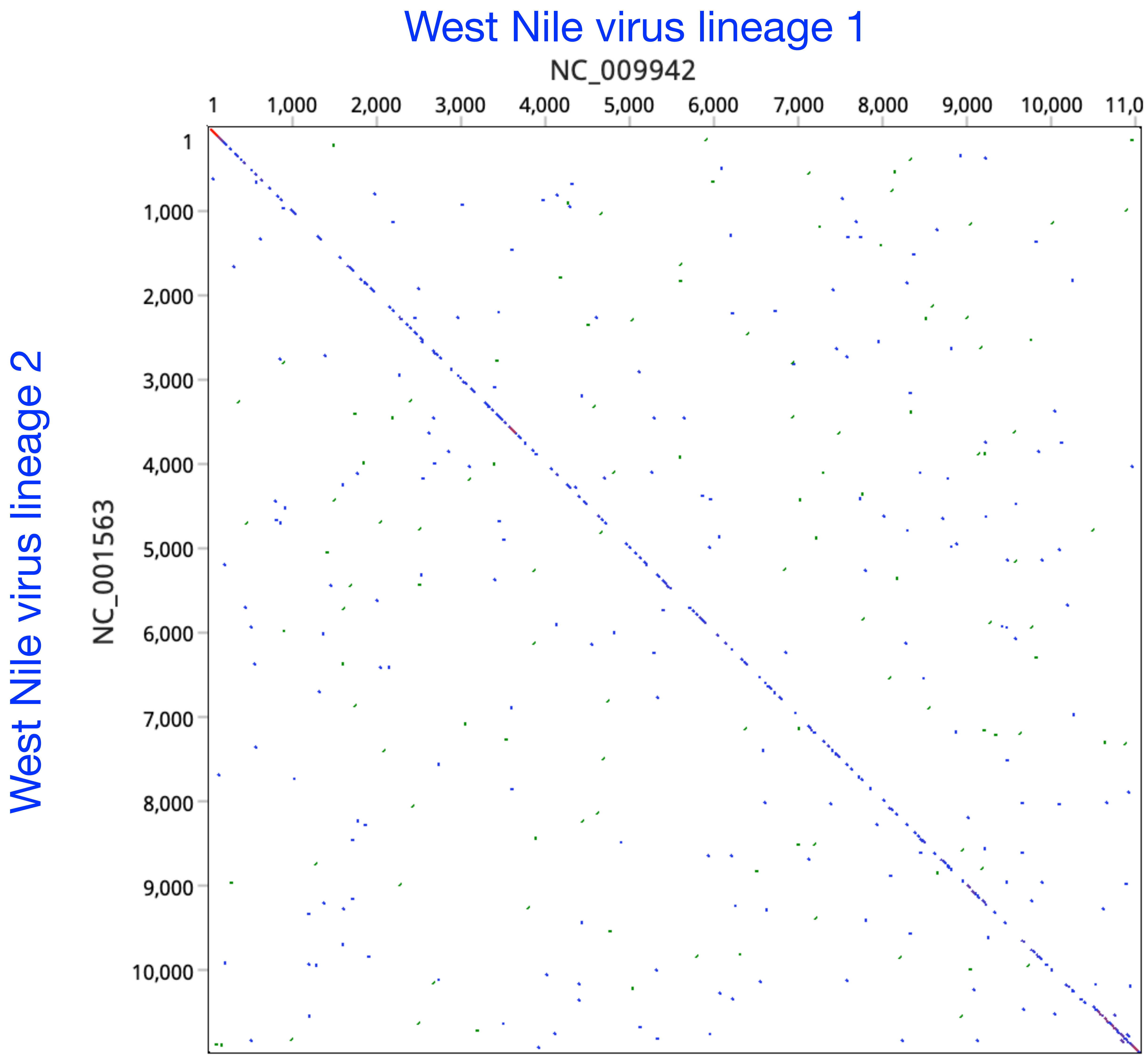
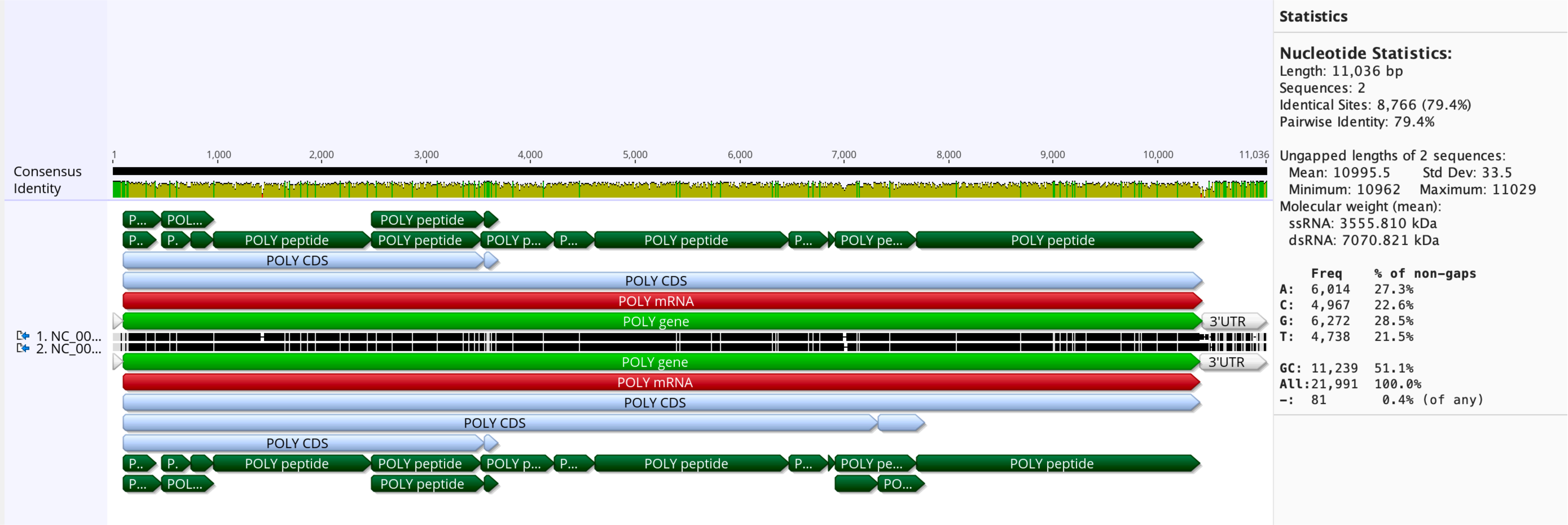
Dotplots!

Mark Stenglein, MIP 280A4

Dotplots are a 2D graphical method for comparing two sequences
They depict overall similarity between two sequences



Dotplots are a complementary alternative to sequence alignments



Statistics

Nucleotide Statistics:

Length: 11,036 bp
Sequences: 2
Identical Sites: 8,766 (79.4%)
Pairwise Identity: 79.4%

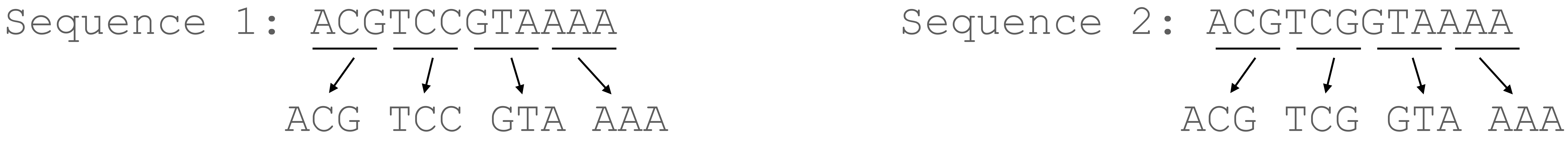
Ungapped lengths of 2 sequences:
Mean: 10995.5 Std Dev: 33.5
Minimum: 10962 Maximum: 11029
Molecular weight (mean):
ssRNA: 3555.810 kDa
dsRNA: 7070.821 kDa

```
     Freq     % of non-gaps
A:   6,014    27.3%
C:   4,967    22.6%
G:   6,272    28.5%
T:   4,738    21.5%

GC: 11,239    51.1%
All:21,991    100.0%
-:   81        0.4% (of any)
```

