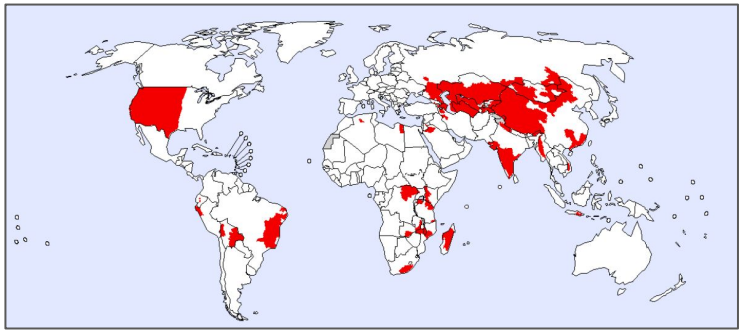
Rascovan et al. (2019)

2019

We never really got rid of the plague. 3 people in China just caught it.

The plague is still a problem around the world — including in the US.

By Sigal Samuel | Updated Nov 20, 2019, 2:30pm EST



WHO/PED (2016)

3 - 7 days
incubation period

The case-fatality ratio of 30%-100% if left untreated

who.int

Why Do Genome Sequencing?

- Zoonoses of rodents.
- Impossible to eradicate.
- Difficult to observe.

- Surveillance work is a forefront concern.
- Anxiety about *disease invisibility*.
- Genomics renders the invisible, visible.

- 2010-2020: 20→ 1500 genomes

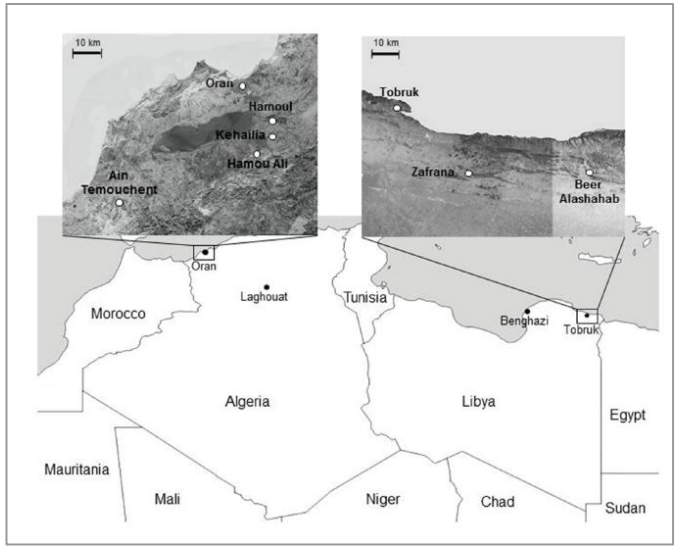
Institut Pasteur

Country / year of last outbreak	Duration of "silence"
Botswana 1989	45 years
Kenya 1990	10 years
Madagascar 1994	60 years
Zambia 1997	33 years
Algeria 2003	50 years

