

# Personal Genomic Toolchain

# Why this matters

- Genetics can increase risk factors for heart disease, one of the world's leading causes of illness and death.

# Why this matters

- Genetics can increase risk factors for heart disease, one of the world's leading causes of illness and death.
- Personal Genomics is already a \$10 billion industry as of 2013.

# Why this matters

- Genetics can increase risk factors for heart disease, one of the world's leading causes of illness and death.
- Personal Genomics is already a \$10 billion industry as of 2013.
- Potentially increase mobility of diagnostic tools

# Project Goals

- Process raw sequence data in minutes

# Project Goals

- Process raw sequence data in minutes
- Reduce resources required

# Project Goals

- Process raw sequence data in minutes
- Reduce resources required
- Increase confidence

# Project Goals

- Process raw sequence data in minutes
- Reduce resources required
- Increase confidence
- Identify common structures



# Project Distinctions

- Use standard tools to improve existing methods

# Project Distinctions

- Use standard tools to improve existing methods
- Takes advantage of locality

# Project Distinctions

- Use standard tools to improve existing methods
- Takes advantage of locality
- Use Python data structures intelligently

# Project Distinctions

- Use standard tools to improve existing methods
- Takes advantage of locality
- Use Python data structures intelligently
- Better algorithms mean faster analysis

# Project Results

- Naive processing: 1Mb (500k calls) in 40 s

# Project Results

- Naive processing: 1Mb (500k calls) in 40 s
- Multiprocessing MR: 10 Mb (5M calls) in 30 s

# Project Results

- Naive processing: 1Mb (500k calls) in 40 s
- Multiprocessing MR: 10 Mb (5M calls) in 30 s
- Further optimizations: 1 Gb (500M calls) in 10 s

# Project Results

- Naive processing: 1Mb (500k calls) in 40 s
- Multiprocessing MR: 10 Mb (5M calls) in 30 s
- Further optimizations: 1 Gb (500M calls) in 10 s
- Memory is now a limiting factor.



# Project Results

- Naive processing: 1Mb (500k calls) in 40 s
- Multiprocessing MR: 10 Mb (5M calls) in 30 s
- Further optimizations: 1 Gb (500M calls) in 10 s
- Memory is now a limiting factor.
- Identifies two behaviors based on confidence

# Other Applications

- Improve ngram intent analysis

# Other Applications

- Improve ngram intent analysis
- Enhance recommendation engines

# Other Applications

- Improve ngram intent analysis
- Enhance recommendation engines
- Aid co-presence cluster identification

# Other Applications

- Improve ngram intent analysis
- Enhance recommendation engines
- Aid co-presence cluster identification
- Inform prediction models

# Read Sequences

- Sequences of bases A, C, G and T

# Read Sequences

- Sequences of bases A, C, G and T
- Sequences consist of billions of base pairs.

# Read Sequences

- Sequences of bases A, C, G and T
- Sequences consist of billions of base pairs.
- Typical raw reads are ~60 Gb.



# Read Sequences

- Sequences of bases A, C, G and T
- Sequences consist of billions of base pairs.
- Typical raw reads are ~60 Gb.
- Human reference genome is ~3 Gb.

