UNIVERSITY OF MICHIGAN

Rapid Indexing and Searching of Genomes

Eli Ben-Michael
Cynthia Chen

Biostatistics Big Data Summer Institute

Symposium on Big Data, Human Health, and Statistics

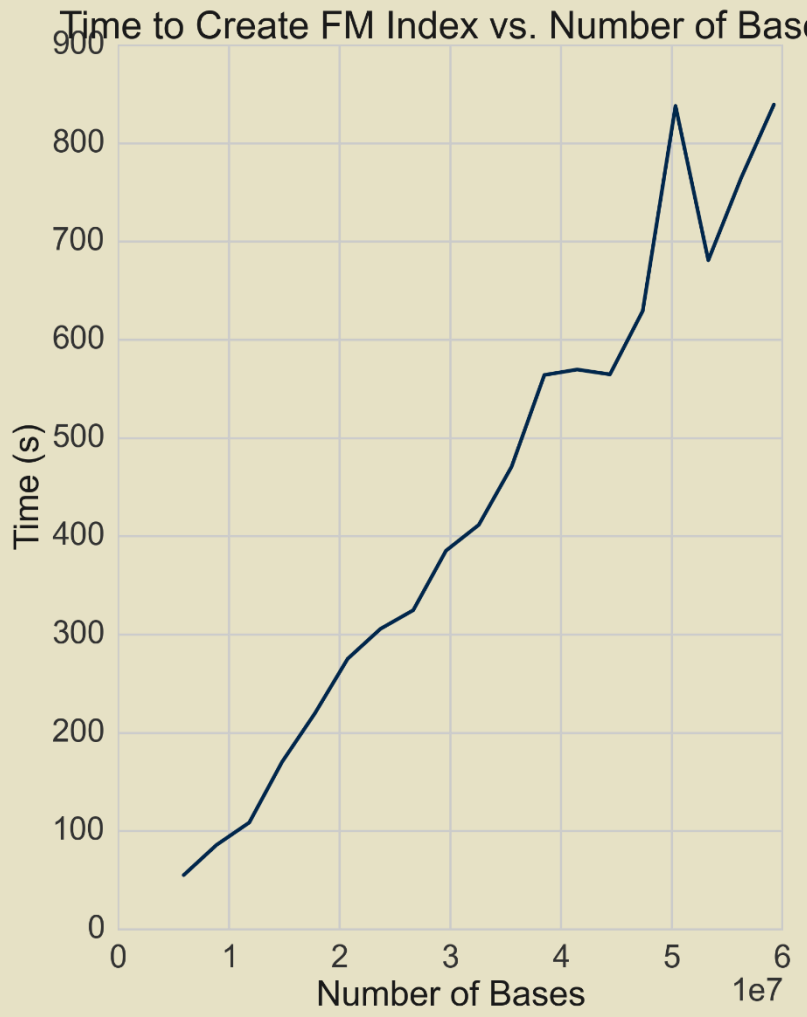
Department of Biostatistics, University of Michigan

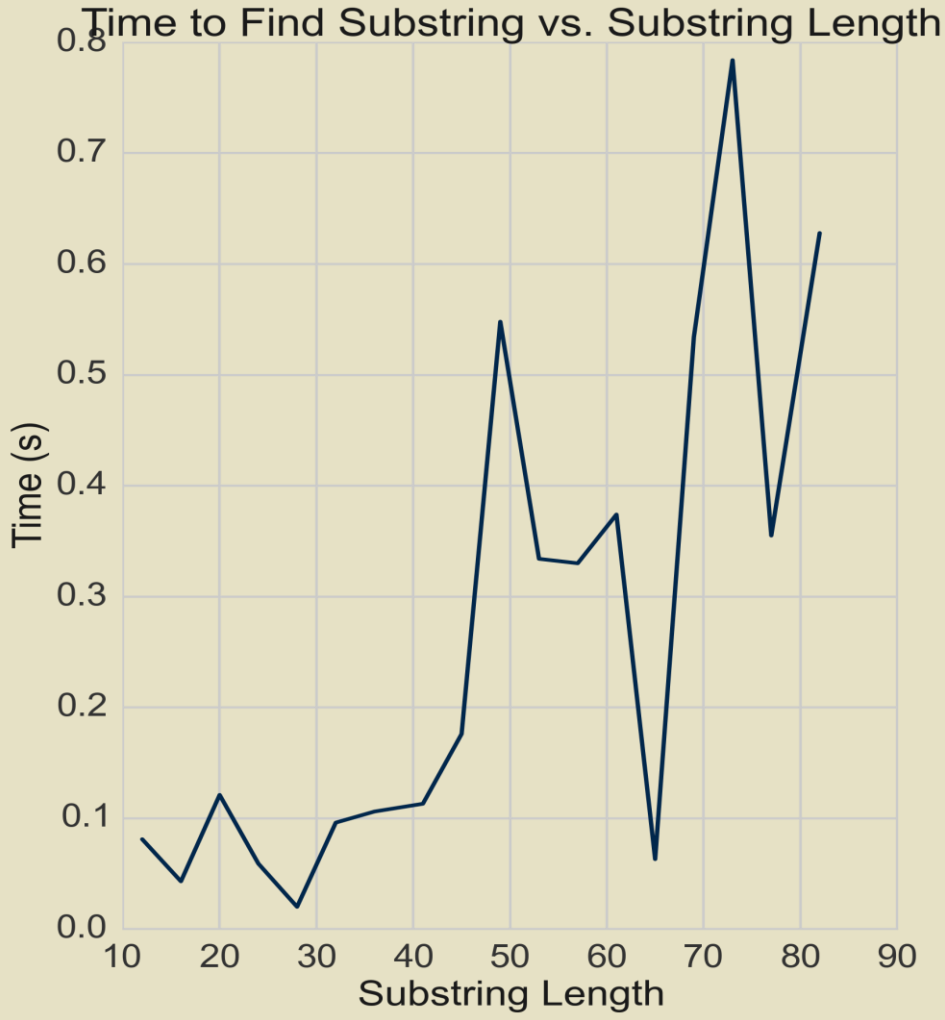
FM Indexes

UNIVERSITY OF MICHIGAN

- Genome is long
  - ~ 3 billion base pairs
  - How can we do efficient search on the genome?
  - How can we compress the genome?
- FM Indexing
  - Places characters with similar right contexts together
  - Efficient O(1) search
  - Good for compression (Huffman Coding)

