Statistical Methods for Genome Wide Regional Analysis with Next Generation Sequencing Data

Hao Wu, Emory University

and

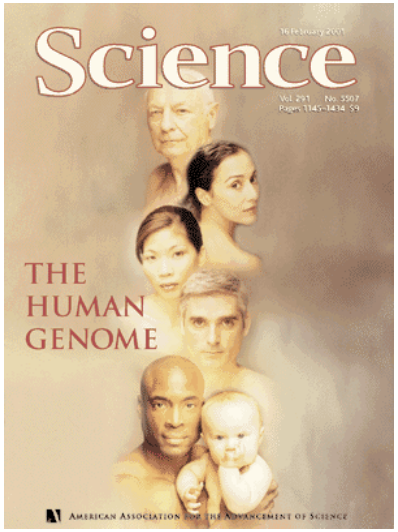
Rafael A Irizarry, DFCI/Harvard

Outline

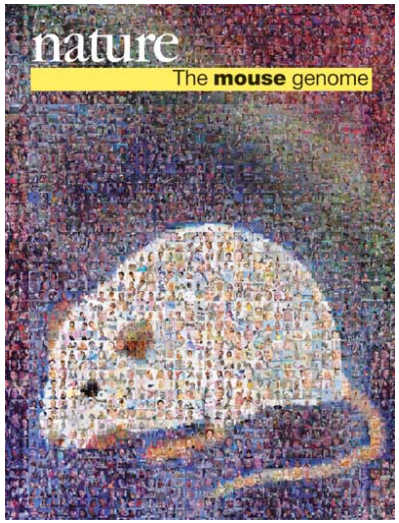
- Introduction to Next Generation Sequencing (NGS)
- Motivation for region finding
- ChIPSeq
- Whole genome bisulfite sequencing (WGBS)
- Computer Lab

D. melanogaster, Science, 2000

H. sapiens, Nature, 2000

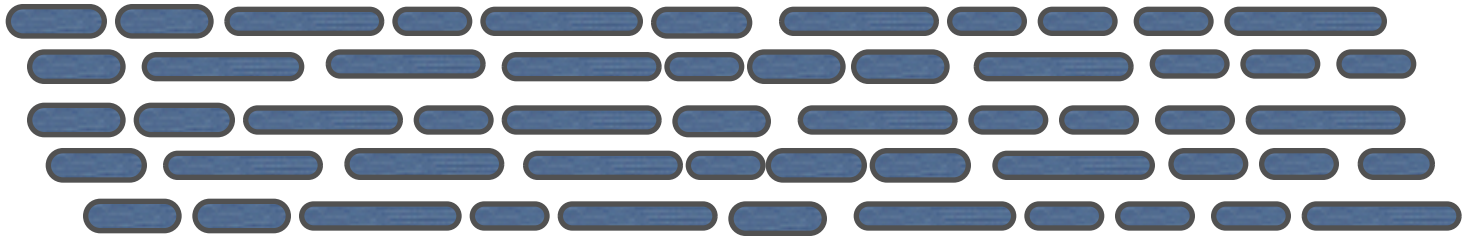


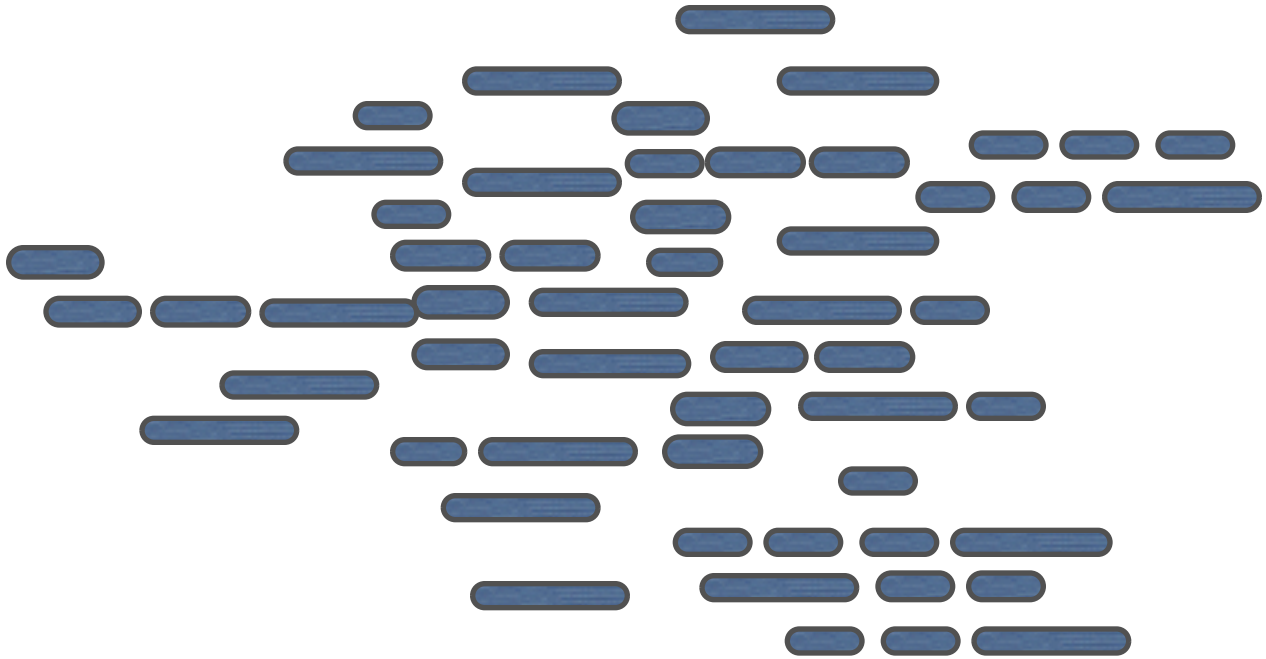
Science, 2000



M. musculus, Nature, 2002

- Back then: millions of clones (thousand bps) in 9 months for billions of dollars
- Today: billion of short reads (35-100 bps) in a week for thousands of dollars
- Claim: Assemble a genome in weeks for less than $100,000





Sequence first 35-400 bps: call them *"reads"*

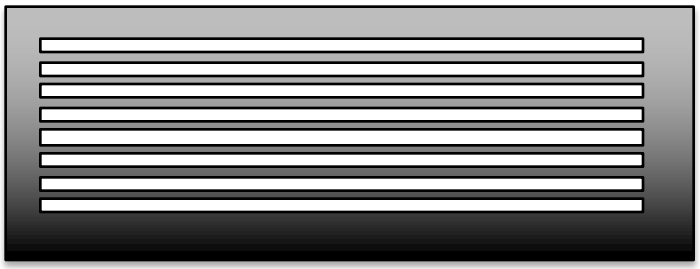
GTTGAGGCTTGCGTTTTTGGTACGCTGGACTTTGT
GTACTCGTCGCTGCGTTGAGGCTTGCGTTTTTGGT
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTTGCGTTTATGGTACGCTGGACTTTGTAGGATAC
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
GCGTTTATGGTACGCTGGACTTTGTAGGATACCCT
GAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGG
GCGTTGAGGCTTGCGTTTATGGTACGCTGGATTTT
CGTTTATGGTACGCTGGACTTTGTAGGATACCCTC
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
GTTTATGGTACGCTGGACTTTGTAGGATACCCTCG
TCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTA
TGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTA
GCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTAC
TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
TCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTTTG
CGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCT
GTTGAGGCTTGCGTTTATGGTACGCTGGGCTTTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC

Available platforms

- Major player:
  - Illumina: HiSeq, MiSeq.
  - LifeTech: SOLiD, IonTorrent.
  - Roche 454.
- Others:
  - Complete Genomics
  - Pacific Bioscience
  - Helicos

illumina®

~7 cm

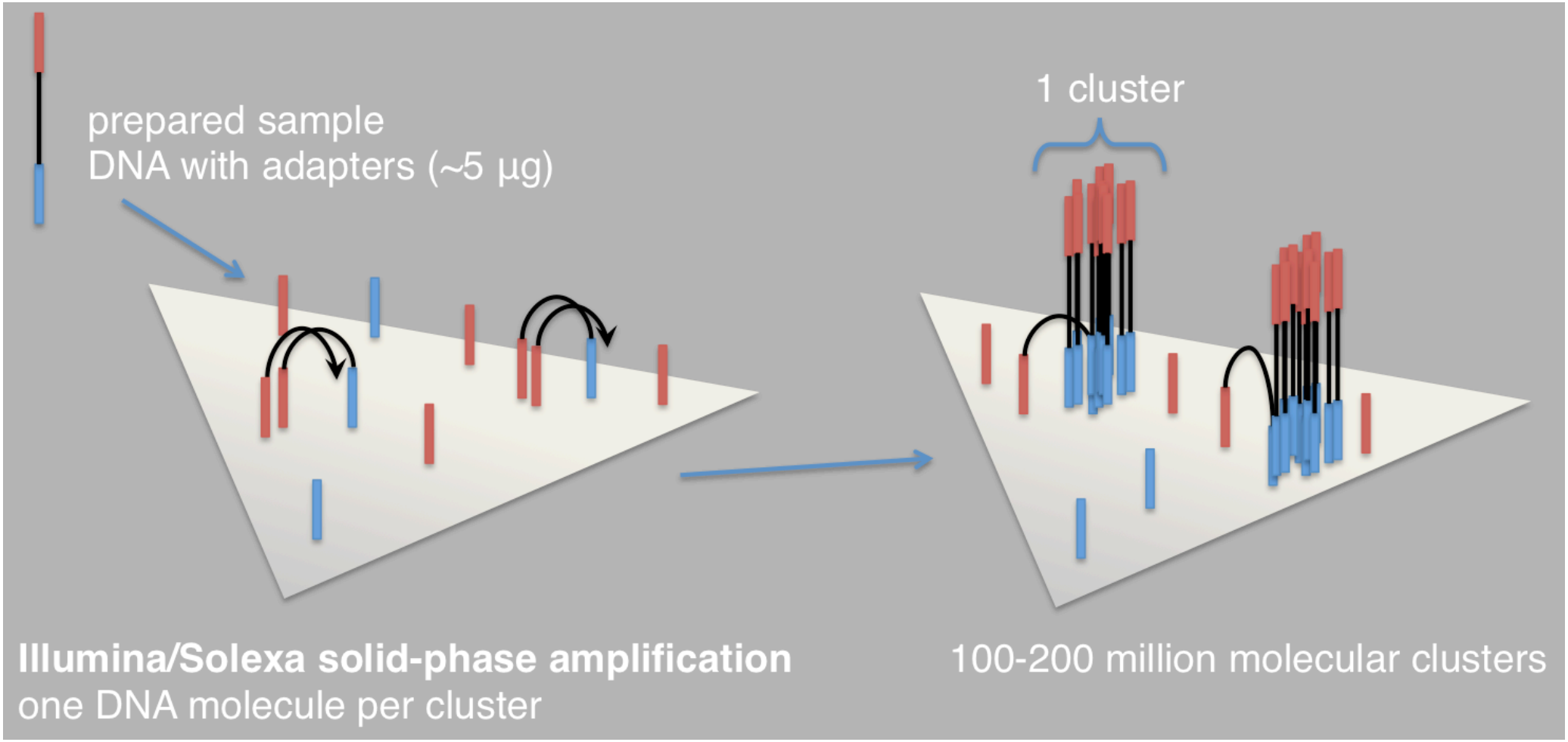


~2 cm

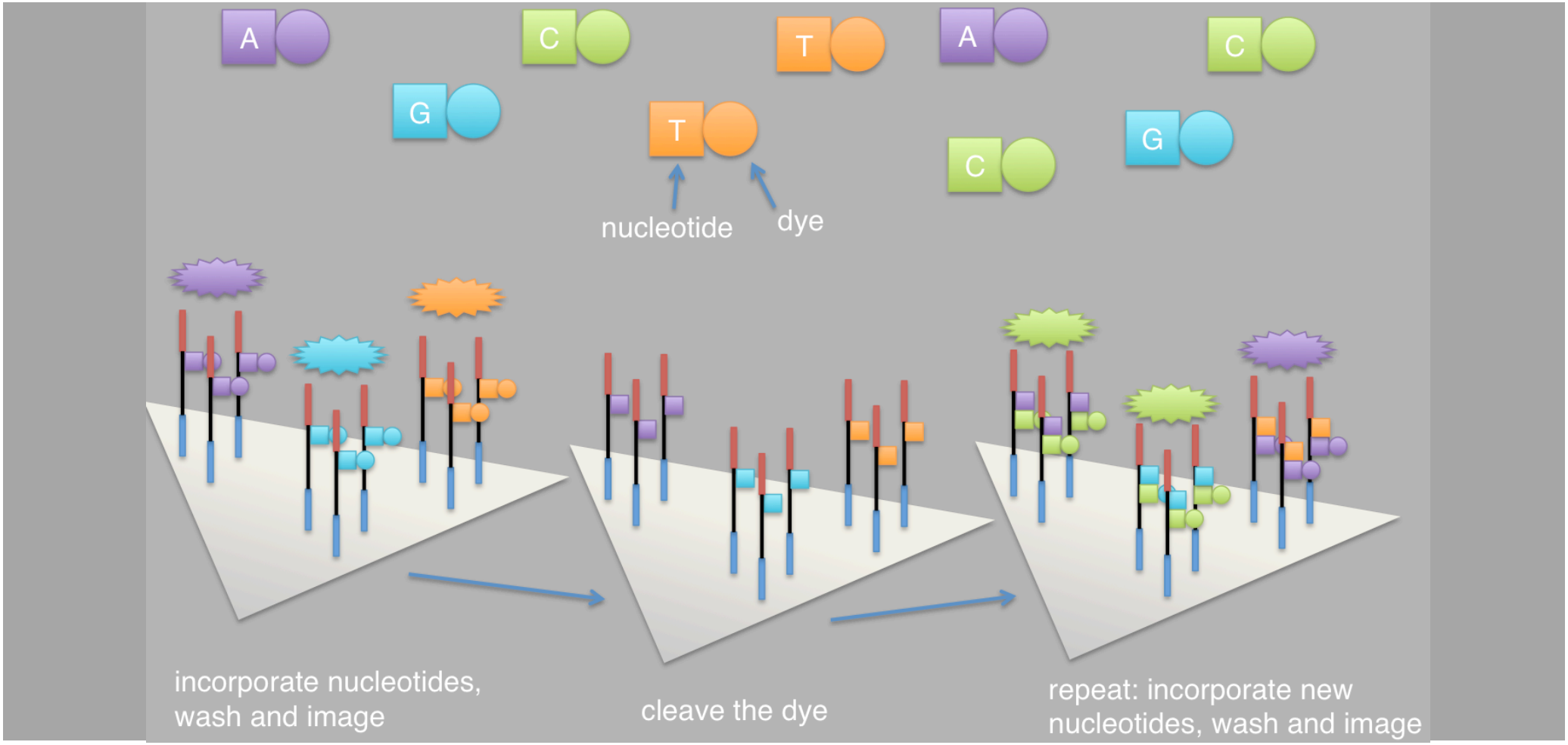
Illumina “flow cell”

- Eight lanes
- ~160M short reads (~50-70 bp) per lane

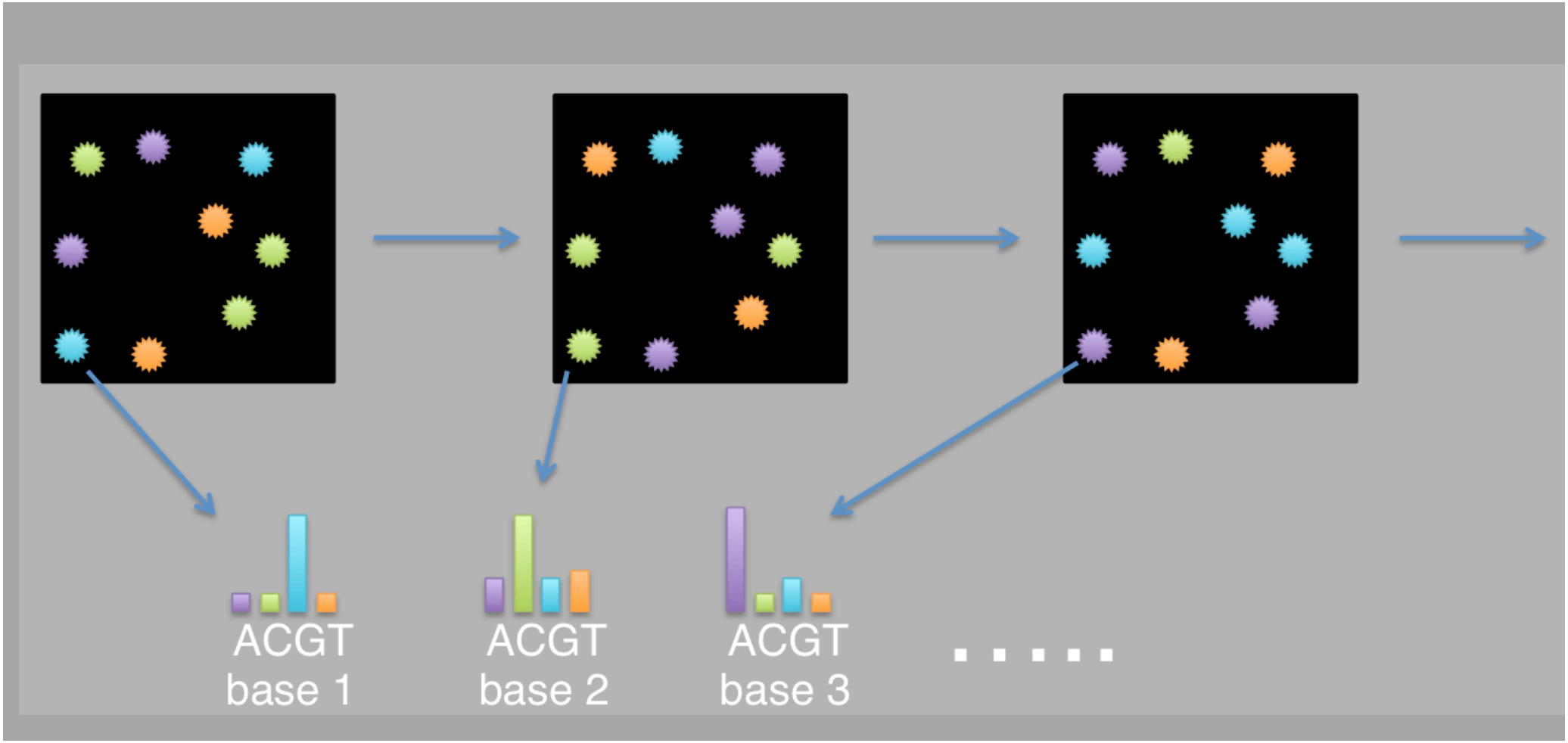
Image: Illumina logo, http://www.illumina.com/



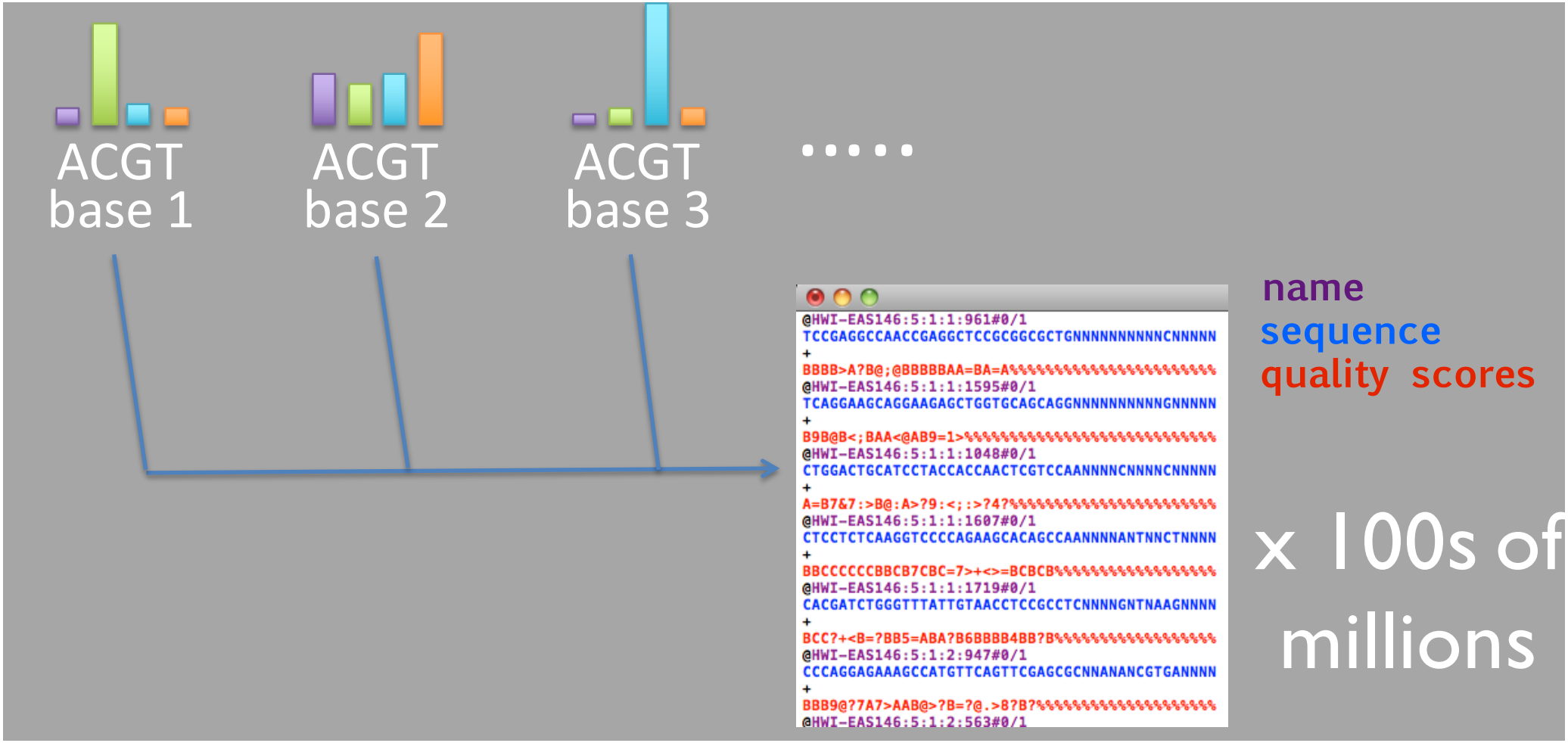
Bridge amplification



Sequencing all bases at once

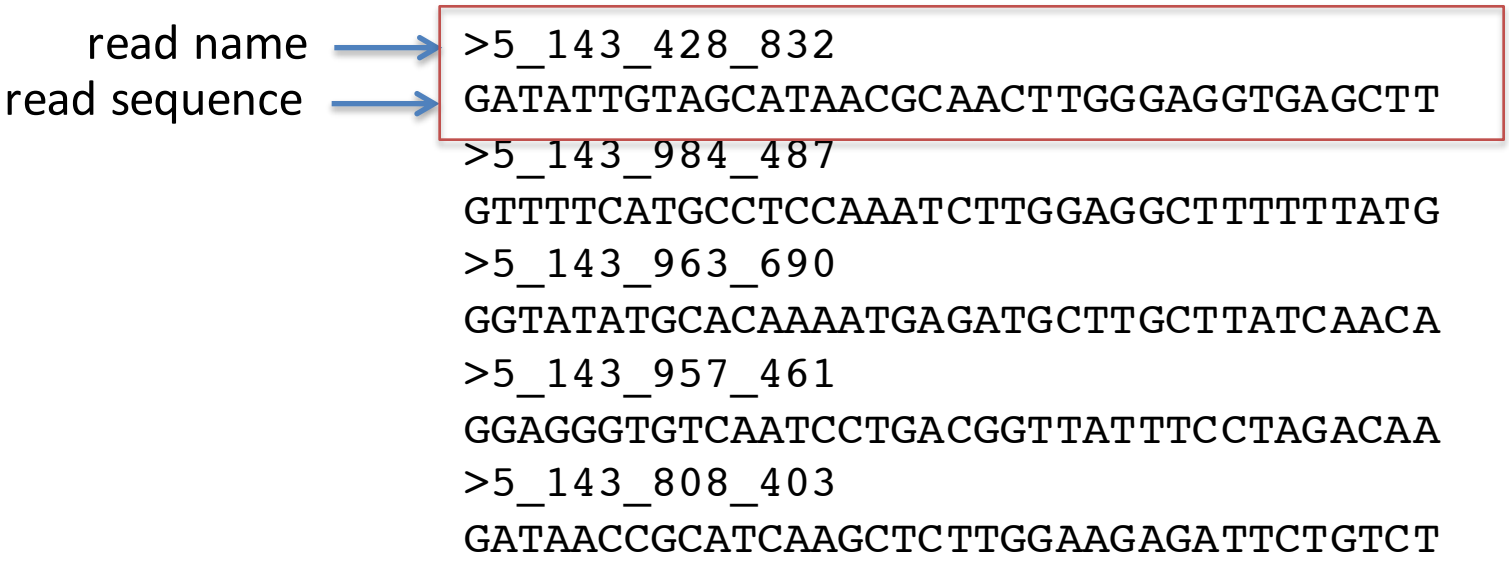


Images from sequencing machine


ACGT
base 1
ACGT
base 2
ACGT
base 3
.....
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNCNNNNN
+
BBBB>A?B@;@BBBBBAA=BA=A%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGGNNNNNNNNNNGNNNNN
+
B9B@B<;BAA<@AB9=1>%%%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNCNNNNCNNNNN
+
A=B7&7:>B@:A>?9:<;:>?4?%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCCAGAAGCACAGCCAANNNNNANTNNCTNNNN
+
BBCCCCCCBBCB7CBC=7>+<>=BCBCB%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCCTCNNNNGNTNAAGNNNN
+
BCC?+<B=?BB5=ABA?B6BBBB4BB?B%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:947#0/1
CCCAGGAGAAAGCCATGTTCAGTTCGAGCGCNNANANCGTGANNNN
+
BBB9@?7A7>AAB@>?B=?@.>8?B?%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:563#0/1
name
sequence
quality scores
x 100s of
millions

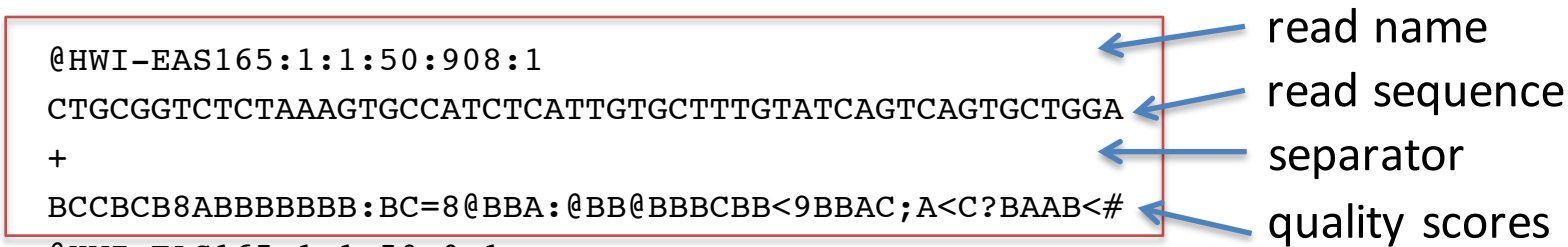
Raw sequence reads from NGS

- Large text file (millions of lines) with simple format.
  - Most frequently used: fasta/fa format for storing the sequences, or fastq format storing both the sequence and corresponding quality scores.
- fasta format:



fastq format

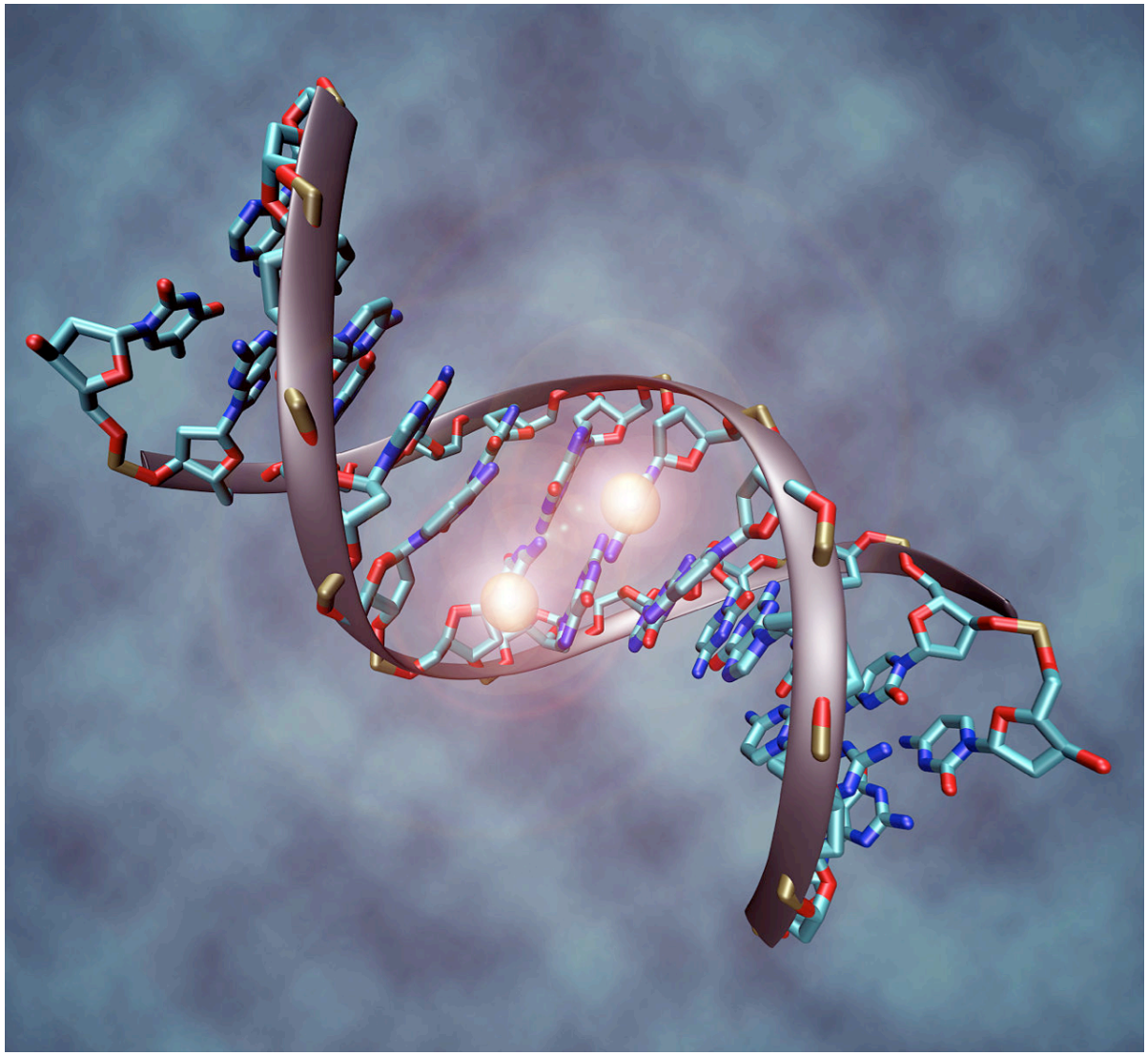
```
@HWI-EAS165:1:1:50:908:1
CTGCGGTCTCTAAAGTGCCATCTCATTGTGCTTTGTATCAGTCAGTGCTGGA
+
BCCBCB8ABBBBBBBB:BC=8@BBA:@BB@BBBCBB<9BBAC;A<C?BAAB<#
@HWI-EAS165:1:1:50:0:1
NCAACCCCCACAGTAATATGTAAAACAAAAACTAAAACCAGGAGCTGAAGGG
+
#BABABBBBBB@08<@?A@7:A@CCBCCCCBBBCCBB=?BBBB@7@B=A>:2
@HWI-EAS165:1:1:50:708:1
GGTCAGCATGTCTTCTGTTAAGTGCTTGCACAAGCTAGCCTCTGCCTATGGG
+
BB@A;B>@A@@=BB=BB?A>@@>B?ABBA=A?@@>@@A:=?>?A@=B8@@AB
@HWI-EAS165:1:1:50:1494:1
CTGGTGTCACACAAGCAGGTCTCCTGTGTTGACTTCACCAGACACTGTCATT
+
BCBB@AB@1ABBBBBBAAB?BBBBAB<A?AA>BB@?1ABBA@BBBA@;B>>:
```



Single-end vs. paired-end sequencing

- Sequence one or both ends of the DNA segments.
- **Single-end** sequencing: sequence one end of the DNA segment.
- **Paired-end** sequencing: sequence both ends of a DNA segments.
  - Result reads are “paired”, separated by certain length (the length of the DNA segments, usually a few hundred bps).
  - Paired-end data can be used as single-end, but contain extra information which is useful in some cases, e.g., detecting structural variations in the genome.
  - Modeling technique is more complicated.

Segment 2 – Applications of NGS in Genomics
(do not include this slide in video)



Not just Assembly

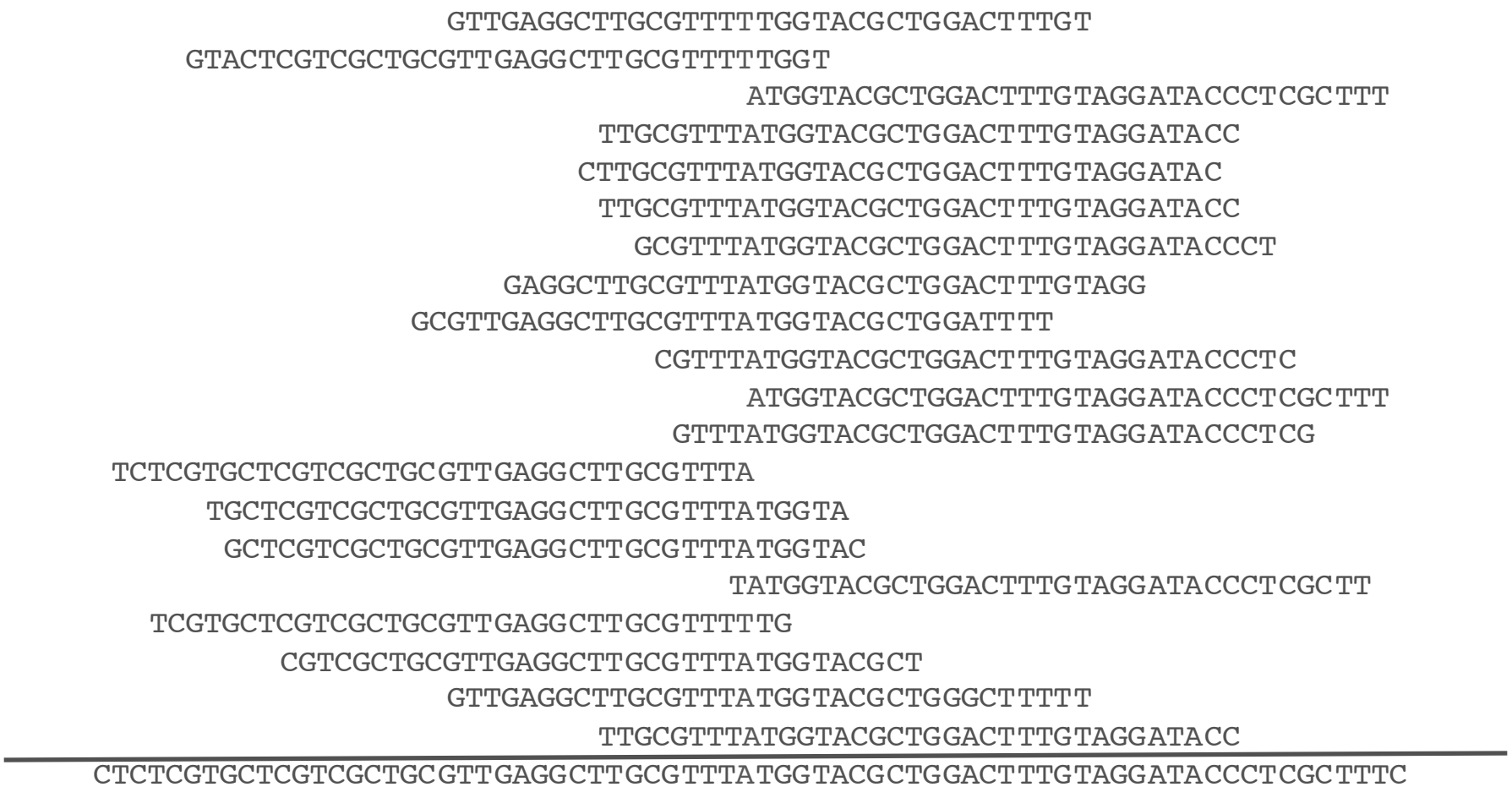
- Resequencing
- SNP discovery and genotyping
- Variant discovery and quantification
- TF binding sites: ChIP-Seq
- Gene expression: RNA-Seq
- Measuring methylation

Image: DNA methylation 18.02.2006 Christoph Bock, http://commons.wikimedia.org/wiki/File:DNA_methylation.jpg CCBY

What to do with all these sequences?

GTTGAGGCTTGCGTTTTTGGTACGCTGGACTTTGT
GTACTCGTCGCTGCGTTGAGGCTTGCGTTTTTGGT
ATGGTACGCTGGACTTTGTAGGATACCTCGCTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTTGCGTTTATGGTACGCTGGACTTTGTAGGATAC
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
GCGTTTATGGTACGCTGGACTTTGTAGGATACCCT
GAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGG
GCGTTGAGGCTTGCGTTTATGGTACGCTGGATTTT
CGTTTATGGTACGCTGGACTTTGTAGGATACCCTC
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
GTTTATGGTACGCTGGACTTTGTAGGATACCCTCG
TCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTA
TGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTA
GCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTAC
TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
TCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTTTG
CGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCT
GTTGAGGCTTGCGTTTATGGTACGCTGGGCTTTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC

Most apps: Start by matching to reference



TTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTT

Reference

Variant detection

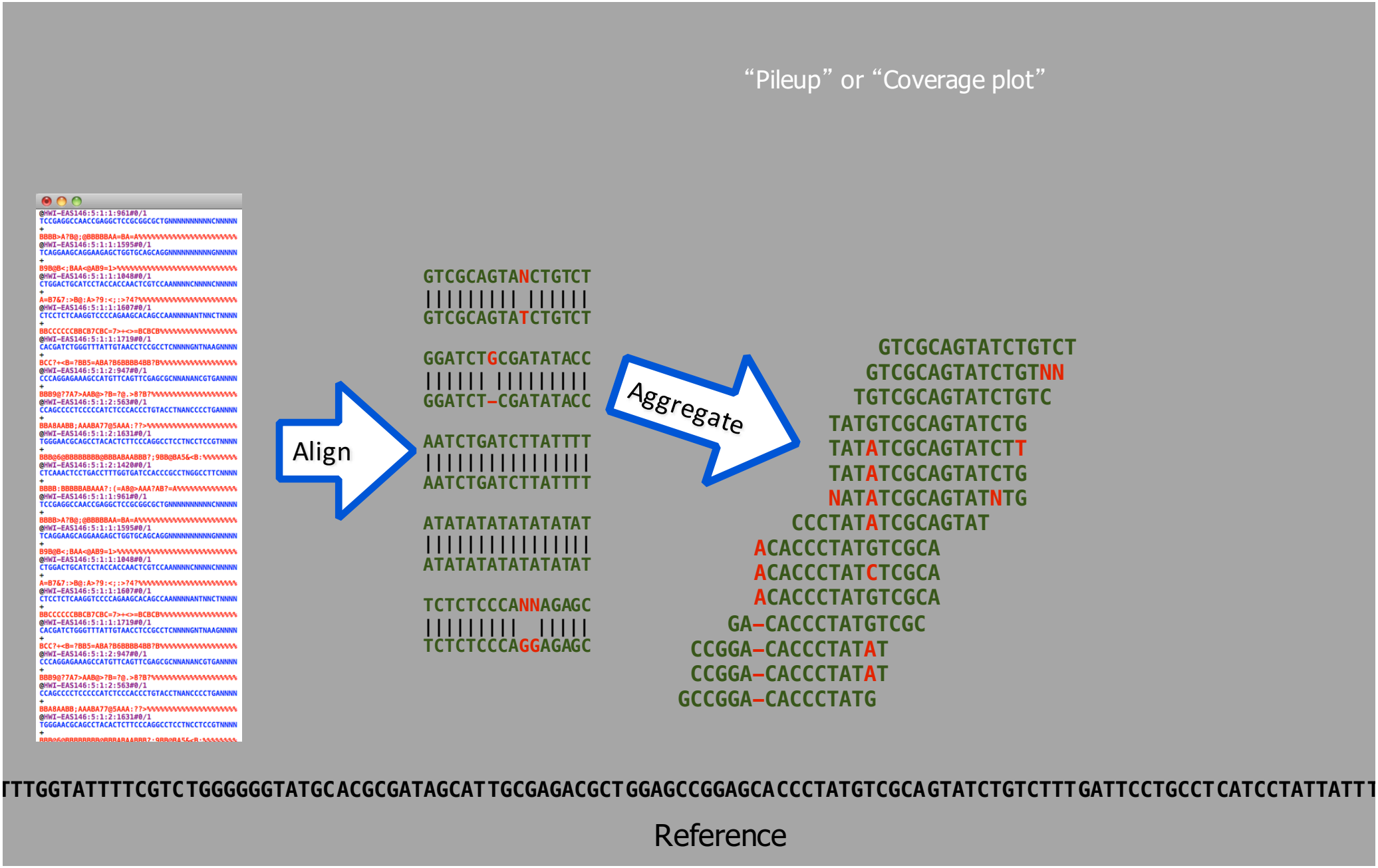
Align

```
GTCGCAGTANCTGTCT
|||||||||| ||||||
GTCGCAGTATCTGTCT

GGATCTGCGATATACC
|||||| |||||||||
GGATCT-CGATATACC

AATCTGATCTTATTTT
||||||||||||||||
AATCTGATCTTATTTT

ATATATATATATATAT
||||||||||||||||
ATATATATATATATAT

TCTCTCCCANNAGAGC
||||||||| |||||
TCTCTCCCAGGAGAGC
```

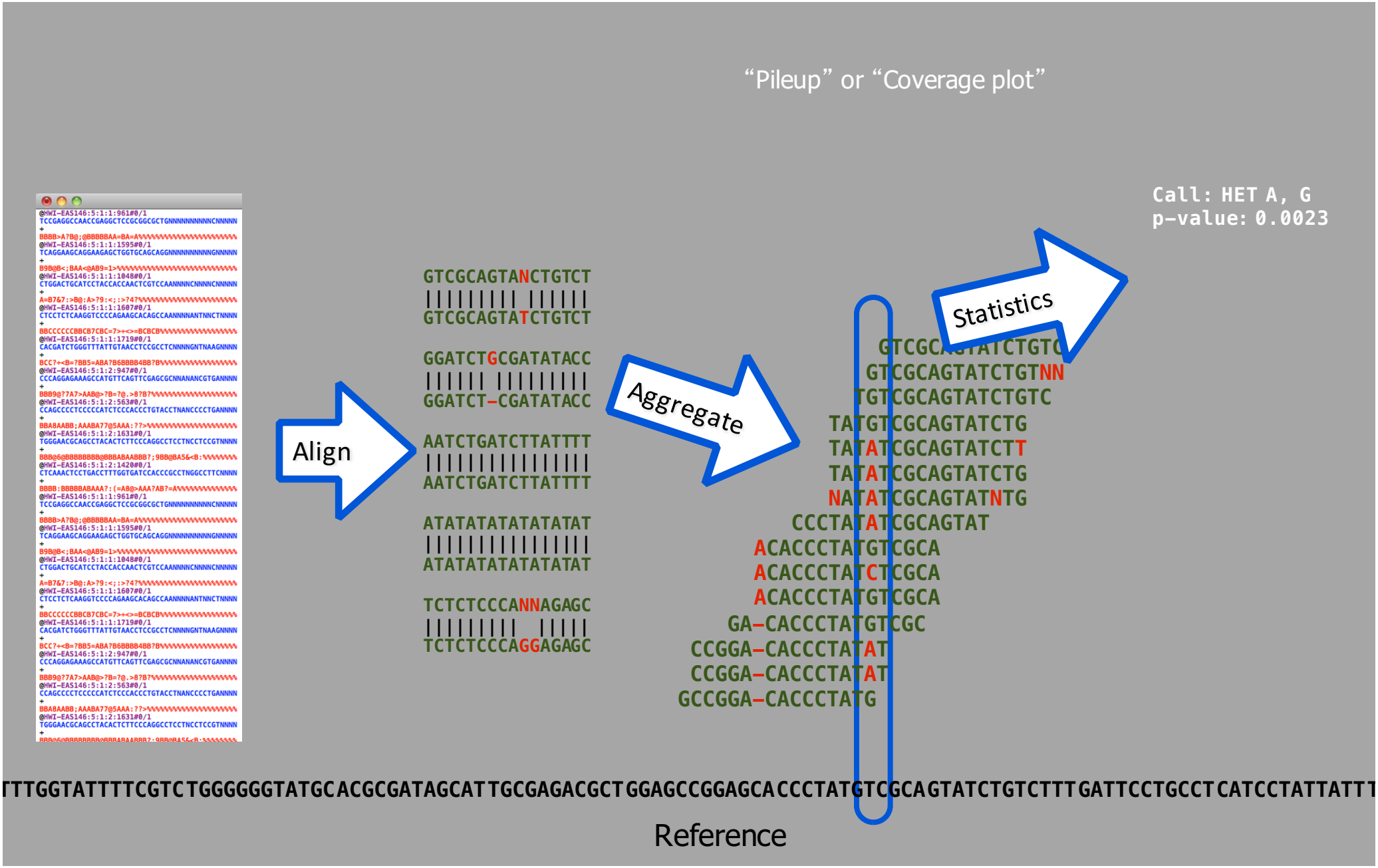
TTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTT

Reference

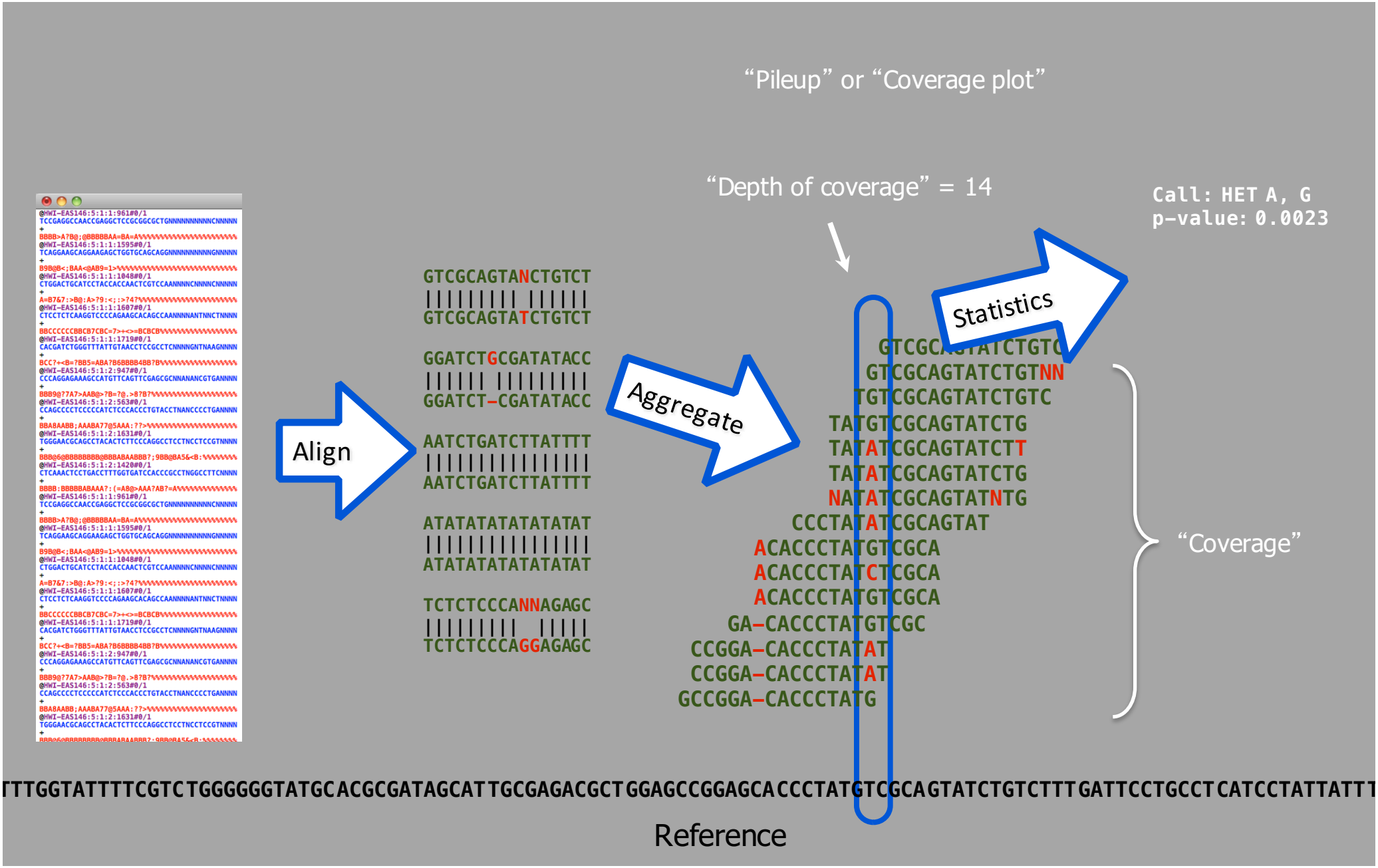
Variant detection



Variant detection



Variant detection



Variant detection

Sample A

GGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGCAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTA

Gene 1

Sample B

RNA-seq differential expression

Sample A

Align

```
GTCGCAGTANCTGTCT
||||||||| ||||||
GTCGCAGTATCTGTCT

GGATCTGCGATATACC
|||||| |||||||||
GGATCT-CGATATACC

AATCTGATCTTATTTT
||||||||||||||||
AATCTGATCTTATTTT
```

GGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGCAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTA

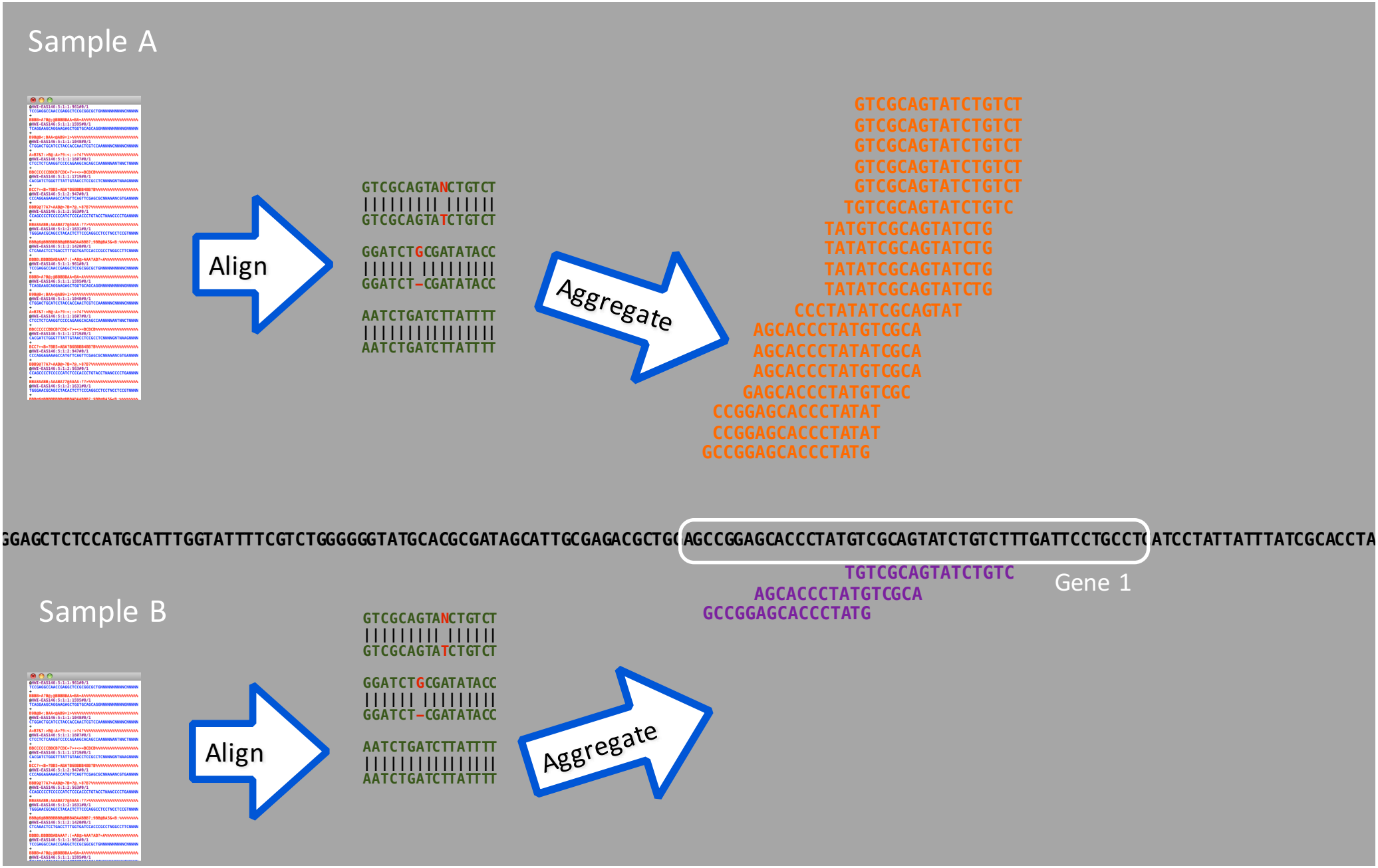
Gene 1

Sample B

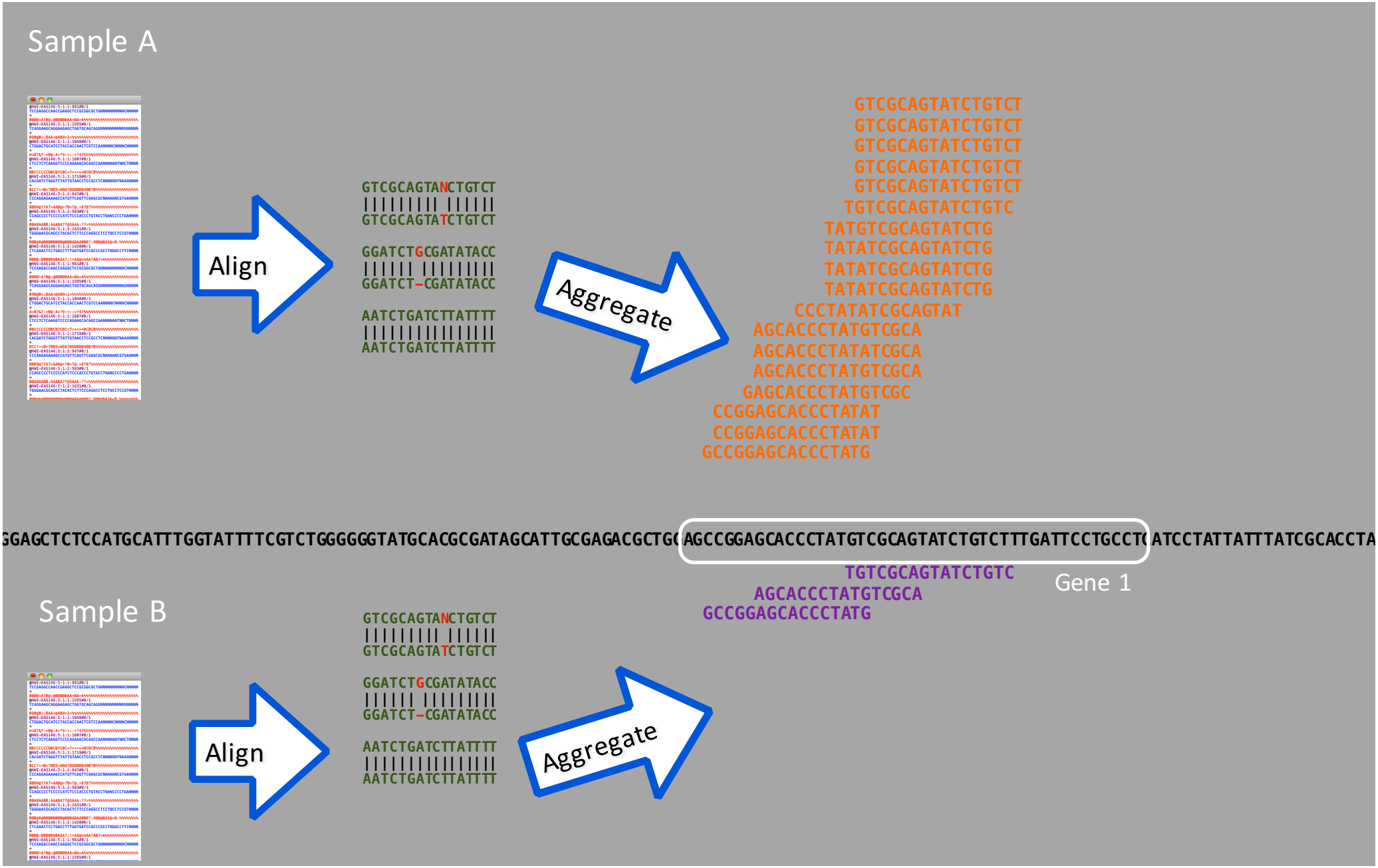
Align

```
GTCGCAGTANCTGTCT
||||||||| ||||||
GTCGCAGTATCTGTCT

GGATCTGCGATATACC
|||||| |||||||||
GGATCT-CGATATACC

AATCTGATCTTATTTT
||||||||||||||||
AATCTGATCTTATTTT
```

RNA-seq differential expression



RNA-seq differential expression



RNA-seq differential expression

TTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCGGGAGCTCTCCA

Reference

ChIP-seq

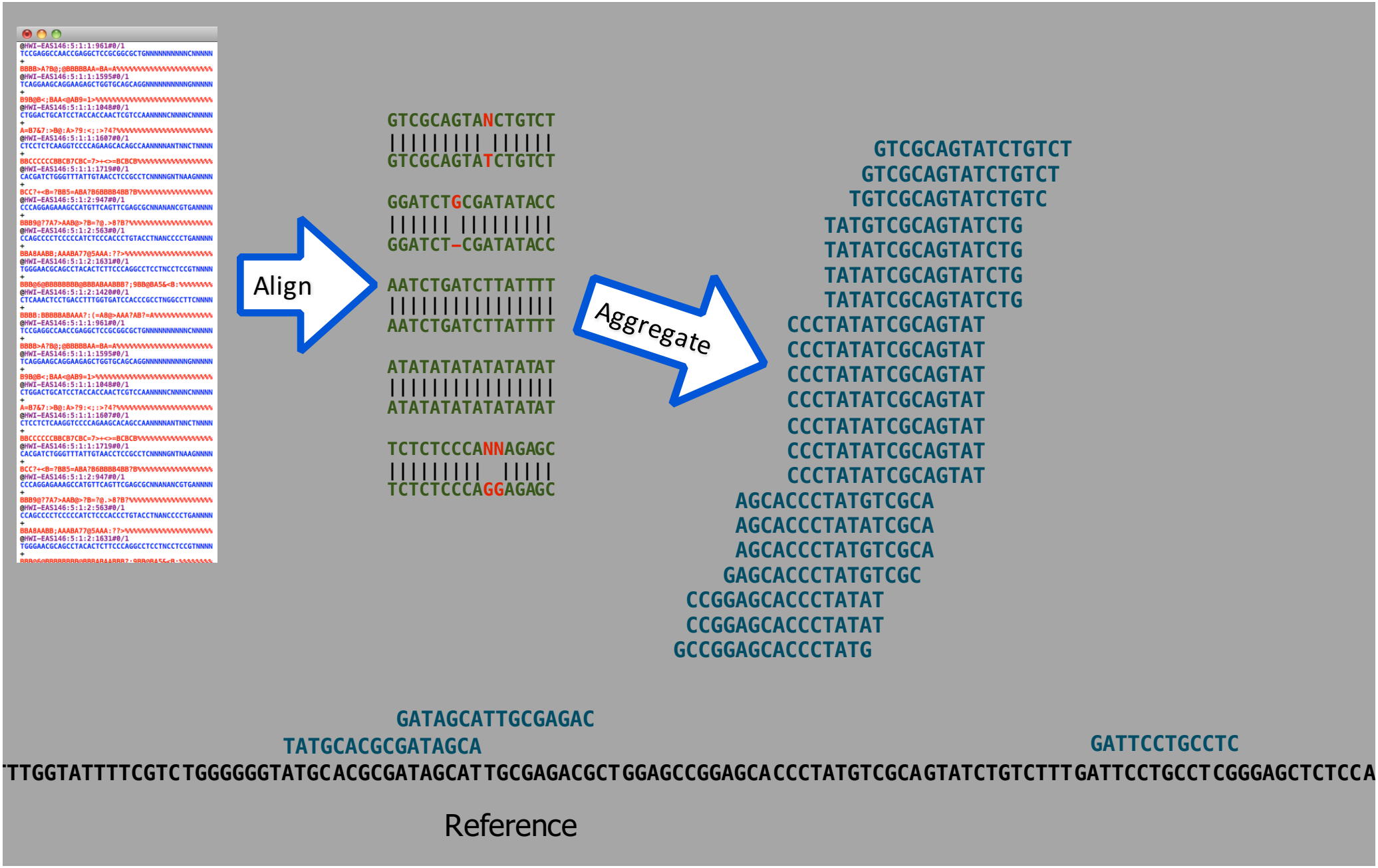
Align

```
GTCGCAGTANCTGTCT
|||||||||| ||||||
GTCGCAGTATCTGTCT

GGATCTGCGATATACC
|||||| |||||||||
GGATCT-CGATATACC

AATCTGATCTTATTTT
||||||||||||||||
AATCTGATCTTATTTT

ATATATATATATATAT
||||||||||||||||
ATATATATATATATAT

TCTCTCCCANNAGAGC
||||||||| |||||
TCTCTCCCAGGAGAGC
```

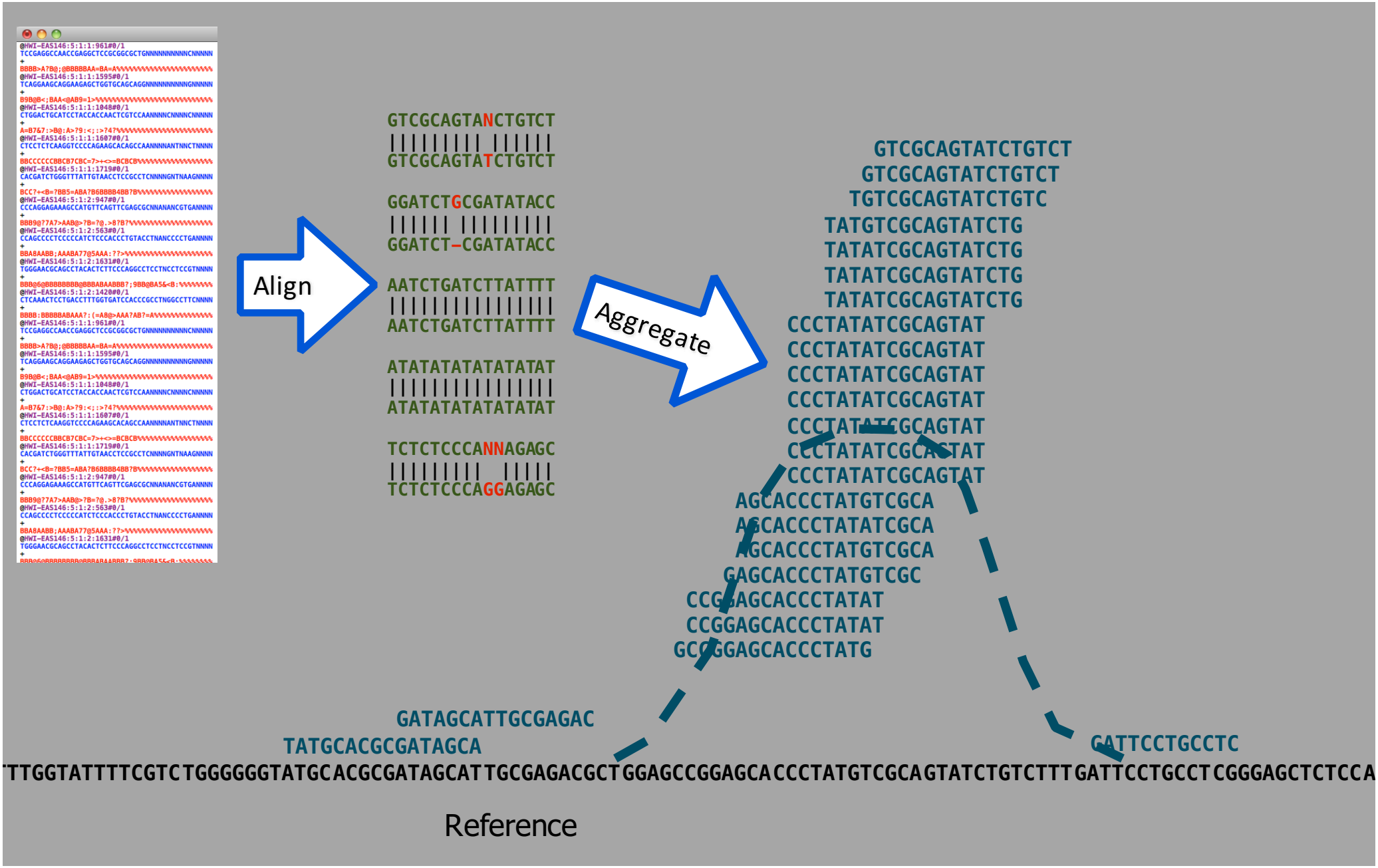
TTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCGGGAGCTCTCCA

Reference

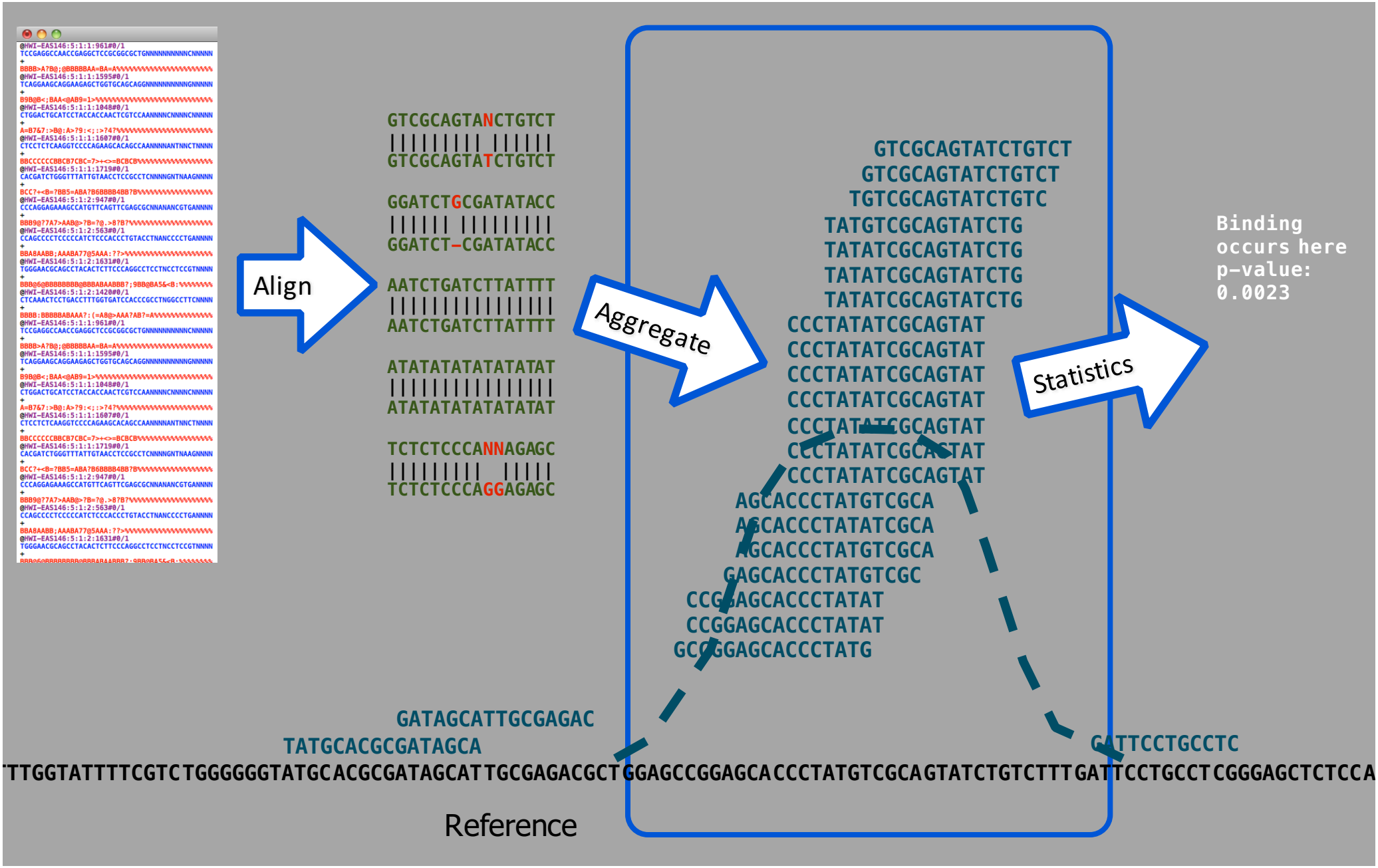
ChIP-seq


Align
Aggregate
GTCGCAGTANCTGTCT
GTCGCAGTATCTGTCT
GGATCTGCGATATACC
GGATCT-CGATATACC
AATCTGATCTTATTTT
AATCTGATCTTATTTT
ATATATATATATATAT
ATATATATATATATAT
TCTCTCCCANNAGAGC
TCTCTCCCAGGAGAGC
GTCGCAGTATCTGTCT
GTCGCAGTATCTGTCT
TGTCGCAGTATCTGTC
TATGTCGCAGTATCTG
TATATCGCAGTATCTG
TATATCGCAGTATCTG
TATATCGCAGTATCTG
CCCTATATCGCAGTAT
CCCTATATCGCAGTAT
CCCTATATCGCAGTAT
CCCTATATCGCAGTAT
CCCTATATCGCAGTAT
CCCTATATCGCAGTAT
CCCTATATCGCAGTAT
AGCACCCTATGTCGCA
AGCACCCTATATCGCA
AGCACCCTATGTCGCA
GAGCACCCTATGTCGC
CCGGAGCACCCTATAT
CCGGAGCACCCTATAT
GCCGGAGCACCCTATG
GATAGCATTGCGAGAC
TATGCACGCGATAGCA
GATTCCTGCCTC
TTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCGGGAGCTCTCCA
Reference

ChIP-seq


Align
Aggregate
GTCGCAGTANCTGTCT
GTCGCAGTATCTGTCT
GGATCTGCGATATACC
GGATCT-CGATATACC
AATCTGATCTTATTTT
AATCTGATCTTATTTT
ATATATATATATATAT
ATATATATATATATAT
TCTCTCCCANNAGAGC
TCTCTCCCAGGAGAGC
GTCGCAGTATCTGTCT
GTCGCAGTATCTGTCT
TGTCGCAGTATCTGTC
TATGTCGCAGTATCTG
TATATCGCAGTATCTG
TATATCGCAGTATCTG
TATATCGCAGTATCTG
CCCTATATCGCAGTAT
CCCTATATCGCAGTAT
CCCTATATCGCAGTAT
CCCTATATCGCAGTAT
CCCTATATCGCAGTAT
CCCTATATCGCAGTAT
CCCTATATCGCAGTAT
AGCACCCTATGTCGCA
AGCACCCTATATCGCA
AGCACCCTATGTCGCA
GAGCACCCTATGTCGC
CCGGAGCACCCTATAT
CCGGAGCACCCTATAT
GCCGGAGCACCCTATG
GATAGCATTGCGAGAC
TATGCACGCGATAGCA
GATTCCTGCCTC
TTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCGGGAGCTCTCCA
Reference

ChIP-seq


Align
Aggregate
Statistics
Binding occurs here
p-value: 0.0023
Reference

ChIP-seq