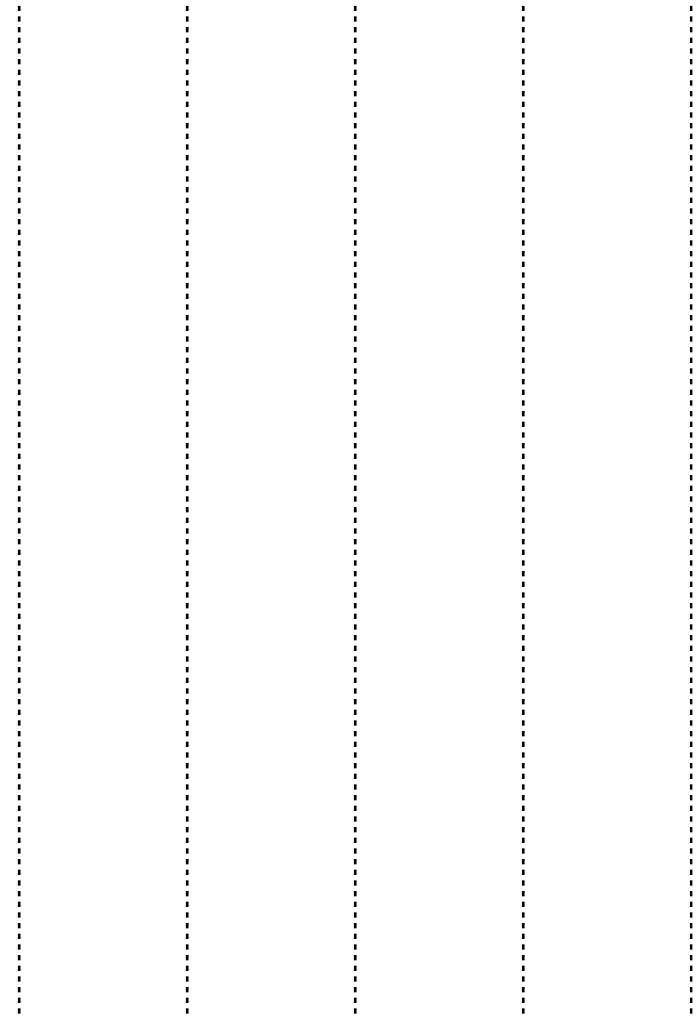
# Read Sequences

- Sequences of bases A, C, G and T
- Sequences consist of billions of base pairs.
- Typical raw reads are ~60 Gb.
- Human reference genome is ~3 Gb.
- Individual bases are difficult to distinguish.

Called word:

Confidence:

Partial word:



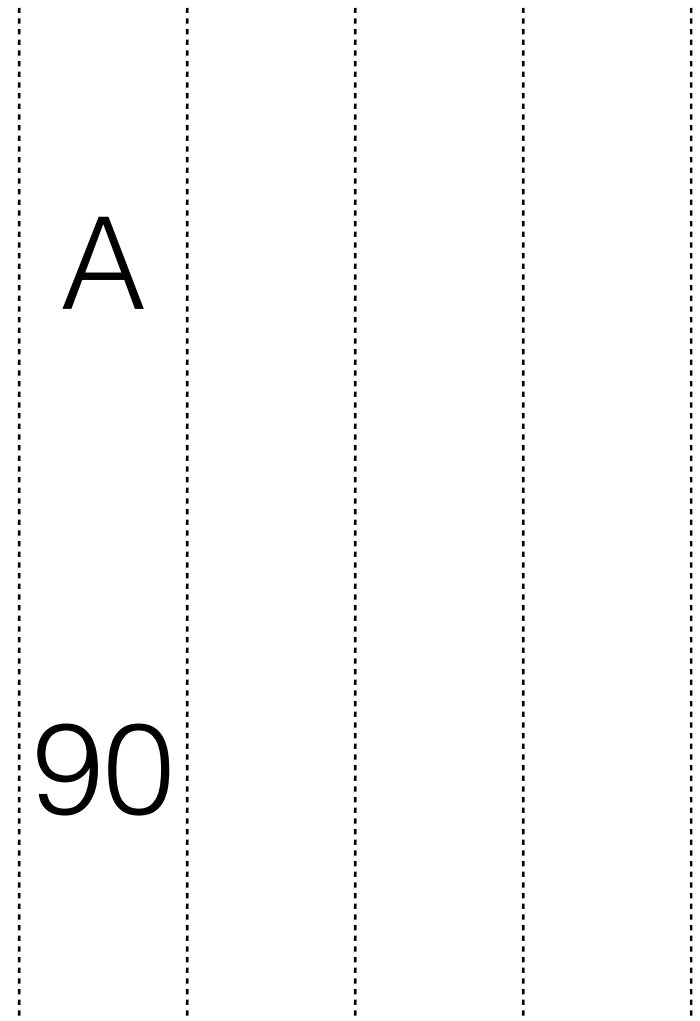
Called word:

A

Confidence:

90

Partial word:



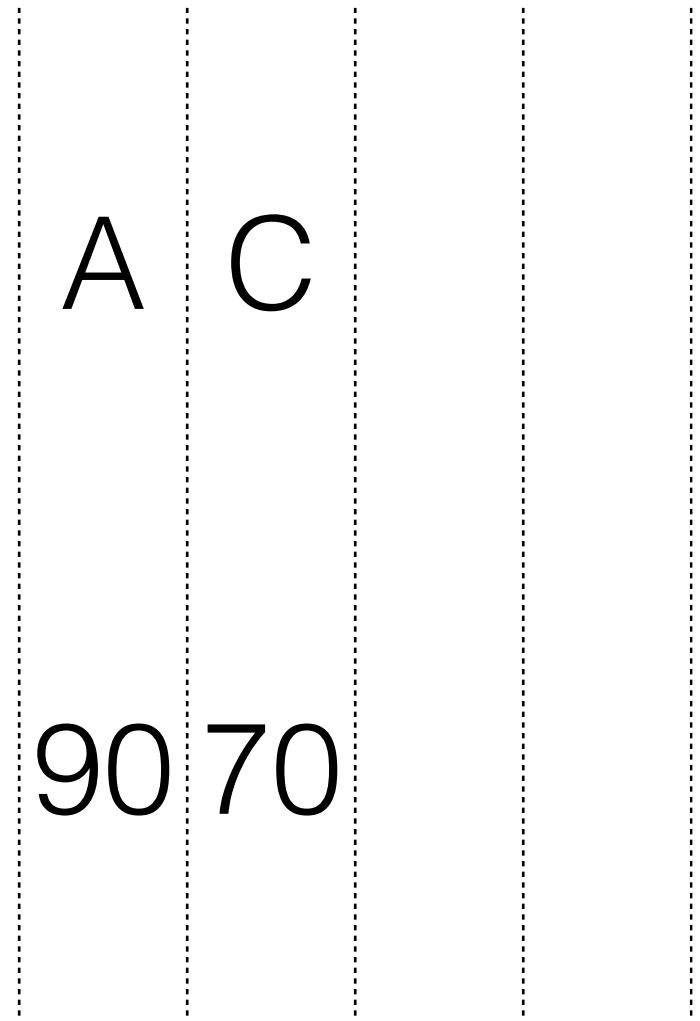
Called word:

A C

Confidence:

90 70

Partial word:



Called word:

A C G

Confidence:

90 70 20

Partial word:

Called word:

A C G T

Confidence:

90 70 20 80

Partial word:

Called word:

A C G T

Confidence:

90 70 20 80

Partial word:



Called word:

A C G T

Confidence:

90 70 20 80

Partial word:

A C \_ T

Called word:

A C G T

Confidence:

90 70 20 80

Partial word:

A C \_ T

# Observations

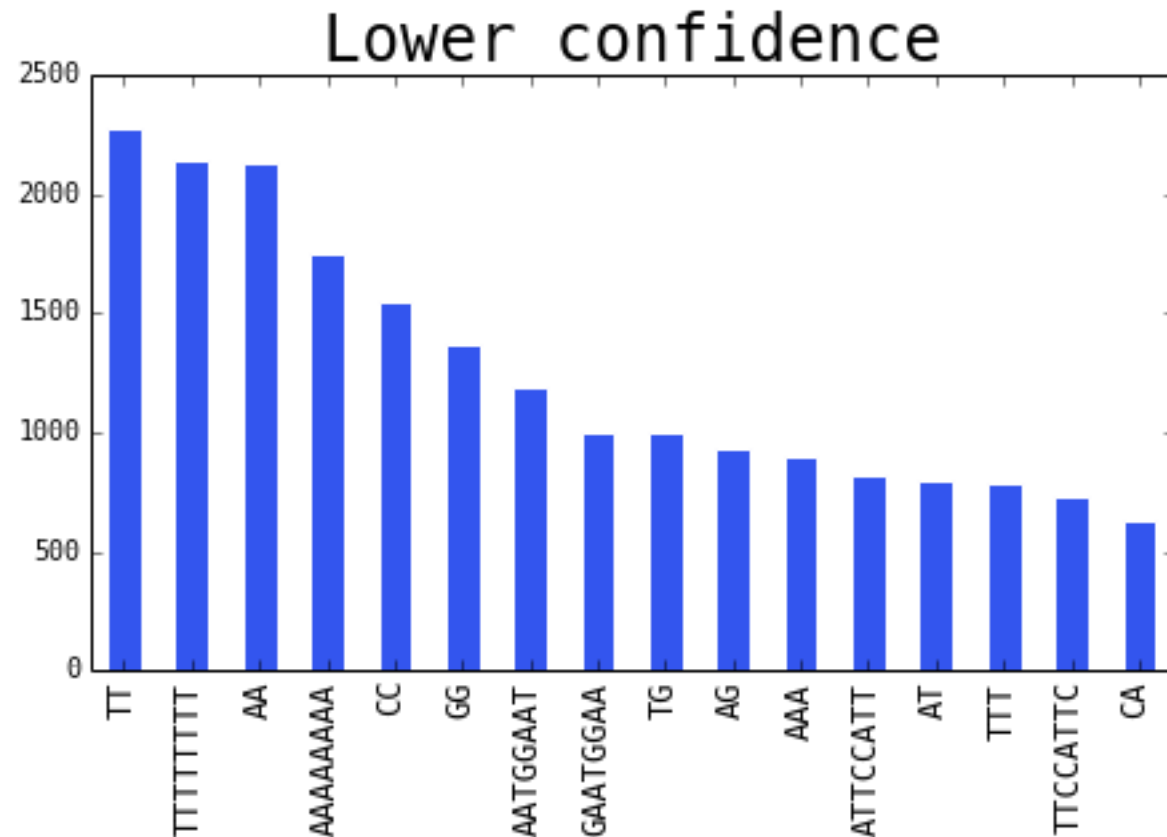
- At lower confidences, longer words are more likely.

# Observations

- At lower confidences, longer words are more likely.
- At higher confidence these words break up, becoming more infrequent.

