

PROJECT 6: SHORT TANDEM REPEATS

Daniel Duan

SHORT TANDEM REPEATS
MICROSATELLITES
SIMPLE SEQUENCE REPEATS

WHAT IS A STR?

TACATGAGATC**ATGATGATGATGATG**GAGCTGTGAGATC

SIGNIFICANCE

- Molecular markers - kinship, population, etc.
- Gene duplication / deletion - diseases
- Marker assisted selection - breeding
- DNA fingerprinting - identification

PROJECT GOAL

- Locate STRs in short reads
- Find the number of repeats at each STR location for an individual

My Genome: 5 repeats

TACATGAGATC**ATGATGATGATGATG**GAGCTGTGAGATC

A Sequence Read:

GATCATGATGATGATGATGG

Human Genome: 5 repeats

TACATGAGATC**ATGATGATG**GAGCTGTGAGATC
GATCATGATGATGATGATGG

DATASET

- Reads - length 30
- Genome - length 1 million
- Downloaded from cm124.herokuapp.com

ALGORITHM

GATCATGATGATGATGATGG

FIND REPEATS IN READ

**GATCATGATGATGATGATGATGG
GATCATGATGATGATGATGATGG**

FIND REPEATS IN READ

GATCATGATGATGATGATGG
GATCATGATGATGATGATGG

The diagram illustrates sequence alignment between two DNA reads. The top read is 'GATCATGATGATGATGATGG' and the bottom read is 'GATCATGATGATGATGATGG'. They are aligned such that the 'GATCAT' prefix of both reads is perfectly matched. The remaining 'GATGATGATGATGG' portion of the top read is aligned with the 'GATGATGATGATGG' portion of the bottom read. This alignment shows a 4-repeat overlap. The overlapping 'GAT' units are highlighted with gray rectangular boxes. The first 'GAT' is highlighted in the top read, and the subsequent three 'GAT' units are highlighted in the bottom read, demonstrating how a single repeat in one read can align with multiple repeats in another.

FIND REPEATS IN READ

GATCATGATGATGATGATGG
GATCATGATGATGATGATGG

FIND REPEATS IN READ

GATCATGATGATGATGATGATGG
GATCATGATGATGATGATGATGG

STORE INFORMATION



STORE INFORMATION



“GATCATGG” : [(5)]

COMPUTE DATA

“GATCATGG” : [(5, 5, 5, 5, 5, 5, 5, 5, 2, 5, 5, 5, 5, 5, 5, 5)]



“GATCATGG” : [(5)]

FIND REPEATS IN GENOME

TACATGAGAT**CATGATGATGG**AGCTGTGAGATC
GATCATG G



“GATCATGG” : (1, 12)

MAP READ TO GENOME

“GATCATGG” : (5)

“GATCATGG” : (1, 12)



1, ATG, 5, 12

PROJECT SETUP

- Javascript
- File In / STD Out
- Node.js
- Google V8 Engine
- Header length 6
- Tail length 2



RESULTS

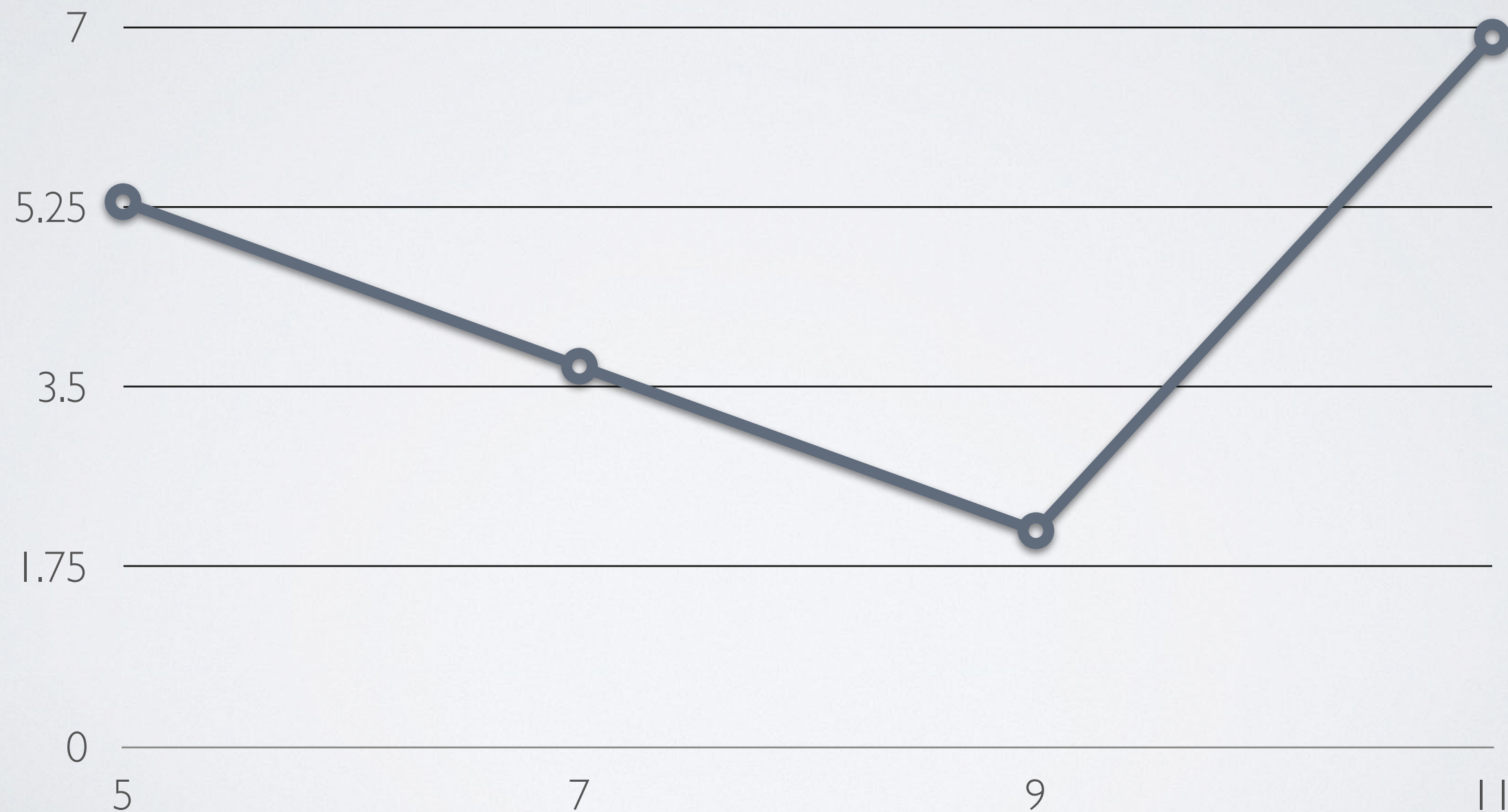
- Don't ever use JS for IO
- 20 min runtime - optimized
- Too long - unoptimized
- 5% unmatched STR
- 250+MB memory footprint
- Bugs



RESULTS

○ % Unmatched STRs

Tag Length vs Unmatched Reads



NEXT STEPS

- Use a different language
- Catch edge read cases
- Reduce memory footprint
- Buffered IO
- Multithreaded processing



“Thank you.”