INFERENCES ON EVOLUTIONARY HISTORY WITH APPROXIMATE BAYESIAN COMPUTATION

Ariella Gladstein
Ecology and Evolutionary Biology
University of Arizona
HOW DID HUMANS SPREAD ACROSS THE WORLD?

WHAT DEMOGRAPHIC EVENTS LEAD US TO WHERE WE ARE TODAY AND THE DIVERSITY WE SEE?
Nielsen et al. 2017
Nielsen et al. 2017

Out of Africa 55–65 kyr ago

Neolithic expansion 10 kyr ago

Migration of click-language speakers 50 kyr ago

Polynesian expansion 3–5 kyr ago

Alternative route to America

Northern-southern split 13 kyr ago

Peopling of America across Beringia 15–23 kyr ago

Migration into Sahul 47.5–55 kyr ago

Palaeolithic Eurasians 45–55 kyr ago

Inuit expansion 3–4 kyr ago

Paleo-Eskimo expansion 4–5 kyr ago

Sintashta expansion 2.5–3.5 kyr ago

Yamnaya expansion 4.5 kyr ago

(IP) CA FC PCS

(Nielsen et al. 2017)
Nielsen et al. 2017

Neolithic expansion 10 kyr ago

Out of Africa 55-65 kyr ago

Migration of click-language speakers 50 kyr ago

Neolithic expansion 2.5-3.5 kyr ago

Yamnaya expansion 4.5 kyr ago

Sintashta expansion 2.5-3.5 kyr ago

Inuit expansion 3-4 kyr ago

Paleo-Eskimo expansion 4-5 kyr ago

Peopling of America across Beringia 15-23 kyr ago

Alternative route to America

Polynesian expansion 3-5 kyr ago

Migration into Sahul 47.5-55 kyr ago

Northern-southern split 13 kyr ago

Possible pre-Columbian contact
WHAT ARE “DEMOGRAPHIC EVENTS”?
WHAT ARE “DEMOGRAPHIC EVENTS”?

- Divergence
WHAT ARE “DEMOGRAPHIC EVENTS”?

- Divergence
- Expansion or reduction
WHAT ARE “DEMOGRAPHIC EVENTS”?

- Divergence
- Expansion or reduction
- Gene flow
AIM: INFERR THE DEMOGRAPHIC HISTORY OF THE ASHKENAZI JEWS.
ASHKENAZI JEWS: AN INTERESTING STUDY POPULATION

- High frequency of genetic disorders
- Population isolate
- Complex demographic history
- Well documented historical record
ASHKENAZI JEWS:
AN INTERESTING STUDY POPULATION

- High frequency of genetic disorders
- Population isolate
- Complex demographic history
- Well documented historical record
HYPOTHESIS OF ASHKENAZI ORIGINS

- **Rhine Valley** (900 CE – present)
- **Eastern Europe** (1200 CE – present)
- **Italy** (300 BCE – present)
- **Israel** (1200 BCE – 70 CE)
WESTERN VS. EASTERN ASHKENAZI JEWS

Yiddish Dialects

Approximate Yiddish territories at the end of the 19th century to beginning of the 20th century
- Western dialect
- Eastern dialect
- Mixed Zones
- The Pale in the Russian Empire
- Modern country borders

Germany, 1900's

Cracow, Poland. 1932
WESTERN VS. EASTERN ASHKENAZI JEWS

Reference census data
MOTIVATION

• Numerous genetic studies on the Ashkenazi Jews.
  • All genome-wide studies treat Ashkenazi Jews as one population.
• Preliminary work consistent with genetic differentiation.
  • Not informative of cause of differentiation.
MODELS OF ASHKENAZI HISTORY
APPROXIMATE BAYESIAN COMPUTATION

- Infer parameter values
- Choose among models
1. Define priors of parameters of model

\( t = \text{unif}[10:1000] \)

\( t = \) time (generations) of divergence between Jewish and Middle Eastern populations
1. Define priors of parameters of model
2. Simulate data many times
1. Define priors of parameters of model
2. Simulate data many times
3. Choose model and estimate parameters based on simulations closest to real data
Model parameters

Store genotype sequences in memory

Calculate summaries of sequences

<10 Kb file with parameter values and summaries

SIMULATION
EMBARRASSINGLY PARALLEL!

Model parameters

Store genotype sequences in memory

Calculate summaries of sequences
INHERITED SCRIPT INTENDED FOR SMALL SEQUENCE

1,389 10kb regions

00000110001
00100010000
00000100101
00100000000
00010001010
00100010001
SIMULATE WHOLE CHROMOSOME

~250 million sites on human chromosome 1
## PROBLEM!

### Parameters Average Walltime and Average Memory

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Average Walltime</th>
<th>Average Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>00:21:00</td>
<td>2.7 Gb</td>
</tr>
<tr>
<td>Random</td>
<td>00:55:11</td>
<td>20 Gb</td>
</tr>
<tr>
<td>Maximum</td>
<td>08:02:11</td>
<td>117 Gb</td>
</tr>
</tbody>
</table>

Too much memory!

Over a decade to complete 6000 runs/month w/ UA resources

Each core on UA HPC has 6G - **Need memory < 6G** for each run
EMBARRASSINGLY PARALLEL & RESOURCE LIGHT!