Results (1)


Time to Create FM Index vs. Number of Bases
Time (s)
Number of Bases
1e7
0
100
200
300
400
500
600
700
800
900
0
1
2
3
4
5
6

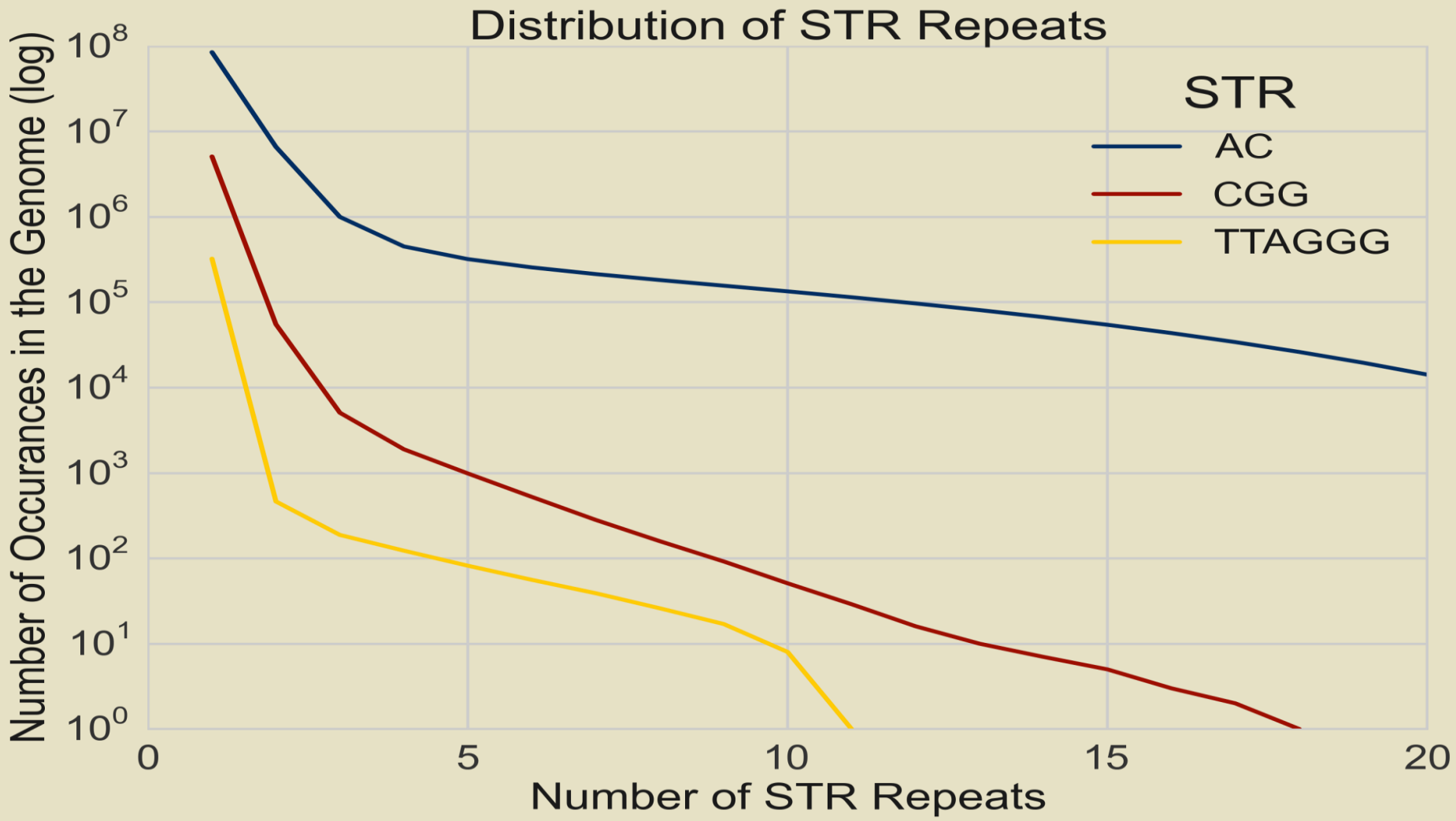

Time to Find Substring vs. Substring Length
Time (s)
Substring Length
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
10
20
30
40
50
60
70
80
90

Short Tandem Repeats (STRs)

UNIVERSITY OF MICHIGAN

- STRs
  - E.g.
    - “AC”: Association with breast cancer
    - “CGG”: Repeat varies based on gender
    - “TTAGGG”: Telomere sequence
  - 2-13 nucleotides repeated consecutively on a strand of DNA
  - Important to understanding genetic variation in humans
  - Little is known about the frequencies

Results (2)


Distribution of STR Repeats
Number of Occurances in the Genome (log)
10^8
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0
0
5
10
15
20
Number of STR Repeats
STR
AC
CGG
TTAGGG

Acknowledgments

- Prof. Goncalo Abecasis
- Prof. Hyun Min Kang
- Sayantan Das
- Prof. Bhramar Mukherjee
- University of Michigan School of Public Health