How to make a dot plot

1) Split sequence 1 and sequence 2 into non-overlapping words (subsequences of a particular length)



Sequence 1: ACGTCCGTAAAA

ACG TCC GTA AAA

Sequence 2: ACGTCGGTAAAA

ACG TCG GTA AAA

Here, words are length 3 (aka “3-mers”)

How to make a dot plot

2) List words from each sequence along the top and side of a matrix

Sequence 1: ACGTCCGTAAAA

Sequence 2: ACGTCGGTAAAA

	ACG	TCC	GTA	AAA
ACG				
TCG				
GTA				
AAA				

How to make a dot plot

3) Color in each cell if the corresponding words from sequence 1 and 2 are identical

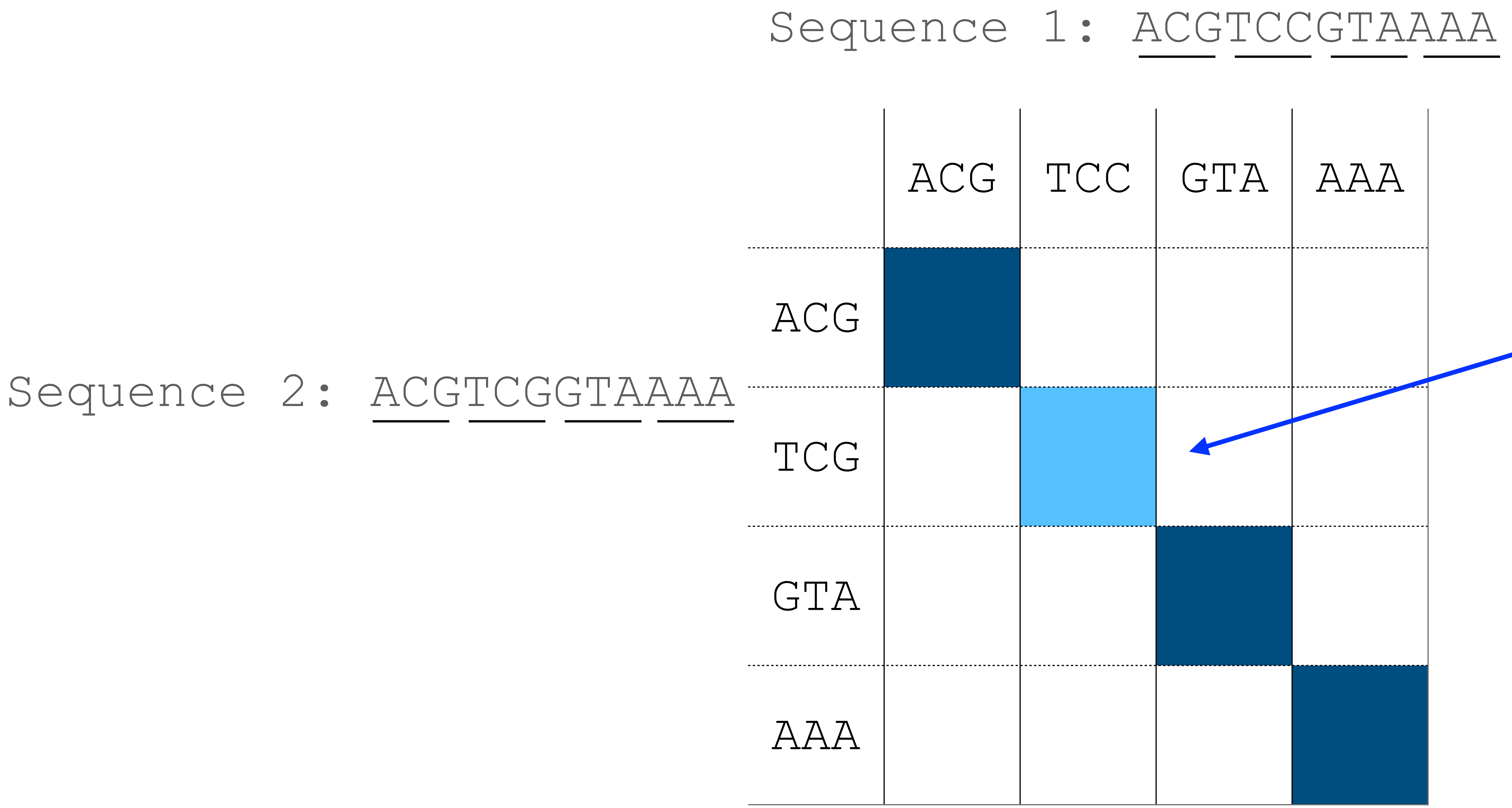
Sequence 1: ACGTCCGTAAAA

Sequence 2: ACGTCGGTAAAA

	ACG	TCC	GTA	AAA
ACG	■			
TCG				
GTA			■	
AAA				■

How to make a dot plot

3) Alternatively, color can reflect the level of identity / similarity between words



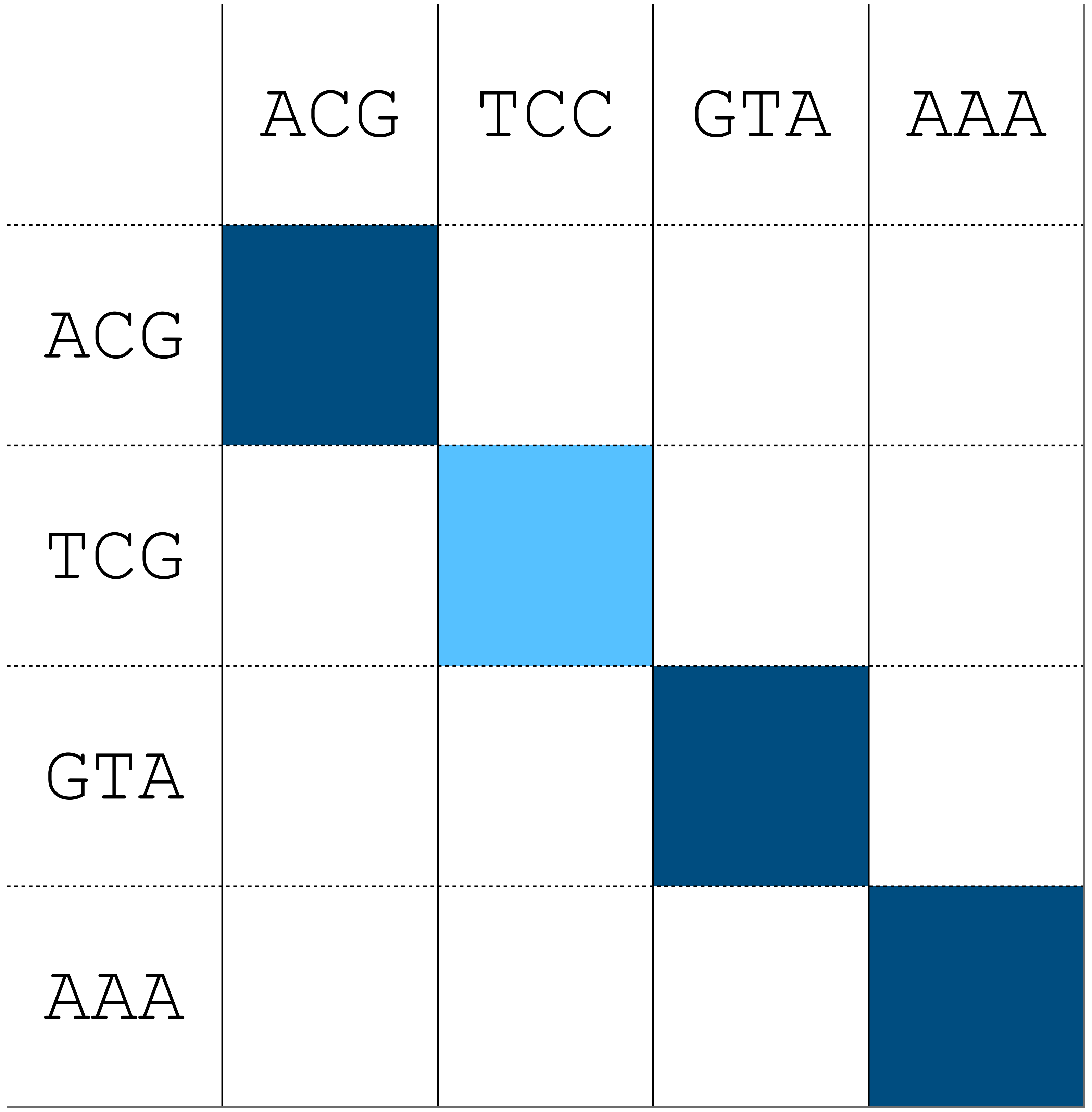
Sequence 1: ACGTCCGTAAAA

Sequence 2: ACGTCGGTAAAA

	ACG	TCC	GTA	AAA
ACG	■			
TCG		▨		
GTA			■	
AAA				■

Here, a lighter color indicates that TCG and TCC are similar but not identical

Dotplots are a different, more visual way to represent sequence similarity



	ACG	TCC	GTA	AAA
ACG	■			
TCG		▨		
GTA			■	
AAA				■

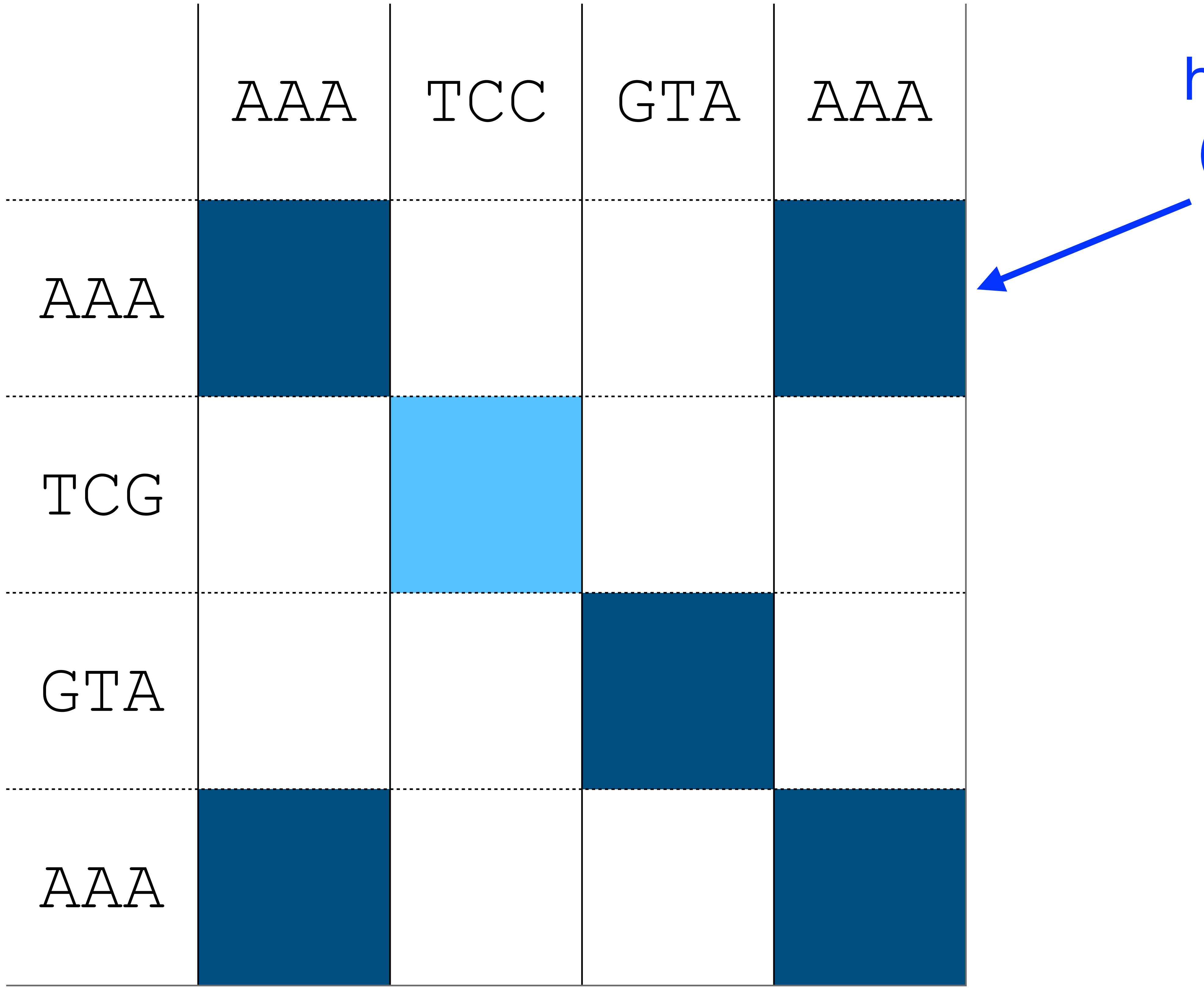
```
Sequence 1: ACGTCCGTAAAA
            ||||| ||||||
Sequence 2: ACGTCGGTAAAA
```

Identity between sequences does not have to be along the diagonal

Sequence 3: AAATCCGTAAAA

Sequence 4:

AAATCGGTAAAA



	AAA	TCC	GTA	AAA
AAA	■			■
TCG		▨		
GTA			■	
AAA	■			■

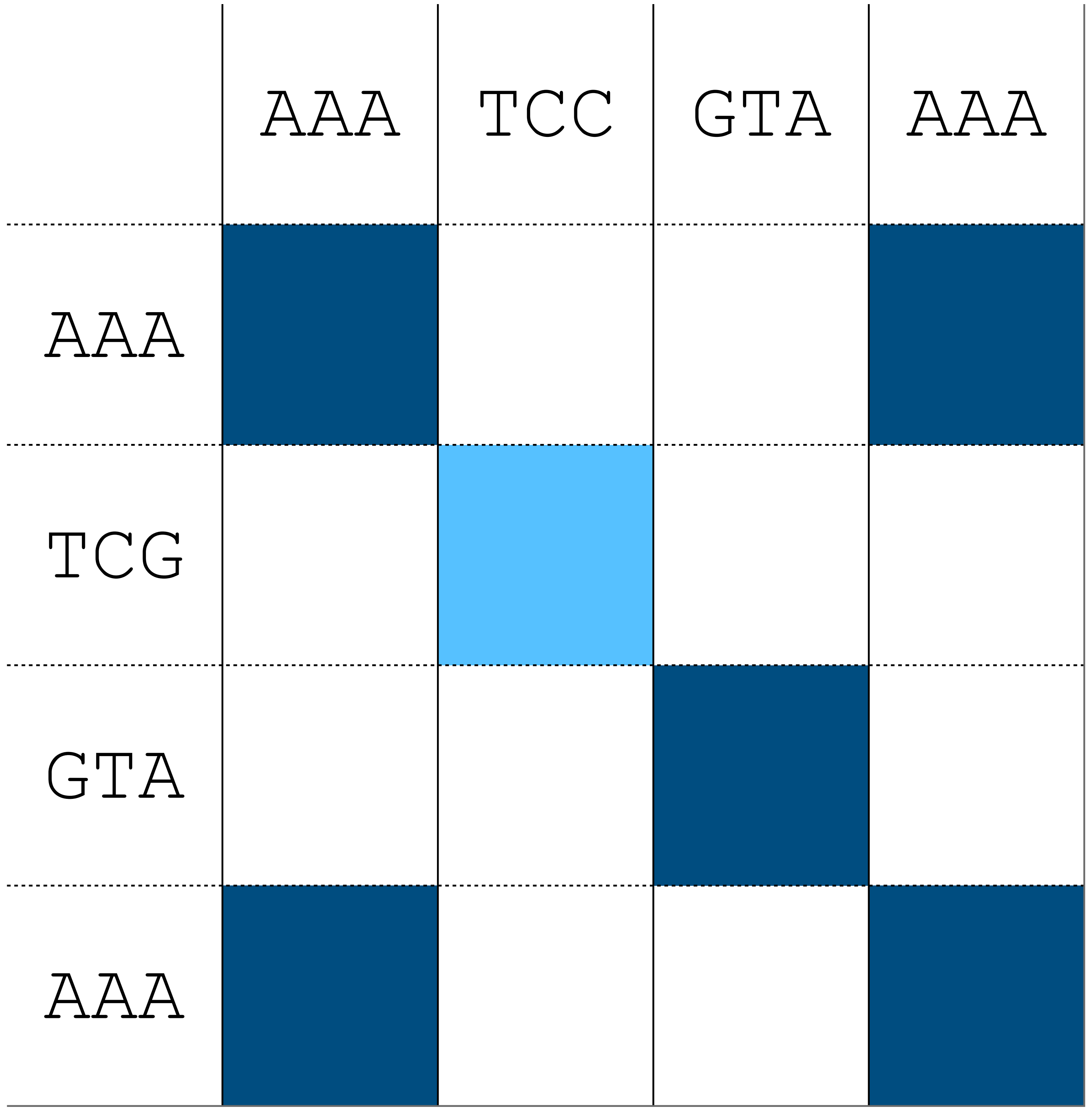
These sequences both have repeated sequence (AAA) at their beginning and end

Dotplots combine global and local alignment information

Sequence 3: AAATCCGTAAAA

Sequence 4:

AAATCGGTAAAA



	AAA	TCC	GTA	AAA
AAA	■			■
TCG		▨		
GTA			■	
AAA	■			■

These sequences both have repeated sequence (AAA) at their beginning and end

This additional level of information is not captured in a single alignment

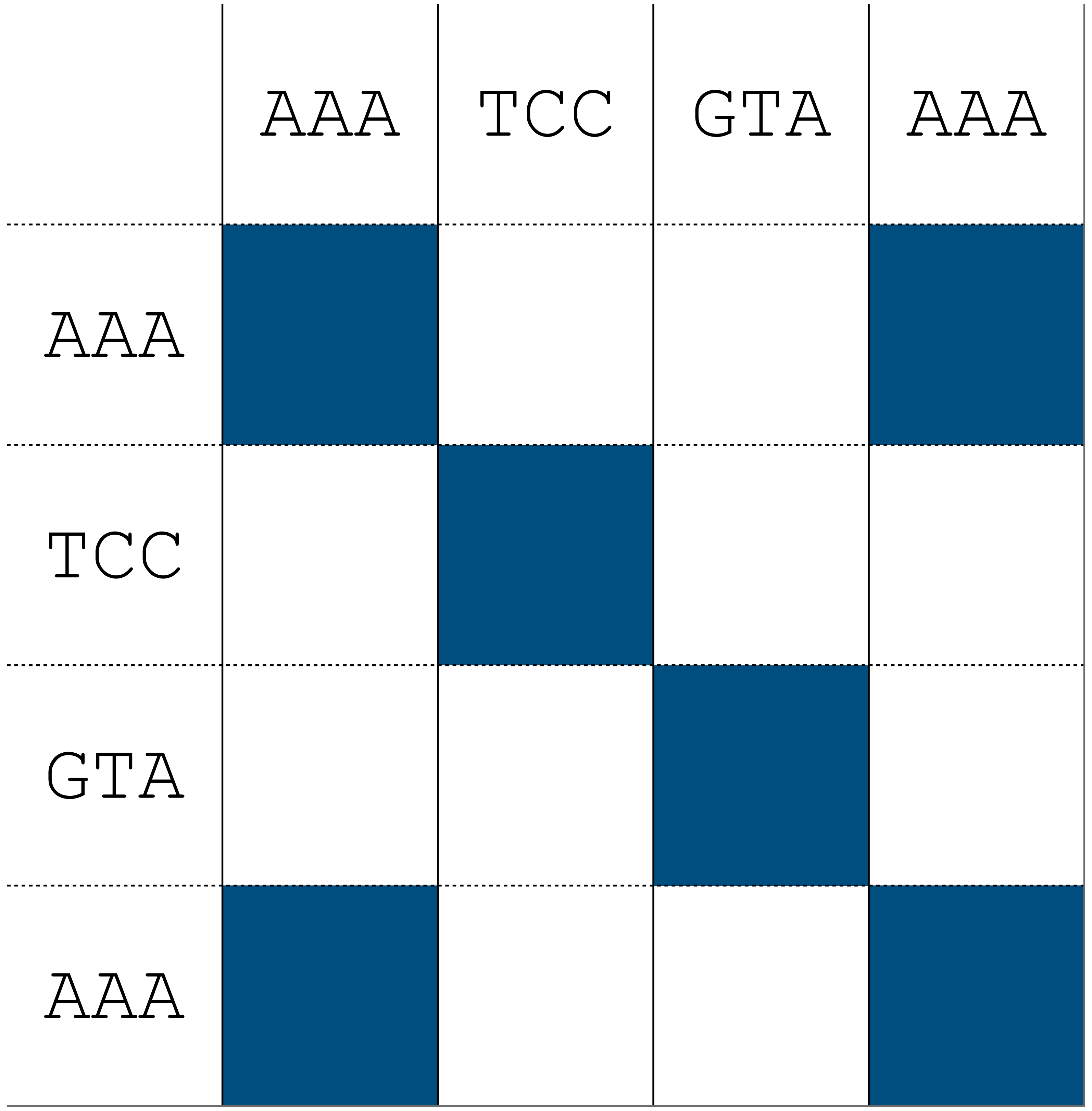
```
Sequence 1: AAATCCGTAAAA
            ||||| ||||||
Sequence 2: AAATCGGTAAAA
```

You can make a dotplot using the same sequence twice

Sequence 3: AAATCCGTAAAA

Sequence 3:

AAATCCGTAAAA

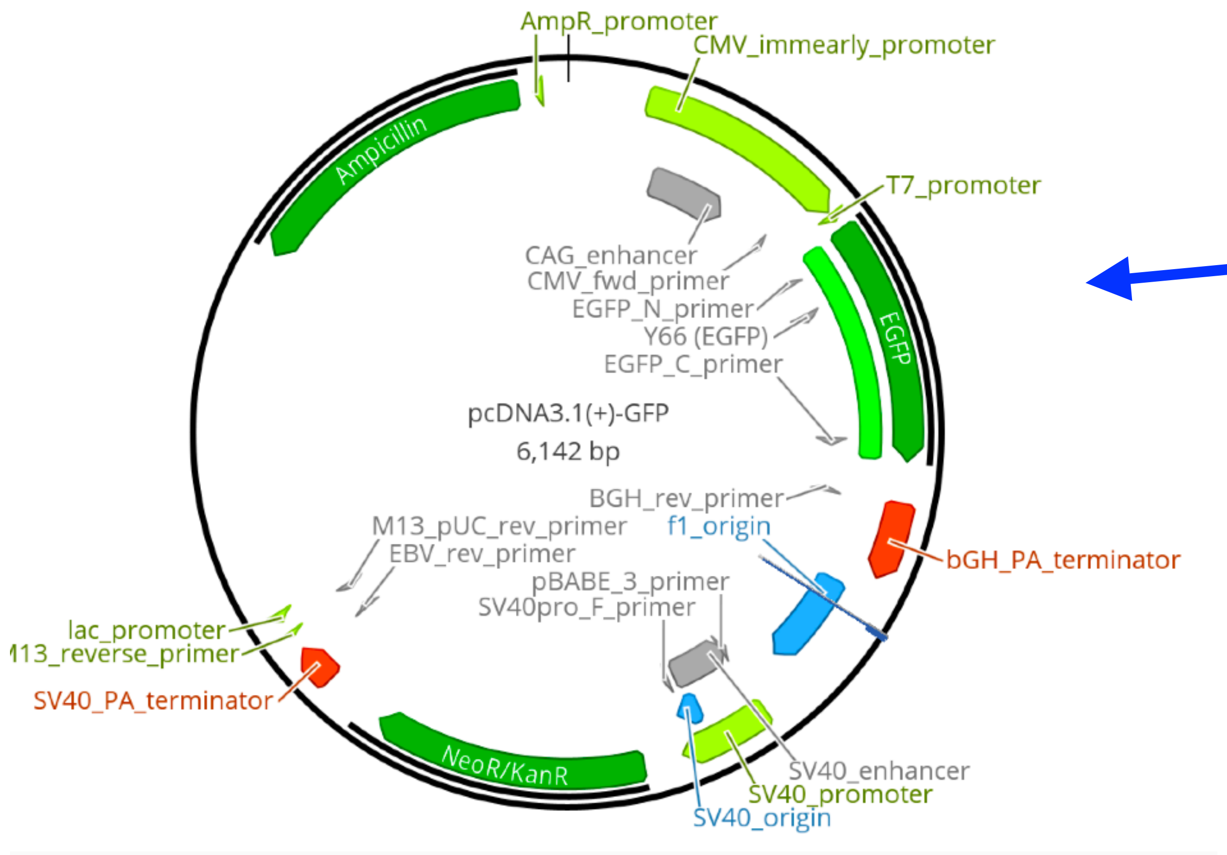


	AAA	TCC	GTA	AAA
AAA	■			■
TCC		■		
GTA			■	
AAA	■			■

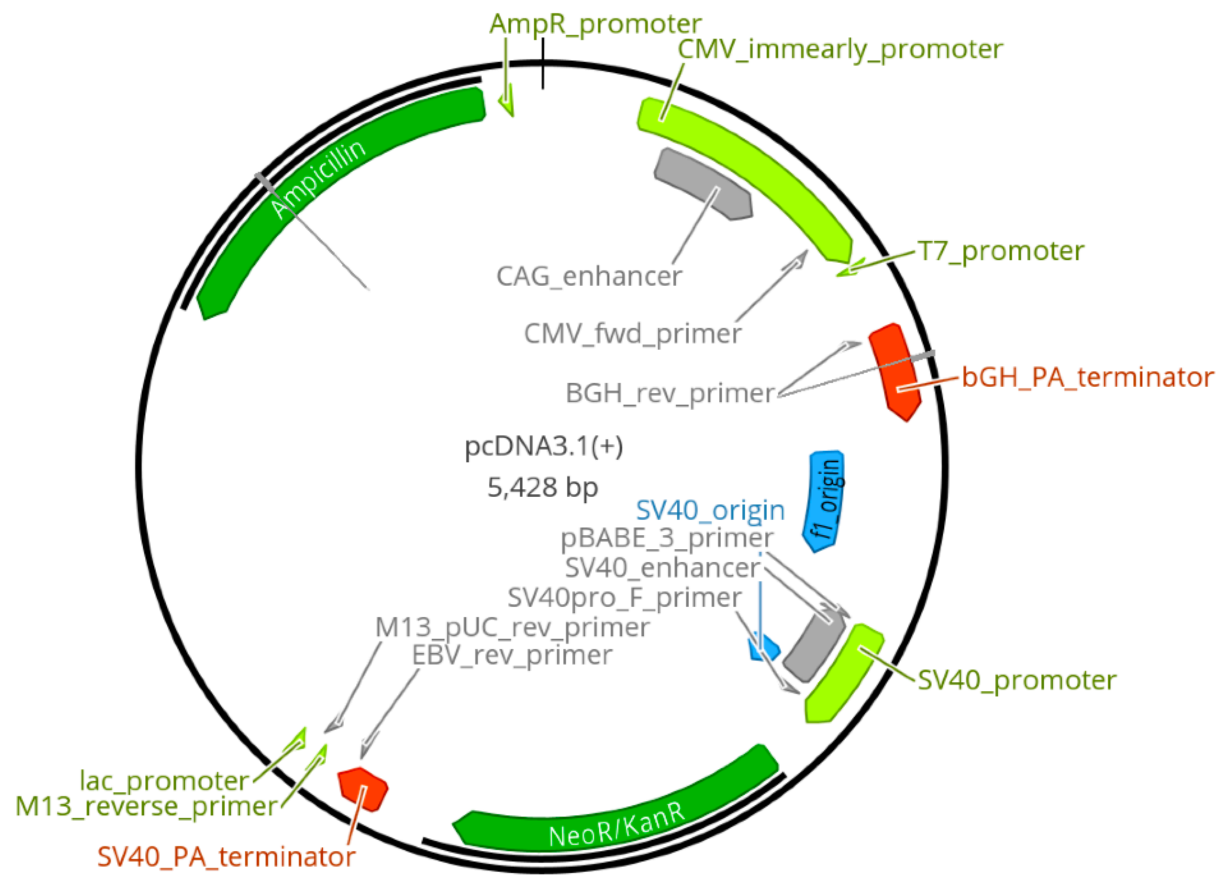
Self dot plots will always have the diagonal filled in

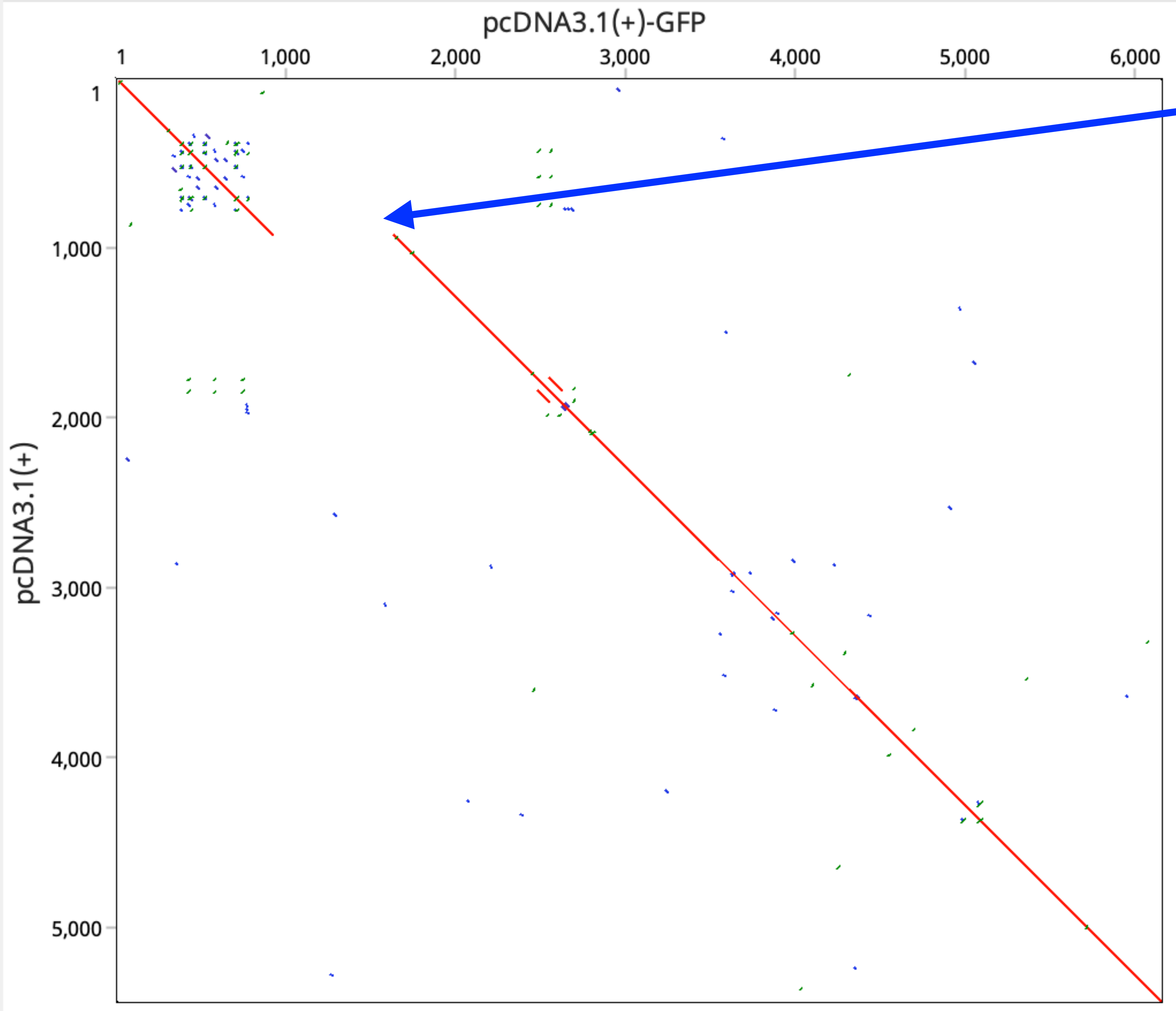
Self dot plots can identify regions of self-similarity within a sequence

Dotplots allow you to visualize structural differences between sequences



The same plasmid, but with a GFP gene cloned into it



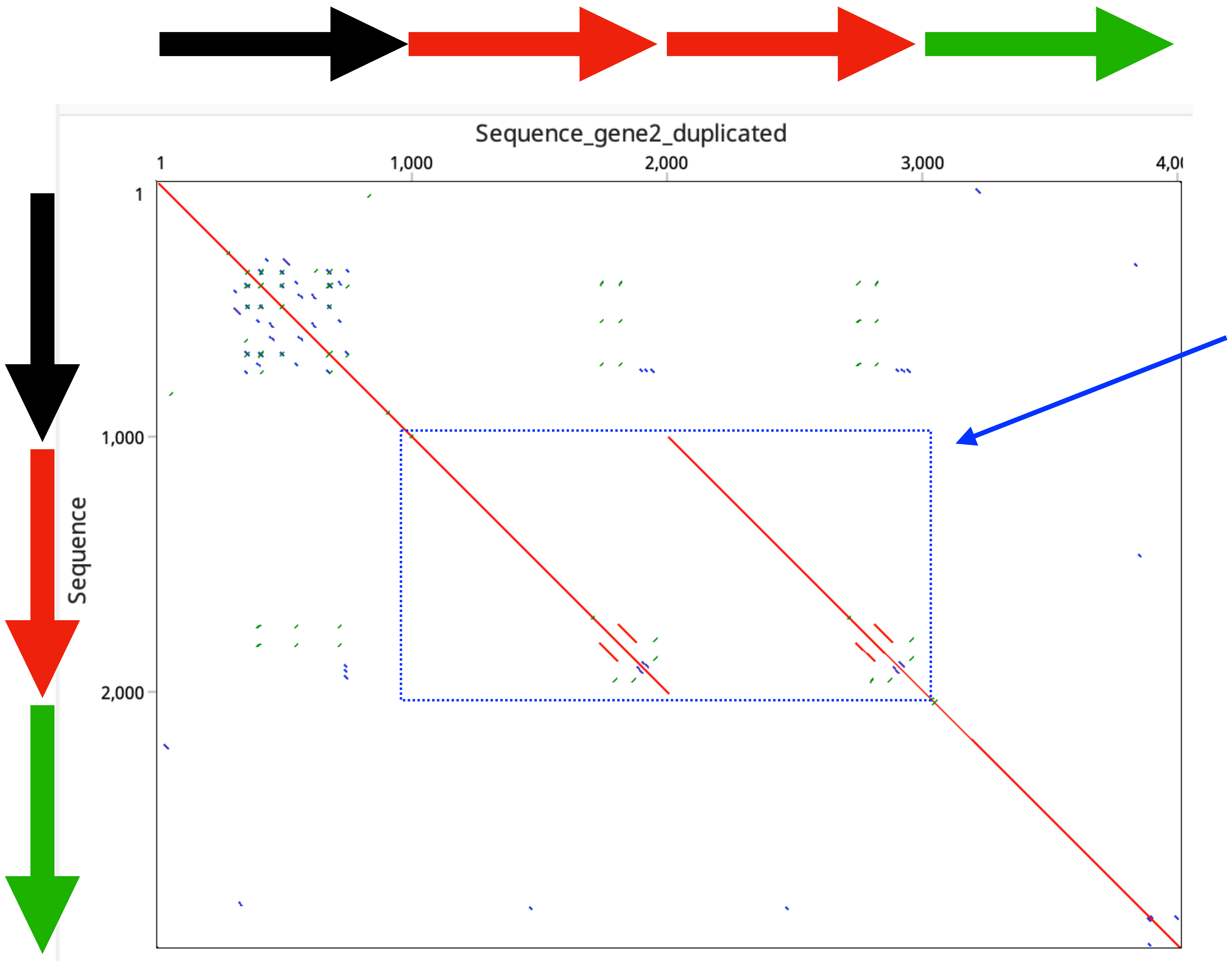


Bases 909-1391 in pcDNA3.1-GFP are absent in pcDNA3.1

This represents an insertion of the GFP gene into pcDNA3.1

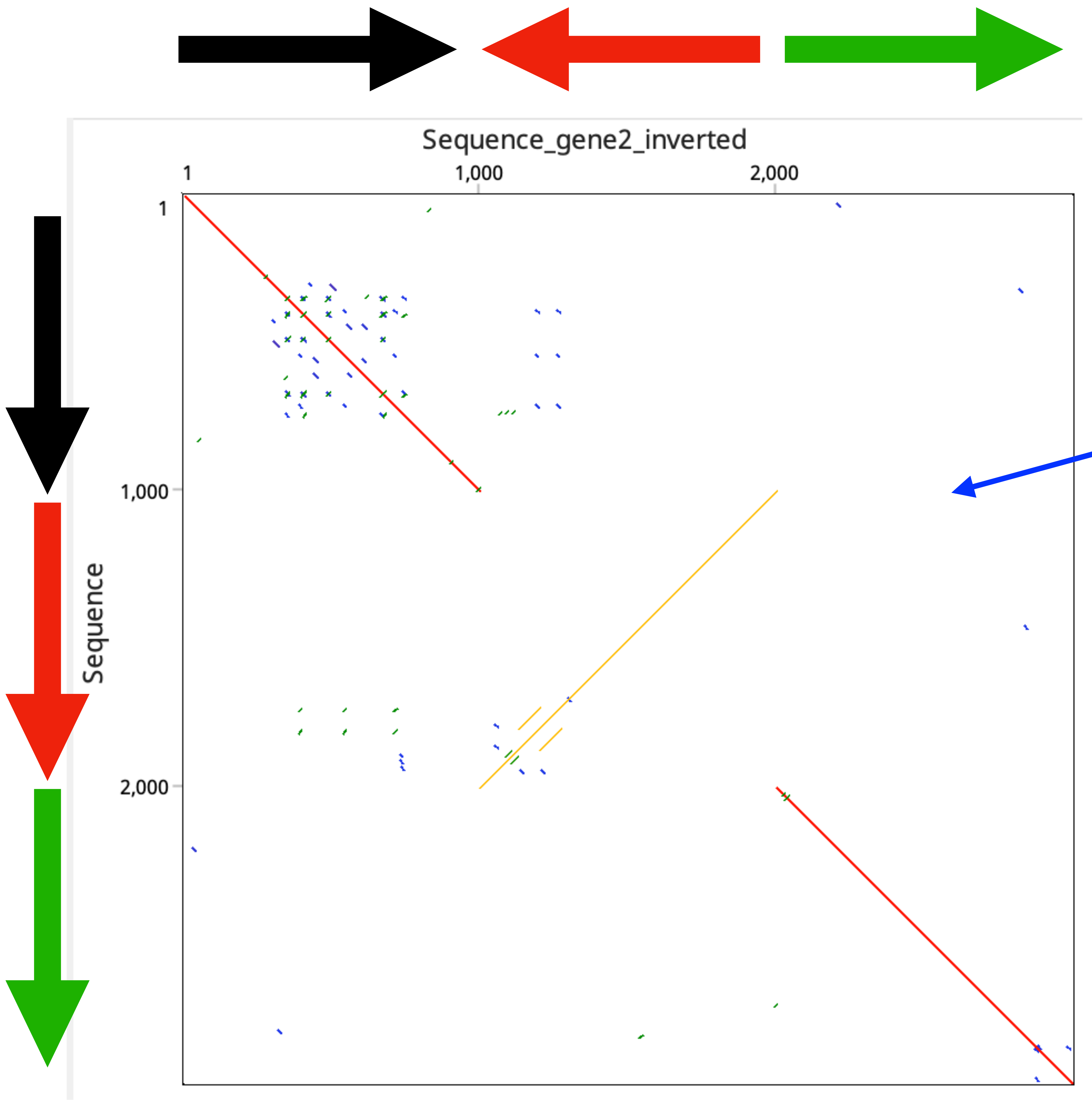
(Or a deletion of the GFP gene from pcDNA3.1-GFP)

Dotplots allow you to visualize structural differences between sequences
For example, a gene duplication



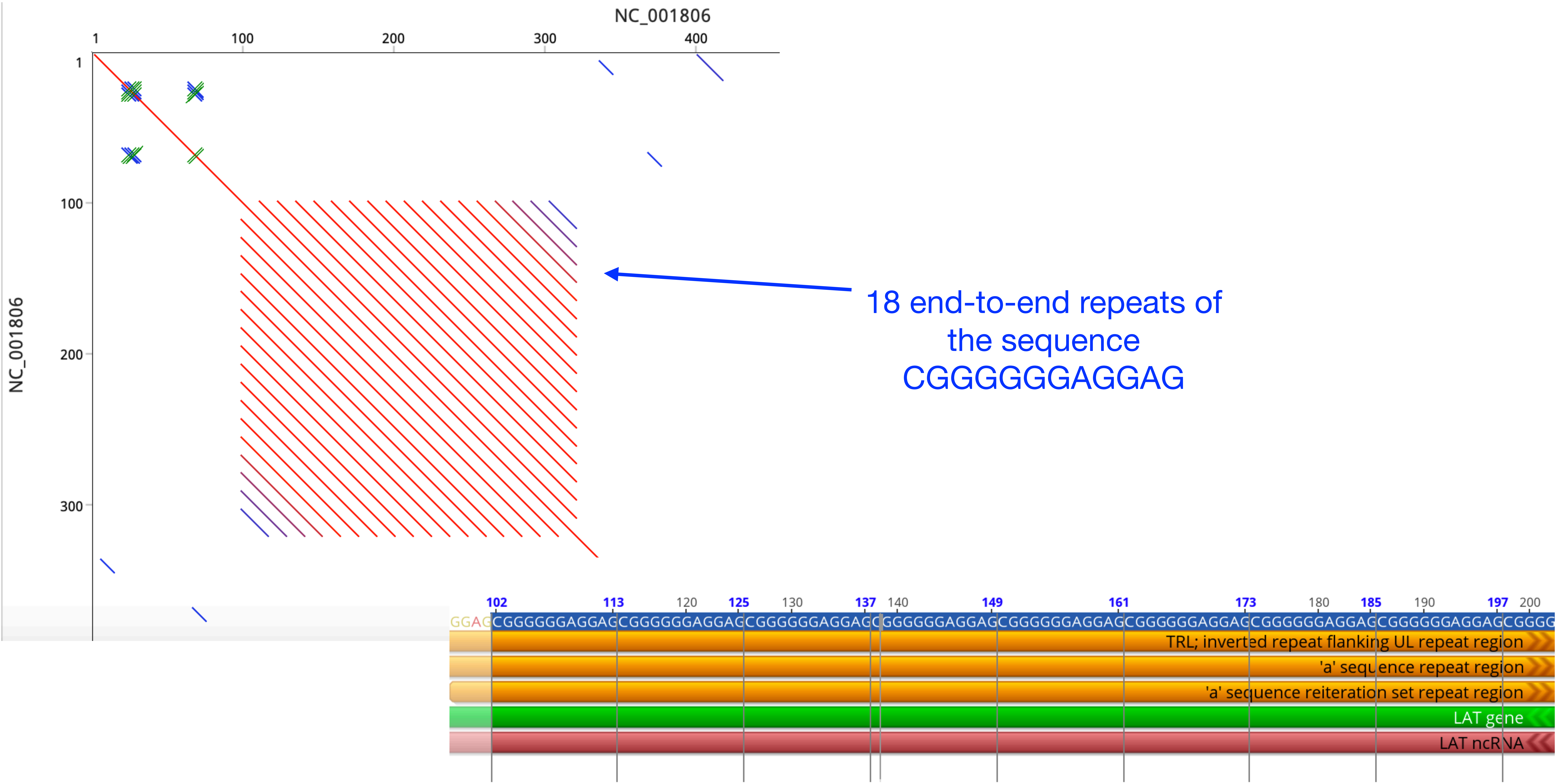
Bases 1000-2000 in the shorter sequence are present in 2 copies in the longer sequence

A gene inversion on a dot plot



Bases 1000-2000 in the two sequences are reverse complements of each other

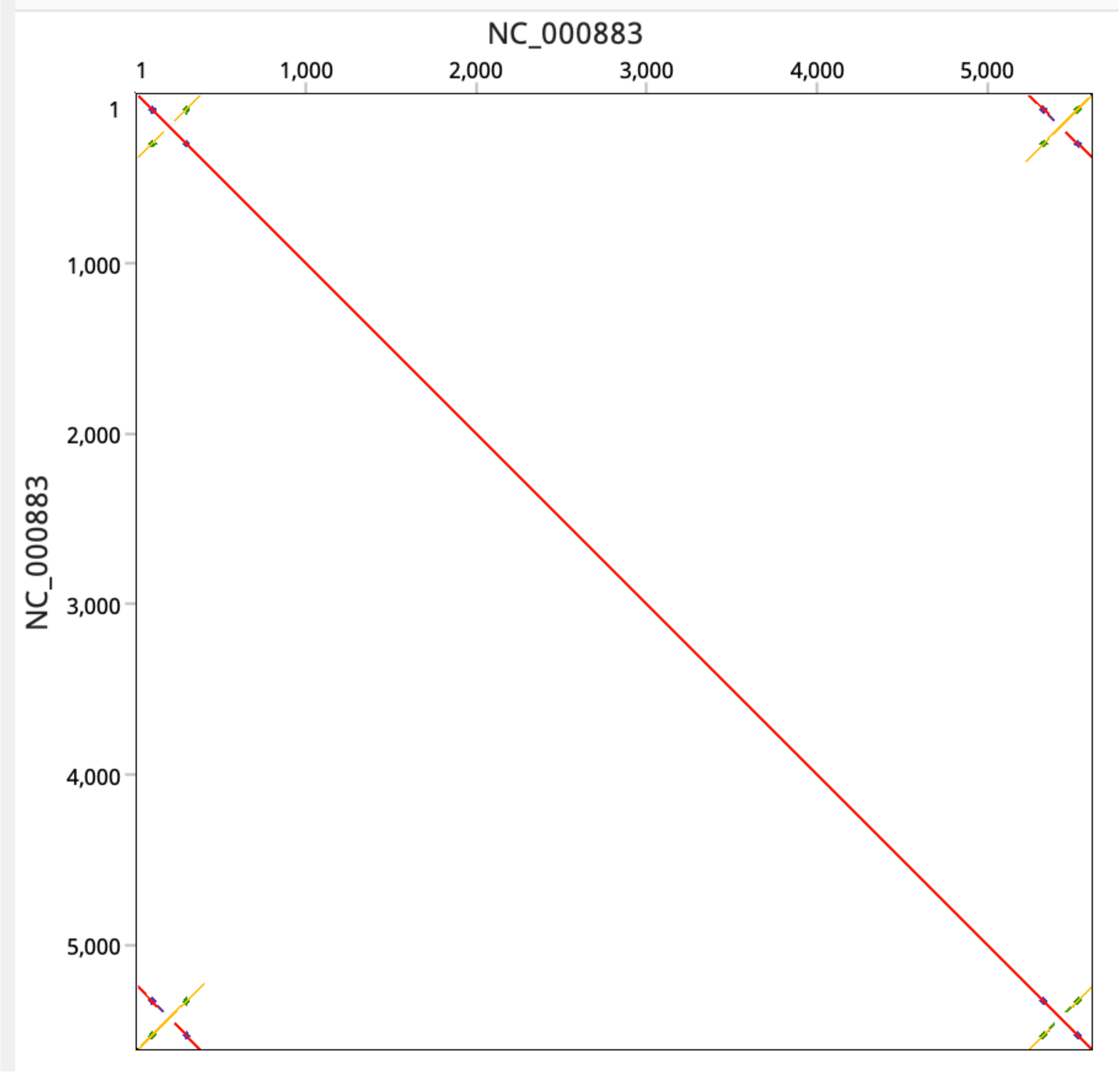
Simple tandem repeat in a herpes simplex virus genome self dotplot

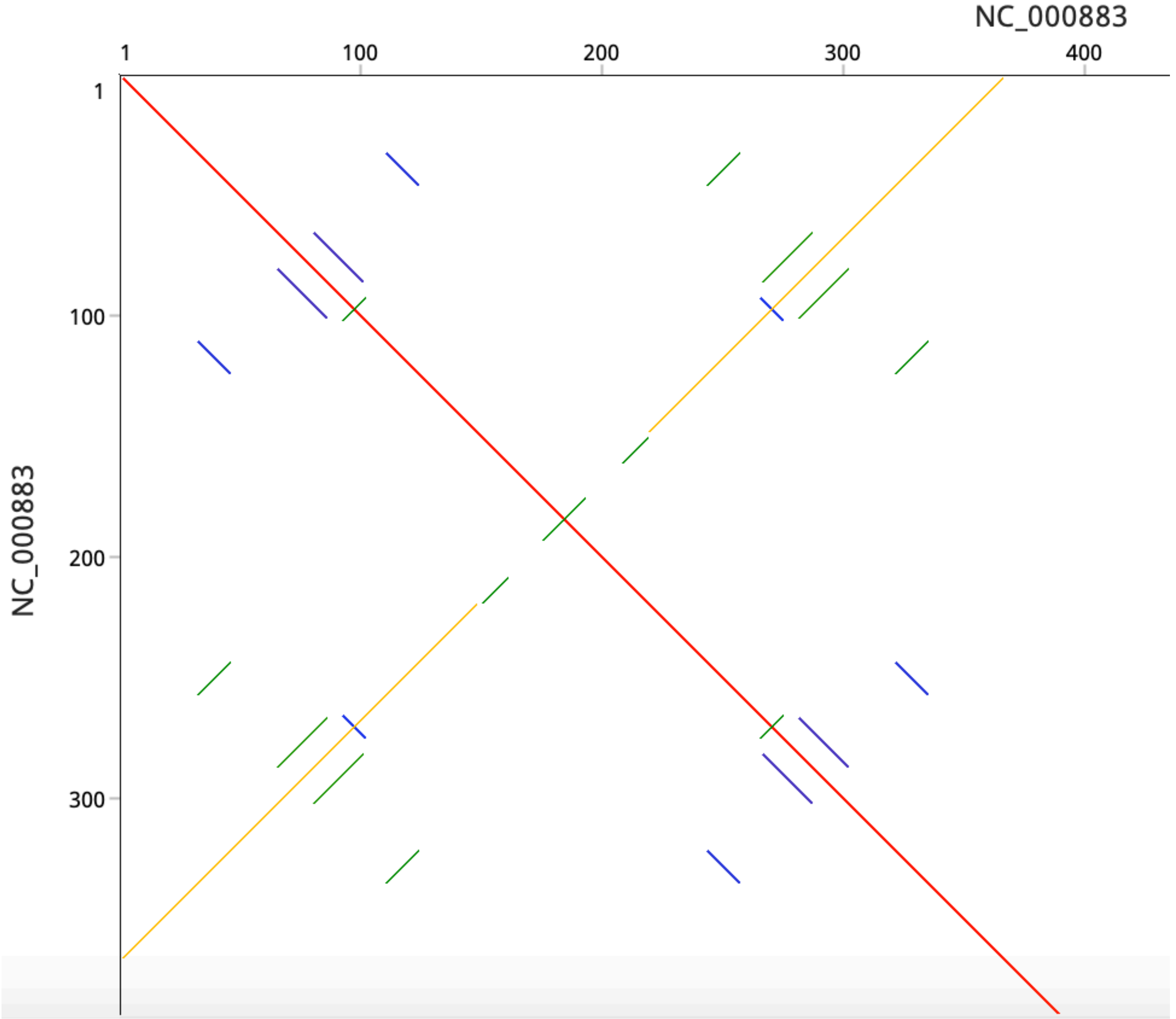


18 end-to-end repeats of the sequence CGGGGGAGGAG

Biological sequence repeats can produce beautiful patterns in dot plots

Human parvovirus B19 self dotplot

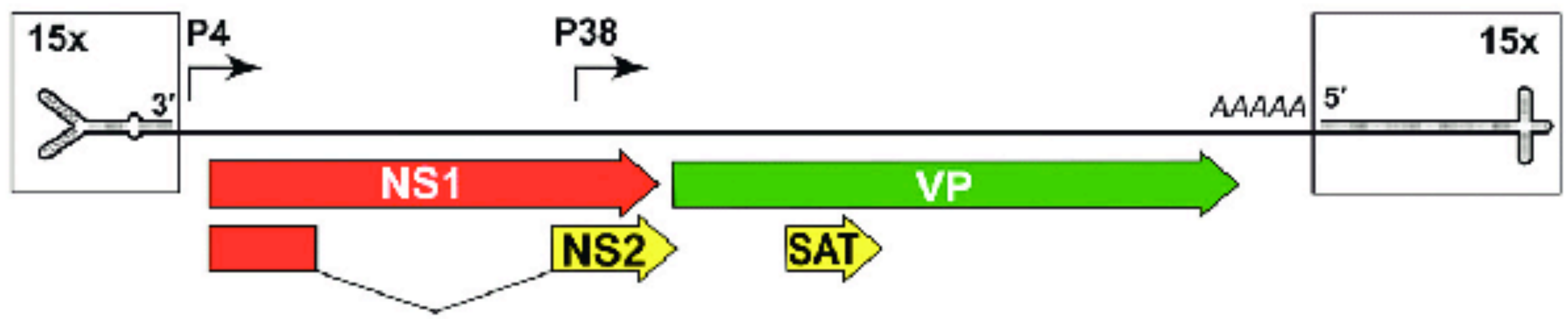




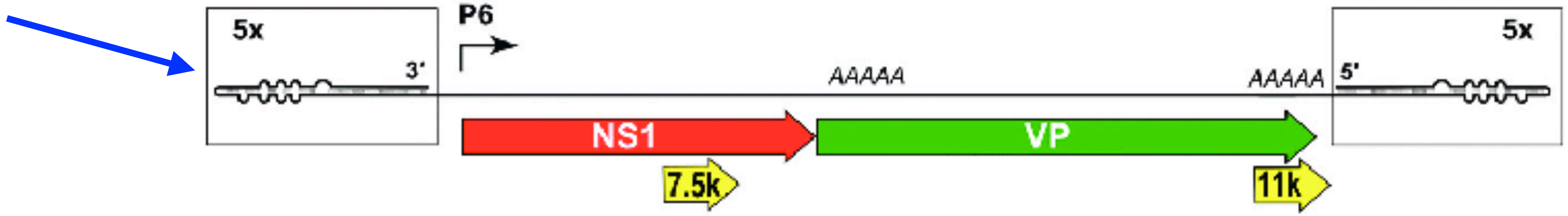
Zoomed in view of upper left corner

Bases ~180-360 are the reverse complement of bases 1-180

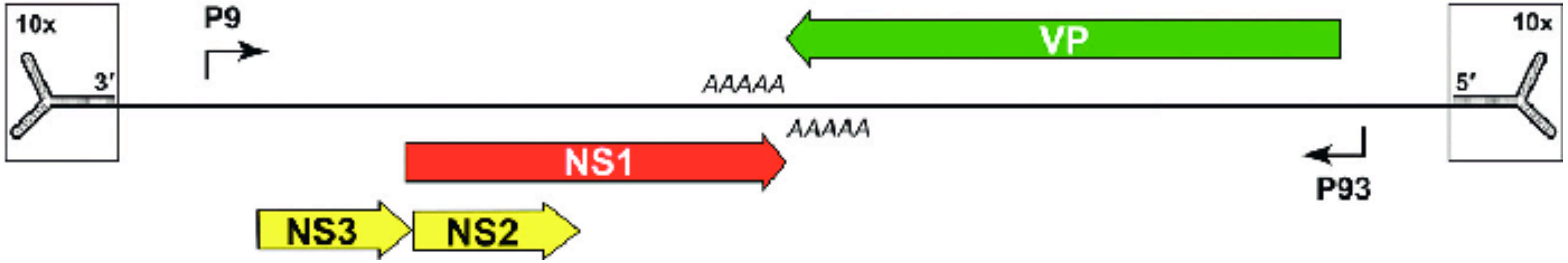
Genus *Protoparvovirus* - minute virus of mice - heterotelomeric - 5148 nt



Genus *Erythroparvovirus* - human parvovirus B19 - homotelomeric - 5596 nt



Genus *Ambidensovirus* - Galleria mellonella densovirus - homotelomeric - 6039 nt



Cotmore et al (2019) J Gen Virol