openwho.org

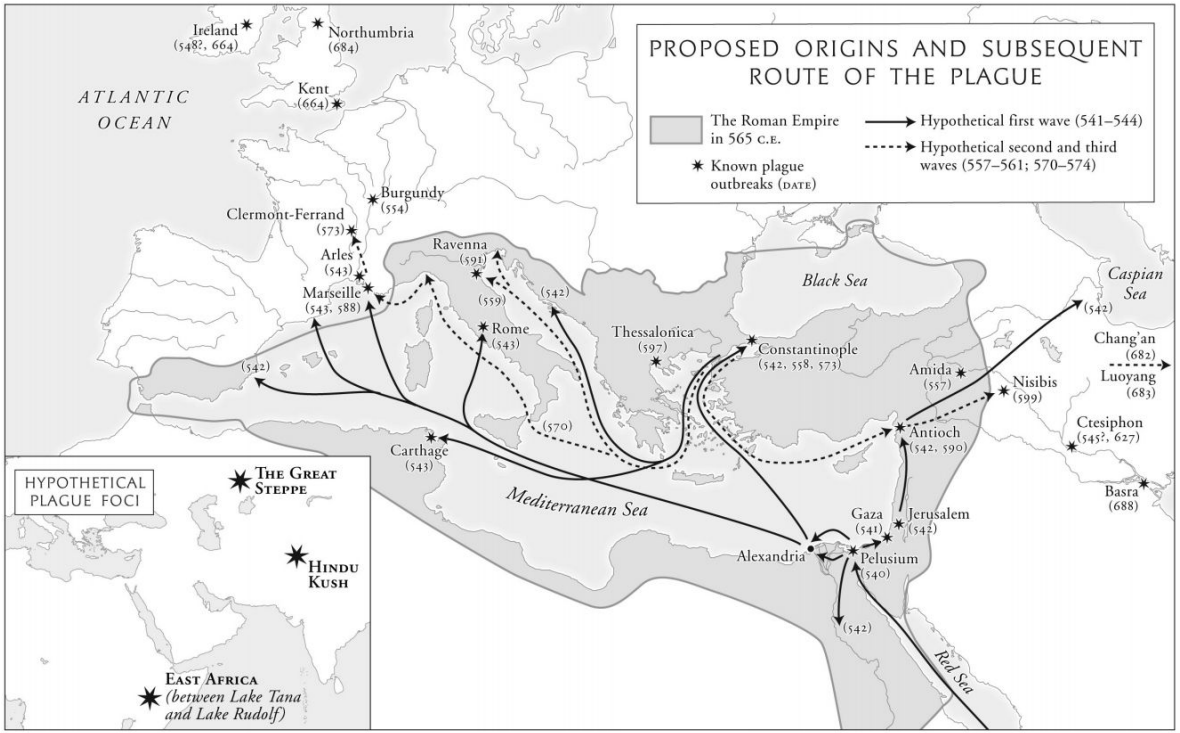
Past Plagues

Algeria (2003), Libya (2009)



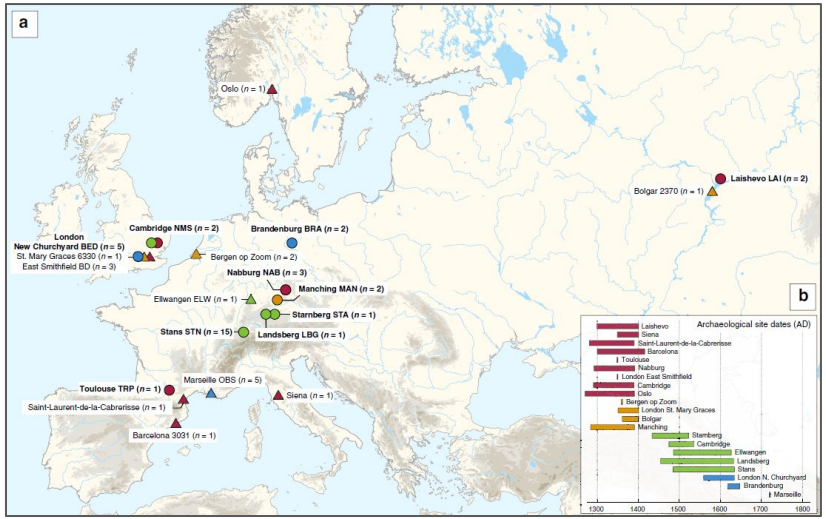
Cabanel et al. (2013)

Roman Empire (6 CE)



Problems

1. **<u>Academic Plague Discourse</u> has been a one-sided *conversation*.**
   Modern → History : Science → Not Science
   - Novices attempting specialist tasks without feedback.
   - Missing out on really interesting questions (What/How vs. Why?)



Spyrou et al. (2019)

Problems

2. **<u>Genomic Data Overload</u>. Methodological and interpretive issues.**

   2010-2020: 20→ 1500 genomes

   - Global phylogeny, stitched together from independent projects.
   - Which regions are over-represented: 80-90% East Asia/China.
   - Which regions are under-represented: Africa.
   - Revealing instability of substrain/clade nomenclature.

