# Low level Processing and Visualization of Genome Sequencing Data

## GSND 5340Q, BMDA

W. Evan Johnson, Ph.D.
Professor, Division of Infectious Disease
Director, Center for Data Science
Rutgers University – New Jersey Medical School

2024-05-01

Section 1

# Motivating Example: X-linked Disease

# Rare X-linked Disease



# Rare X-linked Disease

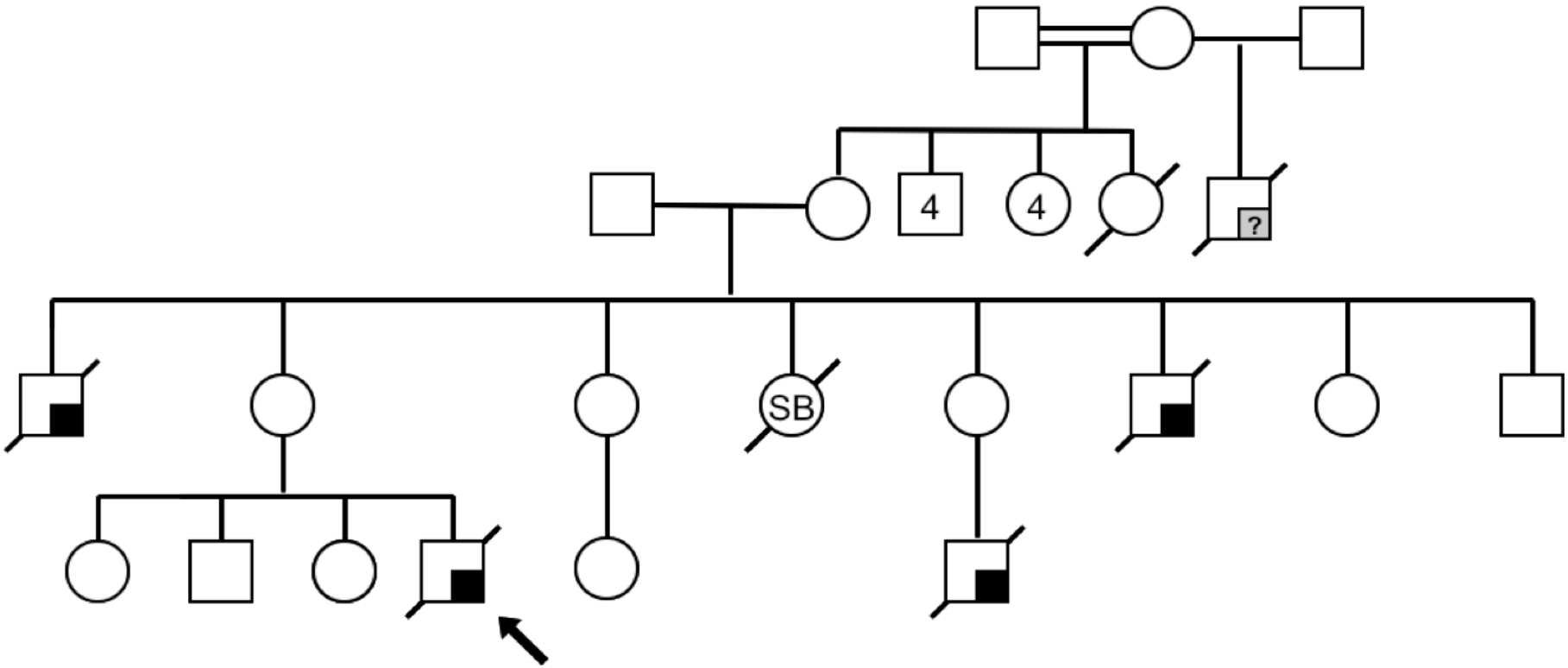
Uncle #1

Uncle #2

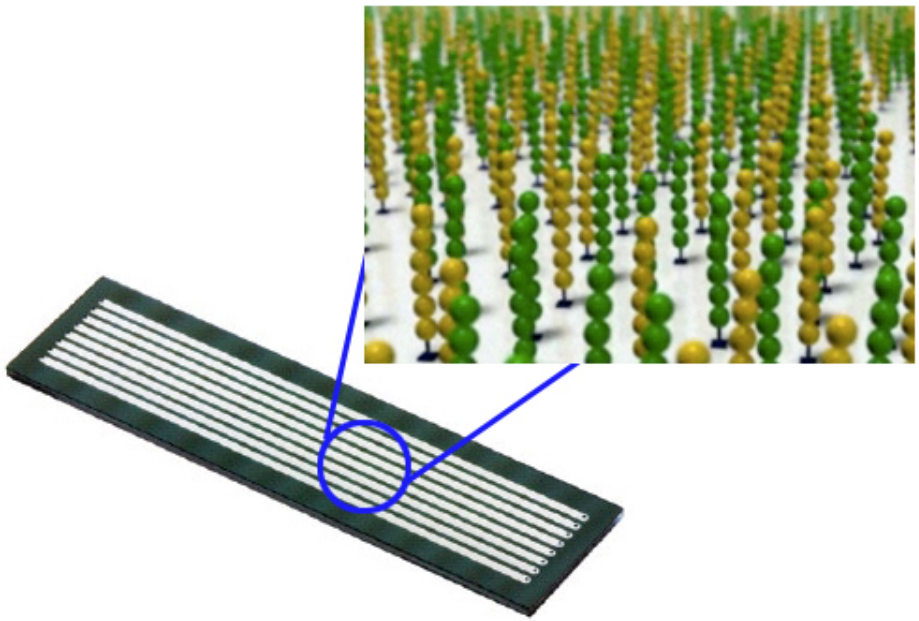
cousin

proband

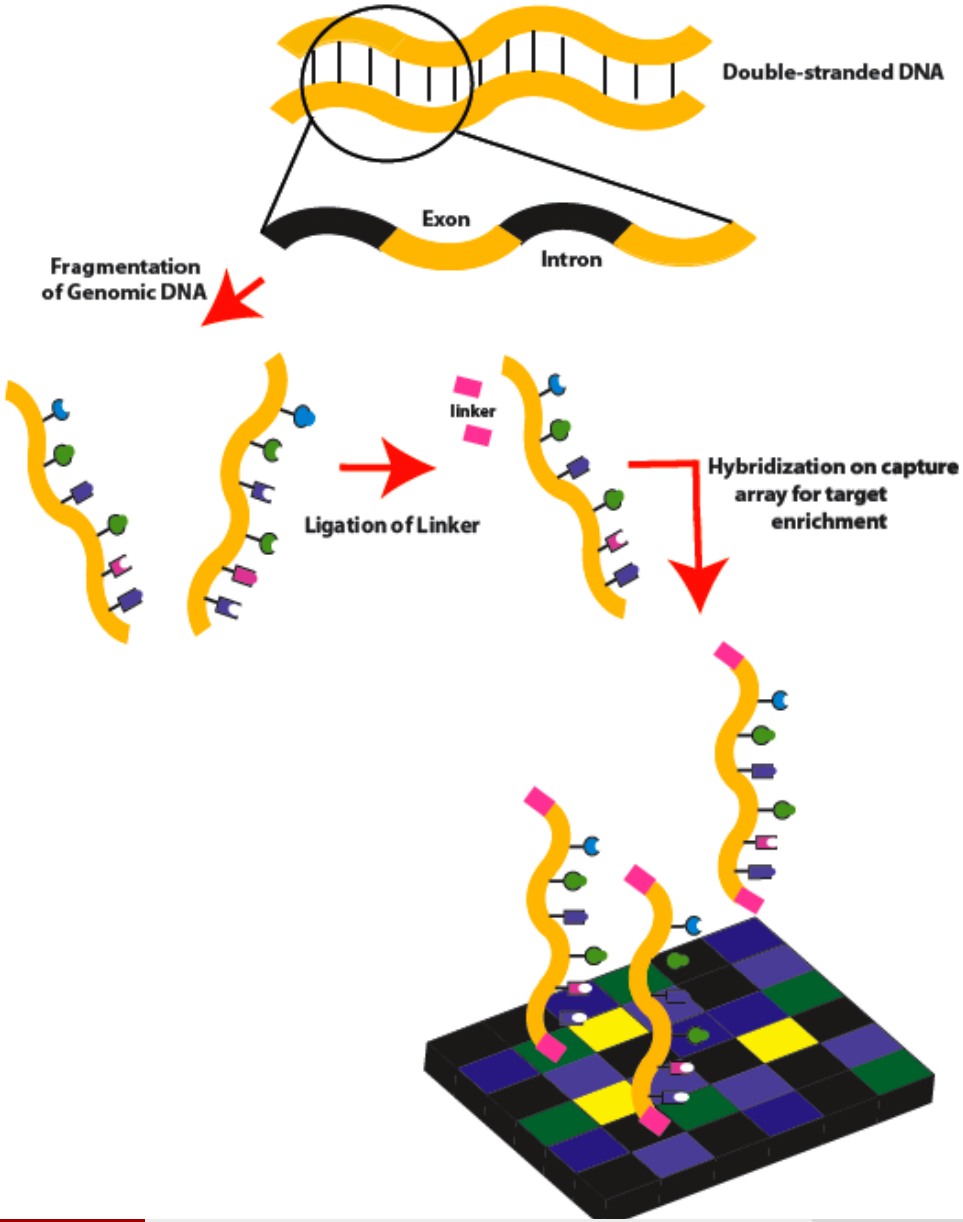
# Rare X-linked Disease



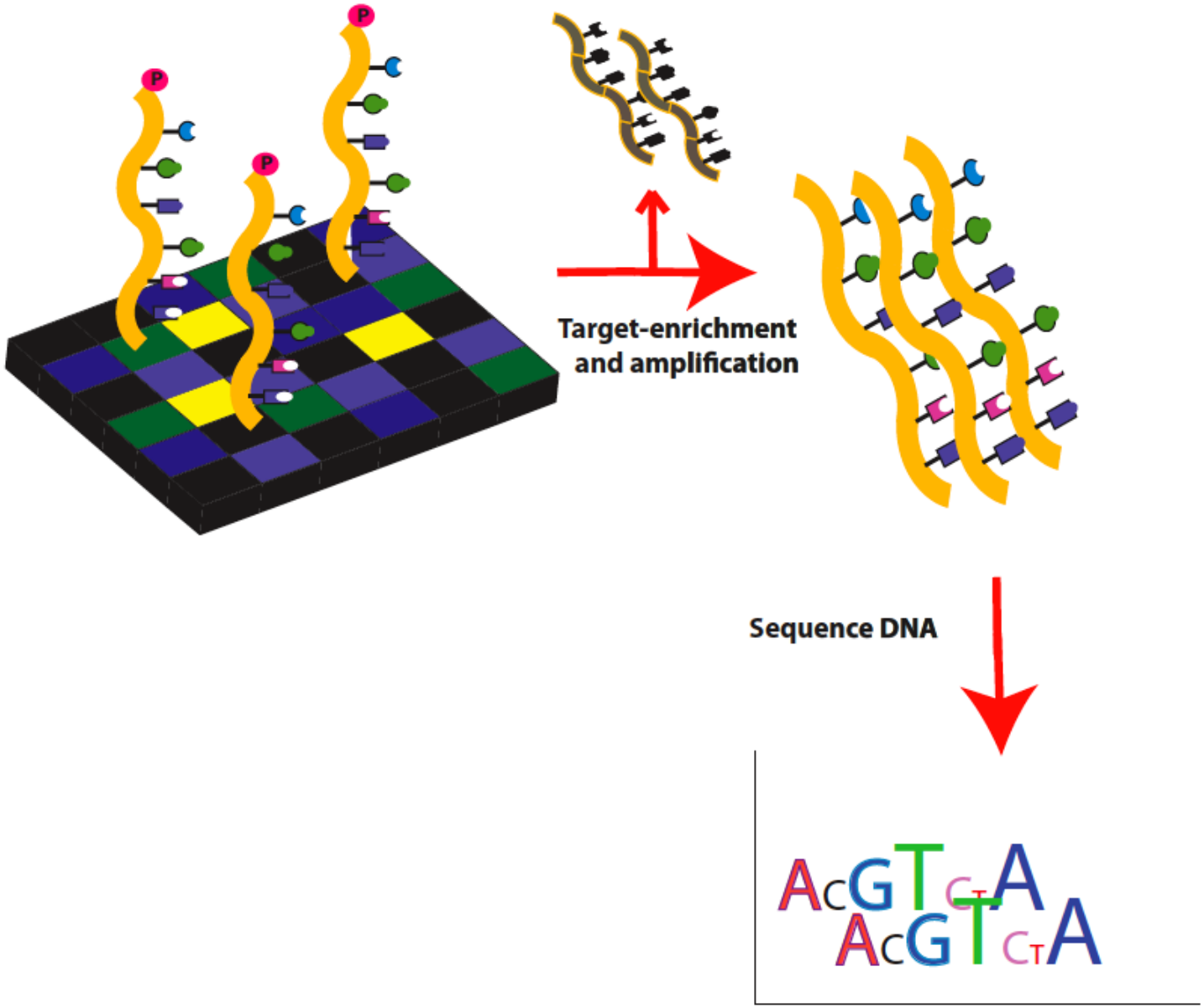
# Exon Capture Sequencing



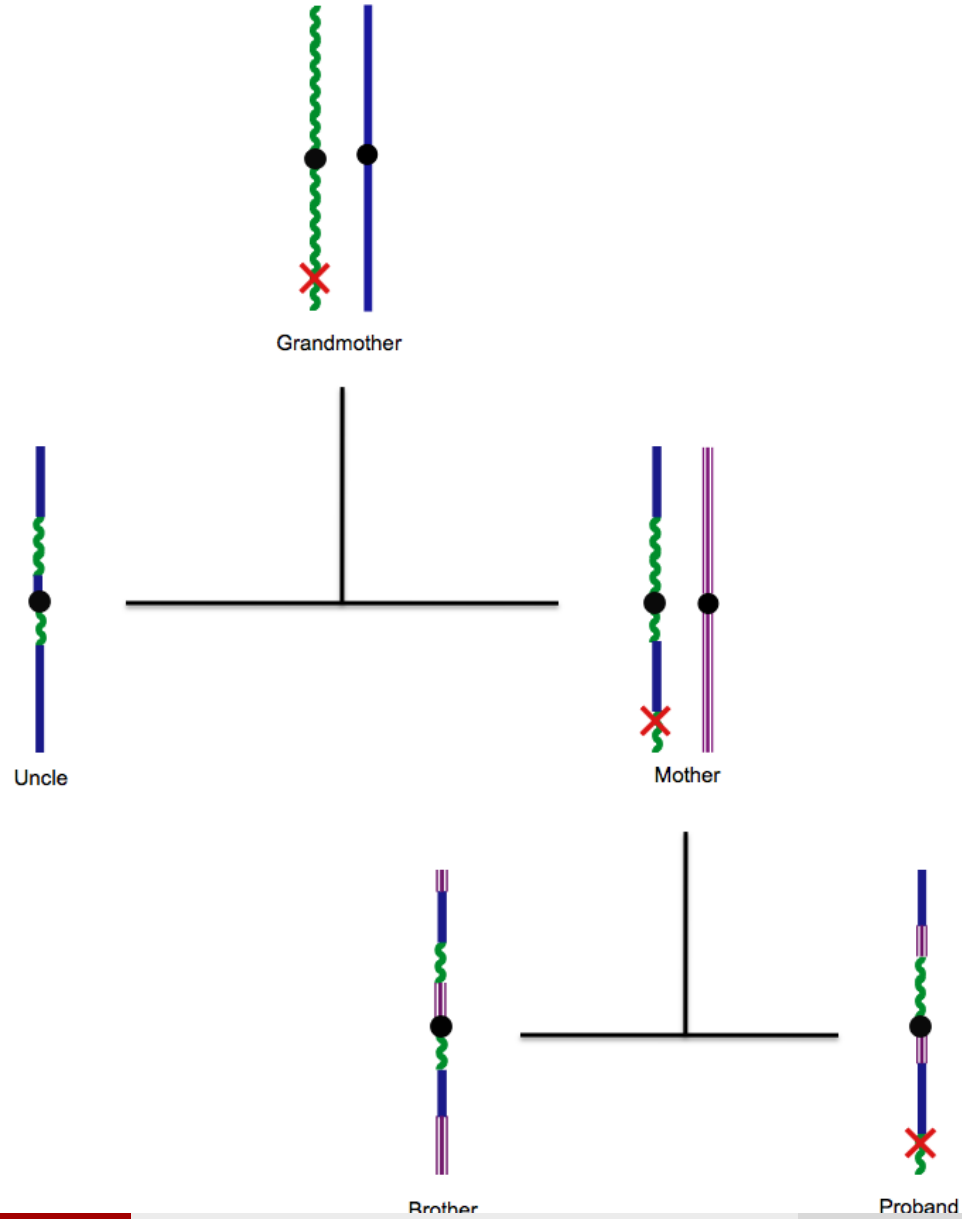
# Exome Capture



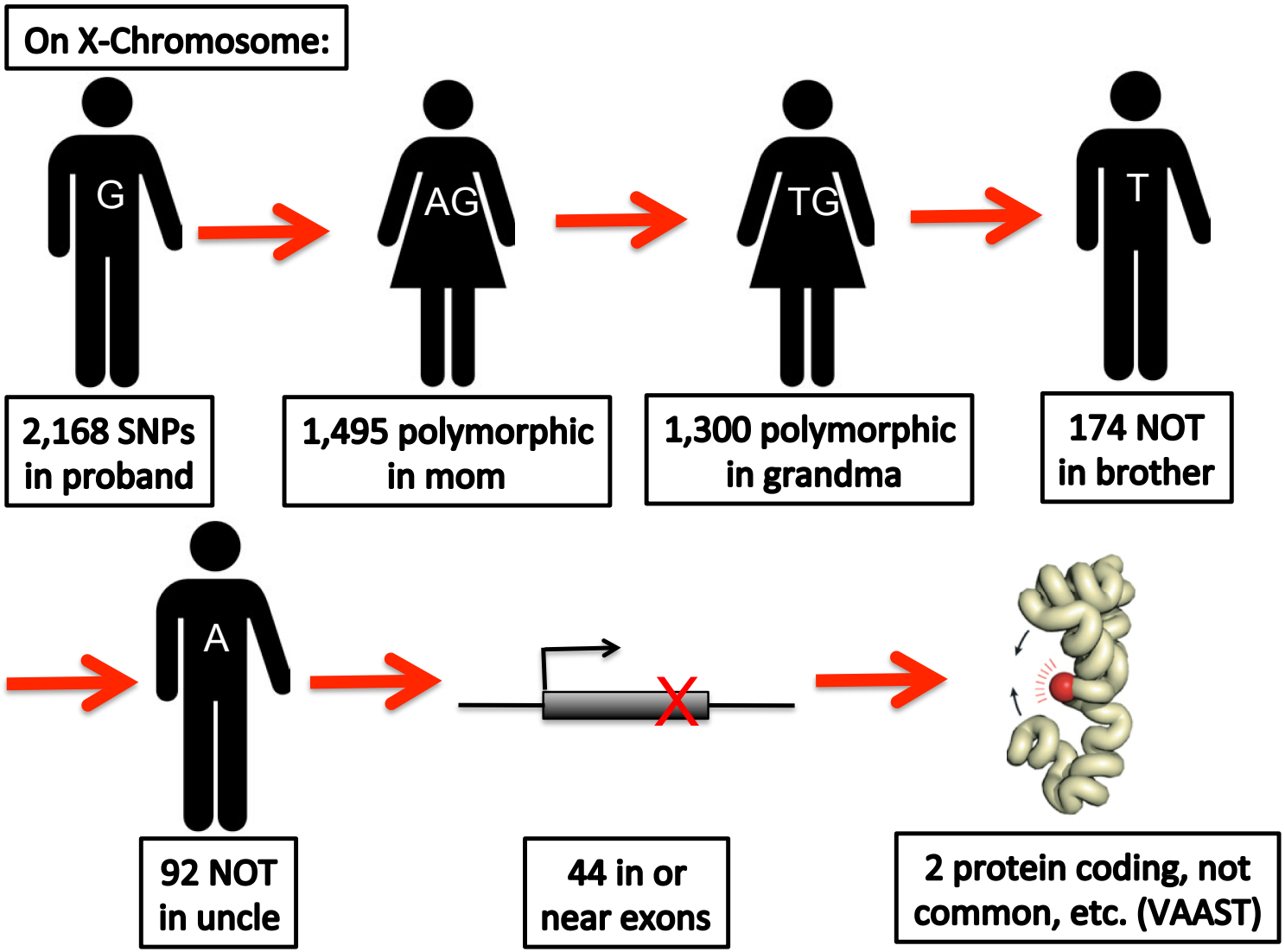
# Exome Capture



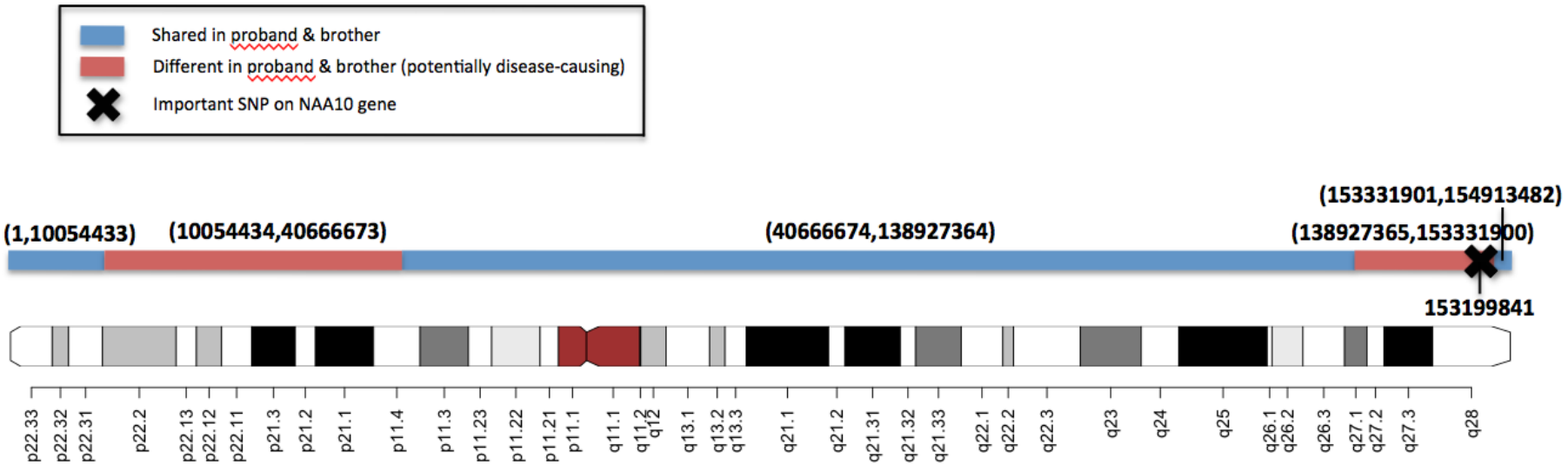
# Rare X-linked Disease



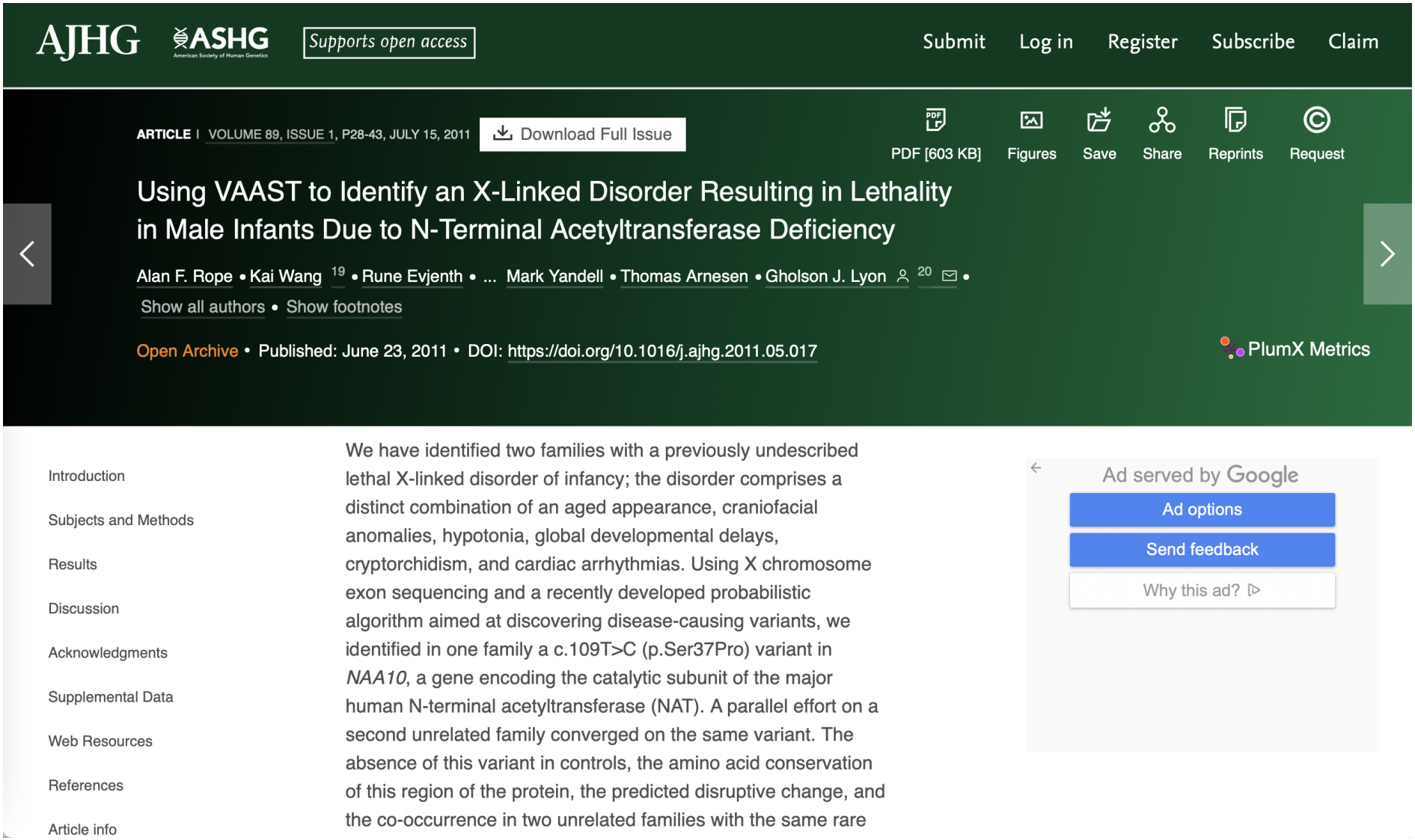
# Rare X-linked Disease}



# Rare X-linked Disease



# Rare X-linked Disease



AJHG ASHG American Society of Human Genetics Supports open access

Submit Log in Register Subscribe Claim

ARTICLE | VOLUME 89, ISSUE 1, P28-43, JULY 15, 2011 Download Full Issue

PDF [603 KB] Figures Save Share Reprints Request

## Using VAAST to Identify an X-Linked Disorder Resulting in Lethality in Male Infants Due to N-Terminal Acetyltransferase Deficiency

Alan F. Rope • Kai Wang [19] • Rune Evjenth • ... Mark Yandell • Thomas Arnesen • Gholson J. Lyon [20] •

Show all authors • Show footnotes

Open Archive • Published: June 23, 2011 • DOI: https://doi.org/10.1016/j.ajhg.2011.05.017

PlumX Metrics

Introduction

Subjects and Methods

Results

Discussion

Acknowledgments

Supplemental Data

Web Resources

References

Article info

We have identified two families with a previously undescribed lethal X-linked disorder of infancy; the disorder comprises a distinct combination of an aged appearance, craniofacial anomalies, hypotonia, global developmental delays, cryptorchidism, and cardiac arrhythmias. Using X chromosome exon sequencing and a recently developed probabilistic algorithm aimed at discovering disease-causing variants, we identified in one family a c.109T>C (p.Ser37Pro) variant in *NAA10*, a gene encoding the catalytic subunit of the major human N-terminal acetyltransferase (NAT). A parallel effort on a second unrelated family converged on the same variant. The absence of this variant in controls, the amino acid conservation of this region of the protein, the predicted disruptive change, and the co-occurrence in two unrelated families with the same rare

# N-terminal acetyltransferase (NAA10)

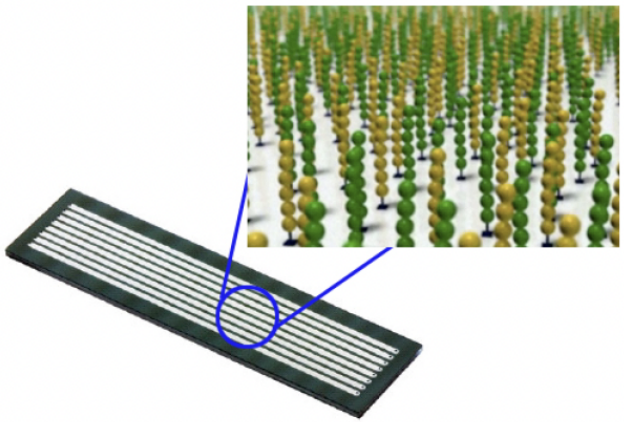
N-terminal acetyltransferase (NAT):

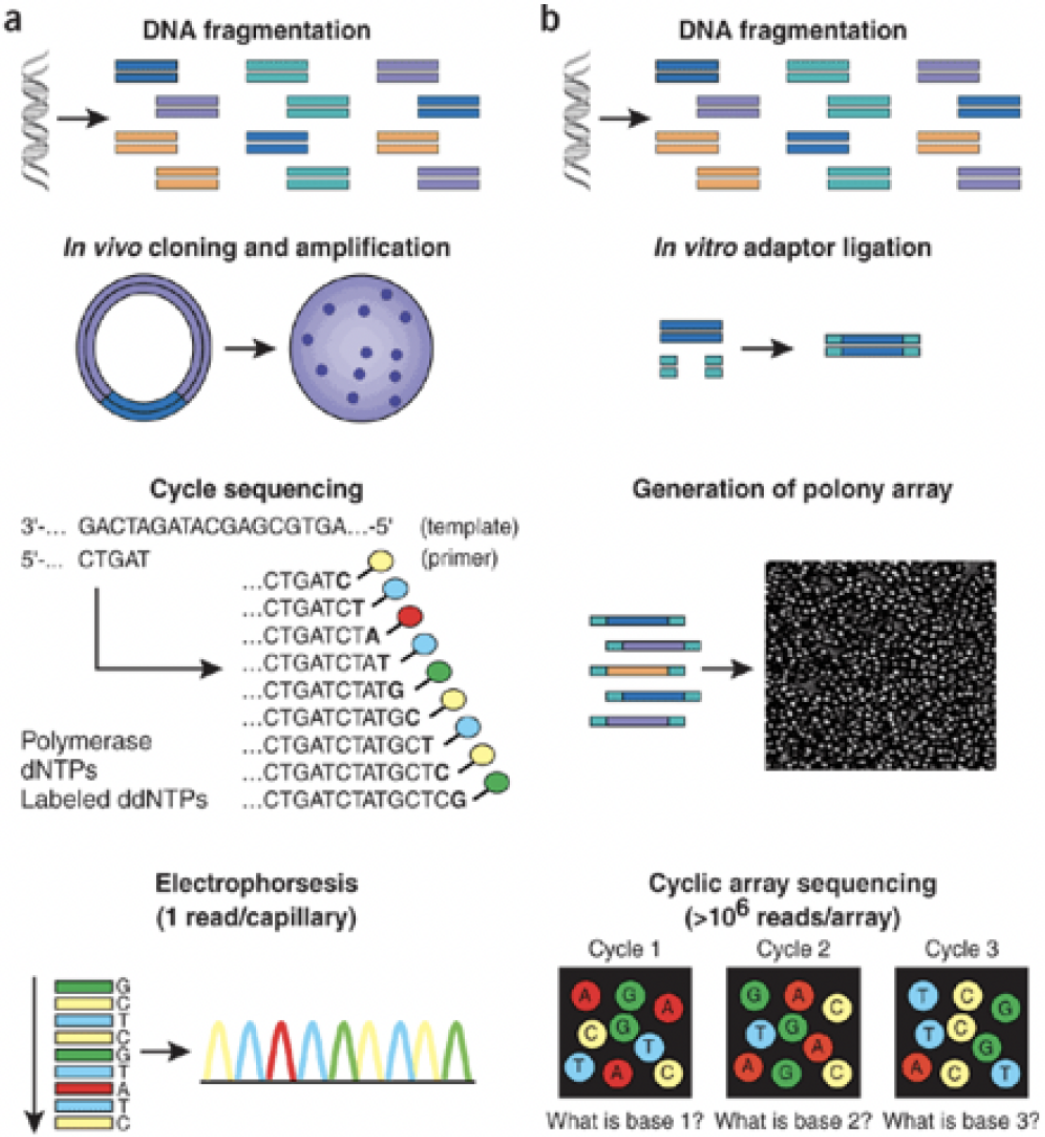
- Common modification (~80-90% of human proteins)
- Depletion from cancer cells linked to cell cycle arrest and apoptosis (Starheim, *BMC Proc* 2009)
- NAT genes directly implicated as cause of genetic disease
- Mutation demonstrated a significantly impaired biochemical activity *in vitro*
- NAA10 lethal if knocked out of Drosophila

Section 2

# Processing Raw Data

# Next-Generation Sequencing






Shendure and Ji, *Nature Biotechnology, 2008*.

# Study Design

Whole Exome

- Less expensive
- Nearly complete ascertainment of variation in the coding ~1% of the genome (i.e. exons)
- Will miss functional variants outside of the coding region

Low Coverage Whole Genome

- Less expensive
- Reasonably good ascertainment of shared variation, but not unique variation

Deep Whole Genome

- More expensive
- Capture most of the genetic information
- Sequence the entire genome of each subject

# History and Evolution of Illumina Data Output

Illumina sequencers have given output in many different formats:

- Illumina .PRB and .INT files
  - Better access to raw data.
  - Base calling algorithms (Bravo and Irizarry, *Biometrics*, 2010)
  - Mapping algorithms (GNUMAP, NOVO)
  - Confusing formats; Large data files
- Illumina .FASTQ files
- Sanger .FASTQ files

# Illumina .INT and .PRB

```
1    1    125 771 1651.8 2189.6 228.1 549.9   219.0 202.5  48.4 3016.8    127.8   6.1 204
1    1    478 16   1050.0 969.9 149.5 311.7    0.0   0.0  39.3 134.1   0.0   0.0   0.0   0.
1    1    780 553  639.6 980.8 555.2 6412.8   1040.1 4408.7 750.3 638.1   946.2 4351.1 7
1    1    123 685 -116.5 341.5 -14.0 -985.9   231.5 1090.0  88.3 -102.2   240.2 513.6
1    1    61   934  40.7  87.3  38.5  21.3  16.4  31.7 100.9  68.8      41.4  29.4  40.4  3
1    1    866 972 2820.5 4698.9 435.8 8502.2 4740.3 4890.2 1491.5 1241.7 2137.5 2505.6 6
```

```
 40 -40 -40 -40  -40 -40 -40  40  -40 -40  40 -40  -40 -40  40 -40  -4
 28 -28 -40 -40   -5  -5  -5  -5    -5  -5  -5  -5    -5  -5  -5  -5    -5
-17 -16 -30  13  -40  26 -26 -40  -21  17 -20 -27   21 -21 -40 -40  -4
-40  40 -40 -40  -40  40 -40 -40   -1   1 -40 -40    -0 -40  -0 -14   26
 -1 -40  -0 -10  -20 -18  16 -35   -1  -8  -2 -13   12 -34 -13 -26   5
-27 -40 -40  27   40 -40 -40 -40  -40 -40  40 -40  -40 -40  40 -40  -4
```

# .fasta or .fa

```
>chrX
ttgaactcctgacctcaggtgatccgcccgccttgacctcccaaagcgct
cctgcctcagcctcccgagtagctgggactacaggtgcctgccaccatgc
....
caggctaatttttgtatttttagtagagacggggtttcaccatgttagc
caggatggtctcaatctcctgacctcatgatccgcctgcctcggcctccc
>chrY
tgtacactttaaatgggtgaattttatggaatgtgaattataCGTGTTGG
CTTGTAAAAAAAAATGATGGAGATGGAGACGTGACTCTAGCGTGAAGGGG
...
GTGGGGAGAGTAGATCTAGAGTGGAGACACCACTTTTAGGAGGTATGATC
cctgccaccatgcctggctaattttgtattttagtagagacagggttt
```

# .fastq or .fq (or .fq.gz)

```
@HWI-EAS240_0001:2:1:1142:17571#0/1
CTCTCTTTTCTCCCCANGTCTCCTCATGACCATATCCNTGTTGTCCATTGTGTANGGNNNCTTC
+HWI-EAS240_0001:2:1:1142:17571#0/1
bababb`abbbbabbYB^[^^`_`_bbbbbbbbbb`_B`_``^`^bbbba^`^^BOOBBBNOOC
```

```
@HWI-EAS240_0001:2:1:1142:17571#0/1
CTCTCTTTTCTCCCCANGTCTCCTCATGACCATATCCNTGTTGTCCATTGTGTANGGNNNCTT
+HWI-EAS240_0001:2:1:1142:17571#0/1
bababb`abbbbabbYB^[^^`_`_bbbbbbbbbb`_B`_``^`^bbbba^`^^BOOBBBNOO
@HWI-EAS240_0001:2:1:1142:6453#0/1
CAGGACGTGCACTATGNCATCCACGATGCAAGTCTTANCATTATTCAGGATACANATANNGGC
+HWI-EAS240_0001:2:1:1142:6453#0/1
_U][aa_aa^^```^[B][][]]]S```Y``a`aaZ]B]]^^^[^```[OVY]XBNNNBB]]]
@HWI-EAS240_0001:2:1:1142:19443#0/1
TCAGAAAACAGAAAGGTCATTTTCTTACTTCTTTGCANGATGCCACCCTCCAGANCAGNNAAT
+HWI-EAS240_0001:2:1:1142:19443#0/1
bbbabb^^bbb^bbabb]bb]^]`]]]bbbbbbbbYYBYYXOOOOOBb`bbVY^HBOOOBBOJO
```

# .FASTQ Comparison

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 .........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                          |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
   (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

# Quality scores (PHRED)

Quality scores (Phred)

- Sanger Phred: Range=(0,40), $P = 1 - 10^{-(ASCII-33)/10}$
- Solexa: Range= (-5,40), $P = \frac{10^{(ASCII-64)/10)}}{1+10^{(ASCII-64)/10}}$
- Illumina 1.3: (0,40), $P = \frac{10^{(ASCII-64)/10)}}{1+10^{(ASCII-64)/10}}$
- Illumina 1.5: Range=(2,40), $P = 1 - 10^{-(ASCII-64)/10}$
- Illumina 1.8: Same as Sanger except Range=(0,41)

# .FASTQ Comparison

> Sanger

| | ASCII | Quality |
|---|---|---|
| ! | 33 | 0.0000 |
| " | 34 | 0.2057 |
| # | 35 | 0.3690 |
| $ | 36 | 0.4988 |
| % | 37 | 0.6019 |
| & | 38 | 0.6838 |
| ' | 39 | 0.7488 |
| ( | 40 | 0.8005 |
| ) | 41 | 0.8415 |
| * | 42 | 0.8741 |
| + | 43 | 0.9000 |
| , | 44 | 0.9206 |
| - | 45 | 0.9369 |
| . | 46 | 0.9499 |
| / | 47 | 0.9602 |
| 0 | 48 | 0.9684 |
| 1 | 49 | 0.9749 |
| 2 | 50 | 0.9800 |

> Solexa

| | ASCII | Quality |
|---|---|---|
| ; | 59 | 0.2403 |
| < | 60 | 0.2847 |
| = | 61 | 0.3339 |
| > | 62 | 0.3869 |
| ? | 63 | 0.4427 |
| @ | 64 | 0.5000 |
| A | 65 | 0.5573 |
| B | 66 | 0.6131 |
| C | 67 | 0.6661 |
| D | 68 | 0.7153 |
| E | 69 | 0.7597 |
| F | 70 | 0.7992 |
| G | 71 | 0.8337 |
| H | 72 | 0.8632 |
| I | 73 | 0.8882 |
| J | 74 | 0.9091 |
| K | 75 | 0.9264 |
| L | 76 | 0.9406 |
| M | 77 | 0.9523 |
| N | 78 | 0.9617 |

> Illumina1.3

| | ASCII | Quality |
|---|---|---|
| @ | 64 | 0.5000 |
| A | 65 | 0.5573 |
| B | 66 | 0.6131 |
| C | 67 | 0.6661 |
| D | 68 | 0.7153 |
| E | 69 | 0.7597 |
| F | 70 | 0.7992 |
| G | 71 | 0.8337 |
| H | 72 | 0.8632 |
| I | 73 | 0.8882 |
| J | 74 | 0.9091 |
| K | 75 | 0.9264 |
| L | 76 | 0.9406 |
| M | 77 | 0.9523 |
| N | 78 | 0.9617 |
| O | 79 | 0.9693 |
| P | 80 | 0.9755 |
| Q | 81 | 0.9804 |
| R | 82 | 0.9844 |
| S | 83 | 0.9876 |

> Illumina1.5

| | ASCII | Quality |
|---|---|---|
| B | 66 | 0.3690 |
| C | 67 | 0.4988 |
| D | 68 | 0.6019 |
| E | 69 | 0.6838 |
| F | 70 | 0.7488 |
| G | 71 | 0.8005 |
| H | 72 | 0.8415 |
| I | 73 | 0.8741 |
| J | 74 | 0.9000 |
| K | 75 | 0.9206 |
| L | 76 | 0.9369 |
| M | 77 | 0.9499 |
| N | 78 | 0.9602 |
| O | 79 | 0.9684 |
| P | 80 | 0.9749 |
| Q | 81 | 0.9800 |
| R | 82 | 0.9842 |
| S | 83 | 0.9874 |
| T | 84 | 0.9900 |
| U | 85 | 0.9921 |

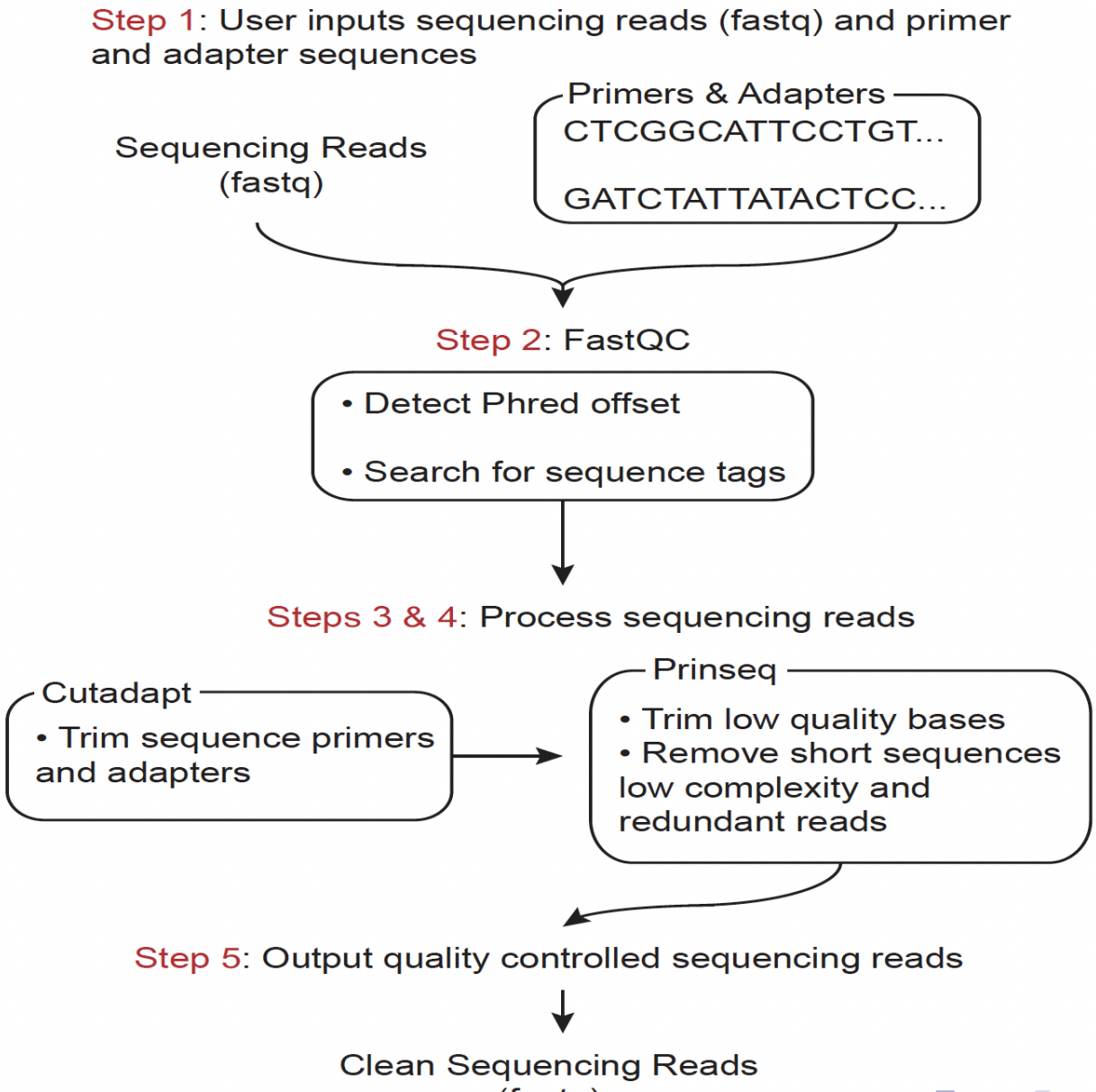
# Quality Control

Need to preprocess the reads:

- Check for quality (FASTQC)
- Trim adapter and (Cutadapt, others)
- Remove duplicate reads, trim low complexity/quality bases/reads (Prinseq)
- Complete pipelines: NCBI Toolkit, QC-Chain, PathoQC (pathoscope.sourceforge.net), others

**Note:** Not comprehensive or updated!

# Read Quality Checks

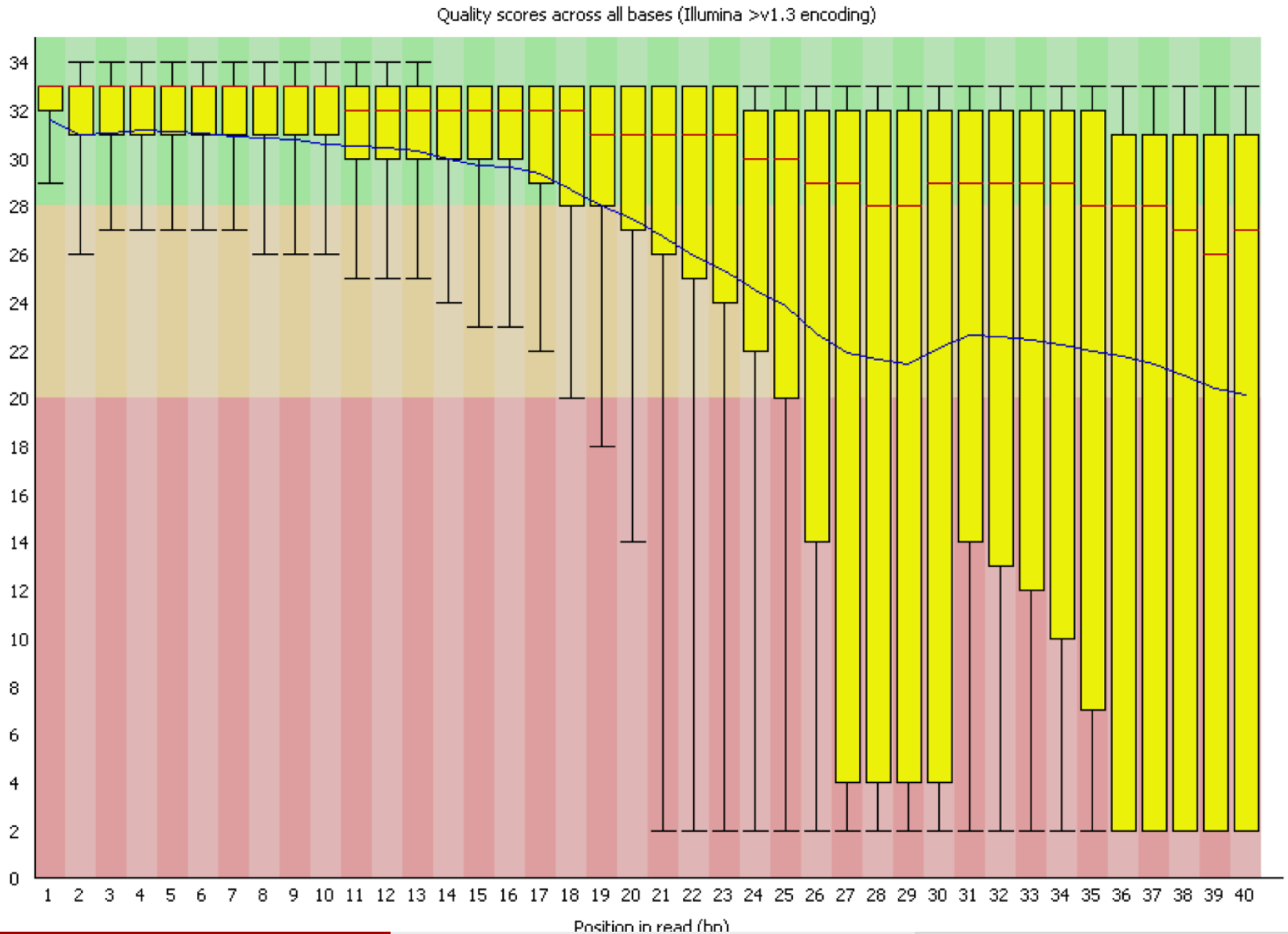


# Read Quality Checks

Comparison of some of the more popular QC pipelines:

| QC Features | QCToolkit | QC-Chain | Cutadapt | Prinseq |
|---|---|---|---|---|
| Parallel computation | X | X | | |
| Phred offset detection | X | X | | |
| Tag sequence removal | X | X | X | |
| Poly-A/T tail trimming | | | X | X |
| PCR duplication filtering | | X | | X |
| Low complexity filtering | | | | X |
| Homopolymer removal | X | | | X |
| GC content filtering | | X | | X |
| N/X content filtering | | | | X |

# FASTQC


Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
Position in read (bp)

# FastQC Example

Running FastQC from command-line:

```
fastqc myfastqfile.fq --outdir=output/
```

Interactive GUI:

```
fastqc
```

Section 3

# Mapping Reads

# Sequencing data alignment

Next-Generation Sequencing (NGS) Data present new challenges:

- Map 'reads' to genome
- Call SNPs and variants from the reads
- Other Applications: RNA-seq, miRNAs, alternative splicing, ChIP-seq, BS-seq, RNA editing

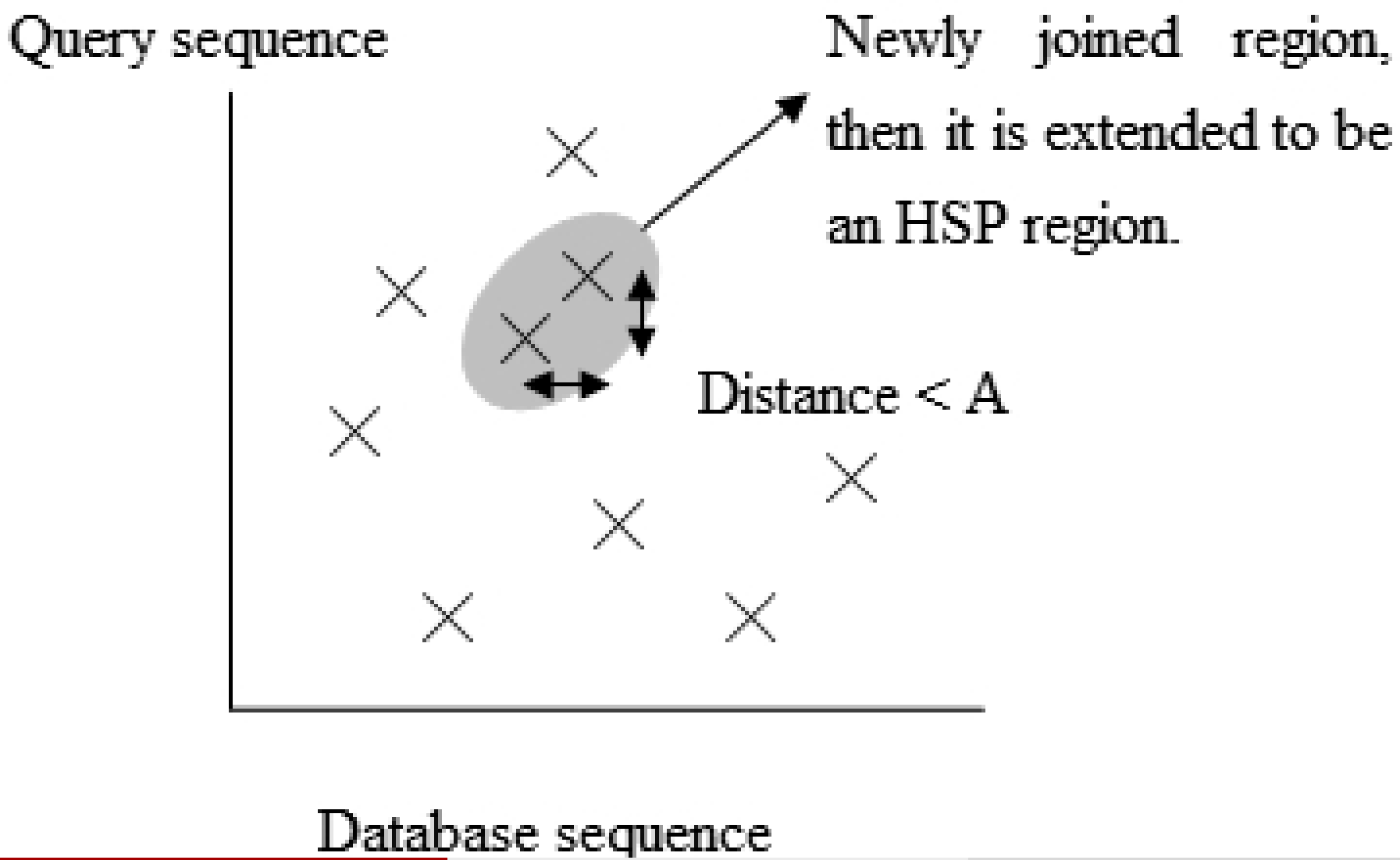
Obstacles:

- Massive data size
- Repeat regions, rare variants

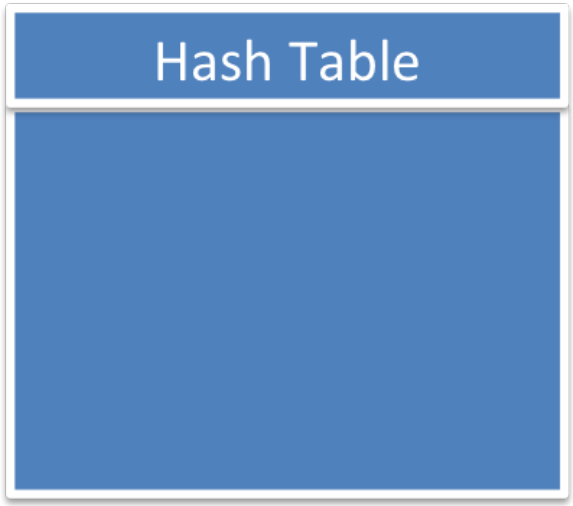
**Goal:** Develop an approach to put the puzzle together!

# Basic Local Alignment Search Tool (BLAST)

BLAST seed extension:

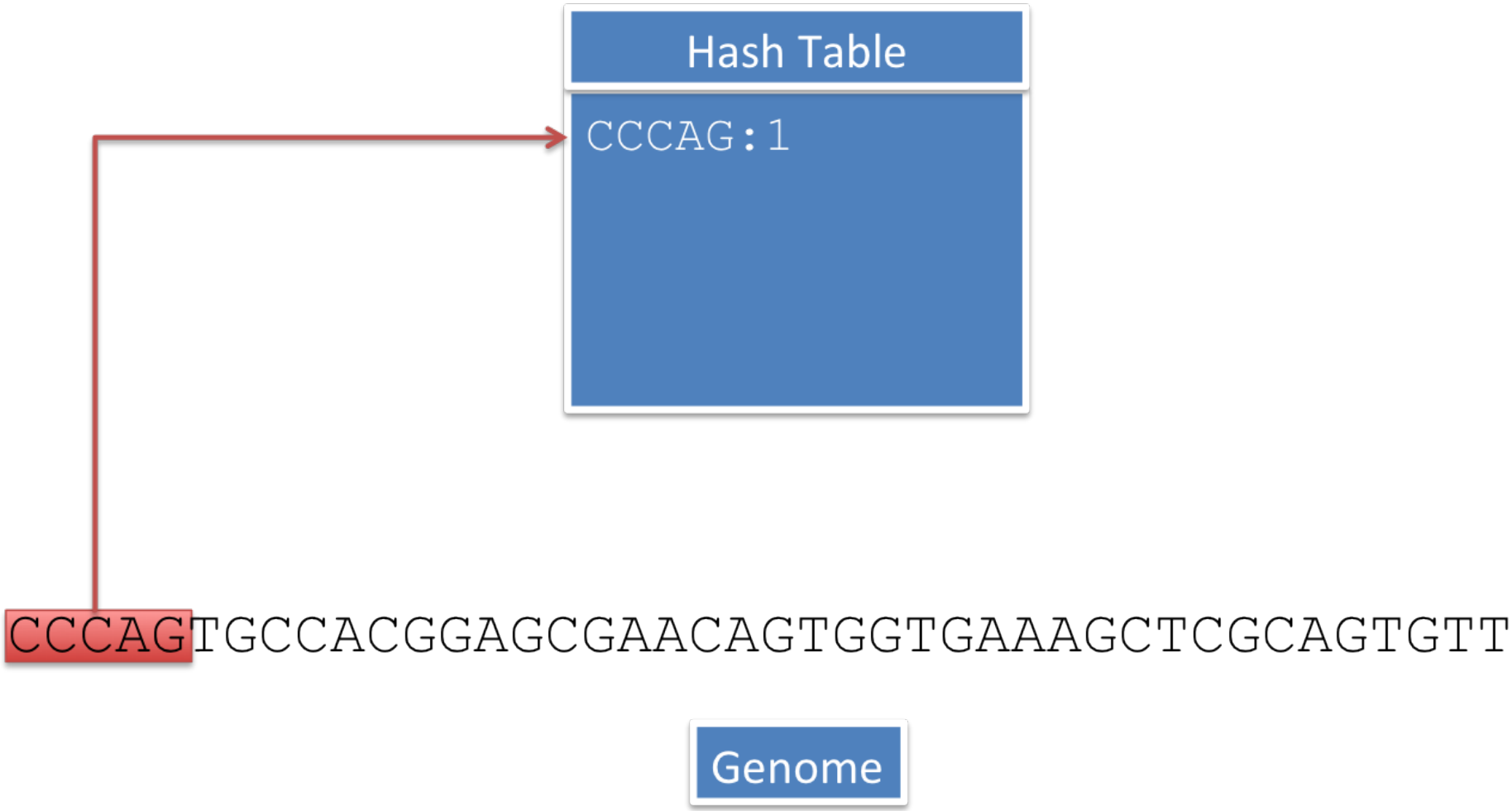


# Creating a Hash Index Table

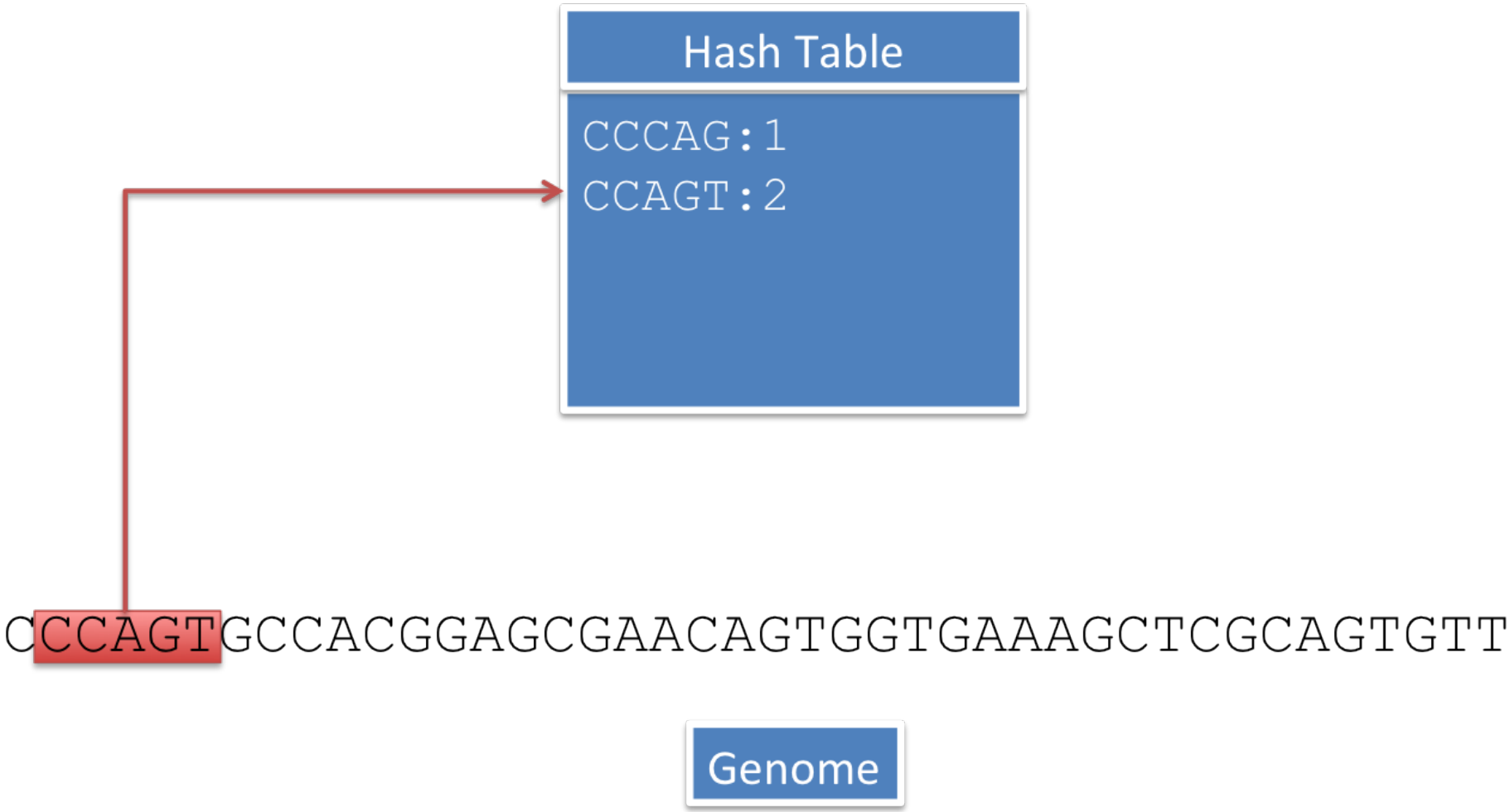



CCCAGTGCCACGGAGCGAACAGTGGTGAAAGCTCGCAGTGTT

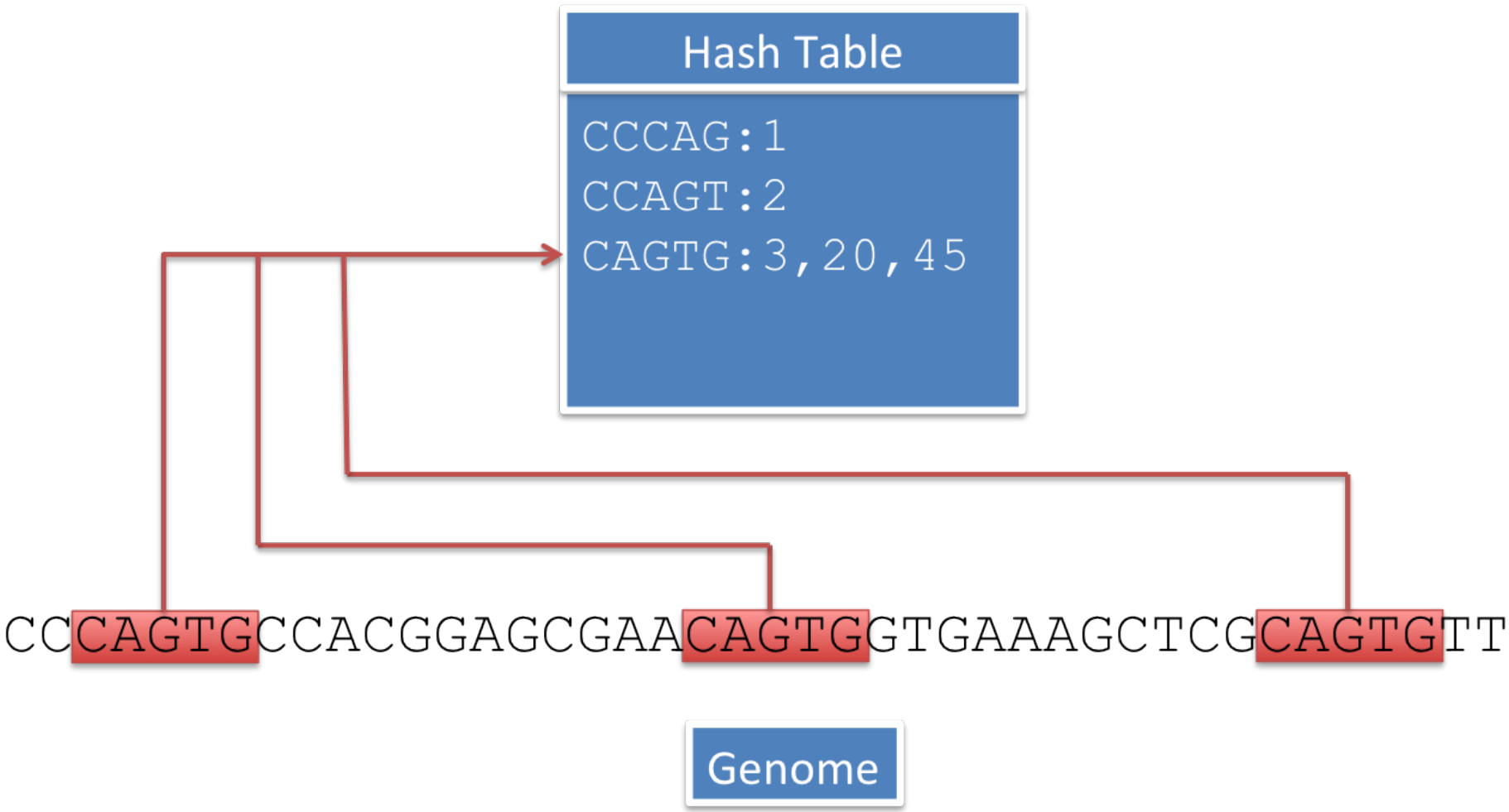
# Creating a Hash Index Table



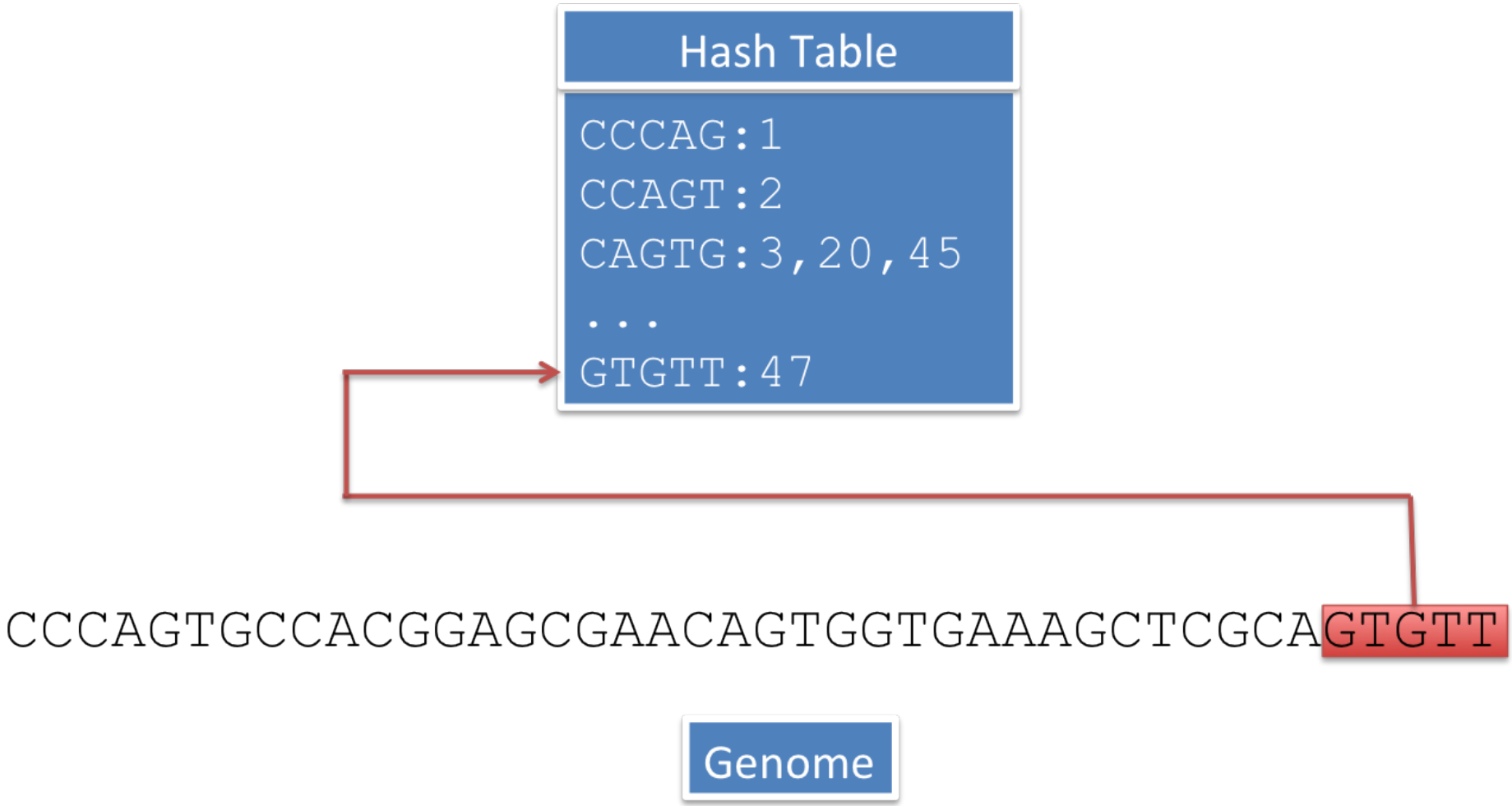
# Creating a Hash Index Table



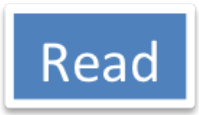
# Creating a Hash Index Table



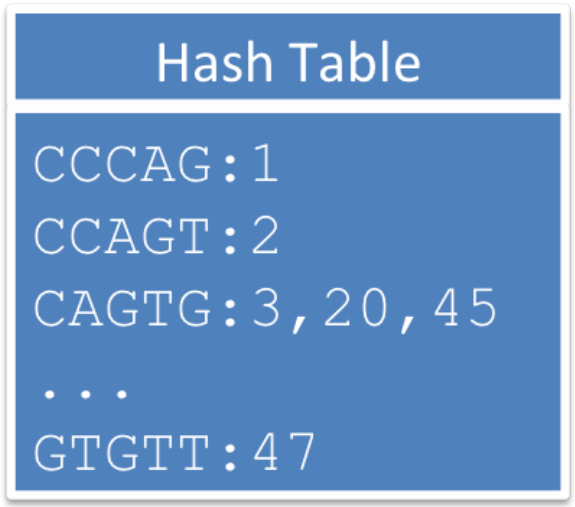
# Creating a Hash Index Table



# Read Lookup

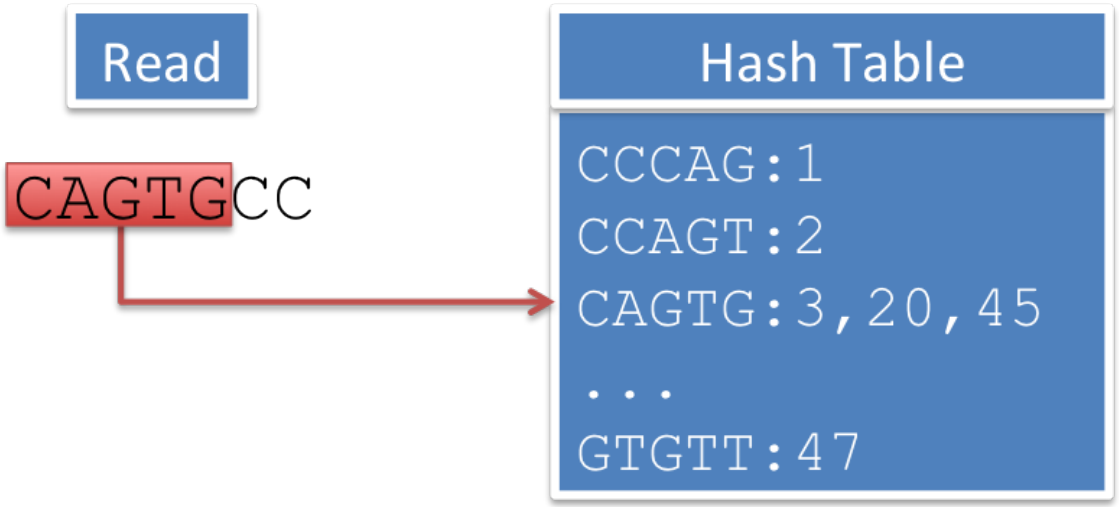



CAGTGCC




CCCAGTGCCACGGAGCGAACAGTGCTGAAAGCTCGCAGTGTT

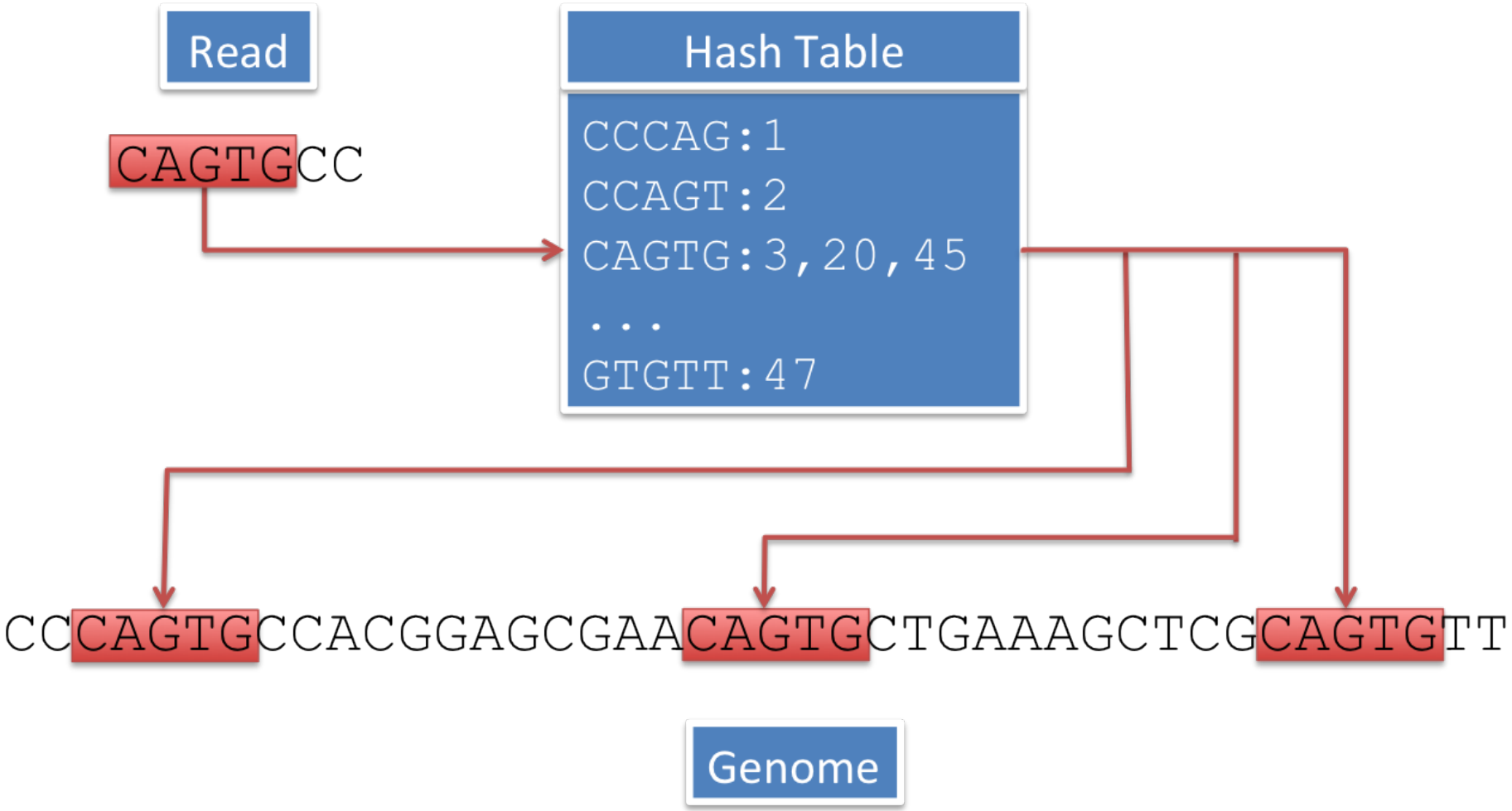
# Read Lookup



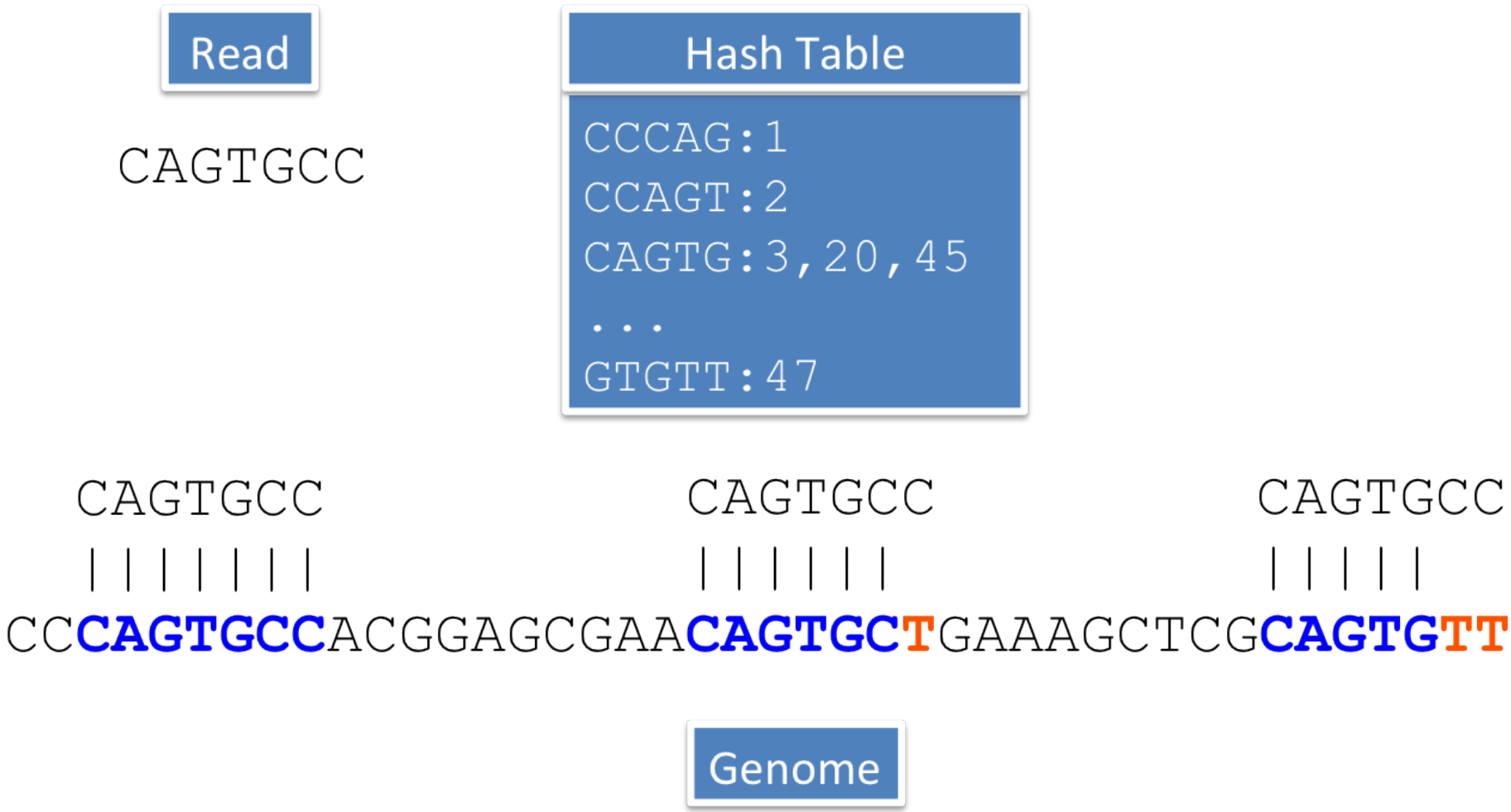

CCCAGTGCCACGGAGCGAACAGTGCTGAAAGCTCGCAGTGTT

Genome

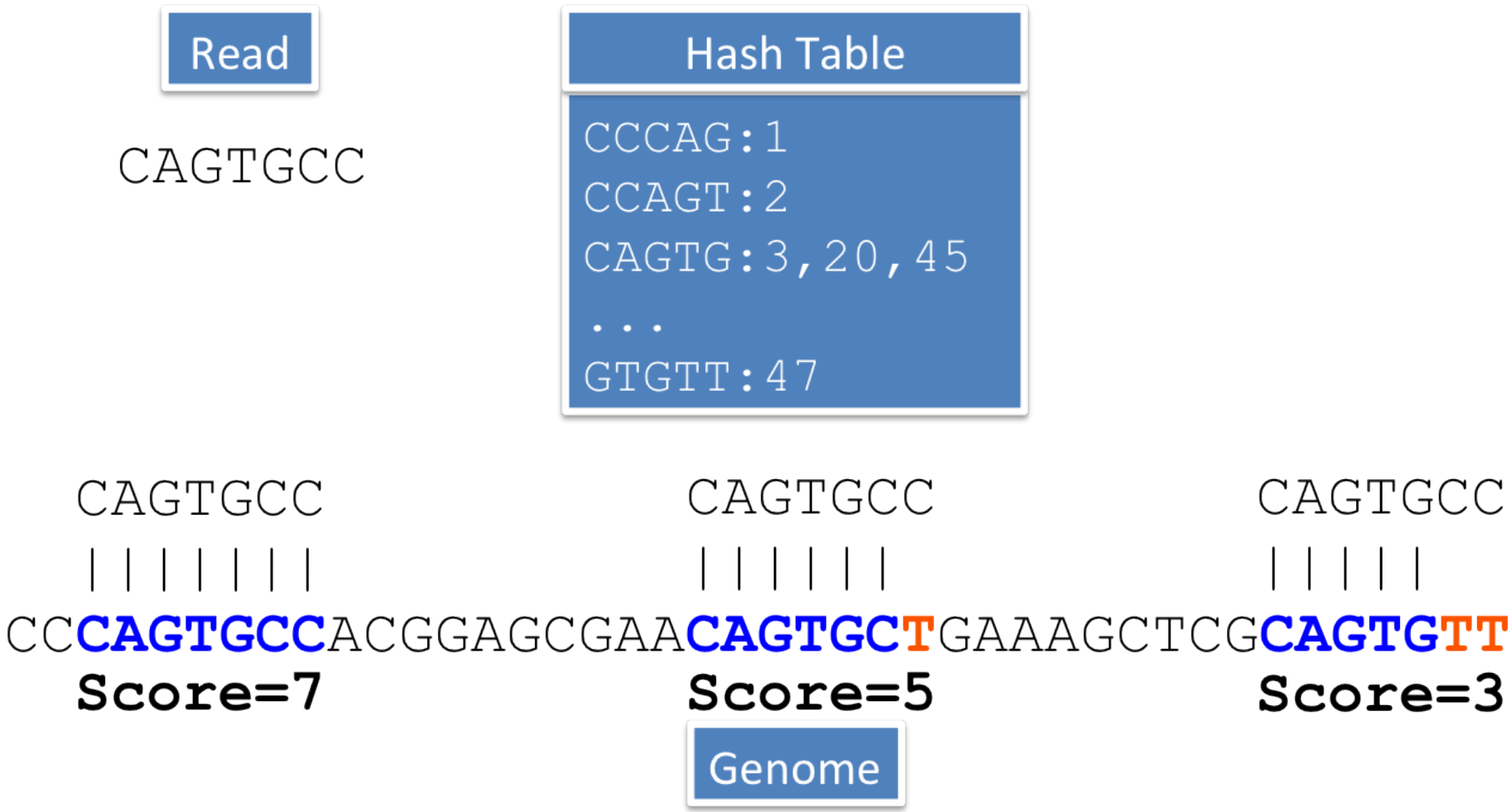
# Read Lookup


Read
CAGTGCC
Hash Table
CCCAG:1
CCAGT:2
CAGTG:3,20,45
...
GTGTT:47
CCCAGTGCCACGGAGCGAACAGTGCTGAAAGCTCGCAGTGTT
Genome

# Read Lookup


Read
CAGTGCC
Hash Table
CCCAG:1
CCAGT:2
CAGTG:3,20,45
...
GTGTT:47
CAGTGCC
CAGTGCC
CAGTGCC
CCCAGTGCCACGGAGCGAACAGTGCTGAAAGCTCGCAGTGTT
Genome

# Read Lookup


Read
CAGTGCC
Hash Table
CCCAG:1
CCAGT:2
CAGTG:3,20,45
...
GTGTT:47
CAGTGCC
CAGTGCC
CAGTGCC
CCCAGTGCCACGGAGCGAACAGTGCTGAAAGCTCGCAGTGTT
Score=7
Score=5
Score=3
Genome

# Session info

```
sessionInfo()
```

```
## R version 4.4.0 (2024-04-24)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sonoma 14.2.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK version 3
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/Denver
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.4.0    fastmap_1.1.1     cli_3.6.2         tools_4.4.0
##  [5] htmltools_0.5.8.1 rstudioapi_0.16.0 yaml_2.3.8        rmarkdown_2.26
##  [9] knitr_1.46        xfun_0.43         digest_0.6.35     rlang_1.1.3
## [13] evaluate_0.23
```