Introduction to Biostrings

INTRODUCTION TO BIOCONDUCTOR IN R

James Chapman
Curriculum Manager, DataCamp

Biostrings

- Algorithms for *fast manipulation* of sequences
- Many Bioconductor packages are dependent on `Biostrings`

```
BiocManager::install("Biostrings")
```

Biological string containers

- *Biostrings* → Memory efficient to store and manipulate sequence of characters
- Containers that can be inherited

For example:

- The BString class comes from *big string*

Strings vs. Sets

- **XString** to store a **single** sequence
  - BString for any string
  - DNAString for DNA
  - RNAString for RNA
  - AAString for amino acids

- **XStringSet** for **many** sequences
  - BStringSet
  - DNAStringSet
  - RNAStringSet
  - AAStringSet

showClass()

```r
showClass("XString")
```

```
Virtual Class "XString" [package "Biostrings"]

Slots:

Name:             shared            offset            length   elementMetadata          metadata
Class:         SharedRaw           integer           integer DataFrame_OR_NULL              list

Extends:
Class "XRaw", directly
Class "XVector", by class "XRaw", distance 2
Class "Vector", by class "XRaw", distance 3
Class "Annotated", by class "XRaw", distance 4
Class "vector_OR_Vector", by class "XRaw", distance 4

Known Subclasses: "BString", "DNAString", "RNAString", "AAString"
```

Biostring alphabets

```
DNA_BASES # 4 DNA bases
RNA_BASES # 4 RNA bases
```

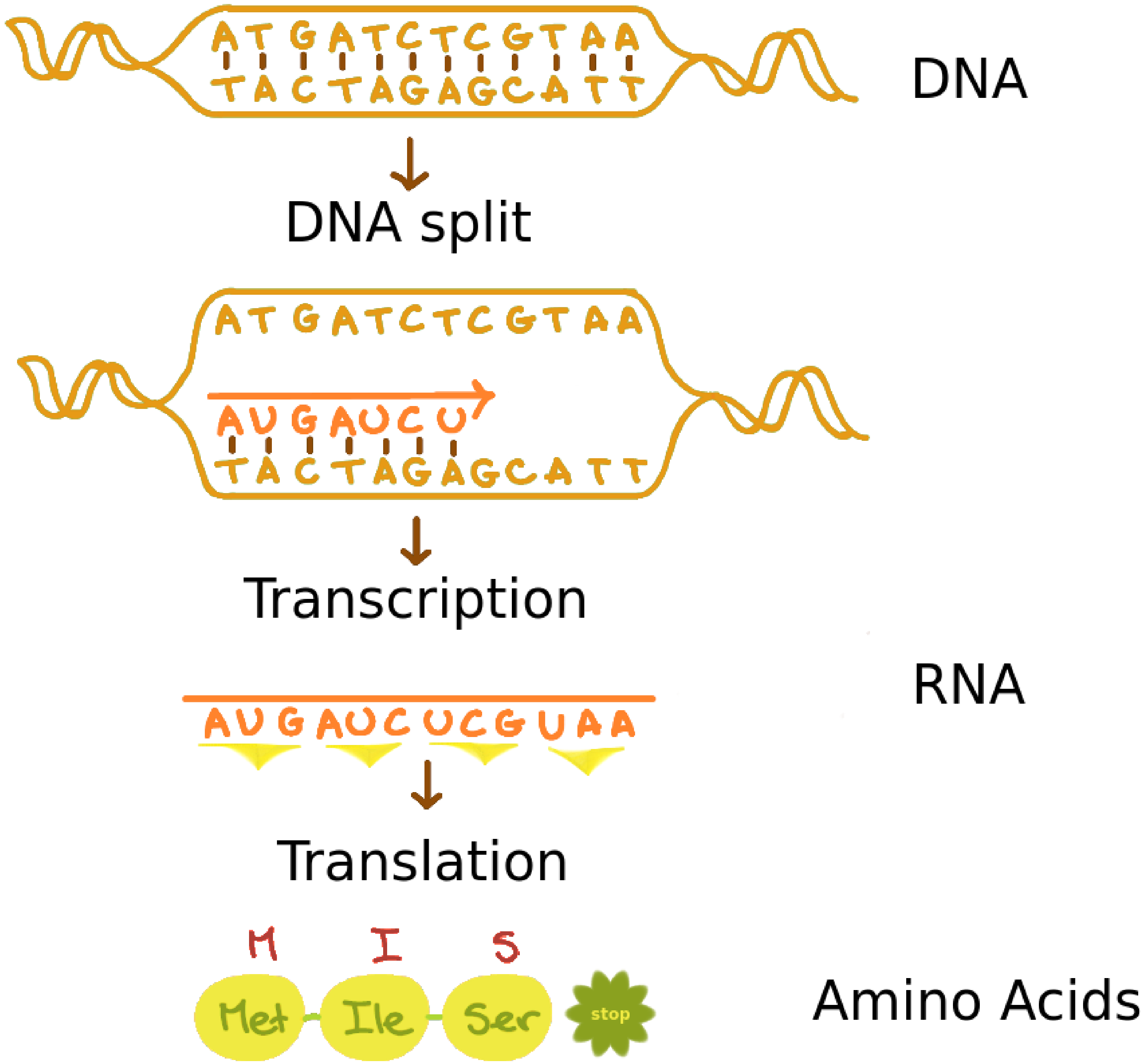
```
"A" "C" "G" "T"
"A" "C" "G" "U"
```

```
AA_STANDARD # 20 Amino acids
```

```
"A" "R" "N" "D" "C" "Q" "E" "G" "H" "I" "L" "K" "M" "F" "P" "S" "T" "W" "Y" "V"
```

```
DNA_ALPHABET # contains IUPAC_CODE_MAP
RNA_ALPHABET # contains IUPAC_CODE_MAP
AA_ALPHABET  # contains AMINO_ACID_CODE
```

[1] For more information IUPAC DNA codes http://genome.ucsc.edu/goldenPath/help/iupac.html



DNA

DNA split

Transcription

RNA

Translation

Amino Acids

Transcription DNA to RNA

```
# DNA single string
dna_seq <- DNAString("ATGATCTCGTAA")
dna_seq
```

```
12-letter DNAString object
seq: ATGATCTCGTAA
```

```
# Transcription DNA to RNA string
rna_seq <- RNAString(dna_seq)
rna_seq
```

```
12-letter RNAString object
seq: AUGAUCUCGUAA
```

Translation RNA to amino acids

```
rna_seq
```

```
12-letter RNAString object
seq: AUGAUCUCGUAA
```

```
# Translation RNA to AA
aa_seq <- translate(rna_seq)
aa_seq
```

Three RNA bases form one AA: `AUG = M, AUC = I, UCG = S, UAA = *`

```
4-letter AAString object
seq: MIS*
```

Shortcut translate DNA to amino acids

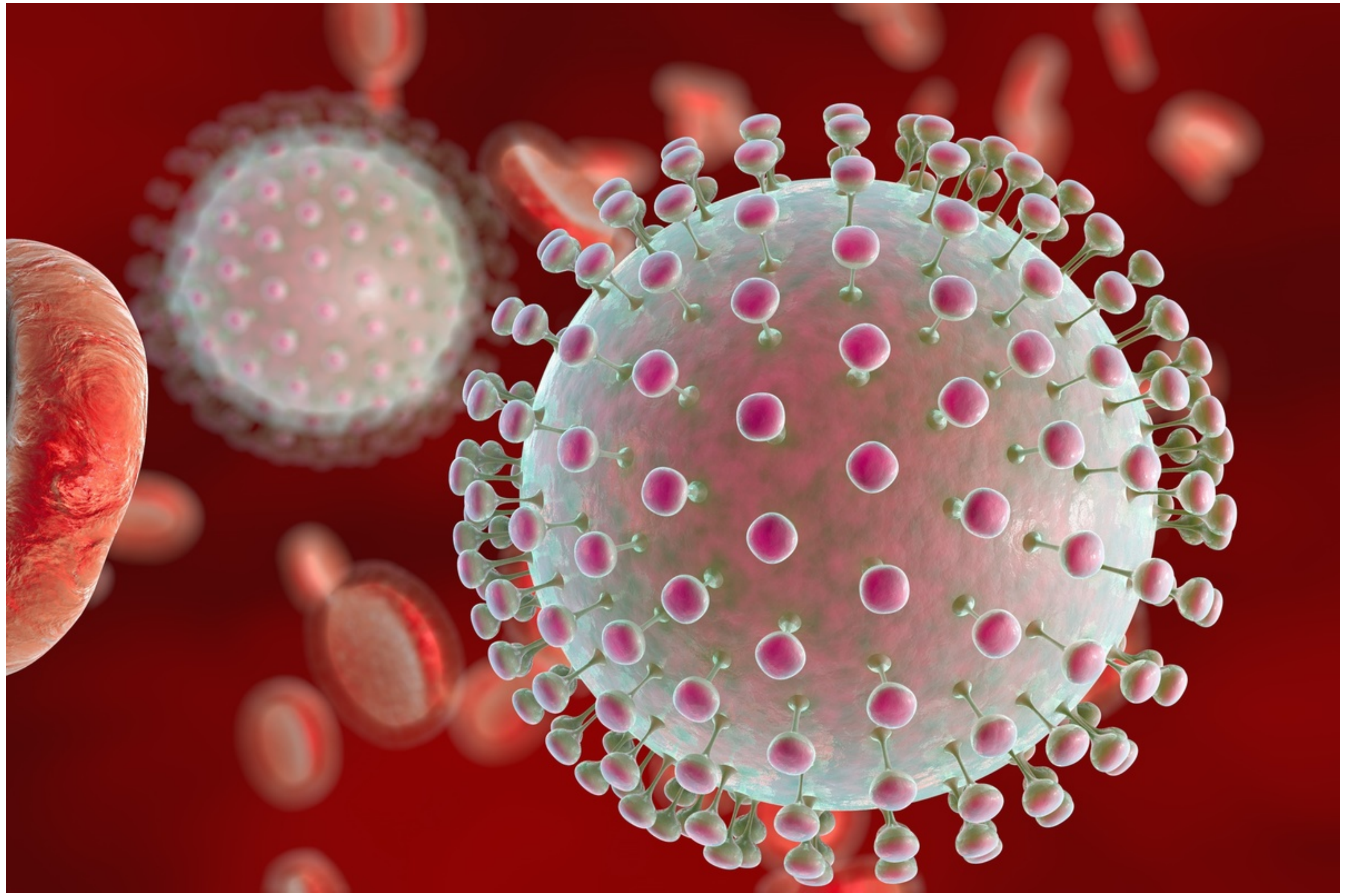
```
dna_seq
```

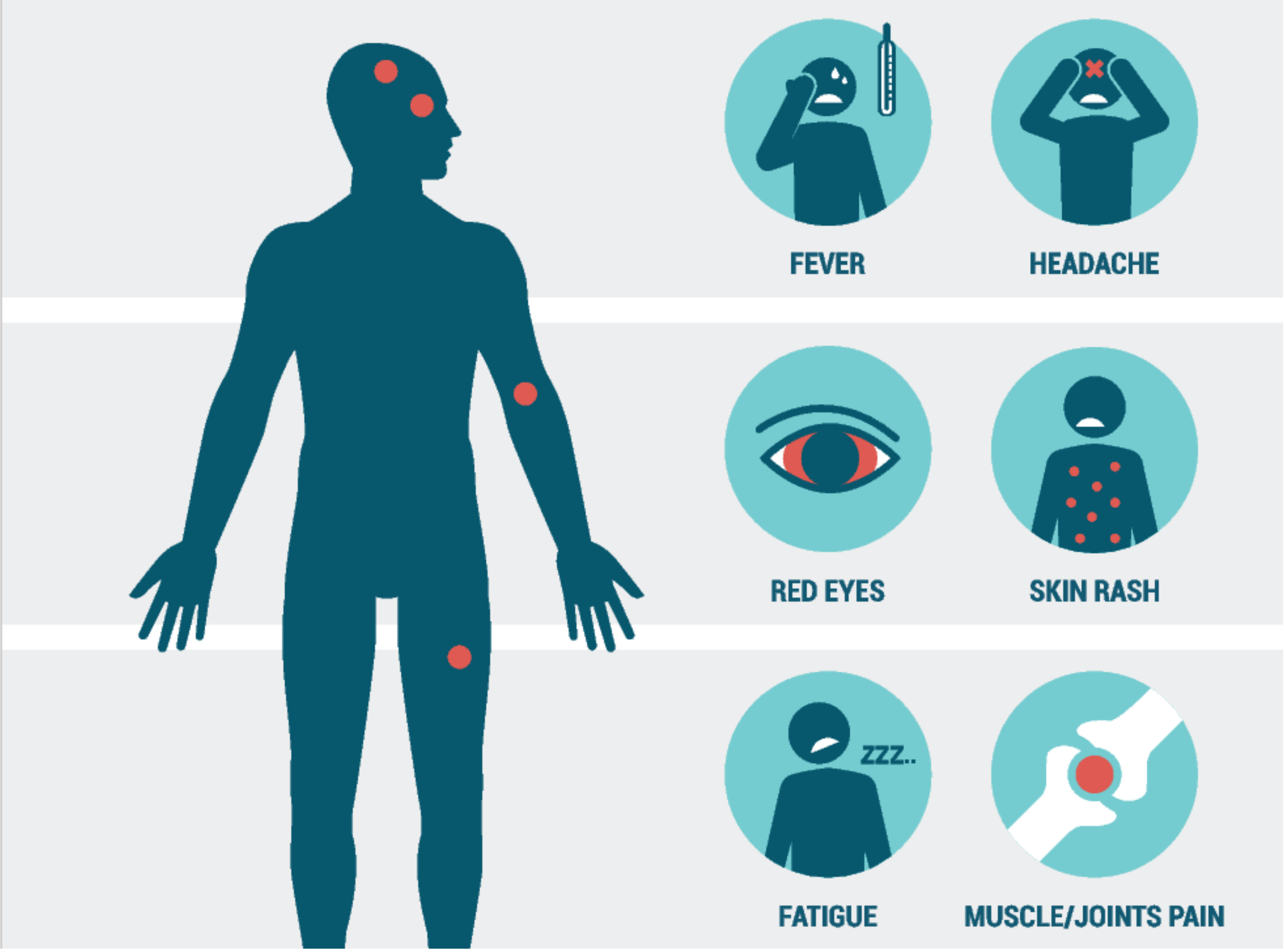
```
12-letter DNAString object
seq: ATGATCTCGTAA
```

```
# translate() also goes directly from DNA to AA
translate(dna_seq)
```

```
4-letter AAString object
seq: MIS*
```

The Zika virus





Let's practice with the Zika virus!

Sequence handling

INTRODUCTION TO BIOCONDUCTOR IN R

James Chapman
Curriculum Manager, DataCamp

Single vs. Set

- **XString** to store a **single** sequence
  - BString for any string
  - DNAString for DNA
  - RNAString for RNA
  - AAString for amino acids

- **XStringSet** for **many** sequences
  - BStringSet
  - DNAStringSet
  - RNAStringSet
  - AAStringSet

Create a StringSet and collate it

```r
# Read the sequence as a set
zikaVirus <- readDNAStringSet("data/zika.fa")
length(zikaVirus)  # the set contains only one sequence
width(zikaVirus)   # and width 10794 bases
```

```
1
10794
```

```r
# Collate the sequence
zikaVirus_seq <- unlist(zikaVirus)

length(zikaVirus_seq)
width(zikaVirus_seq)
```

```
10794
Error in (function (classes, fdef, mtable)  :
  unable to find an inherited method for function ‘width’ for signature ‘"DNAString"’
```

From a single sequence to a set

```
# to create a new set from a single sequence
zikaSet <- DNAStringSet(zikaVirus_seq, start = c(1, 101, 201), end = c(100, 200, 300))
zikaSet
```

```
DNAStringSet object of length 3:
    width seq
[1]   100 AGTTGTTGATCTGTGTGAGTCAGACTGCGACAGTTCGAGTCTGAAG...AACAACAGTATCAACAGGTTTAATTTGGATTTGGAAACGAGAGTTT
[2]   100 CTGGTCATGAAAAACCCCAAAGAAGAAATCCGGAGGATCCGGATTG...CTAAAACGCGGAGTAGCCCGTGTAAACCCCTTGGGAGGTTTGAAGA
[3]   100 GGTTGCCAGCCGGACTTCTGCTGGGTCATGGACCCATCAGAATGGT...TACTAGCCTTTTTGAGATTTACAGCAATCAAGCCATCACTGGGCCT
```

```
length(zikaSet)
width(zikaSet)
```

```
3
100 100 100
```

Complement sequence

ATGATCTCGTAA
TACTAGAGCATT

```
a_seq <- DNAString("ATGATCTCGTAA")
a_seq
```

```
12-letter DNAString object
seq: ATGATCTCGTAA
```

```
complement(a_seq)
```

```
12-letter DNAString object
seq: TACTAGAGCATT
```

Rev a sequence

```
zikaShortSet
```

```
DNAStringSet instance of length 2
width seq                         names
[1]    18 AGTTGTTGATCTGTGTGA       seq1
[2]    18 CTGGTCATGAAAAACCCC       seq2
```

```
rev(zikaShortSet)
```

```
 A DNAStringSet instance of length 2
width seq                         names
[1]    18 CTGGTCATGAAAAACCCC       seq2
[2]    18 AGTTGTTGATCTGTGTGA       seq1
```

Reverse a sequence

```
zikaShortSet
```

```
 A DNAStringSet instance of length 2
width seq                          names
[1]    18 AGTTGTTGATCTGTGTGA        seq1
[2]    18 CTGGTCATGAAAAACCCC        seq2
```

```
reverse(zikaShortSet)
```

```
 A DNAStringSet instance of length 2
width seq                          names
[1]    18 AGTGTGTCTAGTTGTTGA        seq1
[2]    18 CCCCAAAAAGTACTGGTC        seq2
```

Reverse complement

```
# Original rna_seq sequence
8-letter RNAString object
seq: AGUUGUUG
```

```
reverseComplement(rna_seq)
```

```
8-letter RNAString object
seq: CAACAACU
```

```
# Using two functions together
reverse(complement(rna_seq))
```

```
8-letter RNAString object
seq: CAACAACU
```

	Single sequence XString ATCGGTAC	Set of sequences XStringSet ATCGGTAC CCGTAACTT CTTATCGAA
unlist()		*
length()	*	*
width()		*
complement()	*	*
rev()	*	*
reverse()	*	*
reverseComplement()	*	*

Let's practice sequence handling!

INTRODUCTION TO BIOCONDUCTOR IN R

Why are we interested in patterns?

INTRODUCTION TO BIOCONDUCTOR IN R

James Chapman
Curriculum Manager, DataCamp

AGATGGTTGGAGGAGAGAGGATATCTGCAGCCCTATGGGAAGGTTGTTGACCTCGGATGTGGCAGAGGGGGCTGGAGCTA
TTATGCCGCCACCATCCGCAAAGTGCAGGAGGTGAGAGGATACACAAAGGGAGGTCCCGGTCATGAAGAACCCATGCTGG
TGCAAAGCTATGGGTGGAACATAGTTCGTCTCAAGAGTGGAGTGGACGTCTTCCACATGGCGGCTGAGCCGTGTGACACT
CTGCTGTGTGACATAGGTGAGTCATCATCTAGTCCTGAAGTGGAAGAGACACGAACACTCAGAGTGCTCTCTATGGTGGG
GGACTGGCTTGAAAAAAGACCAGGGGCCTTCTGTATAAAGGTGCTGTGCCCATACACCAGCACTATGATGGAAACCATGG
AGCGACTGCAACGTAGGCATGGGGGAGGATTAGTCAGAGTGCCATTGTGTCGCAACTCCACACATGAGATGTACTGGGTC
TCTGGGGCAAAGAGCAACATCATAAAAAGTGTGTCCACCACAAGTCAGCTCCTCCTGGGACGCATGGATGGCCCCAGGAG
GCCAGTGAAATATGAGGAGGATGTGAACCTCGGCTCGGGTACACGAGCTGTGGCAAGCTGTGCTGAGGCTCCTAACATGA
AAATCATCGGCAGGCGCATTGAGAGAATCCGCAATGAACATGCAGAAACATGGTTTCTTGATGAAAACCACCCATACAGG
ACATGGCCTACCATGGGAGCTACGAAGCCCCCACGCAAGGATCAGCGTCTTCCCTCGTGAACGGGGTTGTTAGACTCCT
GTCAAAGCCTTGGGACGTGGTGACTGGAGTTACAGGAATAGCCATGACTGACACCACACCATACGGCCAACAAAGAGTCT
TCAAAGAAAAAGTGGACACCAGGGTGCCAGATCCCCAAGAAGGCACTCGCCAGGTAATGAACATAGTCTCTTCCTGGCTG
TGGAAGGAGCTGGGGAAACGCAAGCGGCCACGCGTCTGCACCAAAGAAGAGTTTATCAACAAGGTGCGCAGCAATGCAGC
ACTGGGAGCAATATTTGAAGAGGAAAAAGAATGGAAGACGGCTGTGGAAGCTGTGAATGATCCAAGGTTTTGGGCCCTAG
TGGATAGGGAGAGAGAACACCACCTGAGAGGAGAGTGTCACAGCTGTGTGTACAACATGATGGGAAAAAGAGAAAAGAAG
CAAGGAGAGTTCGGGAAAGCAAAAGGTAGCCGCGCCATCTGGTACATGTGGTTGGGAGCCAGATTCTTGGAGTTTGAAGC
CCTTGGATTCTTGAACGAGGACCATTGGATGGGAAGAGAAAACTCAGGAGGTGGAGTCGAAGGGTTAGGATTGCAAAGAC
TTGGATACATTCTAGAAGAAATGAATCGGGCACCAGGAGGAAAGATGTACGCAGATGACACTGCTGGCTGGGACACCCGC
ATTAGTAAGTTTGATCTGGAGAATGAAGCTCTGATTACCAACCAAATGGAGGAAGGGCACAGAACTCTGGCGTTGGCCGT
GATTAAATACACATACCAAAACAAAGTGGTGAAGGTTCTCAGACCAGCTGAAGGAGGAAAAACAGTTATGGACATCATTT
CAAGACAAGACCAGAGAGGGAGTGGACAAGTTGTCACTTATGCTCTCAACACATTCACCAACTTGGTGGTGCAGCTTATC

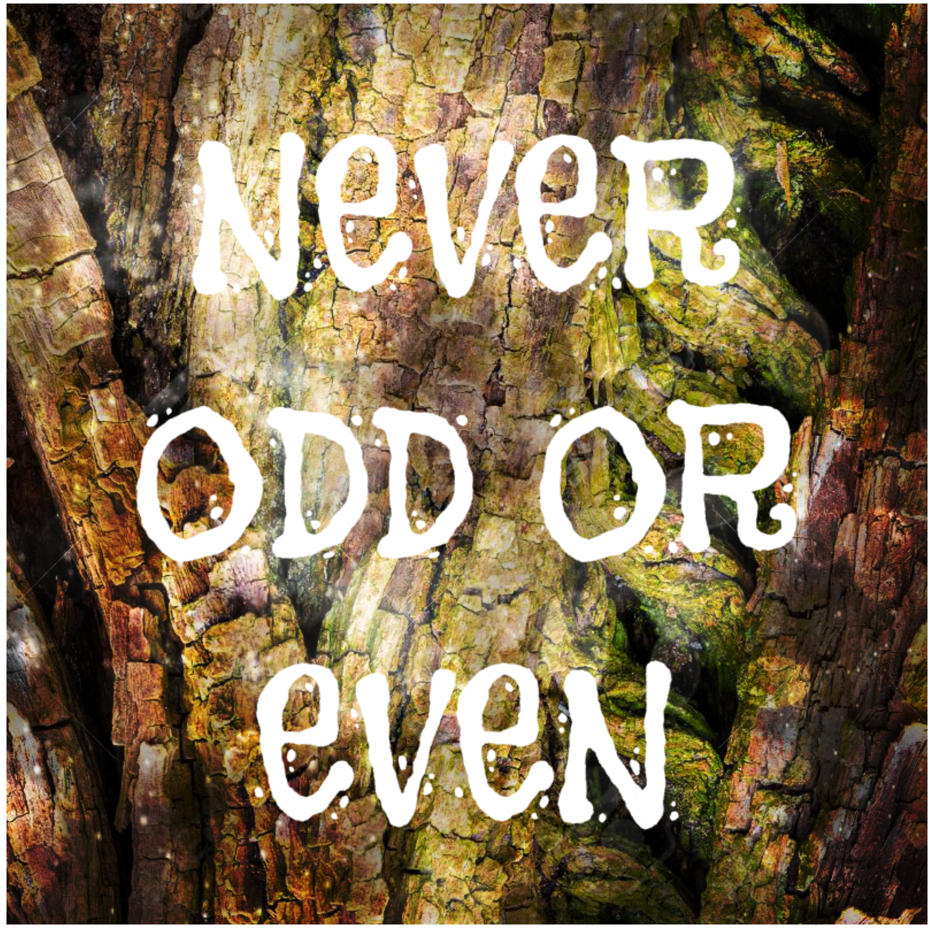
What can we find with patterns?

- Gene start
- Protein end
- Regions that enhance or silence gene expression
- Conserved regions between organisms
- Genetic variation

Pattern matching

- `Biostrings` provides functions for pattern matching
- `matchPattern(pattern, subject)`
  - 1 string to 1 string
- `vmatchPattern(pattern, subject)`
  - 1 set of strings to 1 string
  - 1 string to a set of strings

Palindromes



```
findPalindromes() # find palindromic regions in a single sequence
```

Not new biology

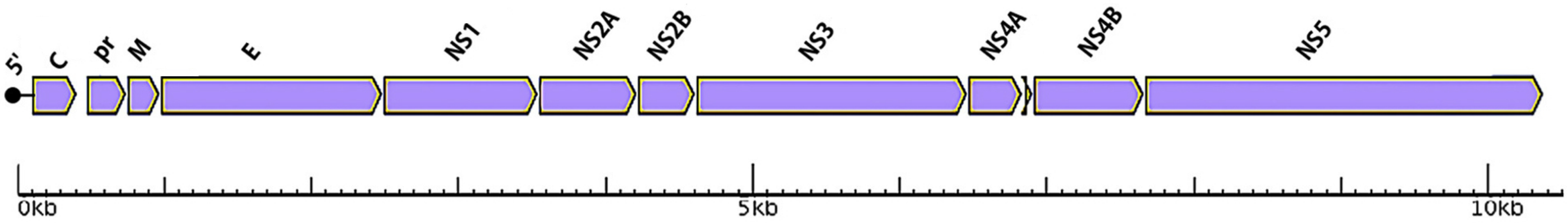
- The Genetic code was first described by Nirenberg in 1963 **On the coding of genetic information** Nirenberg, Marshall et al. Cold Spring Harb Symp Quant Biol 1963, 28
- How translation might differ according to the reading frame, was first described by Streisinger in 1966 **Frameshift Mutations and the Genetic Code** Streisinger, George et al. Cold Spring Harb Symp Quant Biol 1966, 31: 77-84

```
# Original dna sequence
[1]    30 ACATGGGCCTACCATGGGAGCTACGAAGCC
```

```
# 6 possible reading frames, DNAStringSet
[1]    30 ACATGGGCCTACCATGGGAGCTACGAAGCC             + 1
[2]    30 GGCTTCGTAGCTCCCATGGTAGGCCCATGT             - 1
[3]    29  CATGGGCCTACCATGGGAGCTACGAAGCC             + 2
[4]    29  GCTTCGTAGCTCCCATGGTAGGCCCATGT             - 2
[5]    28   ATGGGCCTACCATGGGAGCTACGAAGCC             + 3
[6]    28   CTTCGTAGCTCCCATGGTAGGCCCATGT             - 3
```

```
# 6 possible translations, AAStringSet
[1]    10 TWAYHGSYEA                                 + 1
[2]    10 GFVAPMVGPC                                 - 1
[3]     9 HGPTMGATK                                  + 2
[4]     9 AS*LPW*AH                                  - 2
[5]     9 MGLPWELRS                                  + 3
[6]     9 LRSSHGRPM                                  - 3
```

Conserved regions in the Zika virus



Adapted figure **From Mosquitos to Humans: Genetic Evolution of Zika Virus** Wang, Lulan et al. Cell Host & Microbe 2016, Vol 19 5: 561-565

Facts

- The Zika Virus has a positive strand genome
- It lives in humans, monkeys, and mosquitoes
- The Flaviviruses family and share 11 conserved proteins

Let's practice finding patterns!

INTRODUCTION TO BIOCONDUCTOR IN R