Previous Work

1. **Ancient Plague Discourse:**
    - Jena Plague Researchers (active conversation with historians online).
    - Publish in Science journals, write science papers.
    - End-point integration of historians, only for interpretation?

2. **Genomic Data Overload.**
    - Critique/Self-awareness: aDNA (Spyrou et al. 2019)
    - Proposed practical solutions (Enterobase et al. 2020)
    - Looking to other fields (ex. *M. tuberculosis*)

Questions

1. **How do we move forward in the data revolution?**
   - Methodologically: Data ‘collection’, analysis, visualization.
   - Critically: What biases are present in the data, what are the consequences?

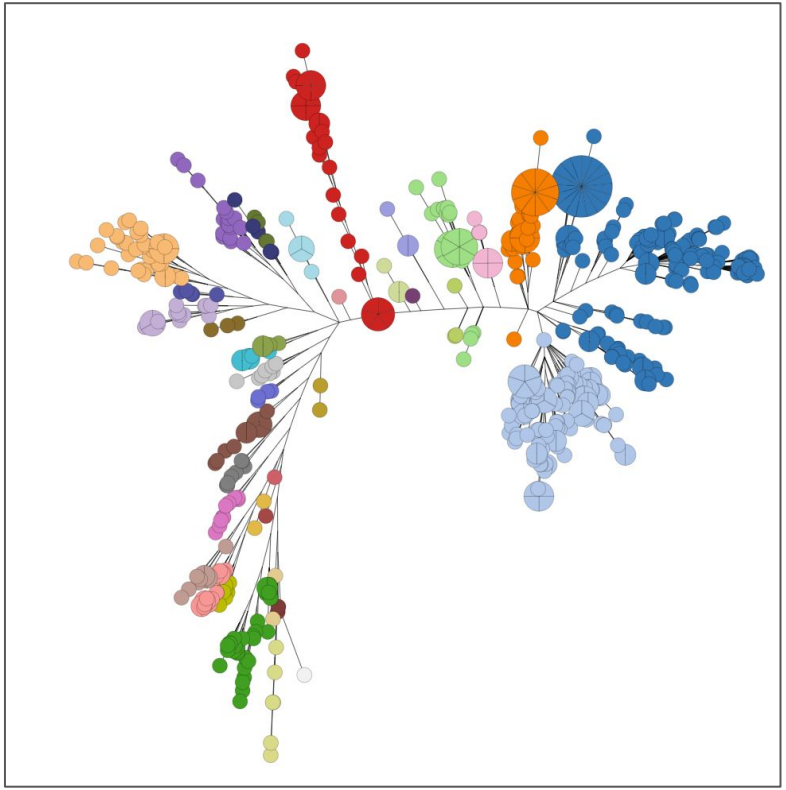
2. **How do we broaden the research potential and utility of phylogenetic studies?**
   - What questions do geneticists ask? Archaeologists? Historians?

Method to the Madness

Data Collection on the internet, Pipes incoming.

Data

- NCBI *Y. pestis* whole genomes
  - assembled and un-assembled
- ~1050 samples (non-laboratory)
- Enterobase → 600 genomes
- Kat's previous work (2018): 340 genomes



Zhou et al. (2020) - Enterobase

Data Collection

1. **Metadata from databases (NCBI, PATRIC, Enterobase, Literature):**
    - Collection Date
    - Geographic Location
    - Host
    - Nomenclature/sub-strain

2. **Download genomic data from NCBI:**
    - Assembled (Button: “Download All Assemblies”)
    - Un-Assembled (Pipeline, Make)

