Exploring Satellite DNA & Tandem Repeats with ModDotPlot

Alex Sweeten

Biodiversity Genomics Academy
October 7th, 2024

@alexsweeten

@alexsweeten.
bsky.social

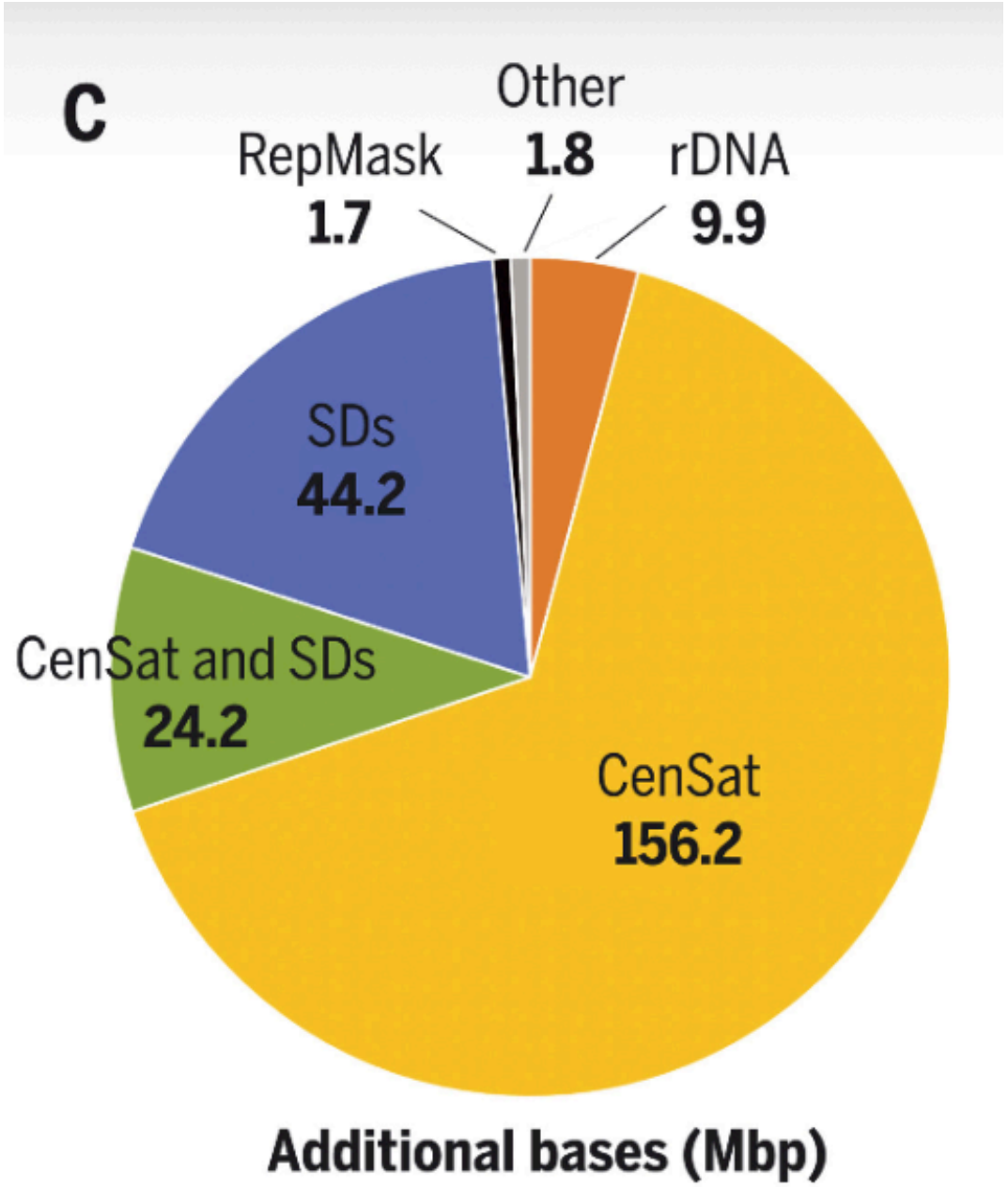
NIH National Human Genome Research Institute

The Forefront of Genomics®

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

A new era

- The first T2T human genome introduced over 200 Mbp of new sequence.
- Overwhelming majority of this are **satellite & segmentally duplicated** sequences.



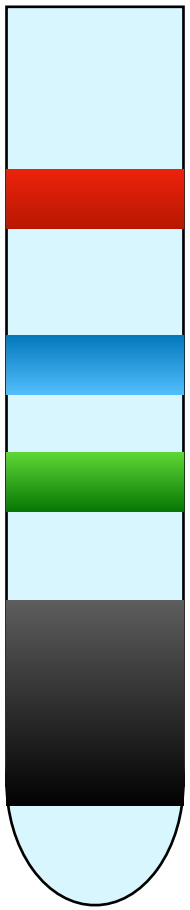
The complete sequence of a human genome.
Nurk, Koren, Rhie, Rautiainen, et al. *Science* (2022)

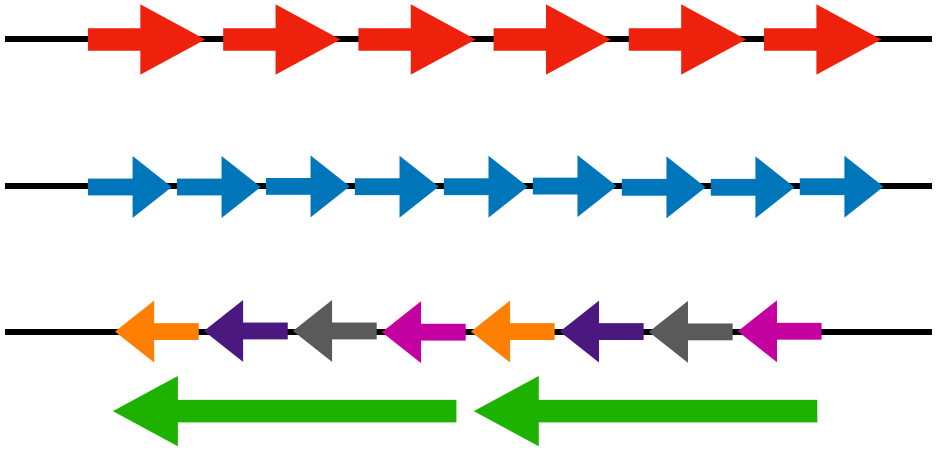
What is Satellite DNA?