- Each job
  - runs ~40 min, and max 50 hrs
  - Uses ~1G, and max 5G memory
  - Uses ~2M in storage

Same input

Combined output
SIMULATIONS ON HTC CLUSTERS, ANALYSES ON VM

- XSEDE
- OSG Connect
- UW HTC
- UA HPC
- CyVerse Atmosphere
- Google Drive
- Data storage, Analyses
- Data backup
- Simulations

Diagram:
- XSEDE connected to OSG Connect
- UW HTC connected to CyVerse Atmosphere
- UA HPC connected to CyVerse Atmosphere
- Data backup connected to CyVerse Data Store
- CyVerse Data Store connected to Google Drive
CHALLENGES: TECHNICAL

• How to handle millions of files?
  • UA HPC has file number limit
  • If there are too many files in a directory simple things take a long time
• How to not overload UA HPC system?
• How to reliably backup data?
• Why do jobs fail?
>1 MILLION SIMULATIONS OF EACH MODEL
MODEL CHOICE

Posterior probability: 0.0065 0.85 0.14
• ~1200 BCE ancestors of Jewish populations diverged from other Middle Eastern populations
  • Experienced extreme population size reduction
• ~1100 CE ancestors of Ashkenazi Jews diverged from other Jewish populations
  • Experienced another population size reduction
  • Experienced gene flow from Europeans (unresolved how much or when)
• ~1500 CE Eastern and Western Ashkenazi Jews diverged
  • Western AJ moderately grew in size
  • Eastern AJ massively grew in size
SIMPRILY: GENERALIZATION OF CODE AND WORKFLOW

- Developed program to simulate any demographic model
- Memory & space efficient
- Use Singularity container
- Pegasus workflow for OSG

https://agladstein.github.io/SimPrily/
THANK YOU!

HAMMER LAB
- Michael Hammer
- Consuelo Quinto-Cortes

OPEN SCIENCE GRID & PEGASUS
- Mats Rynge

UW CENTER FOR HTC
- Lauren Michael
- Christina Koch

OPEN SCIENCE GRID USER SCHOOL
- Tim Cartwright
- Lauren Michael
- Christina Koch

CODING MINIONS
- David Christy
- Logan Gantner
- Mack Skodiak
- Daniel Olson
- Rafael Lopez
- Kayleen Gurrola
- Katie McCready

RESOURCES PROVIDED BY
- University of Arizona HPC
- University of Wisconsin HTC
- CyVerse
- Open Science Grid
- XSEDE
  - Bridges
  - Comet
  - Jetstream

UA HPC CONSULTING
- Mike Bruck
- Dima Shyshlov

OPEN SCIENCE GRID & PEGASUS
- Mats Rynge

UW CENTER FOR HTC
- Lauren Michael
- Christina Koch
CPU HOURS ON THE OPEN SCIENCE GRID
DNA SEQUENCE

Indiv 1
AATCATTTCGGTTTTTAATGCTTTGGGCTGCATTGGGAAA
AATCATATCGGTCTTTAATGCTTTGGGCTGCATTGGGAAA
AATCATATCGGTCTTTAATGCTTTGGGCTGCATTGGGAAA
DNA SEQUENCE, SEGREGATING SITES

Indiv 1

AATCAT TTCCGGTGTTTAATGCTTTGGGCTGCAATGGGAAA
AATCAT ATCGGTCTTAATGCTTTGCGCTGCCCTTGGTAAA
DNA SEQUENCE,  
SEGREGATING SITES

Indiv 1
AATCATTTCGGTTTTAATGCTTGGGCTGCATTGGGAAA  
AATCATATCGGTCTTAATGCTTGCGCTGCCTTGGTAAA

Indiv 2
AAACATTTCCGTCTTTATGGTTGCGCTGCATTGGGAA  
AAACATTTCGGTCTTTAATGCTTGGGCTGCTGCCTTGGTAAA
<p>| Indiv 1 | AATCATTTCGGTTTTAATGCTTTGCGCTGCAATTGGGAATAA | 00000000000000000000000000000100000000 |
| Indiv 2 | AAACATTTCGTCCTTTATGGTTGCGCTGCAATTGGGAATAA | 0010000001001001000100000000000110000000000 |
| Indiv 1 | GATCATTTTCGTTTTTAATGCTTTGCGCTGCAATTGGGAATAA | 00000000000000000000000000000100000000 |
| Indiv 2 | AAACATTTCGTCCTTTATGGTTGCGCTGCAATTGGGAATAA | 0010000001001001000100000000000110000000000 |</p>
<table>
<thead>
<tr>
<th>Indiv 1</th>
<th>0000001010</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0101000100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Indiv 2</th>
<th>0000000100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1011111011</td>
</tr>
</tbody>
</table>
• Inherited from lab mates
• Intended for millions of relatively small simulations
  • 1,389 10kb regions
  • 65 individuals
• Originally took a few minutes to run
• Originally ran parallel on U of A HPC
  • 1 million runs would take approximately 1 month.
PROFILE OF PYTHON SCRIPT

Minimum Simulation Parameters

Maximum Simulation Parameters

*Note different scales
Max memory < 6G goal
Can now run efficiently in parallel

*Note different scales*