Data Processing

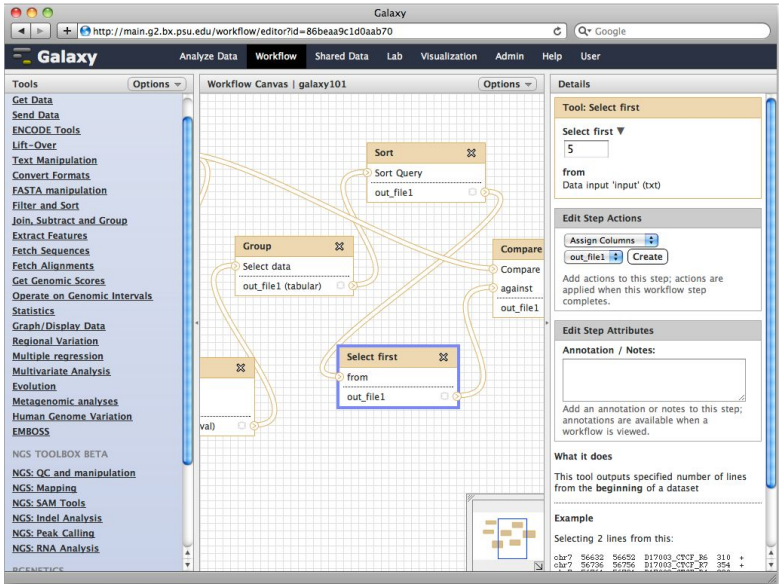
- **Bacterial Genomics Pipeline (“Snippy”)**
  - Whole genome alignments (not just SNPs)
  - Mixed data types as input (fasta contigs, raw reads fastq, mapped bam)

- **Assembled Genomes** → Align to reference.
- **Un-Assembled Genomes** → Pre-processing (trim, merge, align to reference, dedup)
- **Multiple Alignment**
  - Filtering (Low Coverage)
  - Masking (SNP Density, Low Complexity, Repeats)
- **Model Testing**, **Maximum Likelihood Phylogeny**, **Support Testing**
- **Visualization** → Figtree, GrapeTree, R, NextStrain

Pipes

- **Workflow Management System (WMS) / “Pipeline”:**
  - *Execute a series of computational steps*
  - **Error detection, parallelism, reproducibility*
  - **Re-entrancy, dependencies*

    - Galaxy
    - Make
    - Snakemake (GUI: Sequanix)
    - Nextflow (GUI: DolphinNext)

