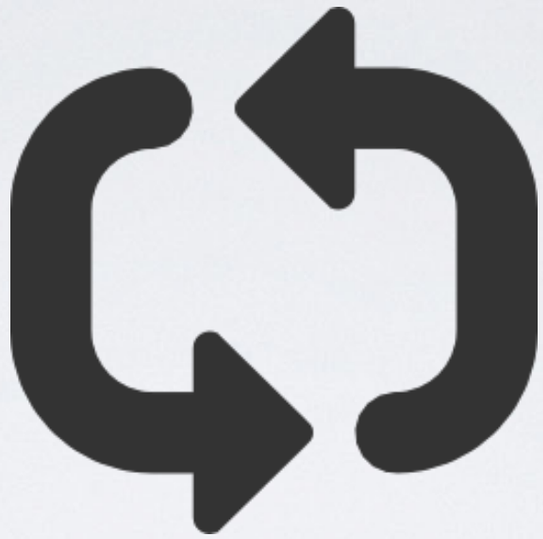


PROJECT 6: SHORT TANDEM REPEATS

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

GATCATGATGATGATGATGATGG

DATASET

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

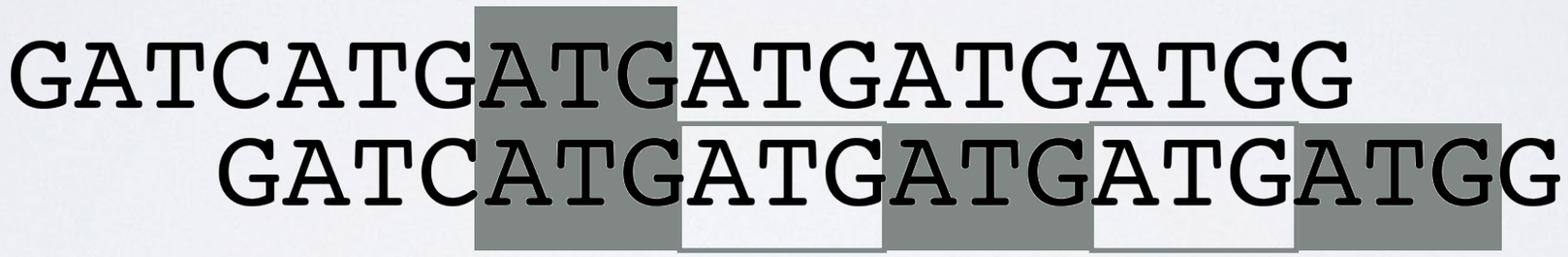
ALGORITHM

GATCATGATGATGATGATGATGG

FIND REPEATS IN READ

```
GATCATGATGATGATGATGG
   GATCATGATGATGATGATGG
```

FIND REPEATS IN READ



FIND REPEATS IN READ

```
GATCATGATGATGATGATGG
    GATCATGATGATGATGATGG
```

FIND REPEATS IN READ

```
GATCATGATGATGATGATGG
      GATCATGATGATGATGATGG
```

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

TACATGAGATC<u>**ATGATGATG**</u>**GAGCTGTGAGATC**

GATCATG **G**

"GATCATGG" : (1, 12)

MAP READ TO GENOME

"GATCATGG" : (5)

"GATCATGG" : (1, 12)

1, ATG, 5, 12

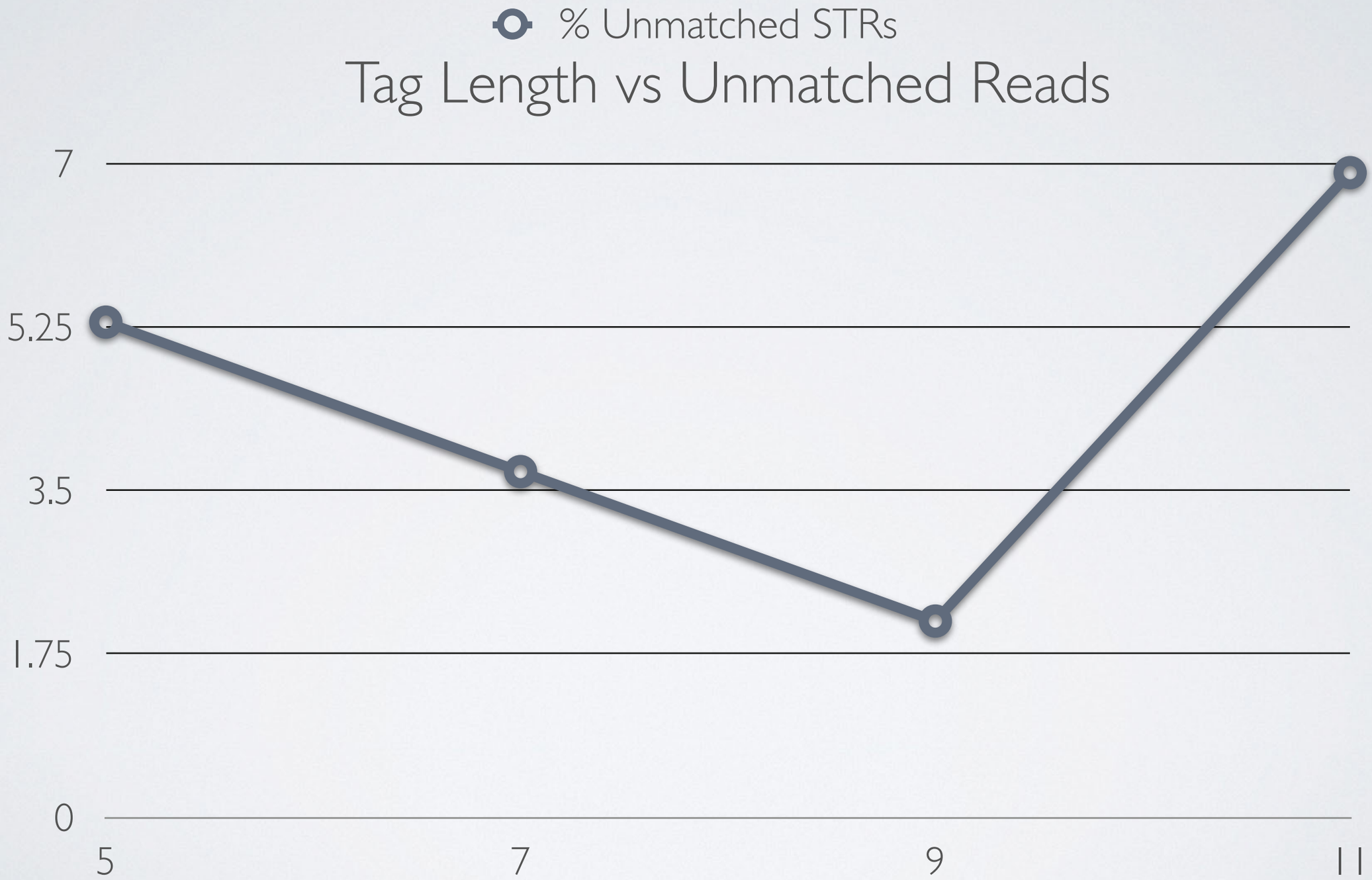
PROJECT SETUP

- 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
7
5.25
3.5
1.75
0
5
7
9
11

NEXT STEPS

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

“Thank you.”