NIH
NHGRI

Satellite DNA

- Consists of long arrays of **tandemly repeating** sequences.
- Primarily found in **centromeres** and **acrocentric short arms**.
- Compared to bulk DNA, satellites differ in sequence complexity, GC-content, and tandem repeat size.





NIH
NHGRI

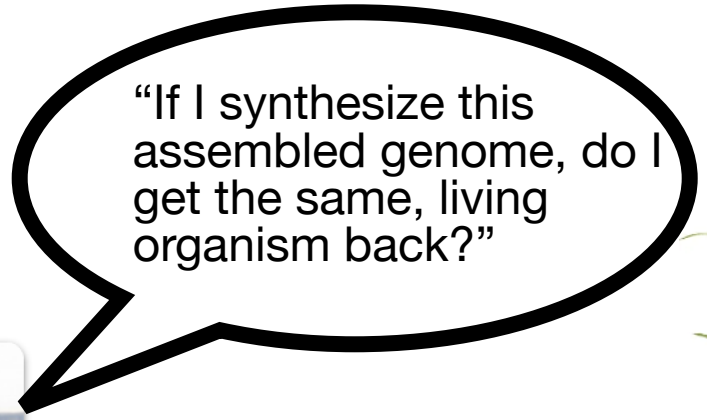
Who cares about satellites?

NIH
NHGRI

Centromeres are Satellites!

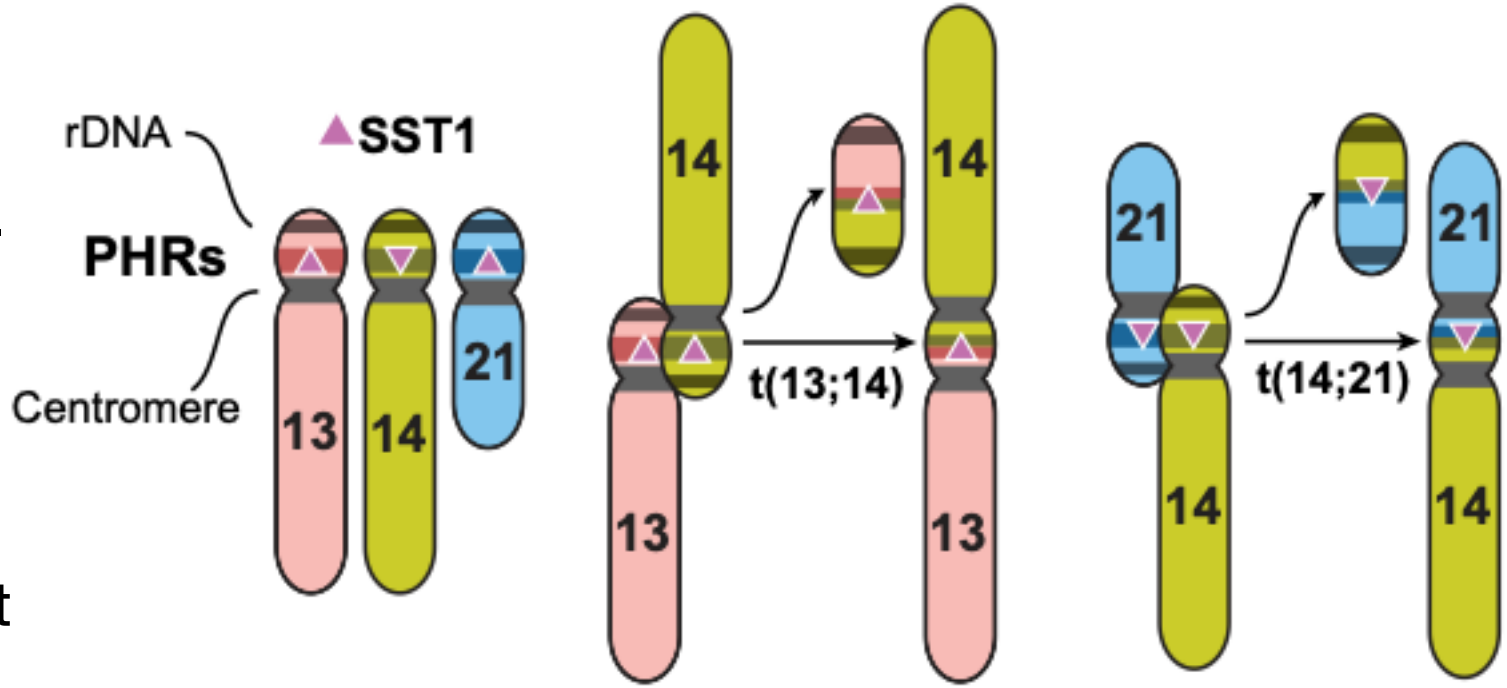
- Without them… you die!

- αSatellites required for kinetochore binding.