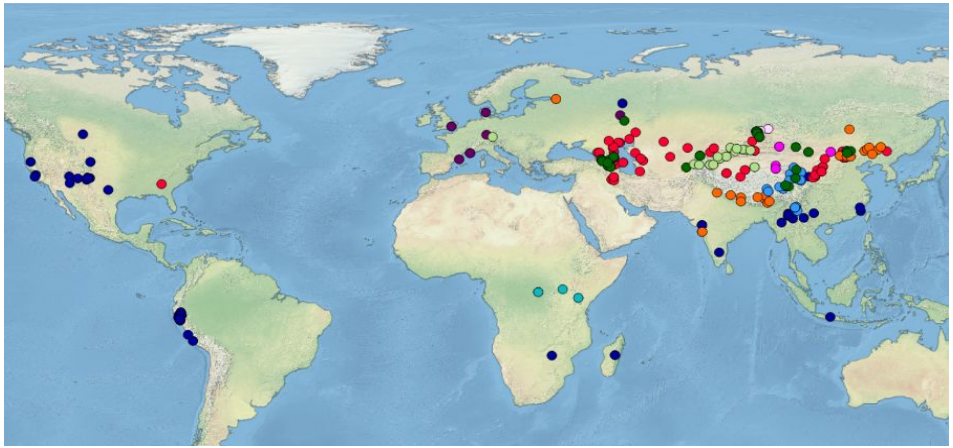


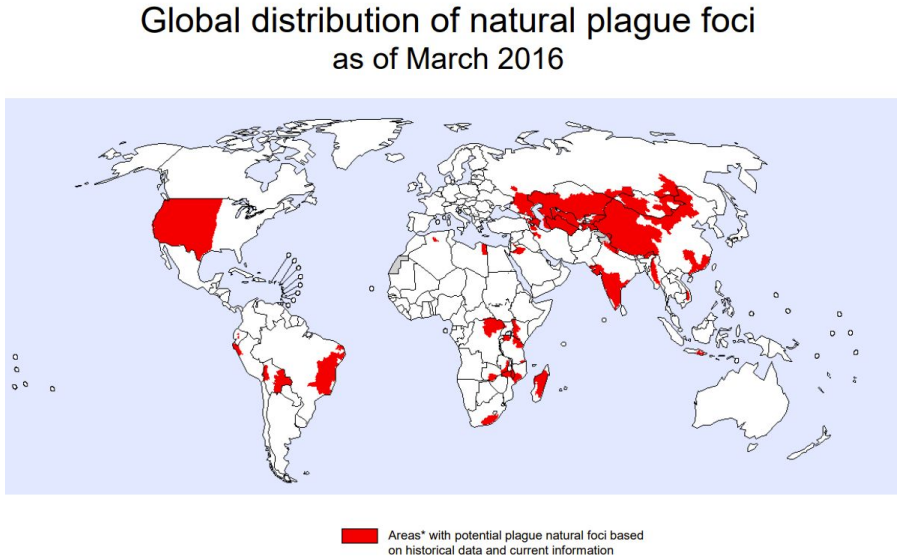
Galaxy

Results: 2018

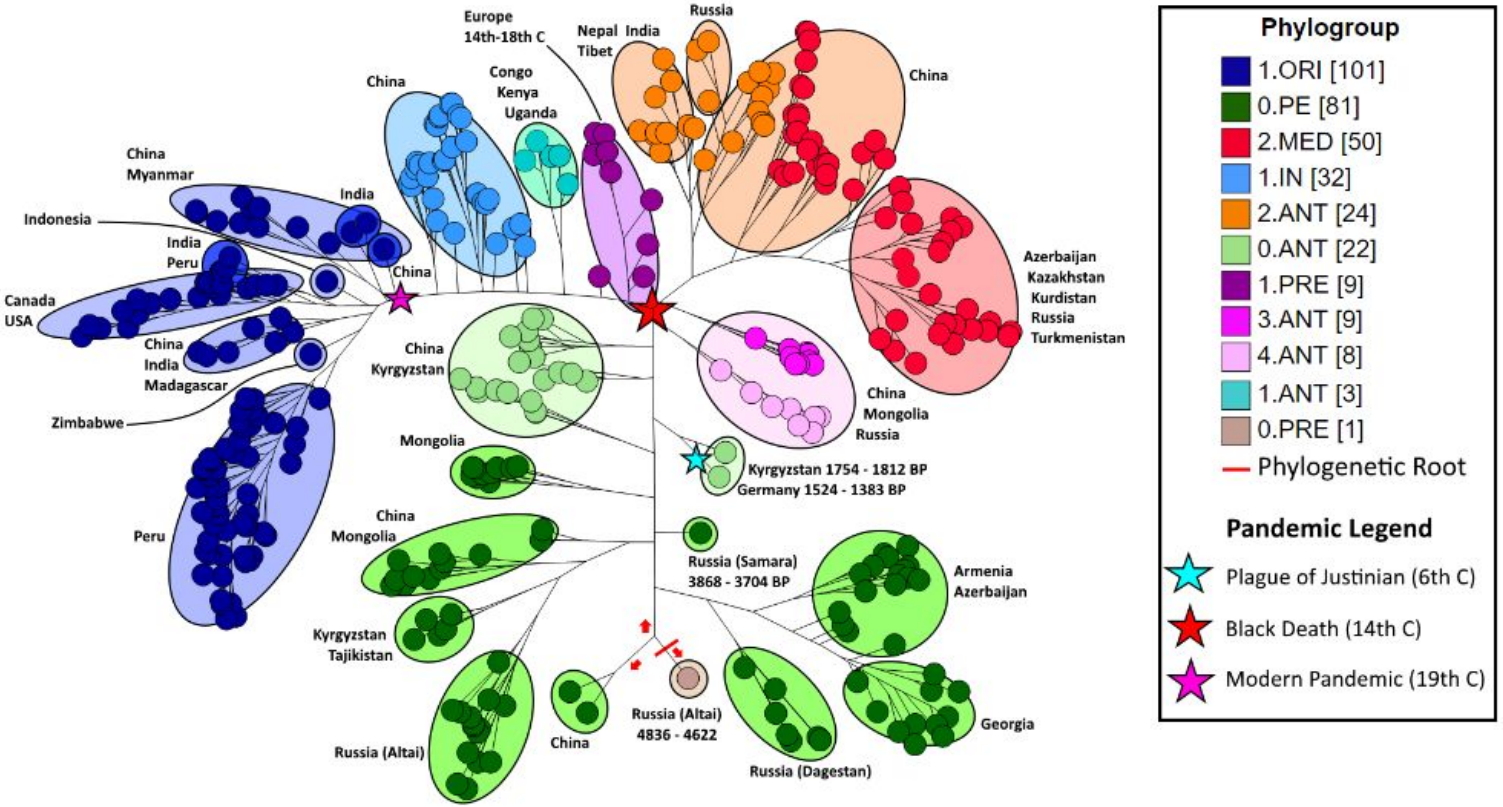
Figtree and R are Fairweather Friends. Always take notes.

