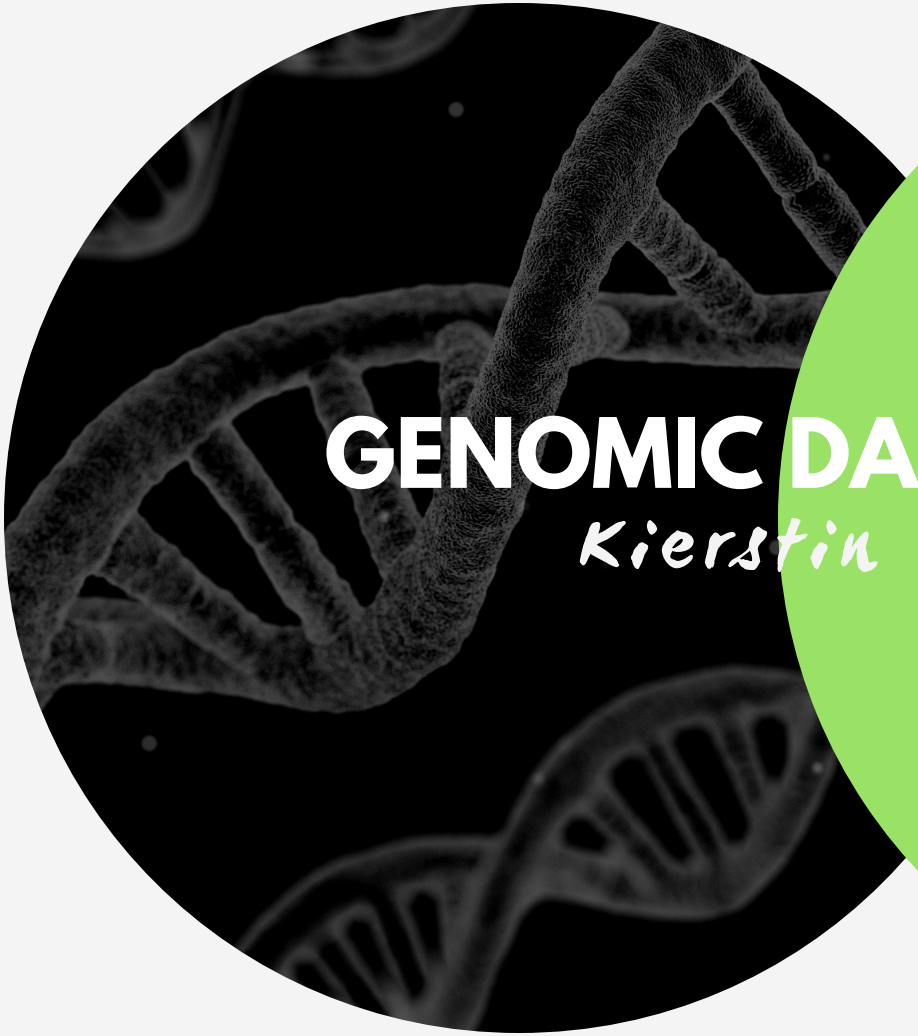


GENOMIC DATA SCIENCE

Kierstin Matsuda

For Dr. Giri | Summer 2018 | CIS 5900 Independent Study

COURSERA GENOMIC DATA SCIENCE SPECIALIZATION

"This specialization covers the concepts and tools to understand, analyze, and interpret data from next generation sequencing experiments. It teaches the most common tools used in genomic data science including how to use the command line, Python, R, Bioconductor, and Galaxy. The sequence is a stand alone introduction to genomic data science or a perfect compliment to a primary degree or postdoc in biology, molecular biology, or genetics."

07 Courses

- Course material primarily instructional and follow-along videos
- Lecture videos and readings included.
- Learn the languages Python and R
- Quiz questions embedded in videos
- Supplemented with YouTube videos, articles and Tedtalks
- Supplemented with Swirl R tutorials
- Supplemented with IBM's Python for Data Science

https://www.coursera.org/specializations/genomic-data-science

INTRODUCTION TO GENOMIC TECHNOLOGIES

”This course introduces you to the basic biology of modern genomics and the experimental tools that we use to measure it. We’ll introduce the Central Dogma of Molecular Biology and cover how next-generation sequencing can be used to measure DNA, RNA, and epigenetic patterns. You’ll also get an introduction to the key concepts in computing and data science that you’ll need to understand how data from next-generation sequencing experiments are generated and analyzed.”

GENOMIC DATA SCIENCE WITH GALAXY

"Learn to use the tools that are available from the Galaxy Project. This is the second course in the Genomic Big Data Science Specialization."

THE R PROGRAMMING ENVIRONMENT

"This course provides a rigorous introduction to the R programming language, with a particular focus on using R for software development in a data science setting."

Completed week 1 as a supplement to Galaxy Lesson

PYTHON FOR GENOMIC DATA SCIENCE

"This class provides an introduction to the Python programming language and the iPython notebook. This is the third course in the Genomic Big Data Science Specialization from Johns Hopkins University."

ALGORITHMS FOR GENOMIC DATA SCIENCE

”We will learn computational methods -- algorithms and data structures -- for analyzing DNA sequencing data. We will learn a little about DNA, genomics, and how DNA sequencing is used. We will use Python to implement key algorithms and data structures and to analyze real genomes and DNA sequencing datasets.”

Longest course of specialization, in-depth course about Python algorithms to analyze sequencing machine data. - IN PROGRESS

COMMAND-LINE FOR GENOMIC DATA SCIENCE

”Introduces to the commands that you need to manage and analyze directories, files, and large sets of genomic data. This is the fourth course in the Genomic Big Data Science Specialization from Johns Hopkins University.”
7 hours of lecture videos included - IN PROGRESS

BIOCONDUCTOR FOR GENOMIC DATA SCIENCE

”Learn to use tools from the Bioconductor project to perform analysis of genomic data. This is the fifth course in the Genomic Big Data Specialization from Johns Hopkins University.”

This course contains 40 videos

This course is taught in R, Bioconductor in RStudio - Half Finished

STATISTICS FOR GENOMIC DATA SCIENCE

"An introduction to the statistics behind the most popular genomic data science projects. This is the sixth course in the Genomic Big Data Science Specialization from Johns Hopkins University."

This course is taught in R, contains 64 instructional videos- IN PROGRESS

part 1 of 3

FROM DNA TO OS AND 1S

part 2 of 3

FUNCTIONS FOR PROCESSING READS

part 3 of 3

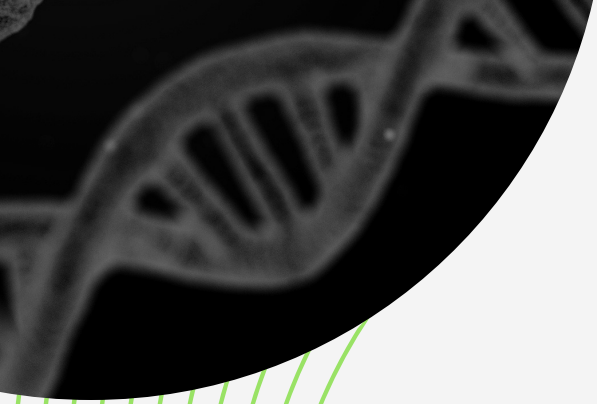
ANALYZING THE GENOME

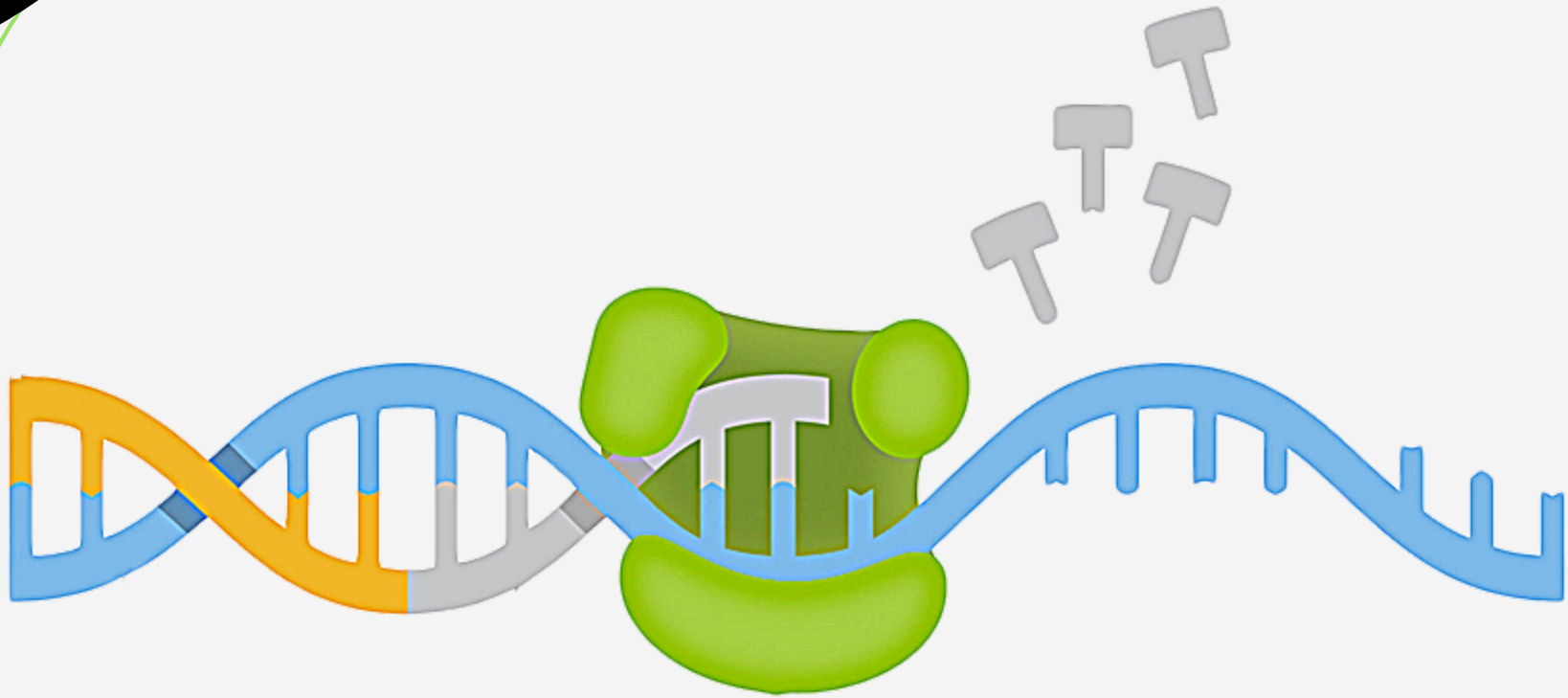
FROM DNA TO OS AND 1S

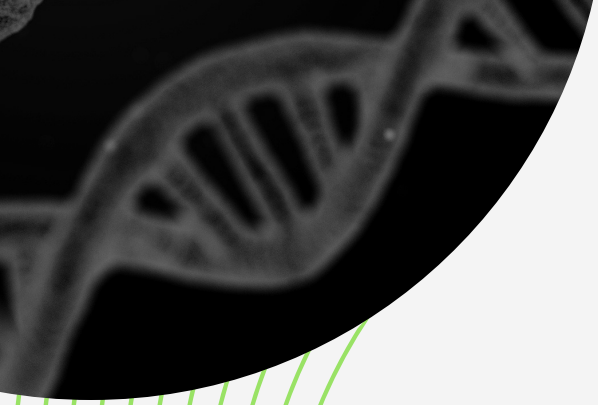
part 1 of 3

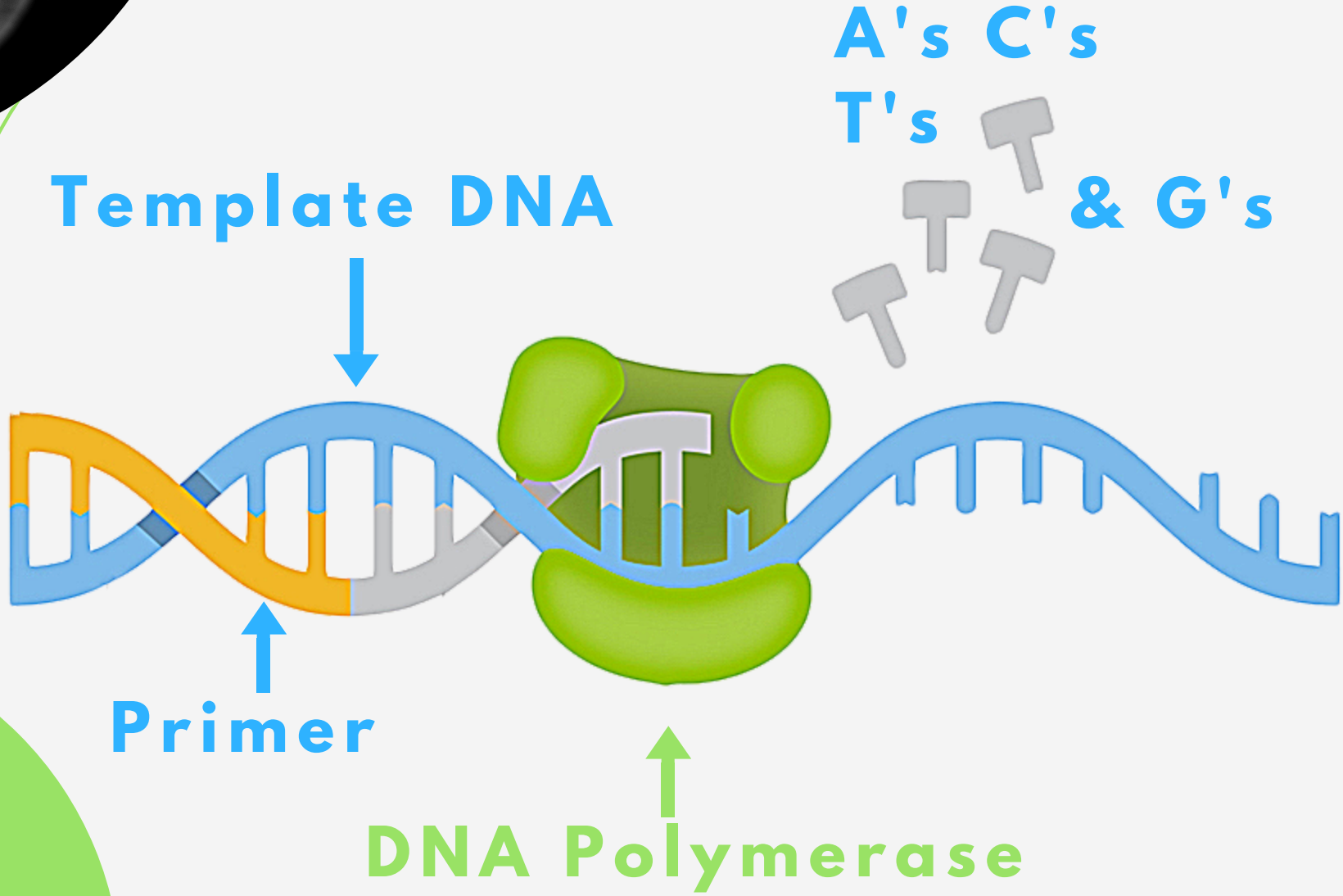
POLYMERASE

Chain Reaction

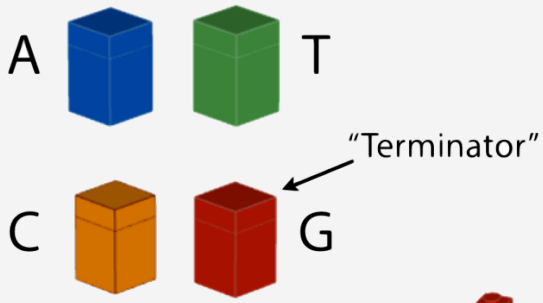


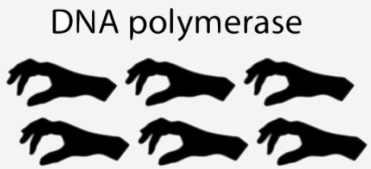





A's C's
T's
& G's
Template DNA
Primer
DNA Polymerase

MANIPULATING THE Polymerase Chain Reaction





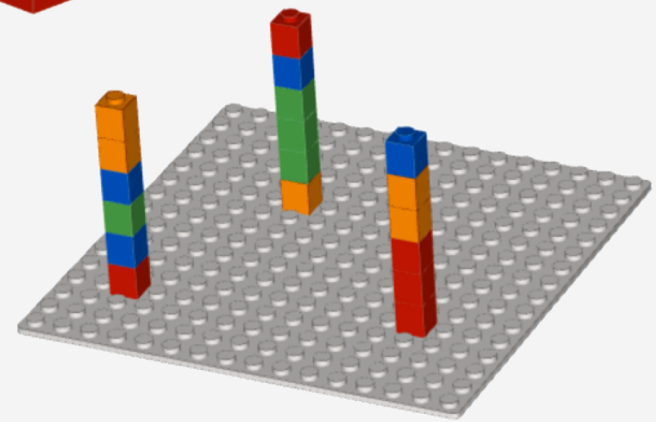


Image courtesy of Johns Hopkins University

MANIPULATING THE Polymerase Chain Reaction

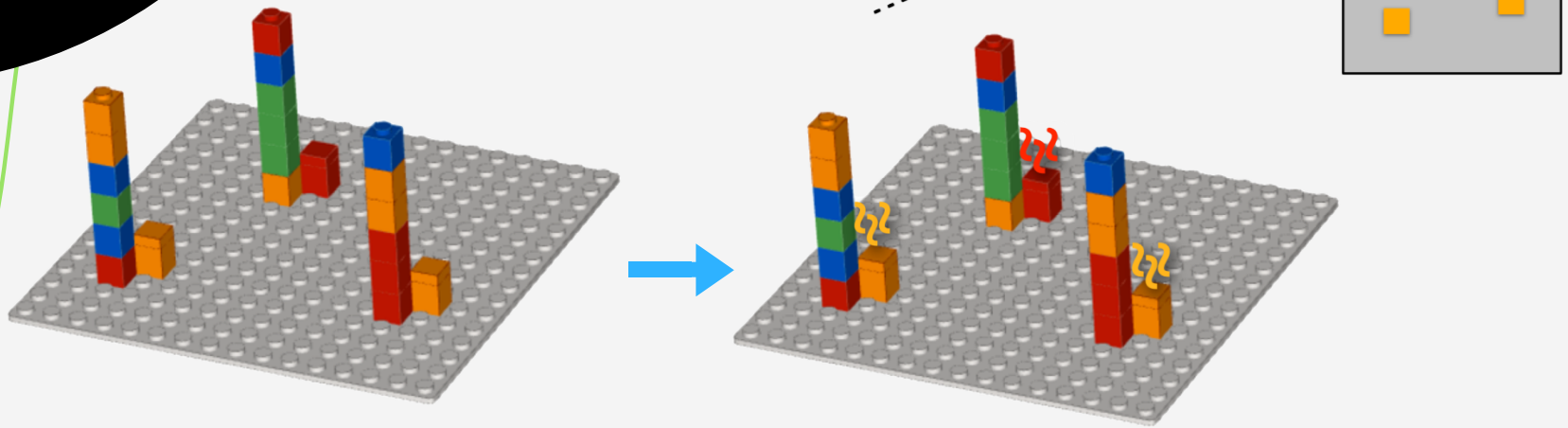


Image courtesy of Johns Hopkins University

MANIPULATING THE Polymerase Chain Reaction

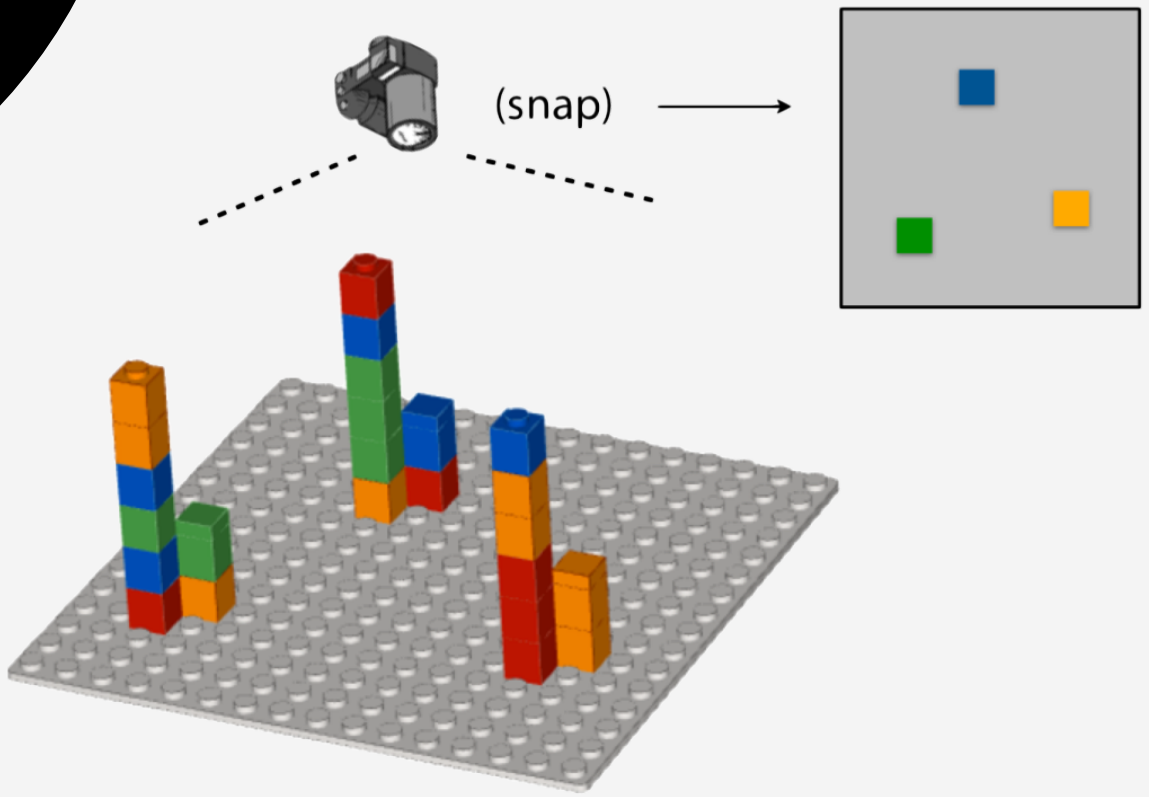
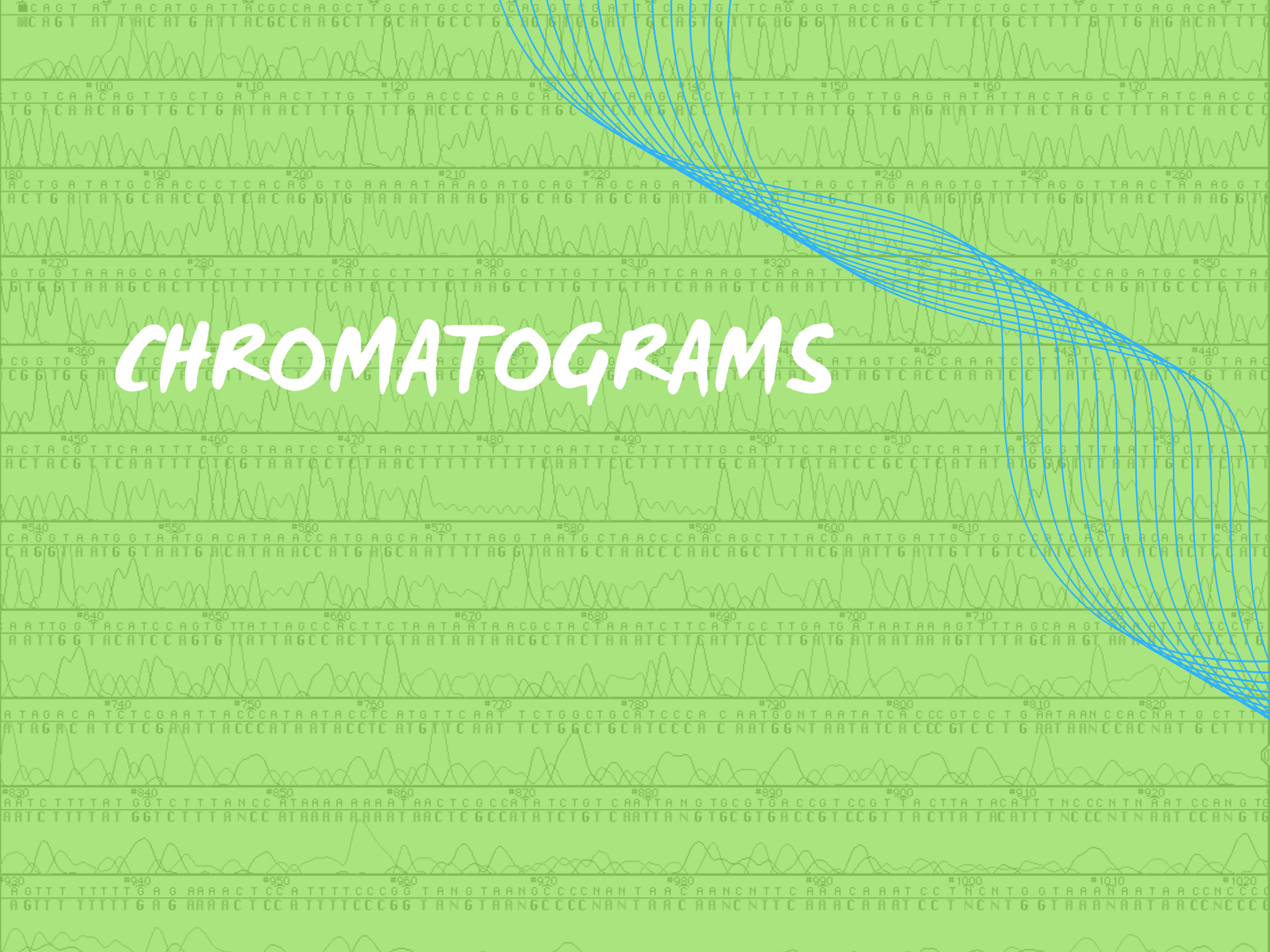
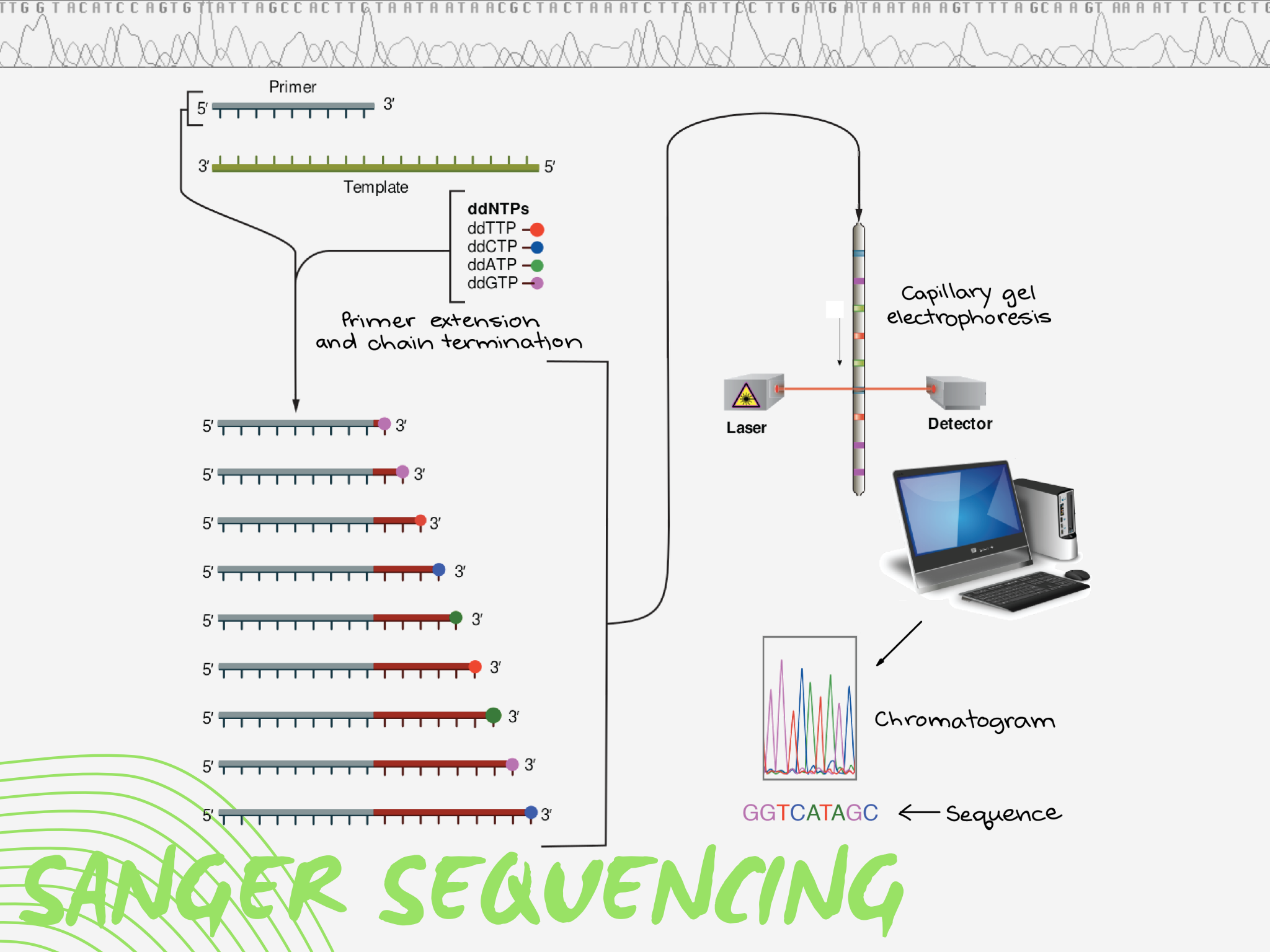
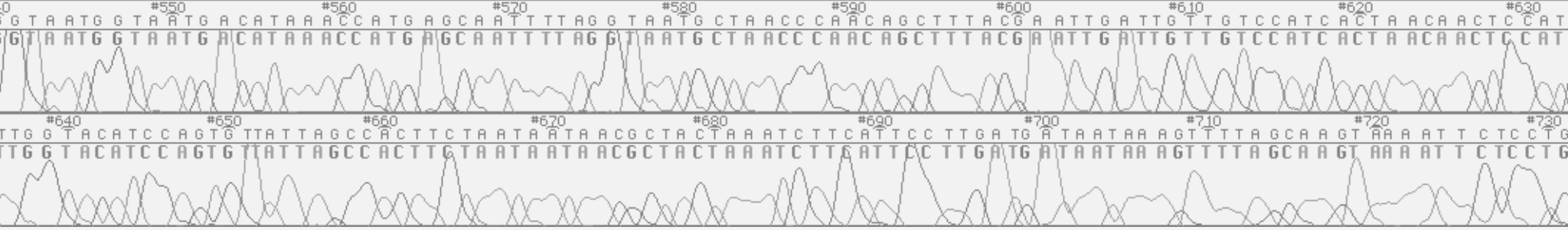


Image courtesy of Johns Hopkins University

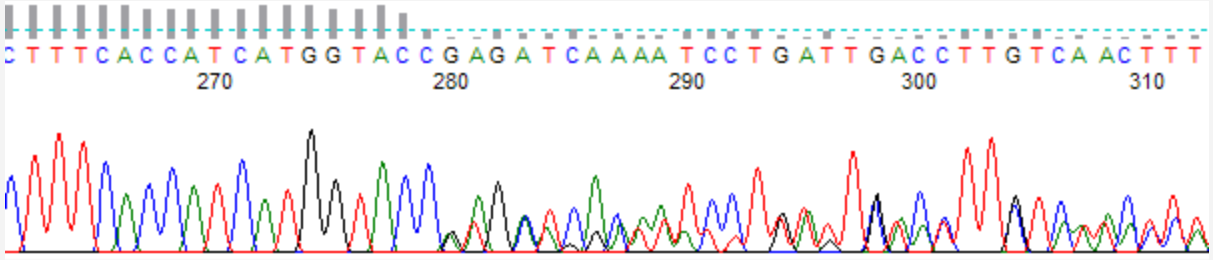

CHROMATOGRAMS

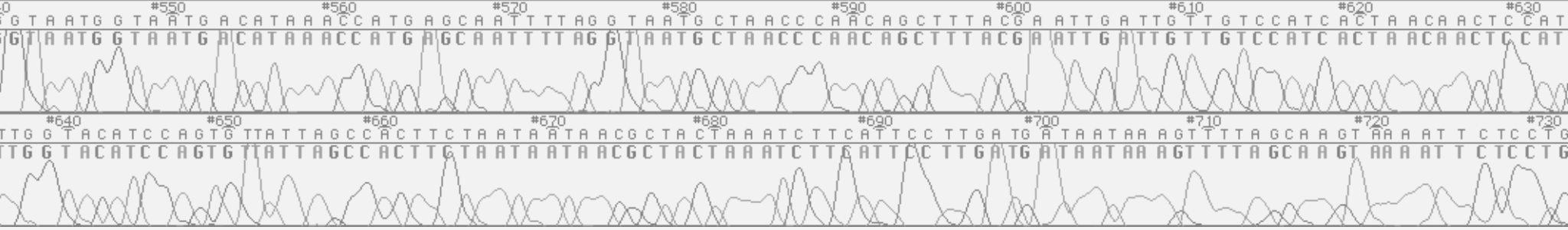

Primer
5′
3′
3′
5′
Template
ddNTPs
ddTTP
ddCTP
ddATP
ddGTP
Primer extension
and chain termination
Capillary gel
electrophoresis
Laser
Detector
Chromatogram
GGTCATAGC
Sequence
SANGER SEQUENCING



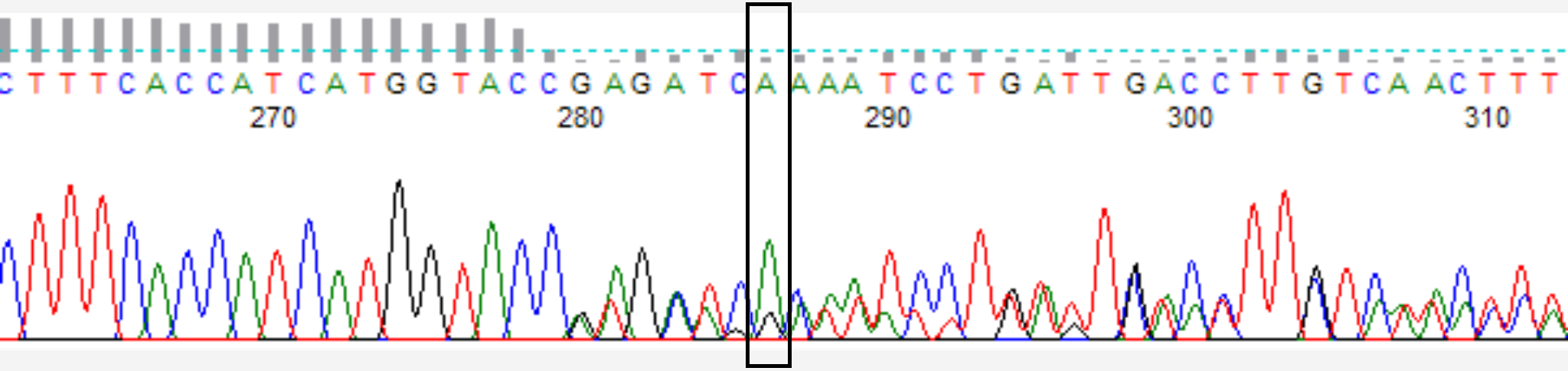
$$Q = -10 * \log_{10} p$$

When Q = 10	1 in 10 chance base is incorrect
When Q = 20	1 in 100 chance base is incorrect
When Q = 30	1 in 1,000 chance base is incorrect



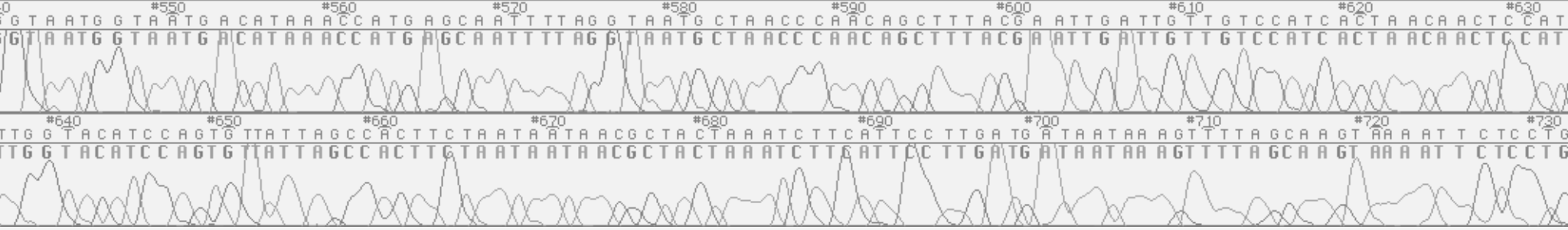


$$Q = -10 * \log_{10} p$$



$$p = 3 \text{ black} / 9 \text{ total}$$

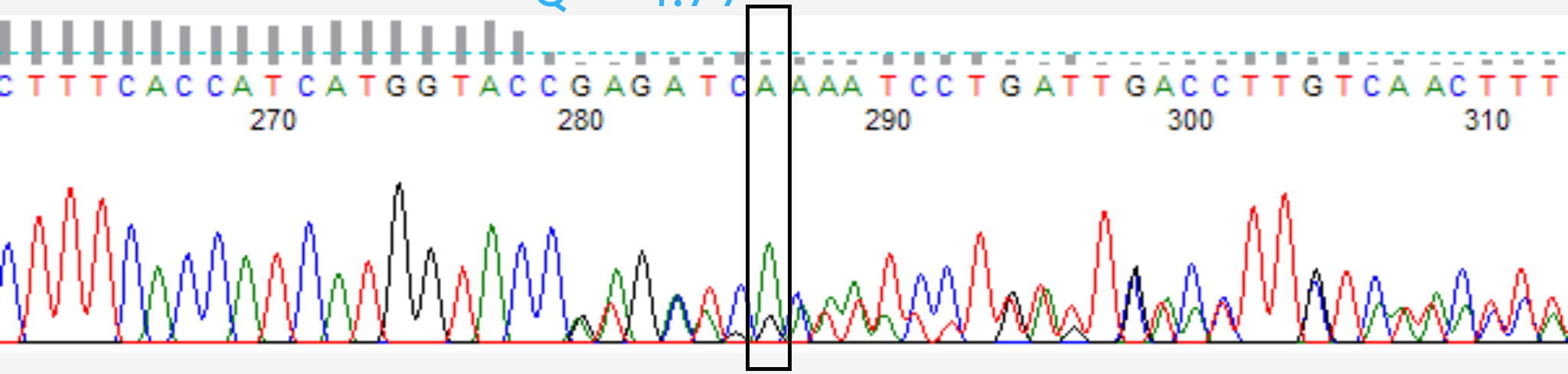
$$p = 1/3$$



$$Q = -10 * \log_{10} p$$

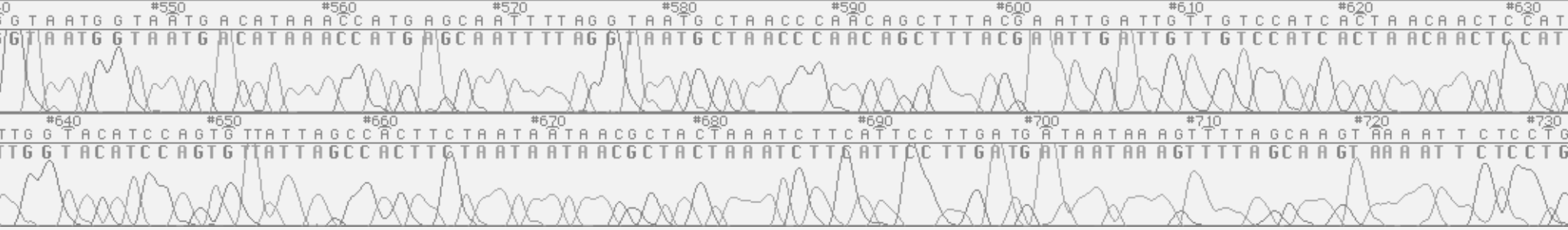
$$Q = -10 * \log_{10} (1/3)$$

$$Q = 4.77$$



p = 3 black / 9 total

p = 1/3



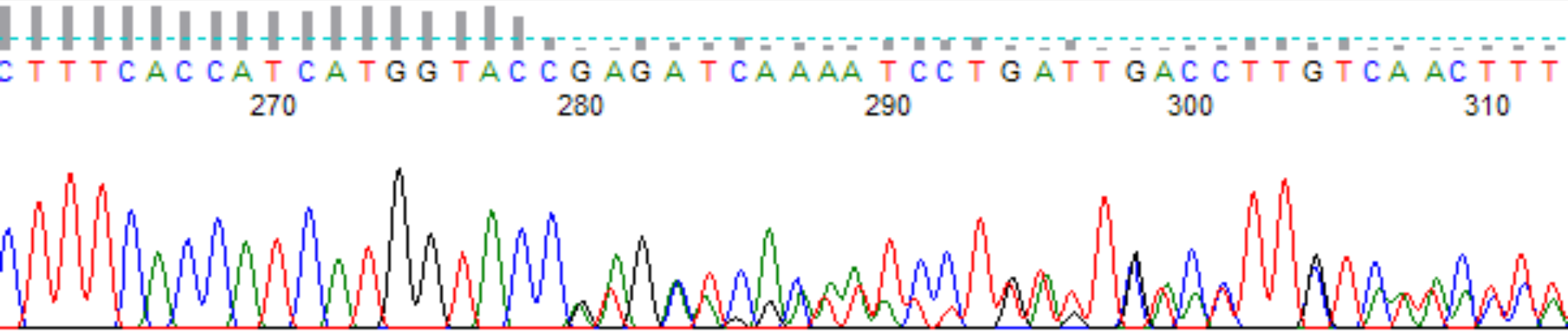
Genome Reads

@ERR037900.3 509.5.41.1218.7494/1

TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCNAACCCTAACCCTAACCCTAACCCTAACCCTAAC

+

HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGHHHHHHHHHHHHHHDHHHDEHHHHFGIHEHEGGGF4#45655366GIGEHAGBG##################



FUNCTIONS FOR PROCESSING READS

part 2 of 3

Genome Reads

A's C's
T's & G's

read metadata

READ DNA

@ERR037900.3 509.5.41.1218.7494/1
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCNAACCCTAACCCTAACCCTAACCCTAACCCTAAC
+
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGHHHHHHHHHHHHHHDHHHDEHHHHFGIHEHEGGGF4#45655366GIGEHAGBG##################

Read Qualities

Genome Reads

```
@ERR037900.1 509.8.8.8903.80024/1
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCNAACCCTAACCCTAACCCTAACCCTAACCCTAAC
+
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGFHHHFHFFHHHHHGHHFHEH@4#55554455HGFBF<@C>7EEF@FBEDDD<=C<E
@ERR037900.2 509.5.68.21343.17610/1
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCNAACCCTAACCCTAACCCTAACCCTNACCCTAAC
+
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCHHHHEHHBA#C>@54455C/7=CGHEGEB;C##########
@ERR037900.3 509.5.41.1218.7494/1
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCNAACCCTAACCCTAACCCTAACCCTAACCCTAAC
+
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGHHHHHHHHHHHHHDHHHDEHHHHFGIHEHEGGGF4#45655366GIGEHAGBG##############
@ERR037900.4 509.3.45.14457.64151/1
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCNAACCCTAACCCTAACCCTAACCCTAACCCTACC
+
HHHHHHHHHHHHHHHHHHHHHHHHHHHIHHHHHHHHHHHHHHHHHHHHHHHIHHHHHIHFHHHIHHHHD#ECA54655GGIBH?BD@+BCBF?5A=::>8?##
@ERR037900.5 509.4.43.16809.33112/1
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTNACCCTAACCCTAACCCTAACCCTAAACCTAACC
+
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIHIHEHHIGHIFFHIIGF6#555:2=7=CB;?3CAACBAC2B##########
@ERR037900.6 509.3.45.14110.19381/1
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTNACCCTAACCCTAACCCTAACCCTAACCCTAACC
+
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIHHHIHIHHHGH:#@@@@9@C@EEGCGGFIFFF9FCAF?EEE4B8>>
@ERR037900.7 509.2.63.5080.119531/1
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTNACCCTAACCCTAACCCTAACCCTAACCCTAACC
+
HHHHHHHHHHHHHHHHHHHHHHHIFHFEGGFHHHHHHGHHHHGHHHHHFHAFGHEHHIHHGBCCDC,#55564565CE:BB44+'5/36,(<<BC<DDBCE
@ERR037900.8 509.1.63.3336.158330/1
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTNAACCTAACCCTAACCCTAACTCTAACCCTAACC
+
HHFHHDHHHHDDGGGDHDHHHHHGHHHHHHHDHHECHHH8GGDEEHHHHEH?3HG<=4>555624/#5/55/555DADA####################
```

Read Qualities

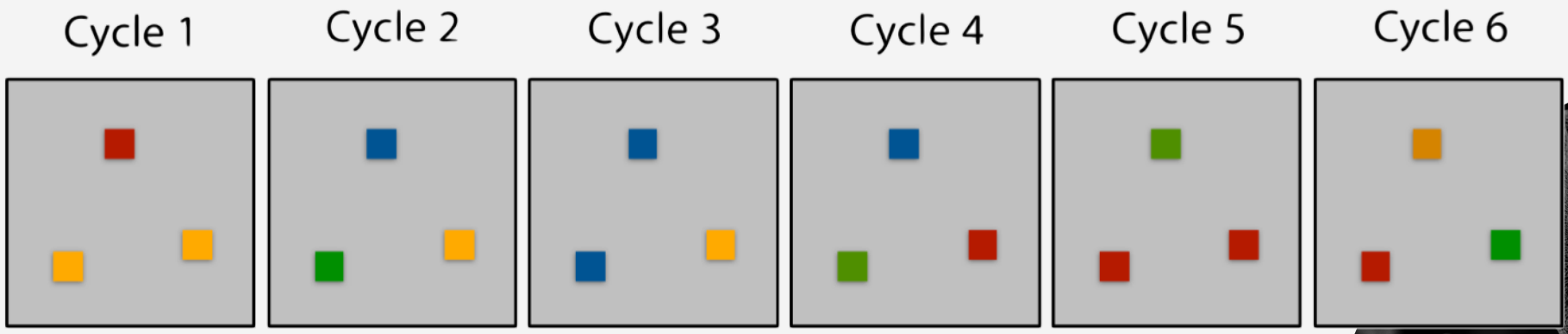
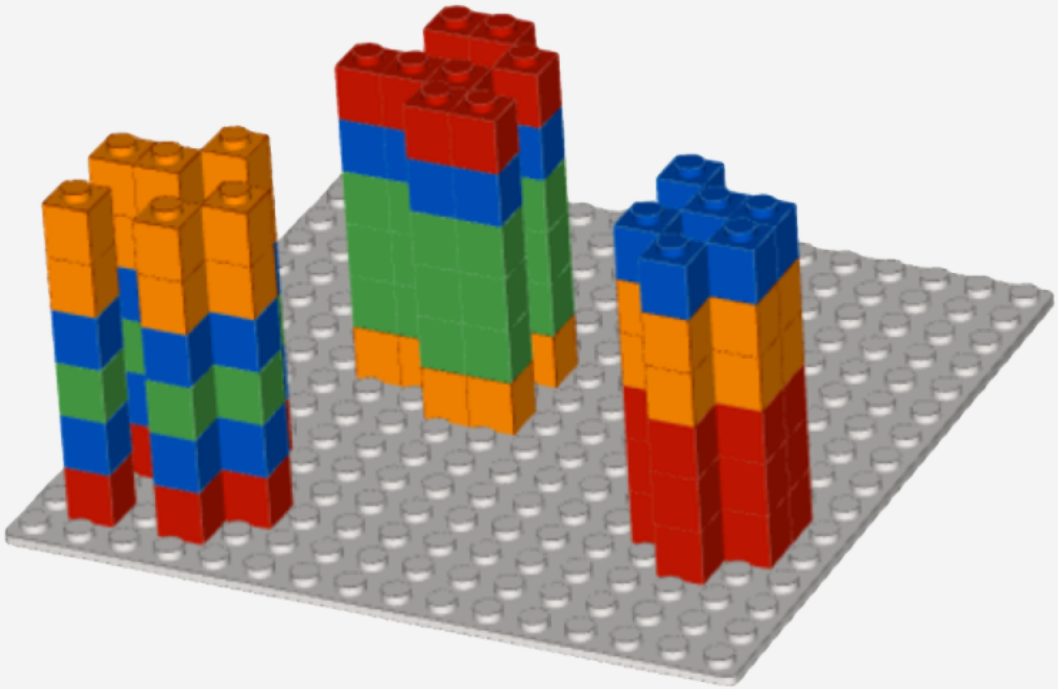


Image courtesy of Johns Hopkins

Read Qualities



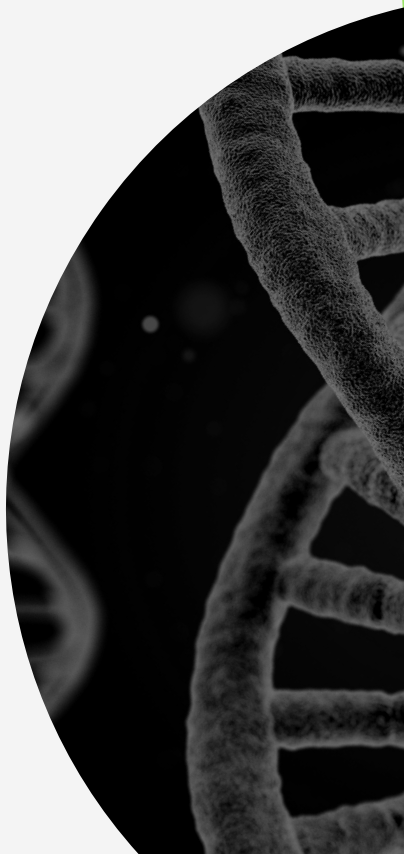
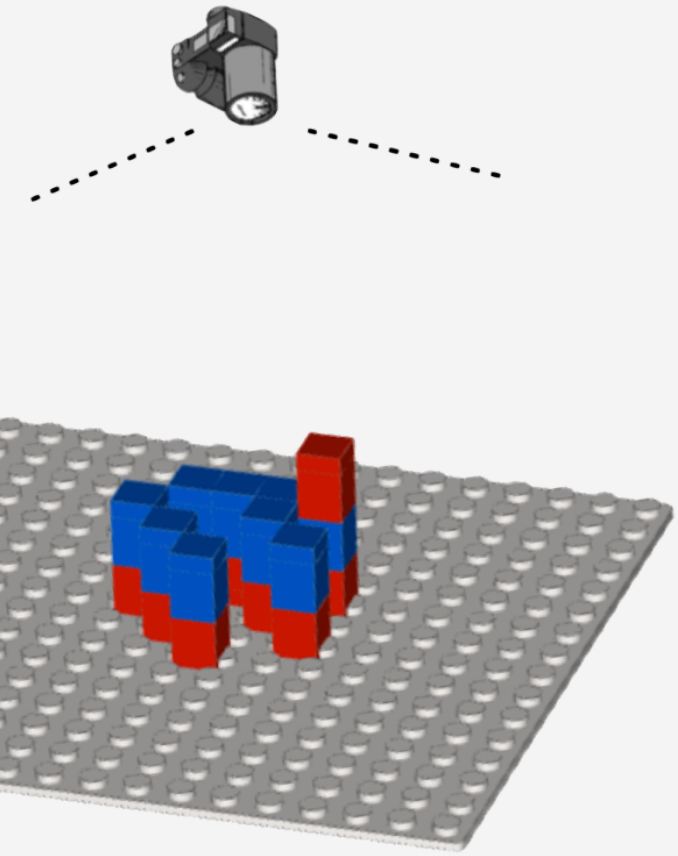
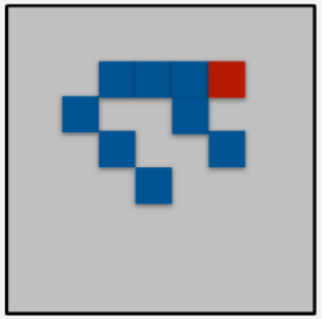


Image courtesy of Johns Hopkins University

Read Qualities





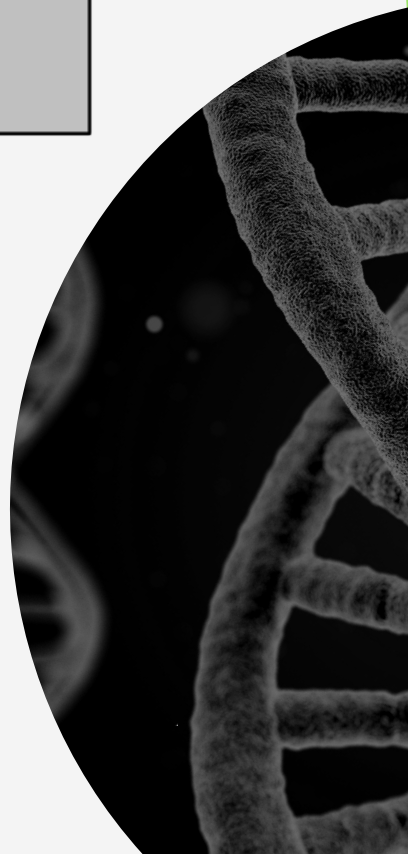
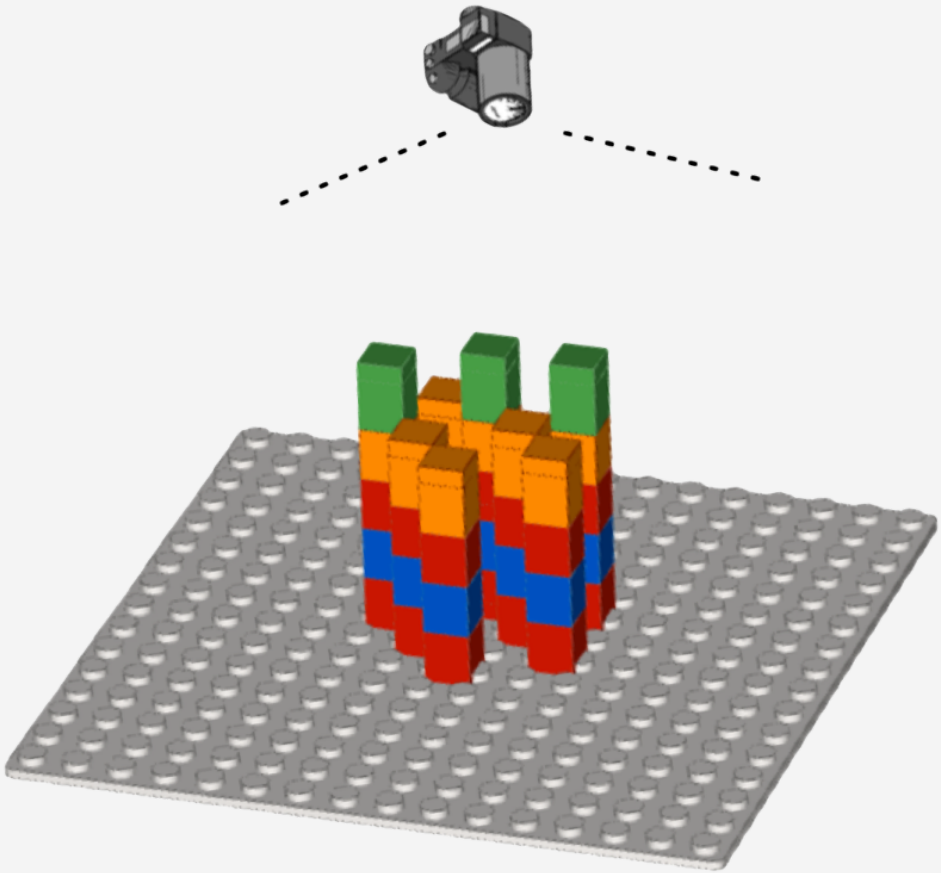
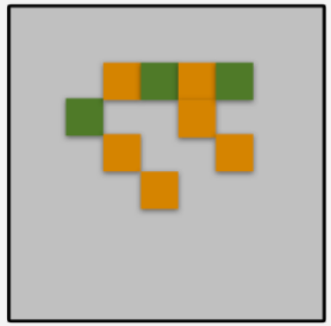


Image courtesy of Johns Hopkins University

Read Qualities





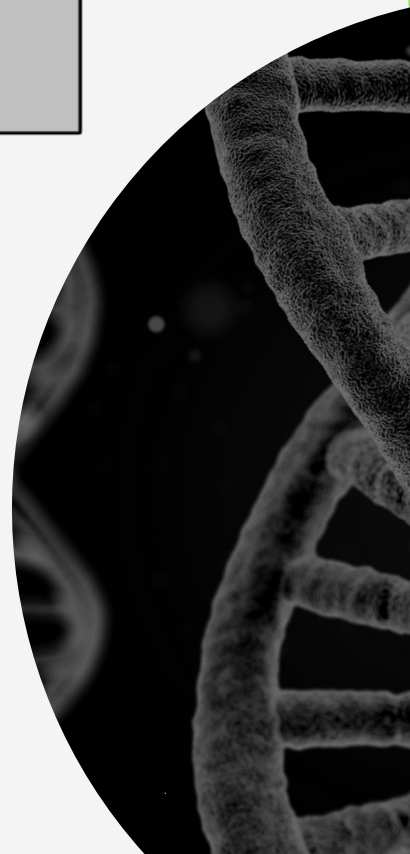


Image courtesy of Johns Hopkins

Phred/Q Score Decoding

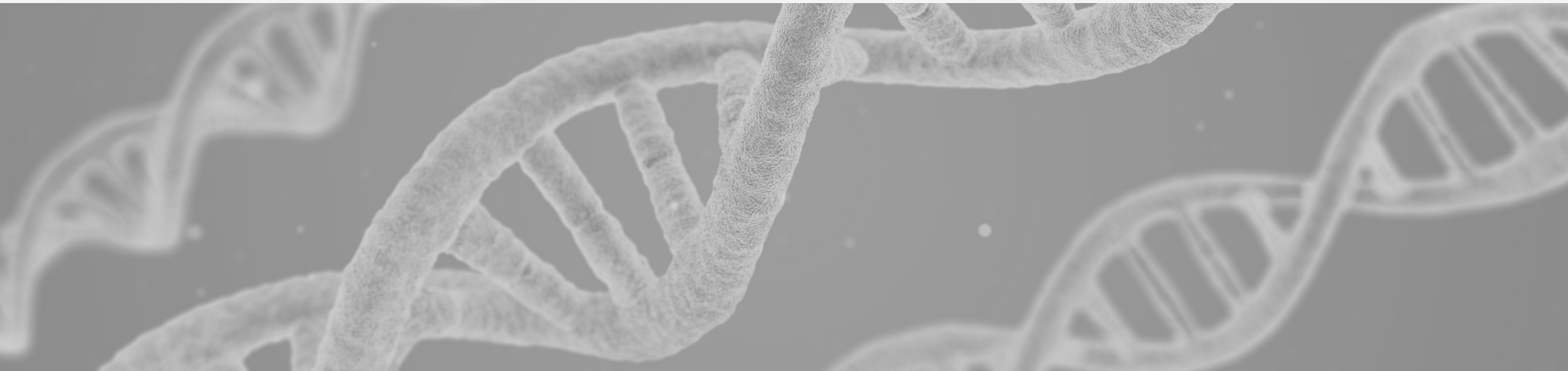
```
#adds 33 then changes it into corresponding ASCII integer
def qToPhred(qValue):
    return chr(qValue + 33)

#changes int to ASCII character then subtracts 33
def phredToQ(phredValue):
    return ord(phredValue) - 33
```

Python Program

EXAMPLE

GENOMES & READS

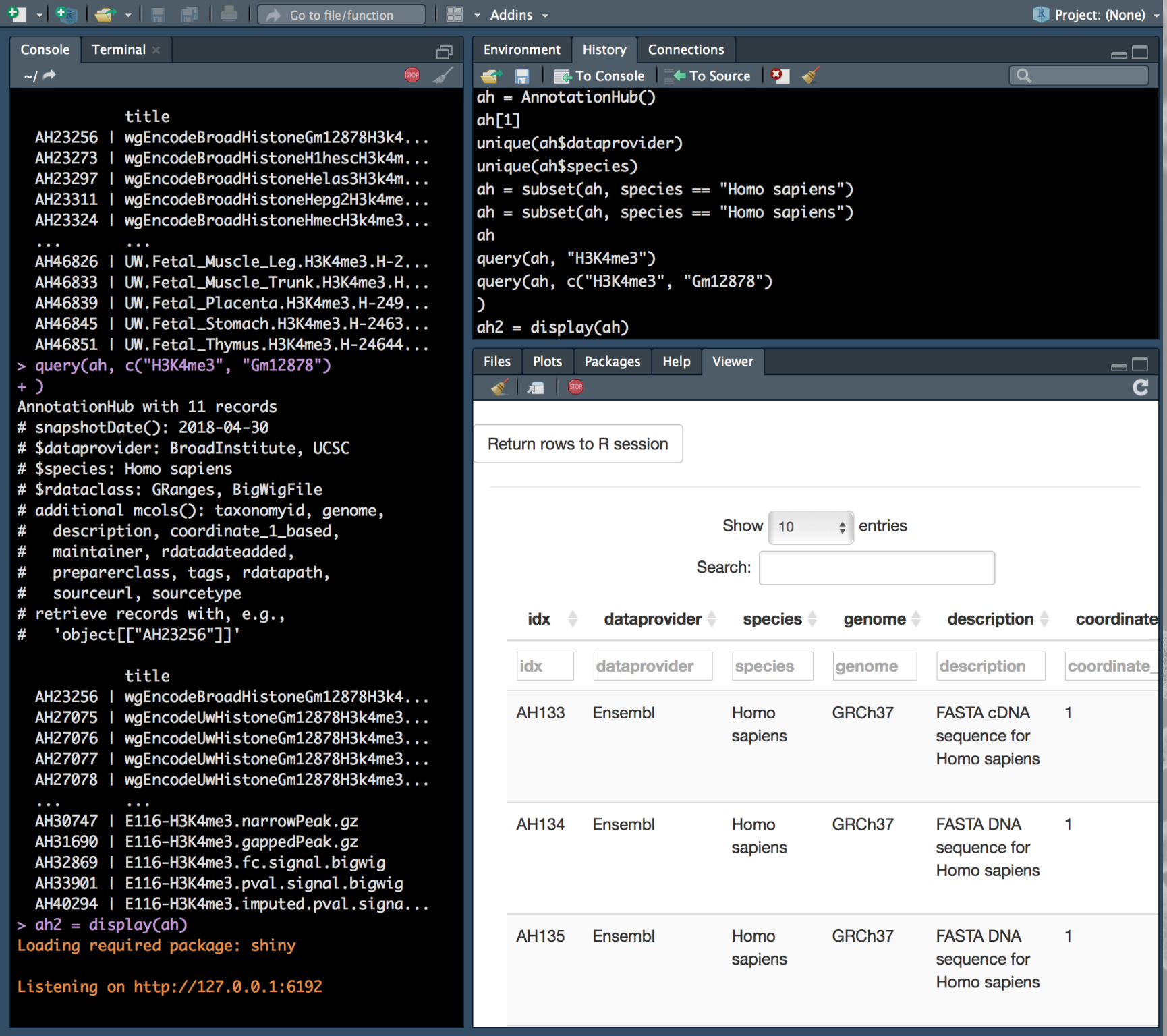


ANALYZING THE GENOME

part 3 of 3

loops

Image courtesy of Rice University

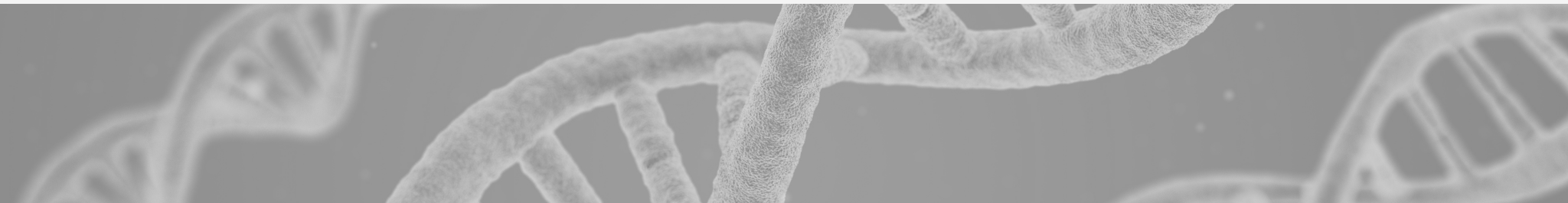

Go to file/function
Addins
Project: (None)
Console
Terminal
~/
title
AH23256 | wgEncodeBroadHistoneGm12878H3k4...
AH23273 | wgEncodeBroadHistoneH1hescH3k4m...
AH23297 | wgEncodeBroadHistoneHelas3H3k4m...
AH23311 | wgEncodeBroadHistoneHepg2H3k4me...
AH23324 | wgEncodeBroadHistoneHmecH3k4me3...
... ...
AH46826 | UW.Fetal_Muscle_Leg.H3K4me3.H-2...
AH46833 | UW.Fetal_Muscle_Trunk.H3K4me3.H...
AH46839 | UW.Fetal_Placenta.H3K4me3.H-249...
AH46845 | UW.Fetal_Stomach.H3K4me3.H-2463...
AH46851 | UW.Fetal_Thymus.H3K4me3.H-24644...
> query(ah, c("H3K4me3", "Gm12878")
+)
AnnotationHub with 11 records
snapshotDate(): 2018-04-30
$dataprovider: BroadInstitute, UCSC
$species: Homo sapiens
$rdataclass: GRanges, BigWigFile
additional mcols(): taxonomyid, genome,
description, coordinate_1_based,
maintainer, rdatadateadded,
preparerclass, tags, rdatapath,
sourceurl, sourcetype
retrieve records with, e.g.,
'object[["AH23256"]]'
title
AH23256 | wgEncodeBroadHistoneGm12878H3k4...
AH27075 | wgEncodeUwHistoneGm12878H3k4me3...
AH27076 | wgEncodeUwHistoneGm12878H3k4me3...
AH27077 | wgEncodeUwHistoneGm12878H3k4me3...
AH27078 | wgEncodeUwHistoneGm12878H3k4me3...
... ...
AH30747 | E116-H3K4me3.narrowPeak.gz
AH31690 | E116-H3K4me3.gappedPeak.gz
AH32869 | E116-H3K4me3.fc.signal.bigwig
AH33901 | E116-H3K4me3.pval.signal.bigwig
AH40294 | E116-H3K4me3.imputed.pval.signa...
> ah2 = display(ah)
Loading required package: shiny
Listening on http://127.0.0.1:6192
Environment
History
Connections
To Console
To Source
ah = AnnotationHub()
ah[1]
unique(ah$dataprovider)
unique(ah$species)
ah = subset(ah, species == "Homo sapiens")
ah = subset(ah, species == "Homo sapiens")
ah
query(ah, "H3K4me3")
query(ah, c("H3K4me3", "Gm12878")
)
ah2 = display(ah)
Files
Plots
Packages
Help
Viewer
Return rows to R session
Show 10 entries
Search:
idx
dataprovider
species
genome
description
coordinate
AH133 Ensembl Homo sapiens GRCh37 FASTA cDNA sequence for Homo sapiens 1
AH134 Ensembl Homo sapiens GRCh37 FASTA DNA sequence for Homo sapiens 1
AH135 Ensembl Homo sapiens GRCh37 FASTA DNA sequence for Homo sapiens 1

BIOCONDUCTOR in R

```
>library(AnnotationHub)
>ahub = AnnotationHub()
>ahub = subset(ahub, species == "Homo sapiens")
```

```
>library("GRanges")
>length(subsetByOverlaps(peaks, promoters, ignore.strand = TRUE)) / len(promoters)

[1] 0.440878
```



BIOCONDUCTOR in R

```
> genes
UCSC track 'refGene'
UCSCData object with 50066 ranges and 5 metadata columns:
                seqnames          ranges strand |
                   <Rle>       <IRanges>  <Rle> |
    [1]             chr1 66999825-67210768      + |
    [2]             chr1   8378145-8404227      + |
    [3]             chr1 48998527-50489626      - |
    [4]             chr1 16767167-16786584      + |
    [5]             chr1 16767167-16786584      + |
    ...              ...               ...    ... .
[50062] chr19_gl000209_random     57209-68123      + |
[50063] chr19_gl000209_random     46646-68123      + |
[50064] chr19_gl000209_random    98135-112667      + |
[50065] chr19_gl000209_random     70071-84658      + |
[50066] chr19_gl000209_random    131433-145745     + |
            name     score     itemRgb
     <character> <numeric> <character>
    [1]   NM_032291       0       <NA>
    [2] NM_001080397      0       <NA>
    [3]   NM_032785       0       <NA>
    [4] NM_001145277      0       <NA>
    [5] NM_001145278      0       <NA>
    ...        ...      ...       ...
[50062]   NM_002255       0       <NA>
[50063] NM_001258383      0       <NA>
[50064]   NM_012313       0       <NA>
[50065] NM_001083539      0       <NA>
[50066]   NM_012312       0       <NA>
              thick                      blocks
          <IRanges>              <IRangesList>
    [1] 67000042-67208778 1-227,91706-91769,98929-98953,...
    [2]   8378169-8404073    1-102,6222-6642,7214-7306,...
    [3] 48999845-50489468    1-1439,2036-2062,6788-6884,...
    [4] 16767257-16785491    1-182,2961-3061,7199-7303,...
    [5] 16767257-16785385    1-104,2961-3061,7199-7303,...
    ...           ...                       ...
[50062]     57249-67717       1-80,280-315,1182-1466,...
[50063]     57132-67717  1-86,10414-10643,10843-10878,...
[50064]     98146-112480      1-46,1523-1557,4002-4301,...
[50065]     70108-83979       1-71,1071-1106,1851-2135,...
[50066]    131468-145120       1-69,862-897,3334-3633,...
-------
seqinfo: 93 sequences (1 circular) from hg19 genome
```

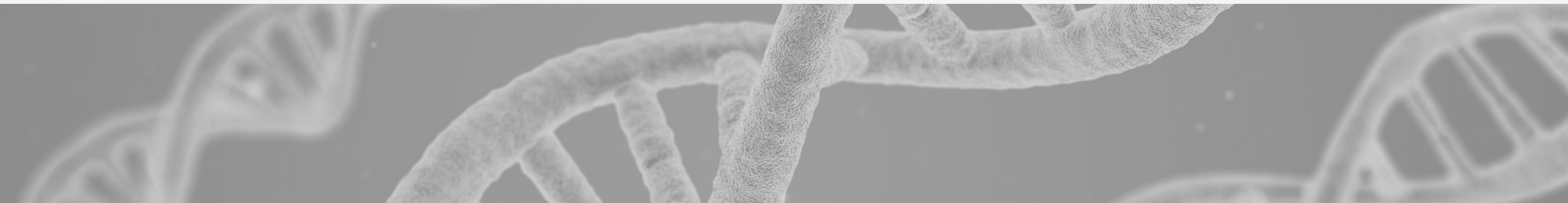
BIOCONDUCTOR in R

```
>> sum(width(reduce(peaks, ignore.strand = TRUE)))/10^6
[1] 11.19044
```

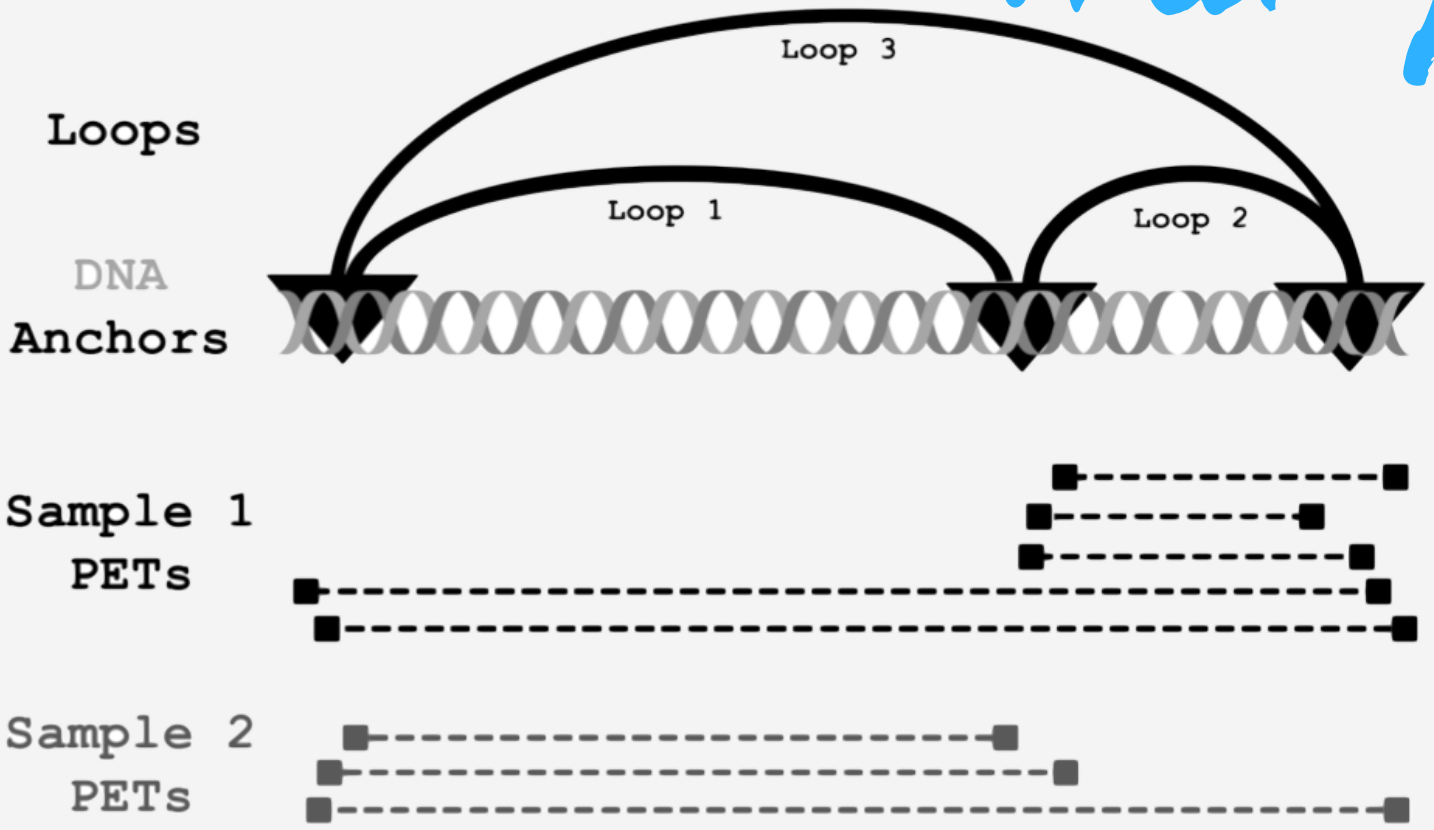
```
          in       out
in      3019608    8170832
out    59259964    0
```

```
[2,2] = 3*10^9 - sum(table)
```

```
          in       out
in      3019608    8170832
out    59259964    2929549596
```



BIOCONDUCTOR Analysis



Counts Matrix

	Sample 1	Sample 2
Loop 1	0	2
Loop 2	3	0
Loop 3	2	1

PET = Paired End Tag

Image courtesy of Massachusetts General Hospital

BIOCONDUCTOR Analysis

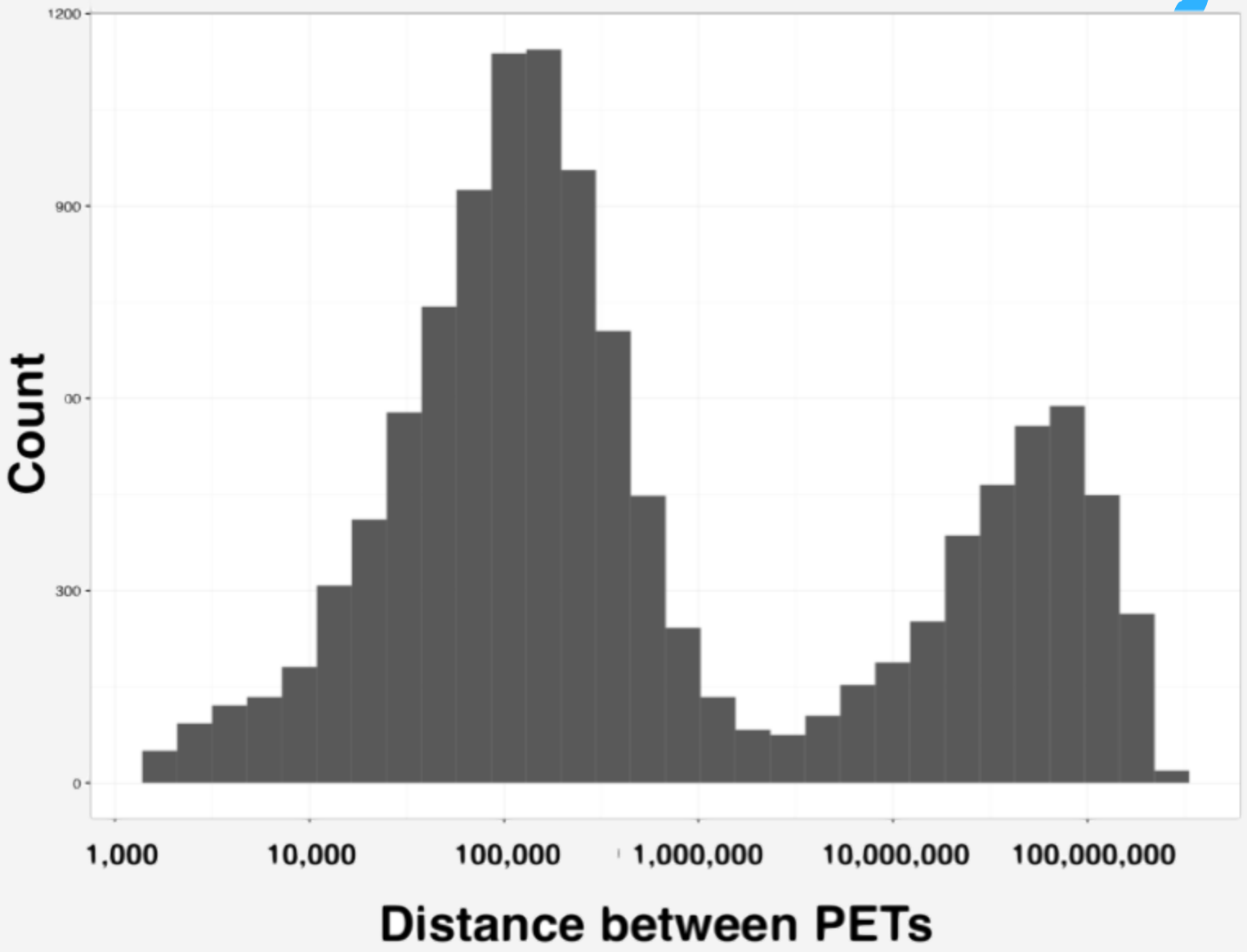
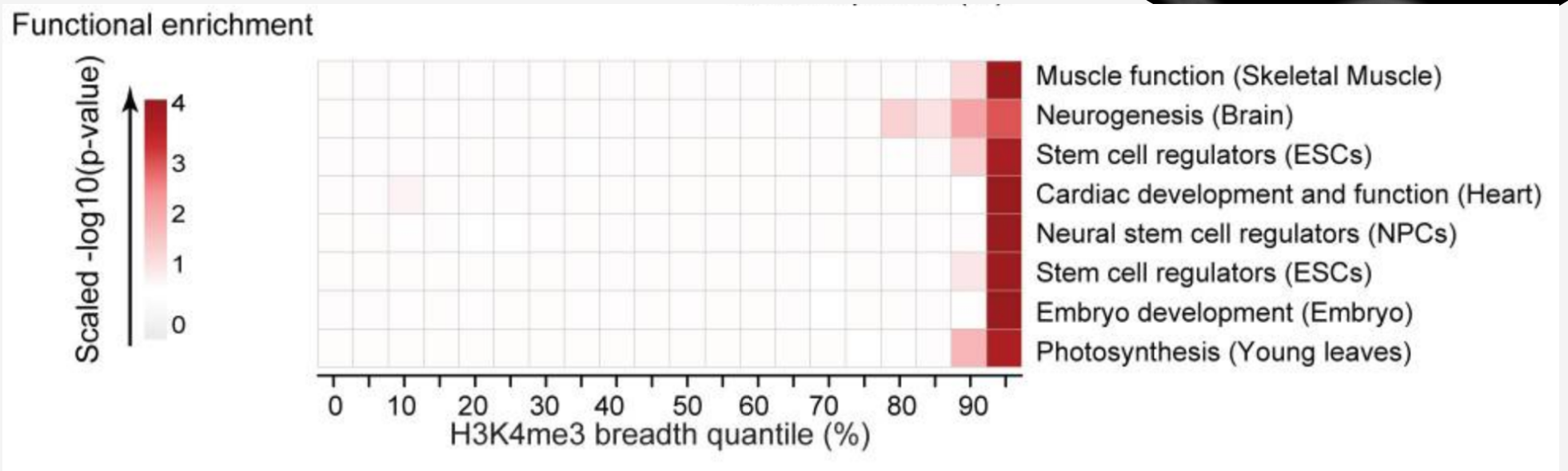


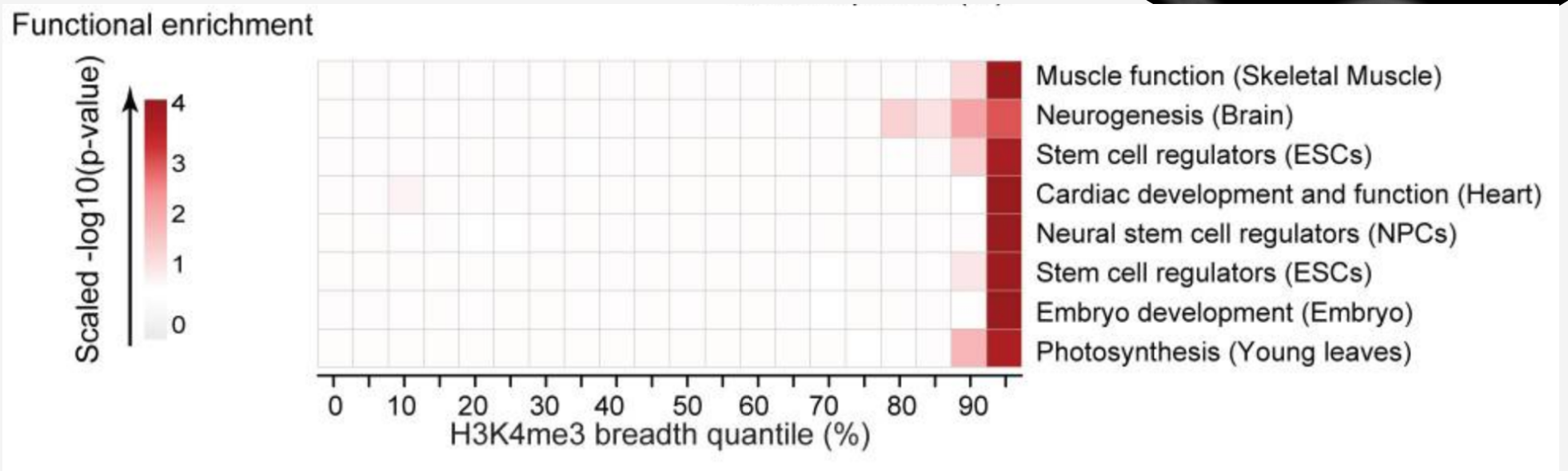
Image courtesy of Massachusetts General Hospital

PEAKS & PROMOTERS



"The top 5% broadest H3K4me3 domains enriches for genes involved in cell/tissue function." (Benayoun, 2014)

PEAKS & PROMOTERS



"The top 5% broadest H3K4me3 domains enriches for genes involved in cell/tissue function." (Benayoun, 2014)

REFERENCES

Wall, J. D., Tang, L. F., Zerbe, B., Kvale, M. N., Kwok, P.-Y., Schaefer, C., & Risch, N. (2014). Estimating genotype error rates from high-coverage next-generation sequence data. Genome Research, 24(11), 1734–1739. http://doi.org/10.1101/gr.168393.113

Ajay, S. S., Parker, S. C. J., Ozel Abaan, H., Fuentes Fajardo, K. V., & Margulies, E. H. (2011). Accurate and comprehensive sequencing of personal genomes. Genome Research, 21(9), 1498–1505. http://doi.org/10.1101/gr.123638.111

Takahashi, H., Kato, S., Murata, M., & Carninci, P. (2012). CAGE- Cap Analysis Gene Expression: a protocol for the detection of promoter and transcriptional networks. Methods in Molecular Biology (Clifton, N.J.), 786, 181–200. http://doi.org/10.1007/978-1-61779-292-2_11

Benayoun, B. A., Pollina, E. A., Uçar, D., Mahmoudi, S., Karra, K., Wong, E. D., … Brunet, A. (2014). H3K4me3 breadth is linked to cell identity and transcriptional consistency. Cell, 158(3), 673–688. http://doi.org/10.1016/j.cell.2014.06.027

RESOURCES

https://www.rdocumentation.org/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3166834/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4216915/

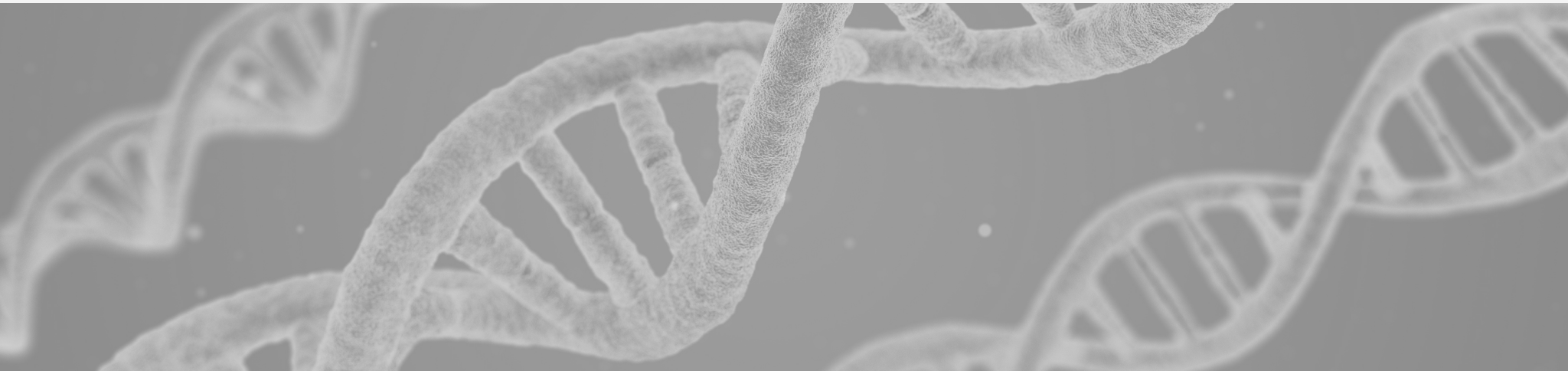
https://www.ted.com/playlists/357/how_does_dna_work

https://www.ted.com/playlists/494/get_into_your_genes

https://www.researchgate.net/figure/H3K4me3-peaks-are-found-at-promoters-where-peak-heights-parallel-gene-expression-levels_fig2_233744684

https://www.sciencedirect.com/science/article/pii/S0092867414014974

ALL HUMANS HAVE GENOMES THAT ARE 99.8-99.9% similar



THANK YOU

7/27/2018