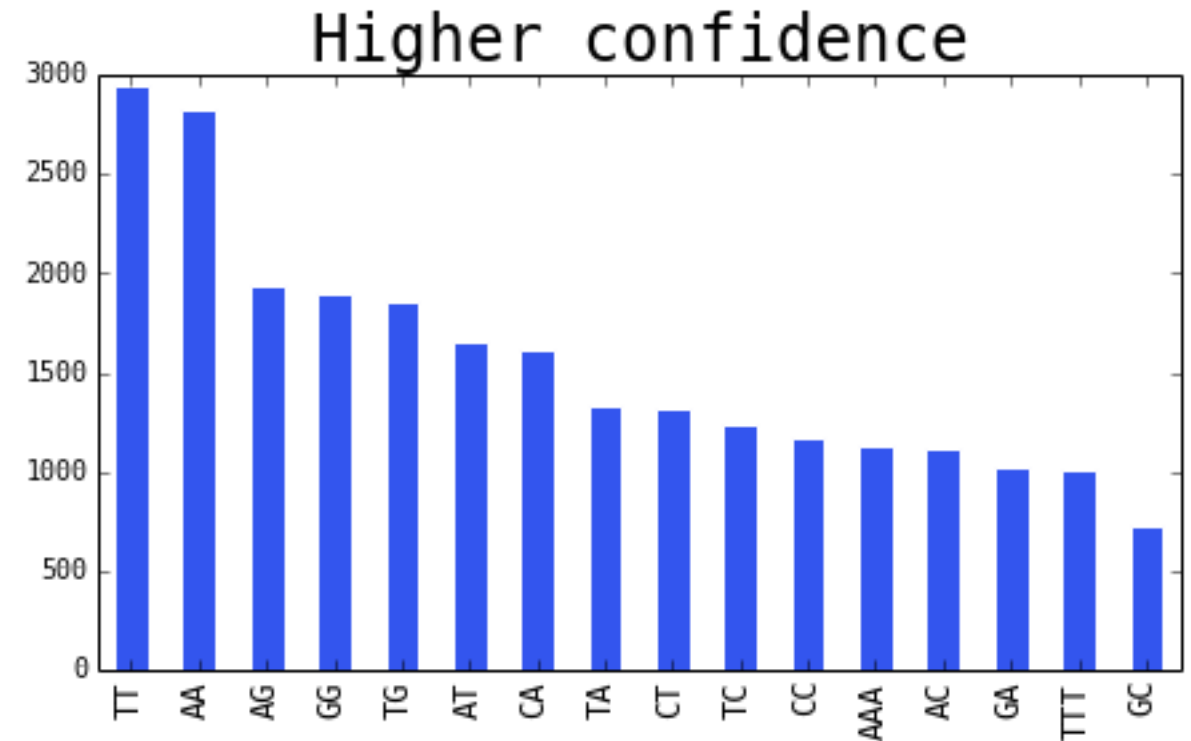
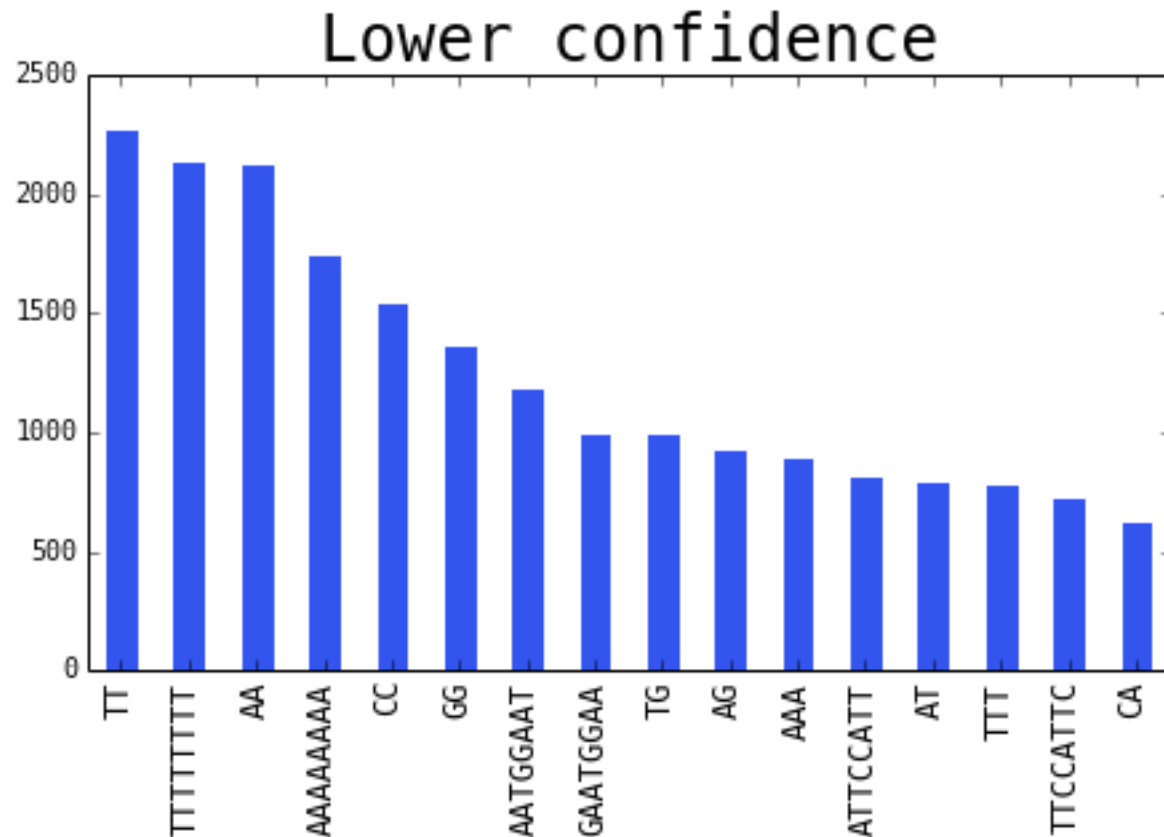
# Observations

- At lower confidences, longer words are more likely.
- At higher confidence these words break up, becoming more infrequent.



# Observations

- At lower confidences, longer words are more likely.
- At higher confidence these words break up, becoming more infrequent.



# Next steps

- Address memory limitations

# Next steps

- Address memory limitations
- Leverage cloud technologies eg AWS EMR



# Next steps

- Address memory limitations
- Leverage cloud technologies eg AWS EMR
- Implement genome alignment

Thank you!