What Biases are Present in the Data?





What Biases are Present in the Data?

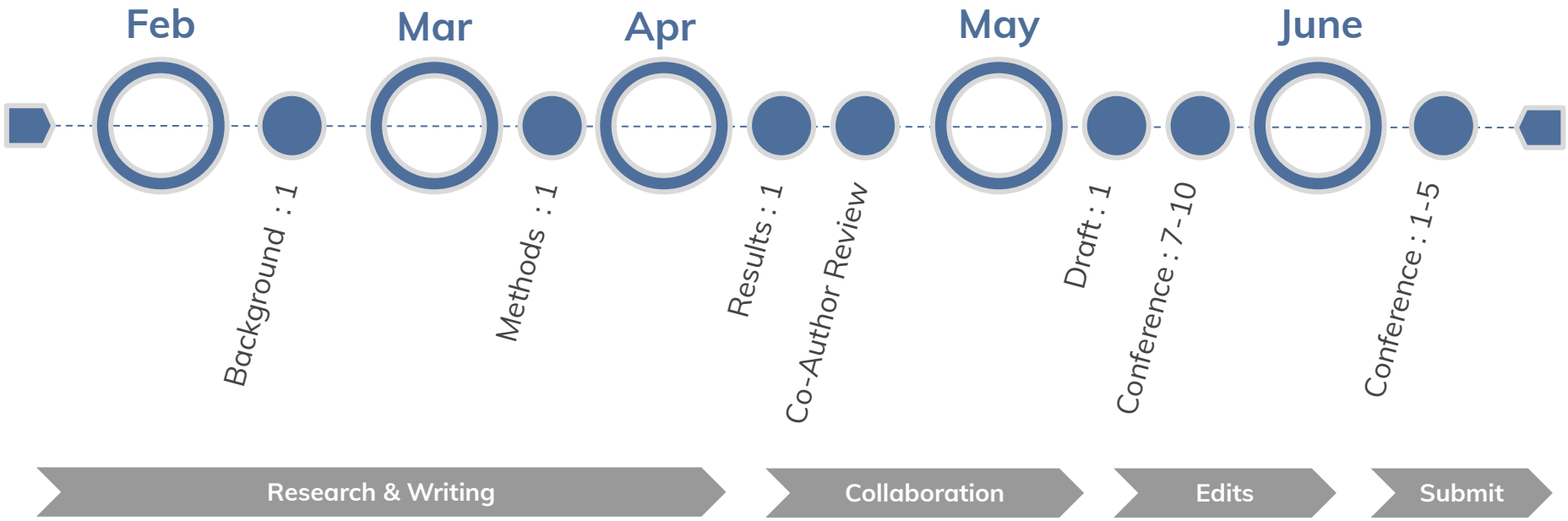


Short Term Goals

February

1. Short background section (500 words)
2. Research and test out a WMS/pipeline language.
3. Redo phylogenetics workflow with the new genome assemblies.
4. Start phylogenetics workflow with some unassembled datasets (ex. ancient).

Winter 2020 Roadmap



Acknowledgements

The Poinar Lab

\+ **Ravneet Sidhu** and **Dirk Hackenberger**

Collaborators

- Brian Golding
- Nukhet Varlik
- Ann Carmichael
- Eddie Holmes

Lewis & Ruth Sherman Centre for Digital Scholarship

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada