Haplotype Assembly

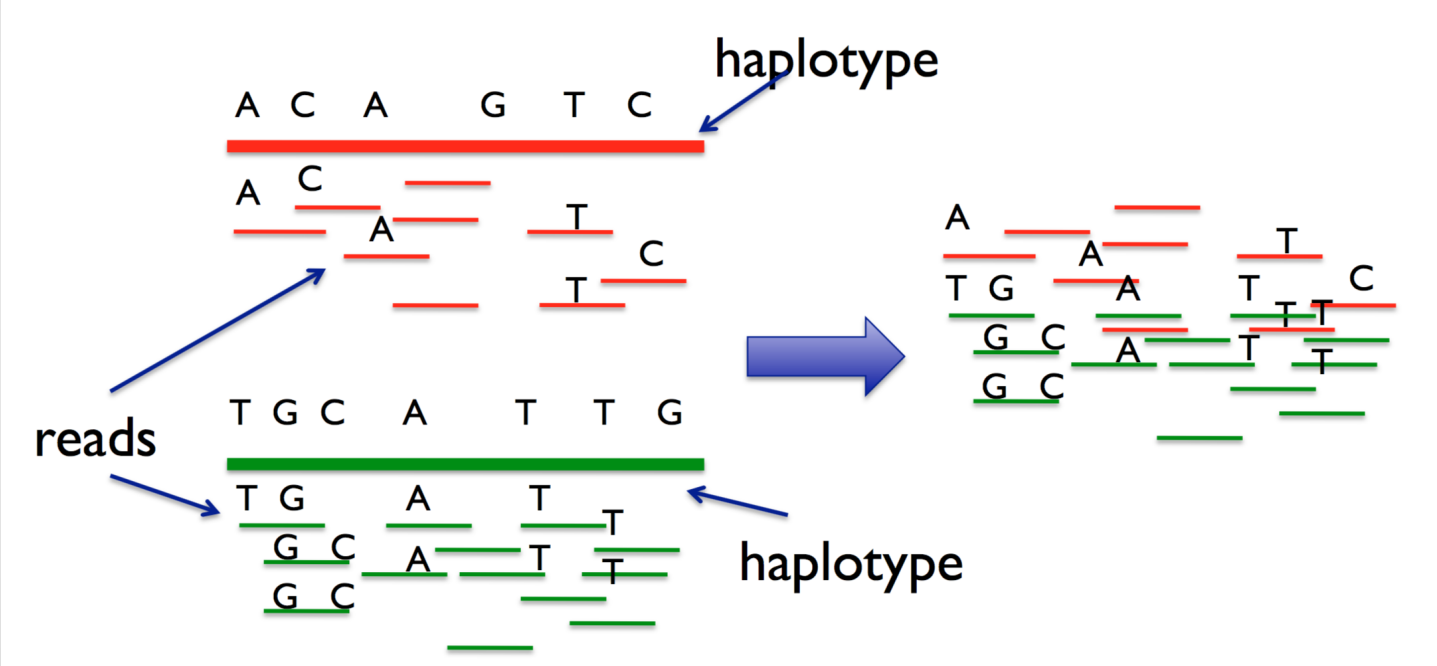
Daniel Norman

Motivate the Problem

- Haplotype information is needed
  - Haplotypes differentiate humans from each other

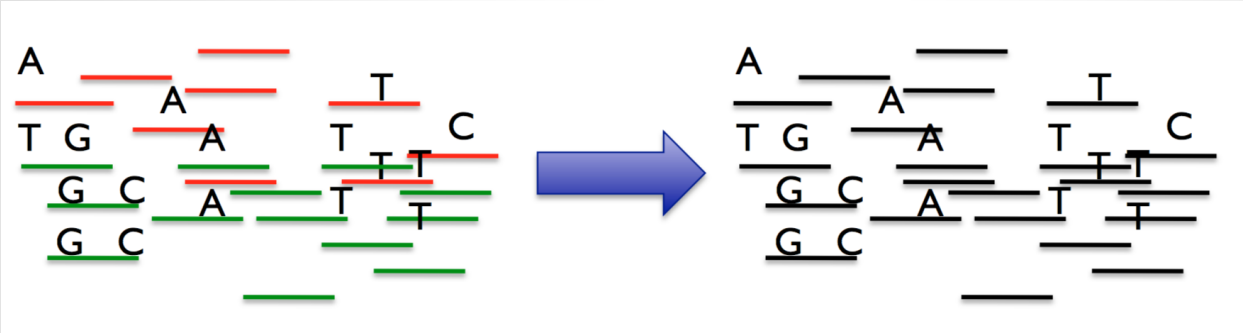
- Machines can read DNA to produce chunks of the haplotypes, but unsure which chromosome (which haplotype) a read is from

Motivate the Problem



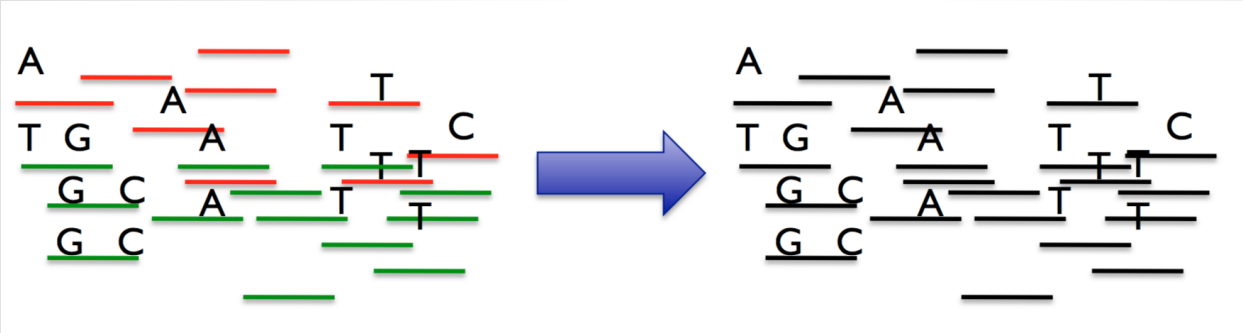
Images from Dr. Eleazar Eskin, UCLA

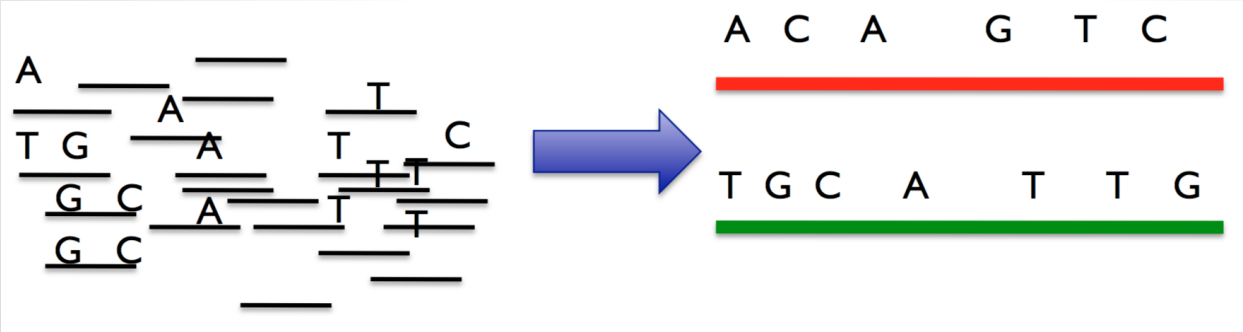
Motivate the Problem



Images from Dr. Eleazar Eskin, UCLA

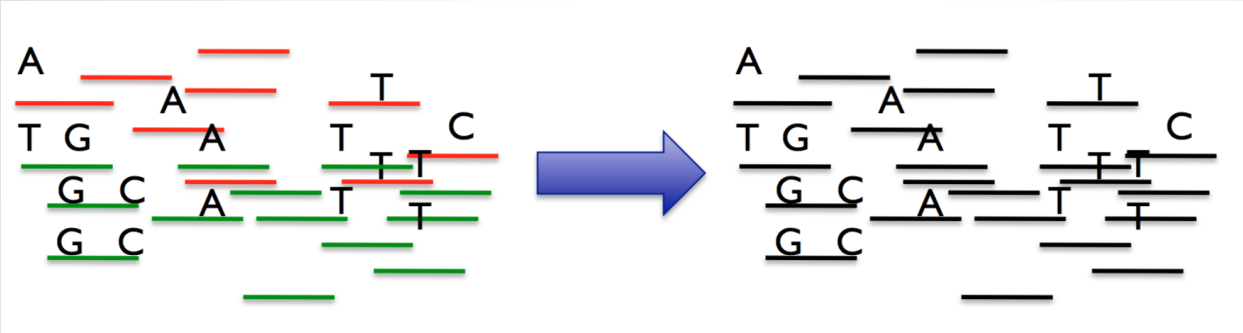
Motivate the Problem

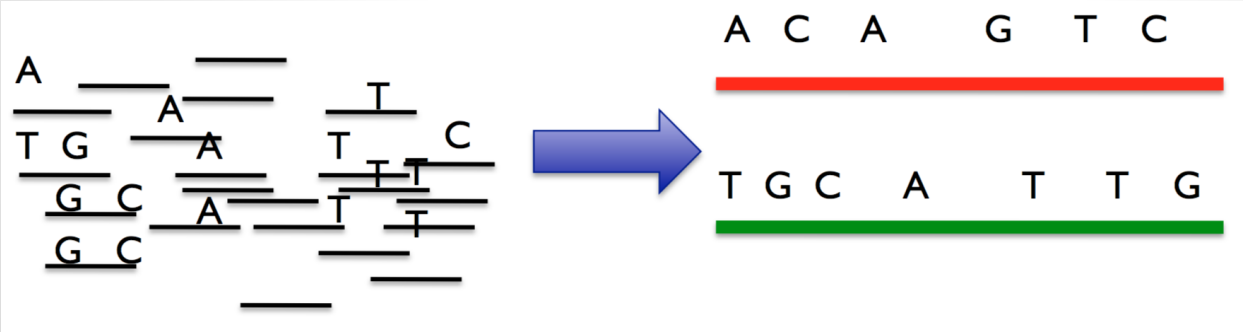




Images from Dr. Eleazar Eskin, UCLA

Motivate the Problem





END GOAL!

Images from Dr. Eleazar Eskin, UCLA

Computational Problem

Homozygous Sites

ATACGGCTAGATTC

ATGCGGTTAGCTTT

Heterozygous Sites

Computational Problem

Homozygous Sites

```
ATACGGCTAGATTC
ATGCGGTTAGCTTT
```

Heterozygous Sites

```
__0___1___1__0
__1___0___0__1
```

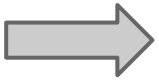
0: Minor allele
1: Major allele

Computational Problem

```
__0___1___1__0
__1___0___0__1
```

Computational Problem

```
__0___1___1__0
__1___0___0__1
```



```
0110…
1001…
```

Not recording homozygous SNPs?

Computational Problem

How do we get from reads...

```
01101
 1101
 00101
  1010
   10100
```

...

Computational Problem

How do we get from reads...

```
01101
 1101
 00101
  1010
   10100
```

...

to haplotypes?

```
01101011...
10010100...
```

Benchmarks

- Speed
- Accuracy (Switch Distance)

Benchmarks

- Speed
- Accuracy (Switch Distance)

Baseline - Easy Project Algorithm

Assume no errors in reads

Baseline - Easy Project Algorithm

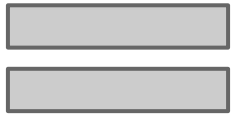
Assume no errors in reads

Go through all reads,
placing them in the
haplotype they match

Baseline - Easy Project Algorithm

Assume no errors in reads

Go through all reads, placing them in the haplotype they match



Simple and fast!

Easy Algorithm Visualization

```
01101                H1:
 1101                H2:
   10110
   01001
     11011
```

Easy Algorithm Visualization

```
01101
 1101
   10110
   01001
     11011
```

```
H1: 01101
H2:
```

Easy Algorithm Visualization

```
01101
 1101
   10110
   01001
     11011
```

H1: 01101

H2: 10010

Easy Algorithm Visualization

```
01101
 1101
   10110
   01001
     11011
```

```
H1: 01101
H2: 10010110
```

Easy Algorithm Visualization

```
01101
 1101
   10110
   01001
     11011
```

```
H1: 01101001
H2: 10010110
```

Easy Algorithm Visualization

```
01101
 1101
   10110
   01001
     11011
```

```
H1: 01101001
H2: 1001011011
```

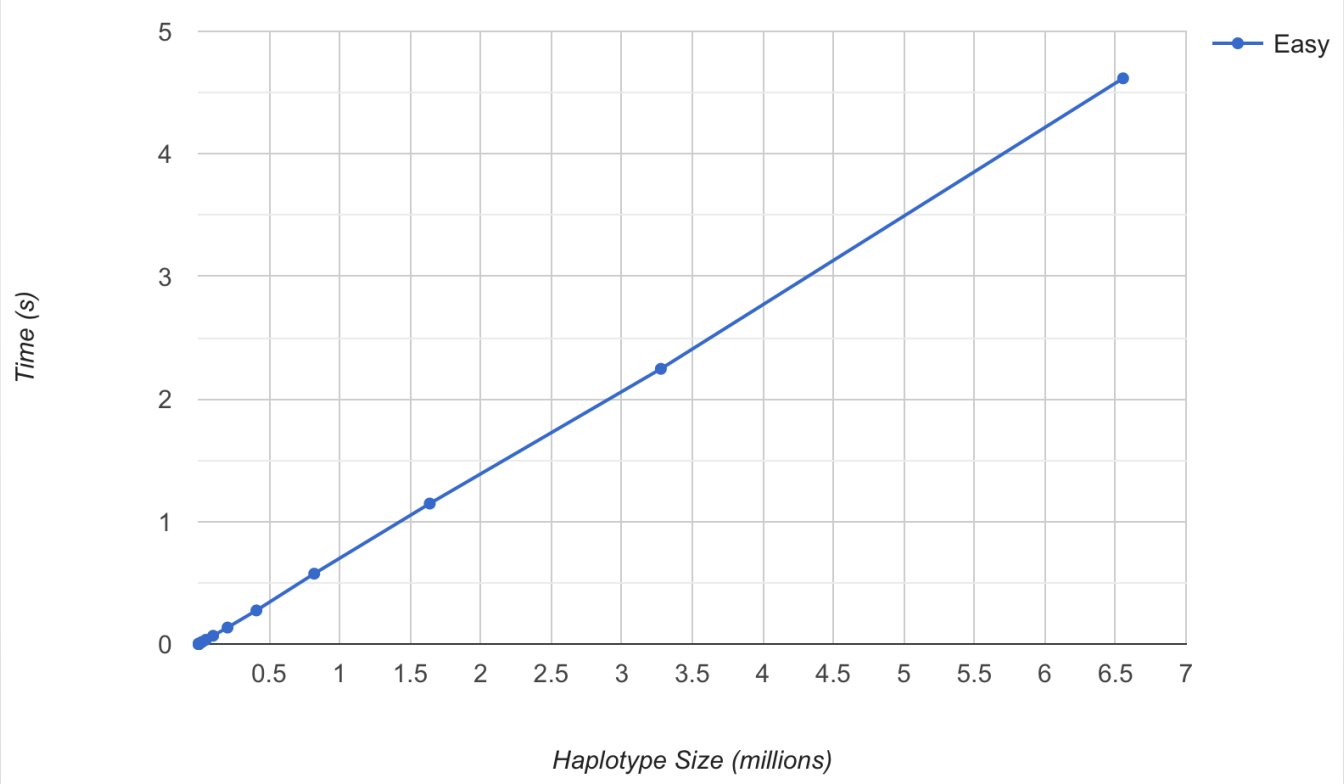
Easy Algorithm Visualization

```
01101
 1101
   10110
   01001
     11011
```

```
H1: 0110100100
H2: 1001011011
```

ALL DONE!

Easy Algorithm Runtime



Medium Project Algorithm

Allow for errors in reads

Medium Project Algorithm

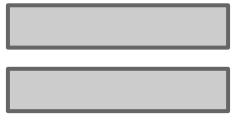
Allow for errors in reads

Partition reads into two distinct groups based on a matching criteria, then reassemble

Medium Project Algorithm

Allow for errors in reads

Partition reads into two distinct groups based on a matching criteria, then reassemble



Accurate!

Medium Algorithm Visualization

```
A: 01101
B:  1100
C:   010110
D:    00000
E:      11011
F:       10101
```

H1:

H2:

Medium Algorithm Visualization

```
A: 01101
B:  1100
C:   010110
D:    00000
E:      11011
F:       10101
```

H1: A

H2:

Medium Algorithm Visualization

```
A: 01101
B:  1100
C:   010110
D:    00000
E:      11011
F:       10101
```

H1: A, B

H2:

Medium Algorithm Visualization

```
A: 01101
B:  1100
C:   010110
D:     00000
E:       11011
F:        10101
```

```
H1: A, B
H2: C
```

Medium Algorithm Visualization

```
A: 01101
B:  1100
C:   010110
D:     00000
E:       11011
F:        10101
```

H1: A, B

H2: C

Medium Algorithm Visualization

```
A: 01101
B:  1100
C:   010110
D:     00000
E:       11011
F:        10101
```

```
H1: A, B
H2: C, E
```

Medium Algorithm Visualization

```
A: 01101
B:  1100
C:   010110
D:     00000
E:       11011
F:        10101
```

```
H1: A, B
H2: C, E, F
```

Medium Algorithm Visualization

Now reassemble one haplotype, SNP by SNP

Example: SNP 5 on Haplotype 1

```
A: 01101          H1, SNP 5:
B:  1100
C:   010110
```

Medium Algorithm Visualization

Now reassemble one haplotype, SNP by SNP

Example: SNP 5 on Haplotype 1

```
A:  01101
B:   1100
C:    010110
```

```
H1, SNP 5:
  = [1+0+flip(0)]/3
```

Medium Algorithm Visualization

Now reassemble one haplotype, SNP by SNP

Example: SNP 5 on Haplotype 1

```
A:  01101          H1, SNP 5:
B:   1100            = [1+0+flip(0)]/3
C:    010110         = [1+0+1]/3
```

Medium Algorithm Visualization

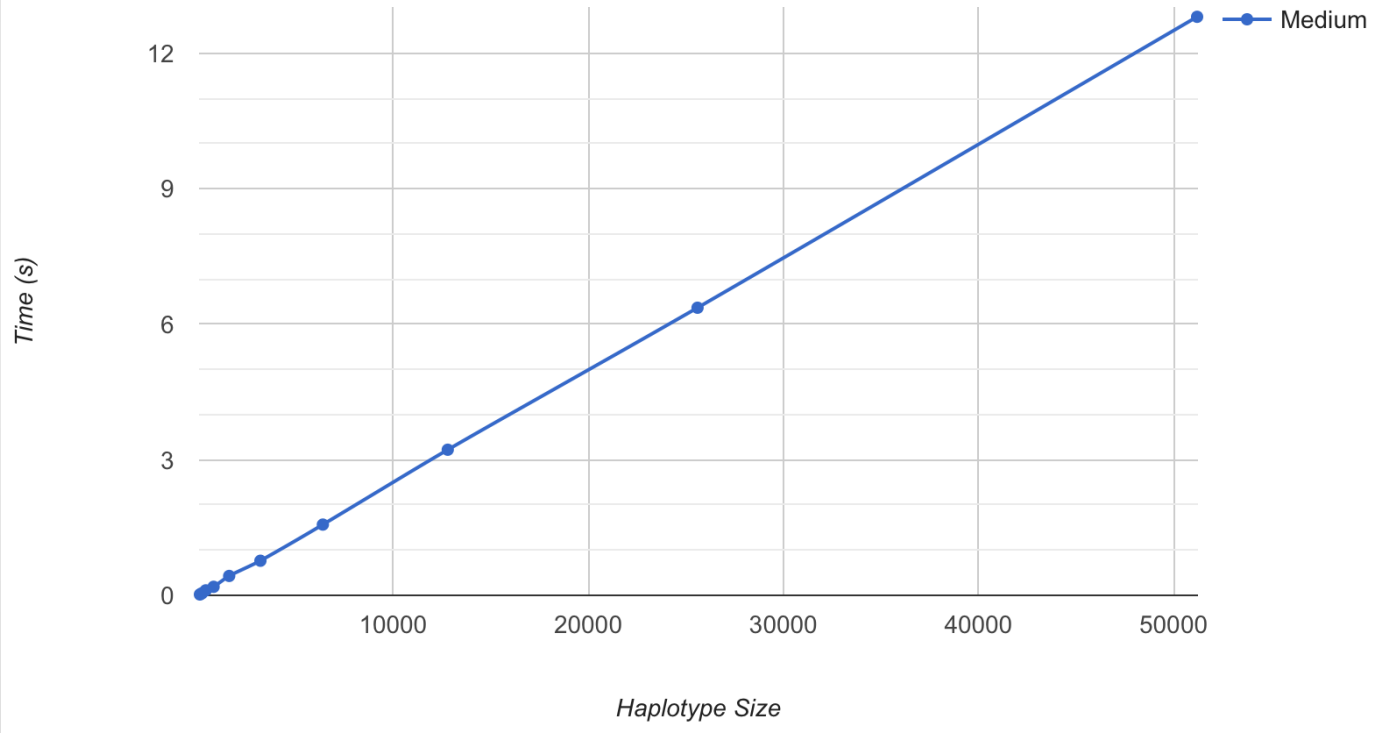
Now reassemble one haplotype, SNP by SNP

Example: SNP 5 on Haplotype 1

```
A:  01101
B:   1100
C:    010110
```

```
H1, SNP 5:
   = [1+0+flip(0)]/3
   = [1+0+1]/3
   = 1
```

Medium Algorithm Runtime



Time for Haplotype of Size 50,000

Easy: 0.035s

Medium: 12.8s

Time for Haplotype of Size 50,000

Easy: 0.035s

Medium: 12.8s

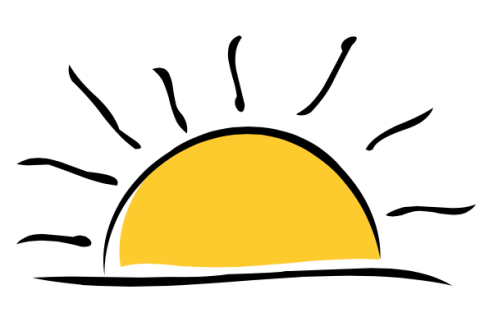
Medium is 365 times as slow as Easy

Time for Haplotype of Size 50,000

Easy: 0.035s

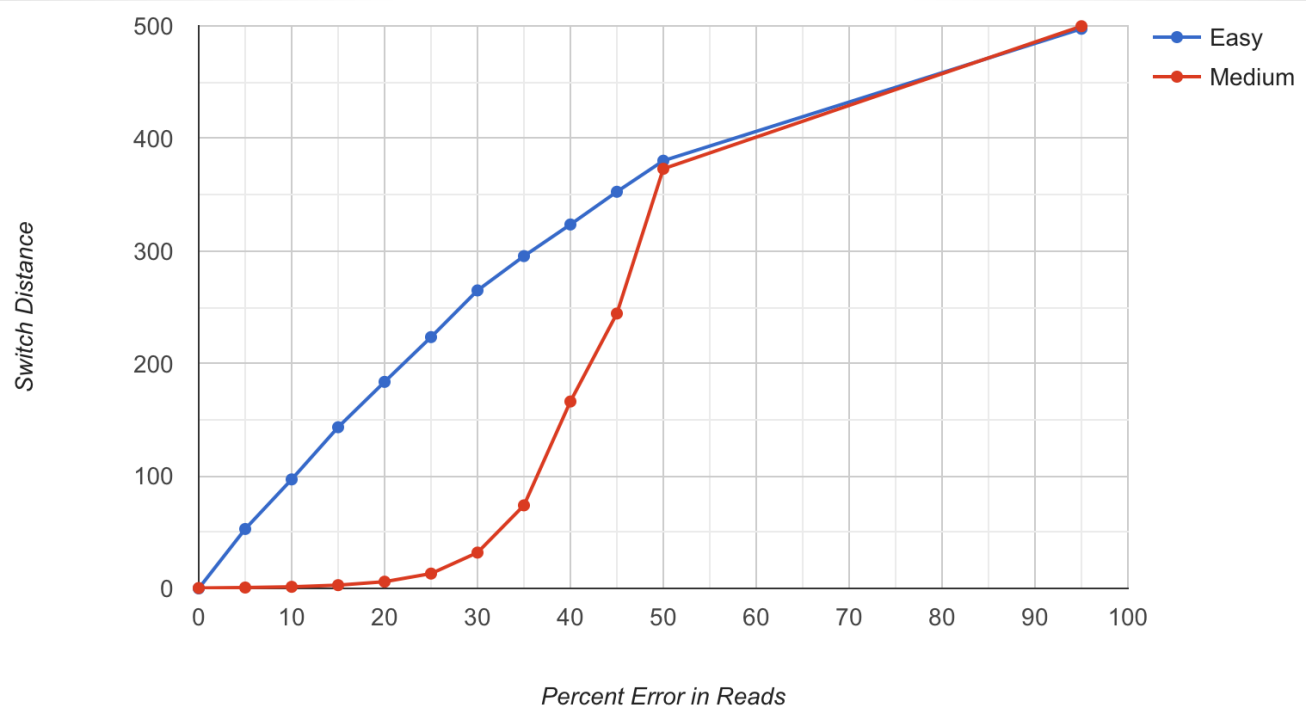
Medium: 12.8s

Medium is 365 times as slow as Easy



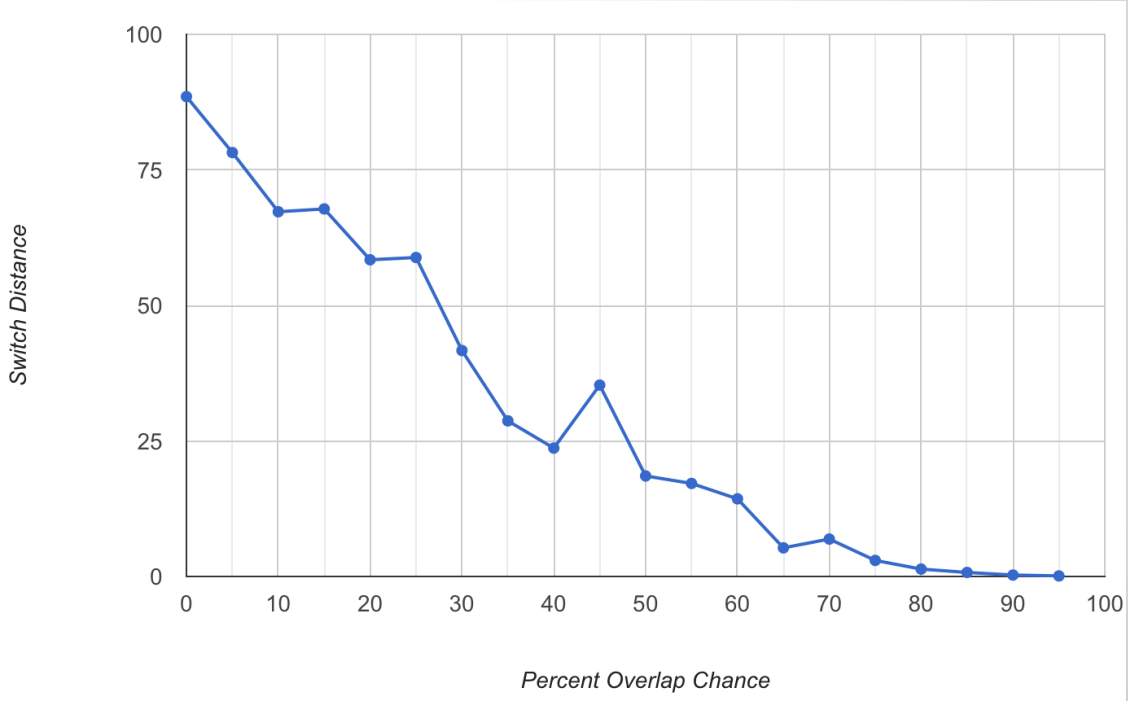
Accuracy - Easy vs Medium

Haplotype length: 1000 SNPs



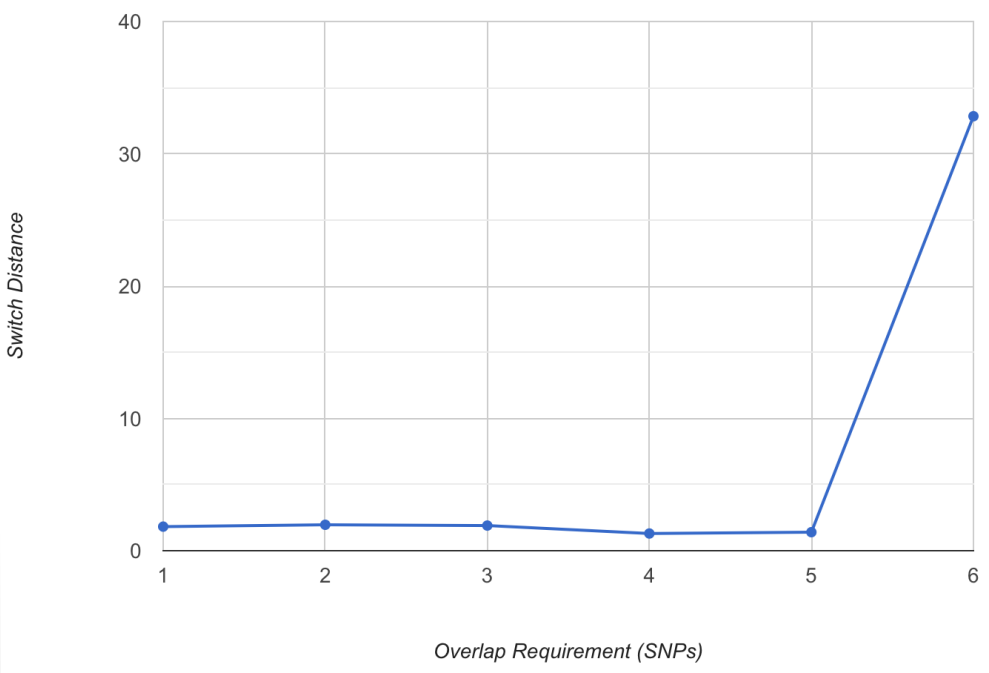
Overlap Chance in Reads

```
101101
0101
  00101
  011011

Overlap Chance
  = 50%
```



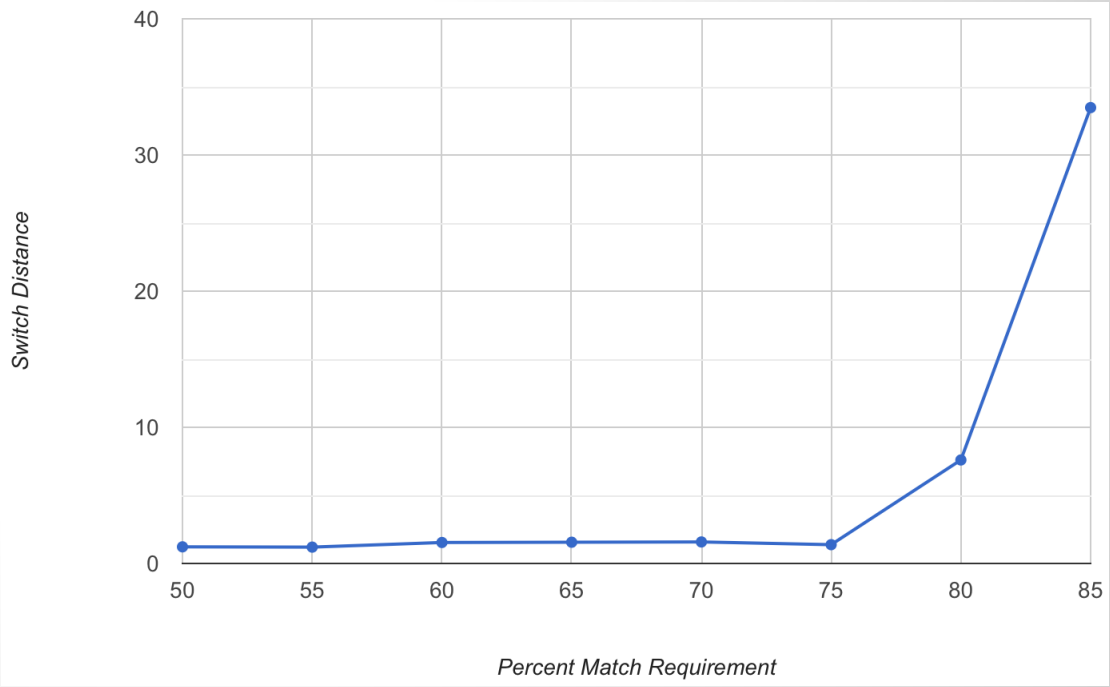
Overlap Requirement

```
101101
  100101

Overlap = 4
```



Match Requirement

```
101101
  100101

Matches = 75%
```



Something Interesting

- Medium gets slightly worse accuracy than Easy at very high error rates
  - Medium requires reads to meet match criteria
  - If few meet criteria, it has to guess often
  - Easy will place reads into a haplotype no matter what

- Hard algorithm working
  - Requires very low error rate and lots of overlap

Thank You!