Robertsonian Translocations

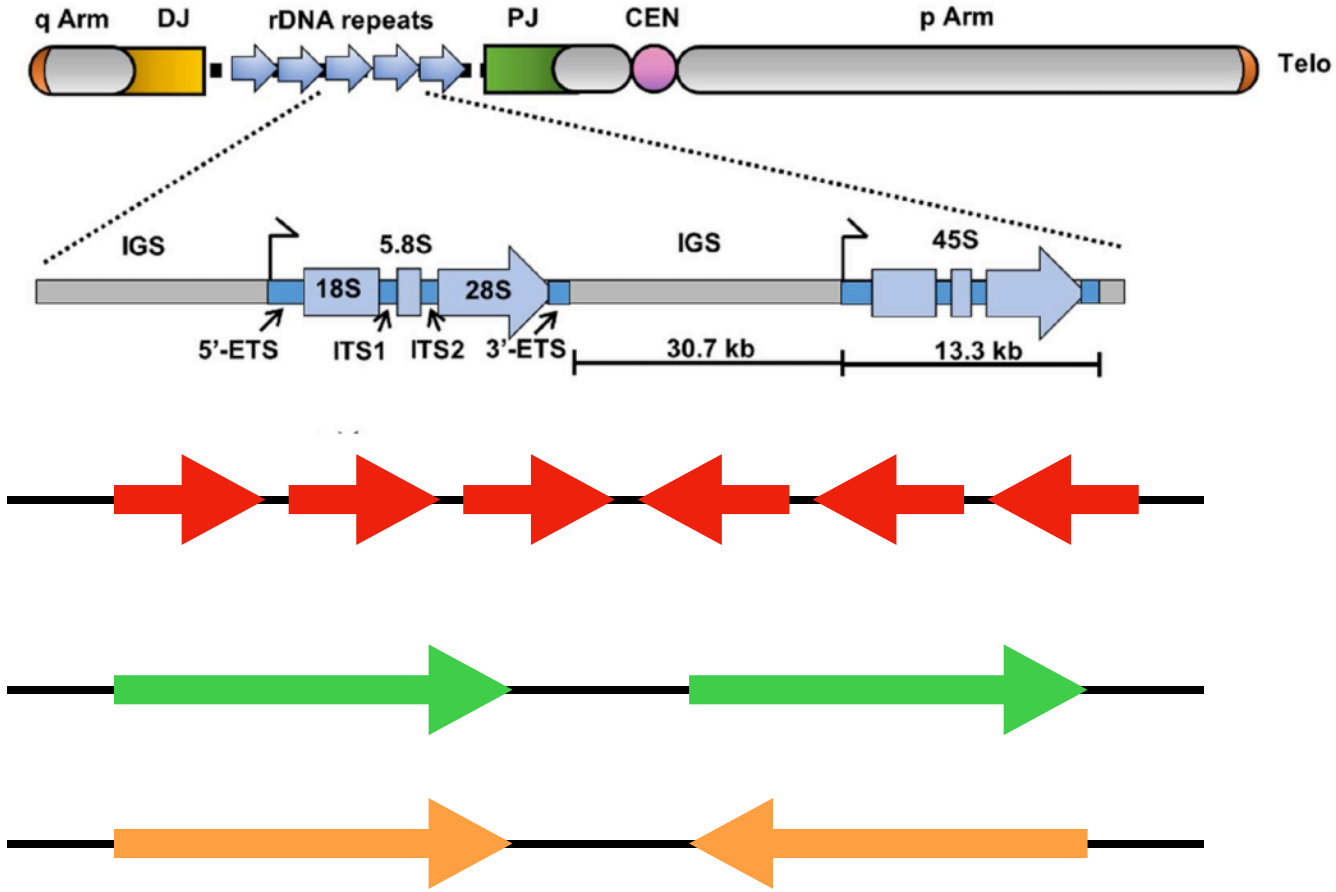
- Present in 1/800 humans.
- Acrocentric short arms recombine during meiosis.
- SST1 satellite array is on an inverted duplication in Chr14. This leads to long-arm fusions & loss of short arms.



The formation and propagation of human Robertsonian chromosomes.
De Lima, Guarracino, Koren, Potapova, et al. *BiorXiv* (2024)

Other Repeats Worth Caring About

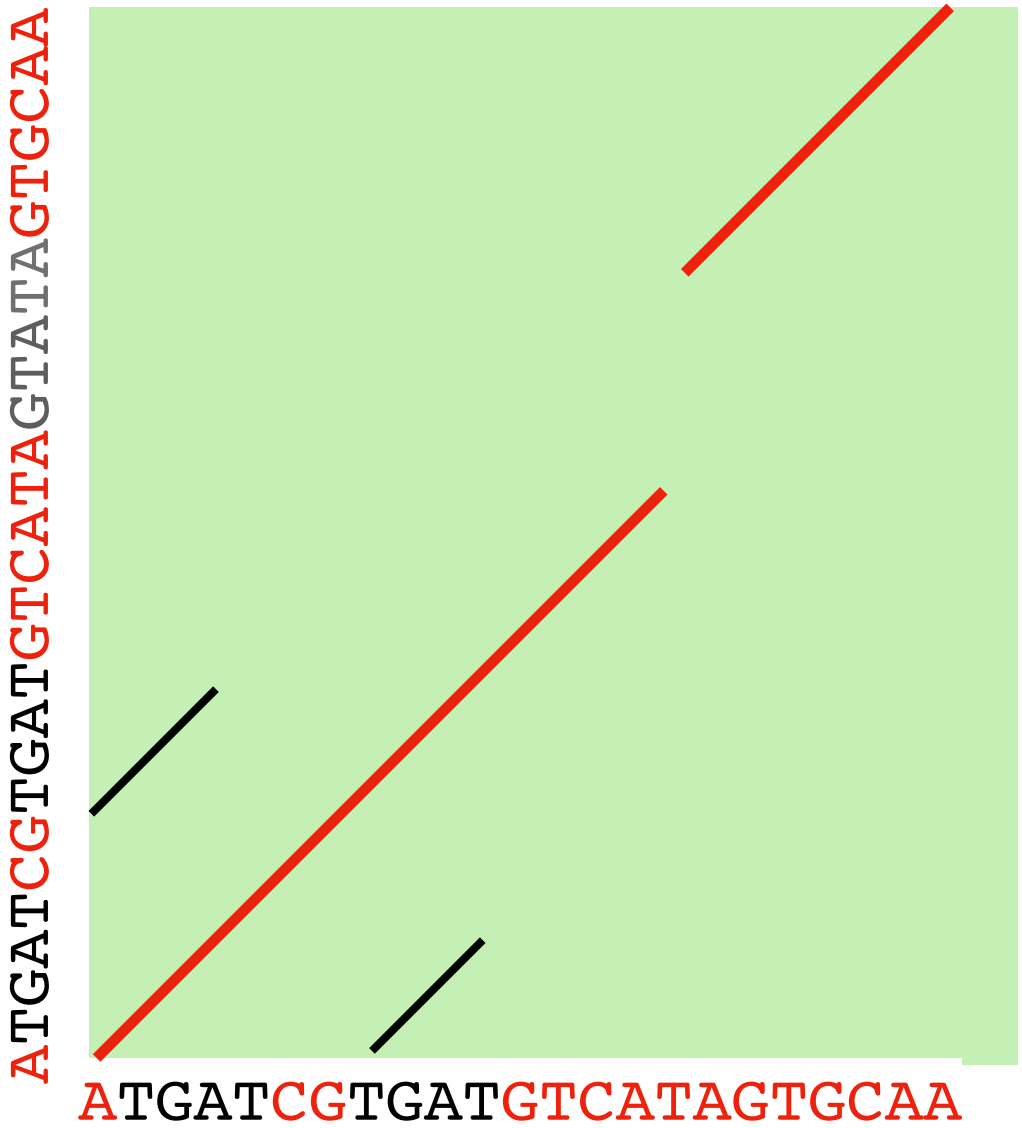
- Tandem gene Arrays:
    - *TSPY* - ChrY
    - *rDNA* - Acrocentrics
- Palindromes
- Large Duplications
- Large Inversions



How can we visualize repeats?

Dotplots

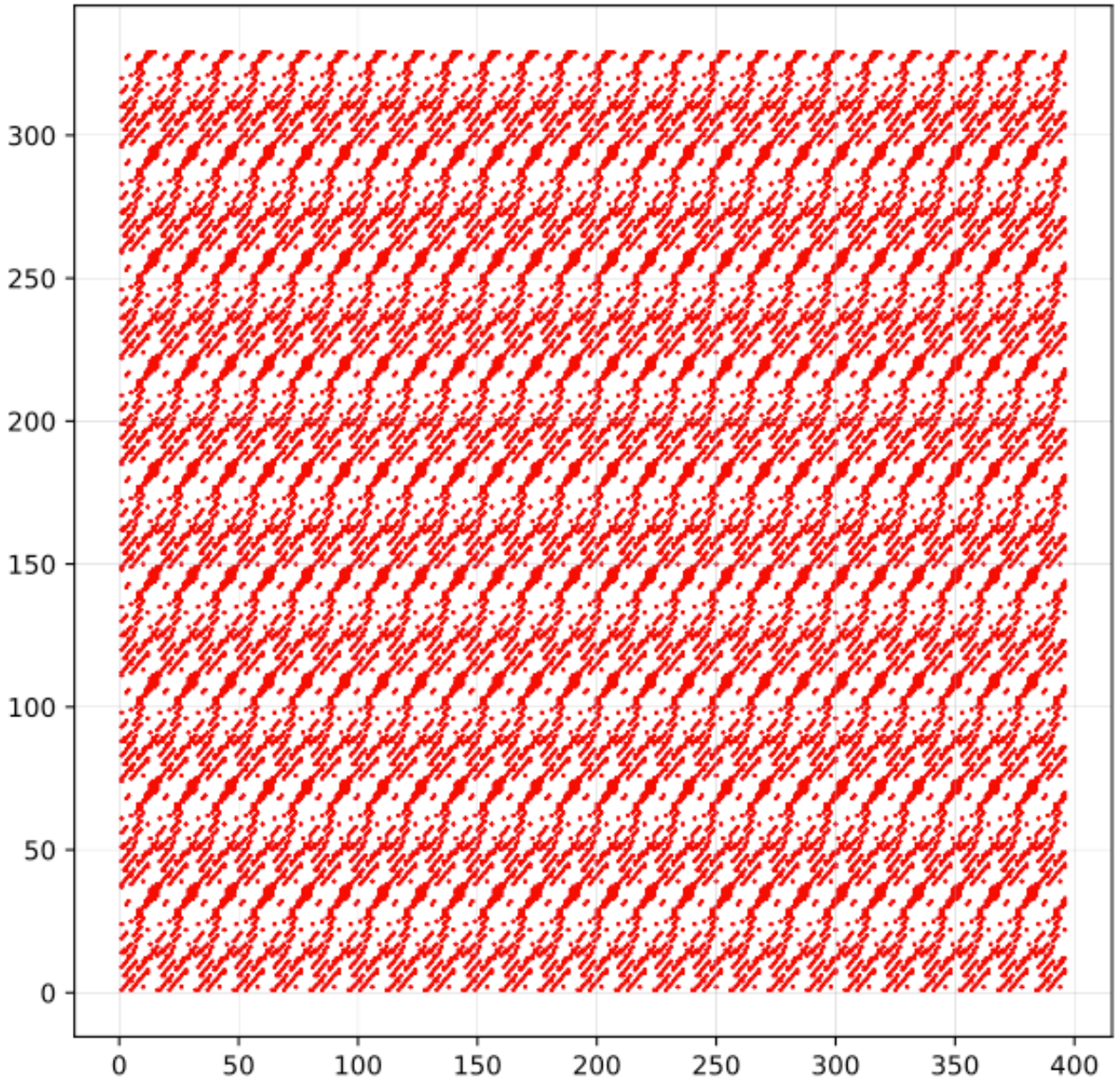
- A matrix consisting of two sequences, drawn with matches of length *l* (4 in this example).
- Shows self-similarity (self-plot) and orthologous relationships (comparative plot).
- “Vectorized” approach.



Dotplots

- Visualizing larger repeats quickly becomes overwhelming!!

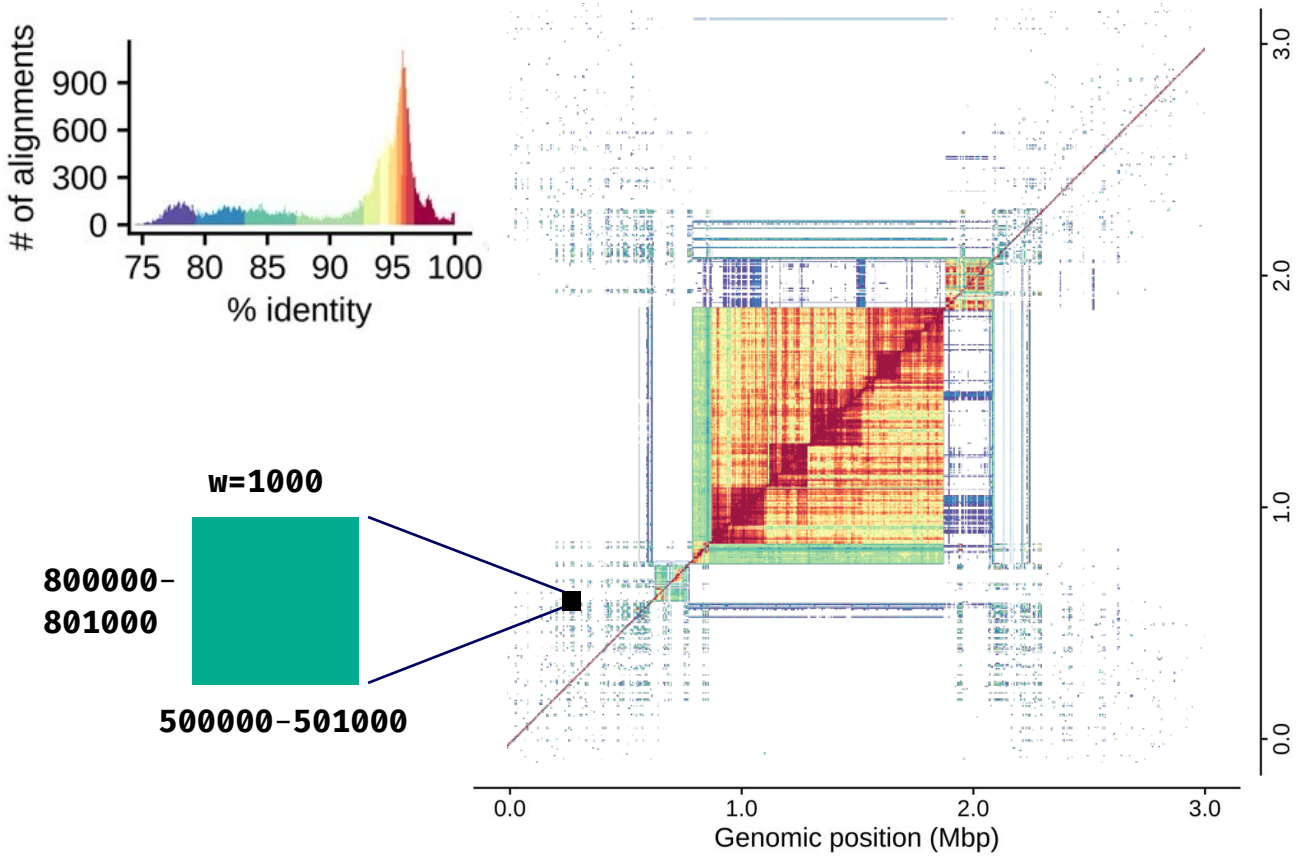
(GATTG)n



NIH
NHGRI

Dotplots

- Convert a vectorized into a rasterized model.
- Each pixel represents a pair of intervals of a fixed size *w.*
- Can compute sequence identity for each pairwise combination of intervals.

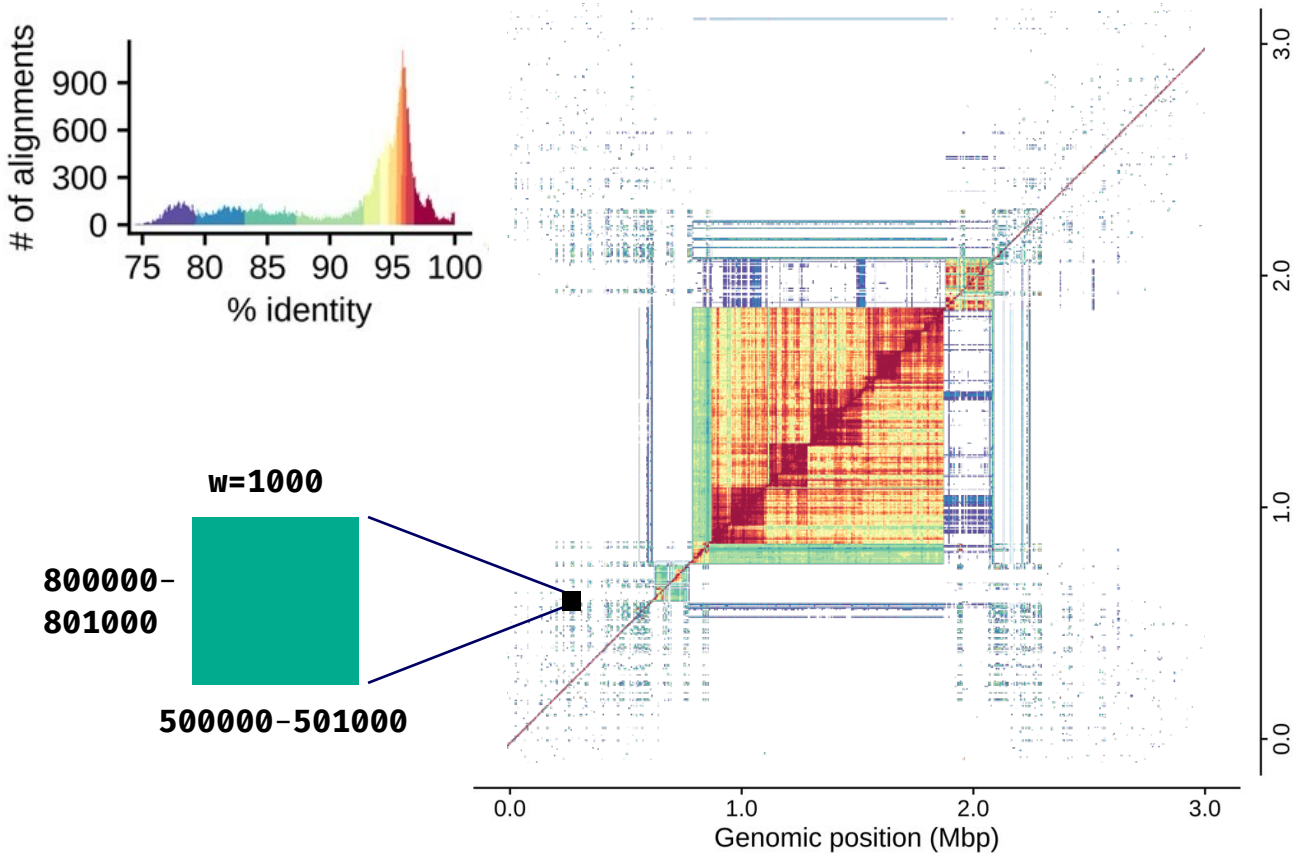


StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps.
Vollger, Kerpedjiev, Phillippy, Eichler. *Bioinformatics* (2022)

Computing Sequence Identity

$$100\left(\frac{M}{M+X+I+D}\right)$$

- Use Minimap2 to estimate identity through matches (M), mismatches (X), insertions (I) and deletions (D)
- Scales quadratically with genome size. We can do better!



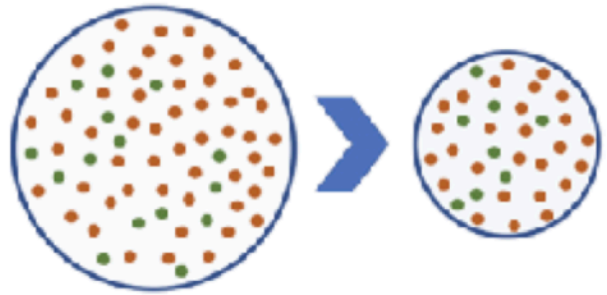
StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps.
Vollger, Kerpedjiev, Phillippy, Eichler. *Bioinformatics* (2022)

We can save time with sketching!

NIH
NHGRI

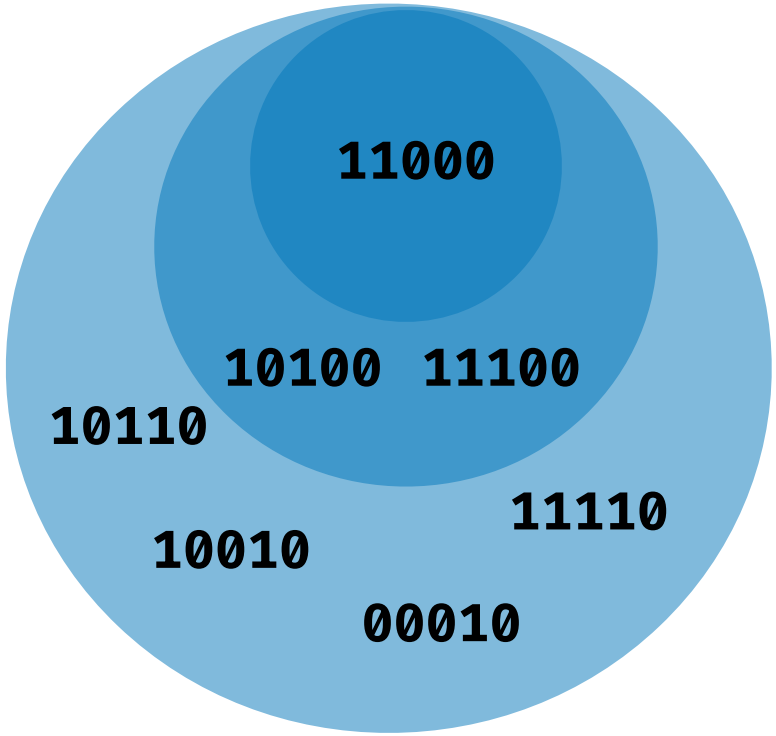
What is sketching?

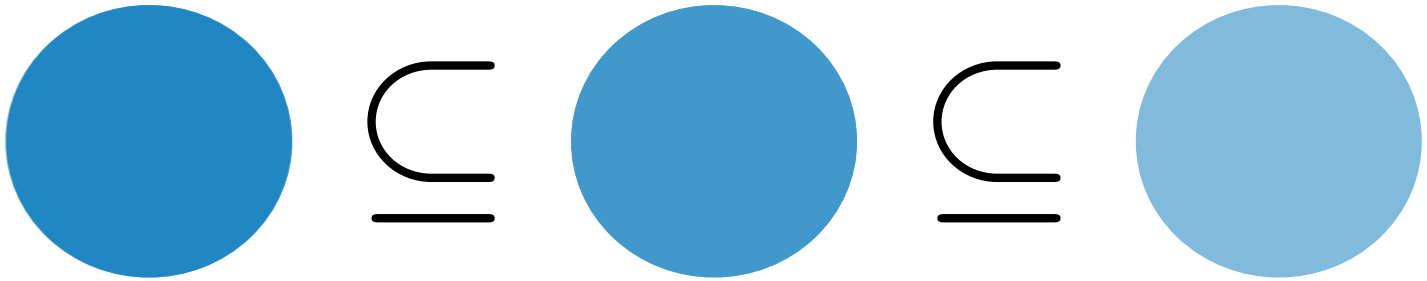
- “Smart” subsample of the data
- Has a guarantee that certain properties are retained.



Modimizer Sketch

- Apply a hash function on all *k*-mers. Check if *k*-mer *mod s = 0.*
- If *s* is a power of 2, this creates a guaranteed hierarchy of *k*-mers
- Efficient: simply look at the last log2(s) digits for 0's.

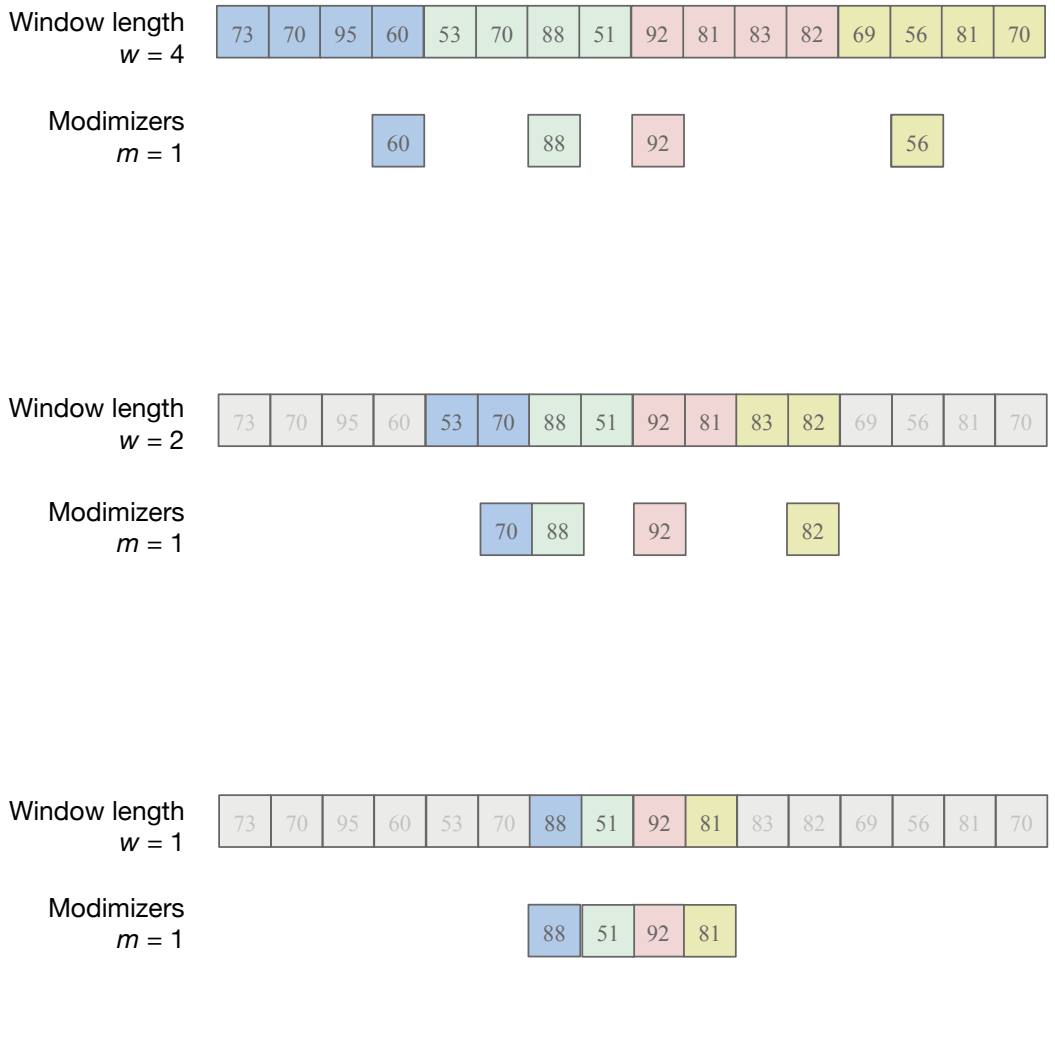


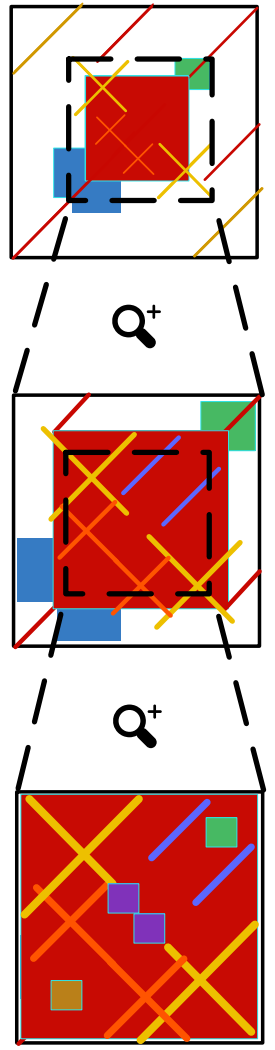


On the Resemblance and Containment of Documents.
Broder. IEEE Complexity of Sequences (1997).

Interactive Plots

- Modimizer hierarchies allow for the construction of dynamically changing dotplots.
- Adapt *s* to the current field of view.
- In this example, 1 modimizer is used

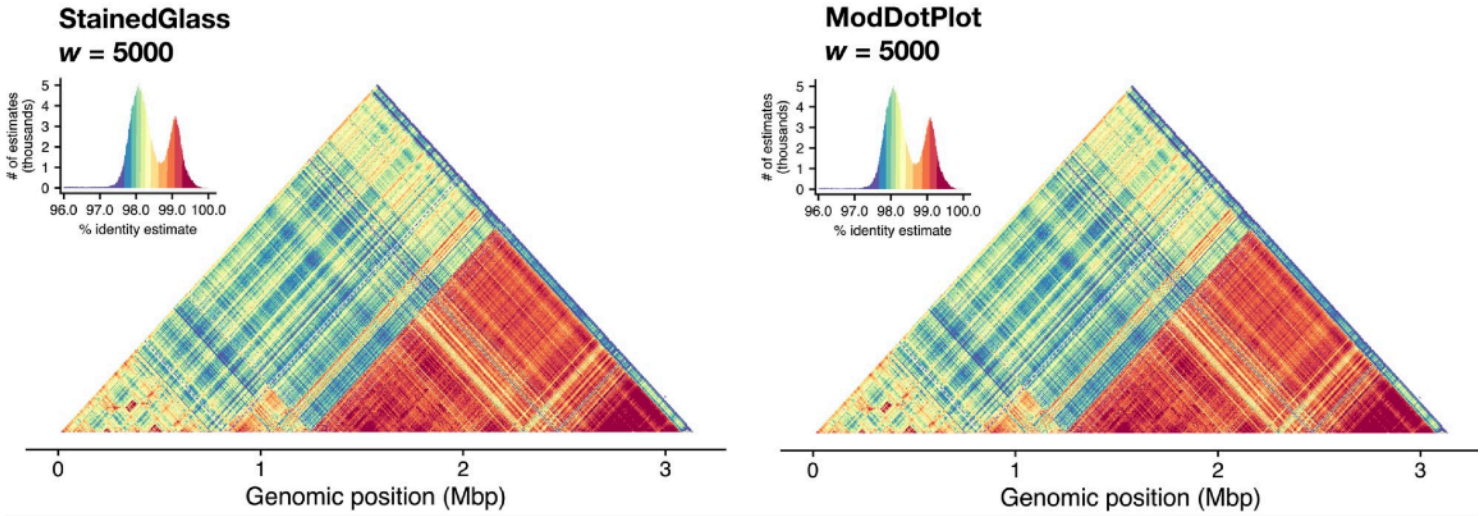


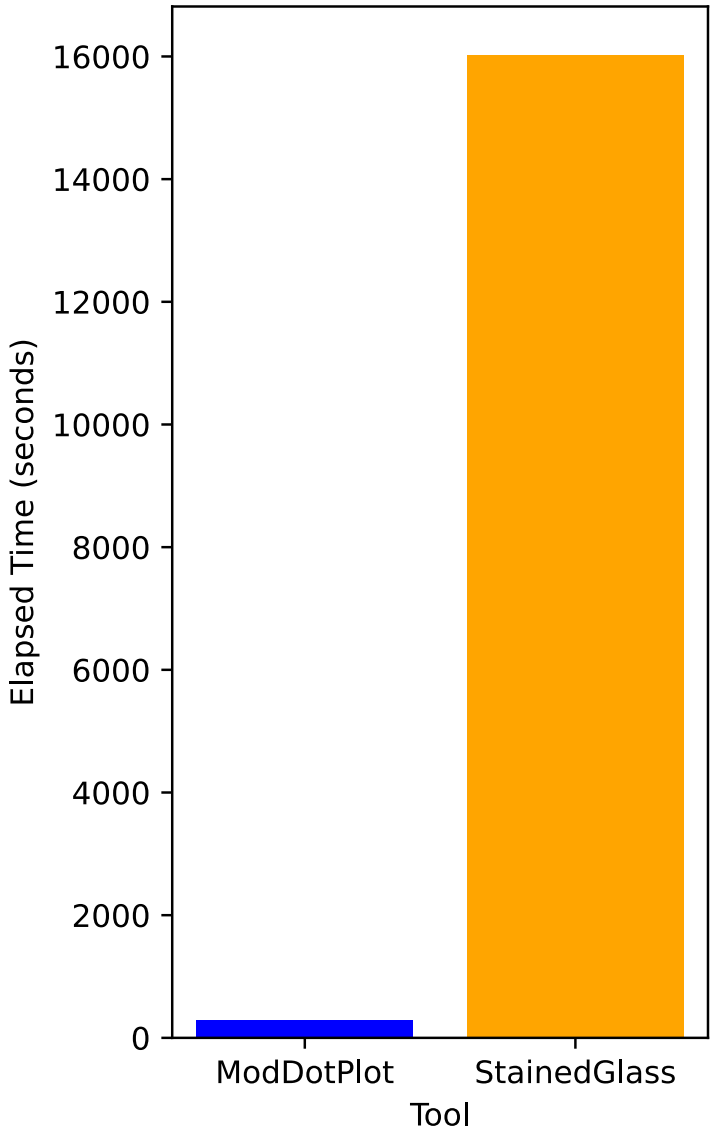


NIH NHGRI

Speed Comparison

- Computes identity for the *Arabadopsis* whole genome in ~5 minutes, while accurately representing the data!





NIH NHGRI

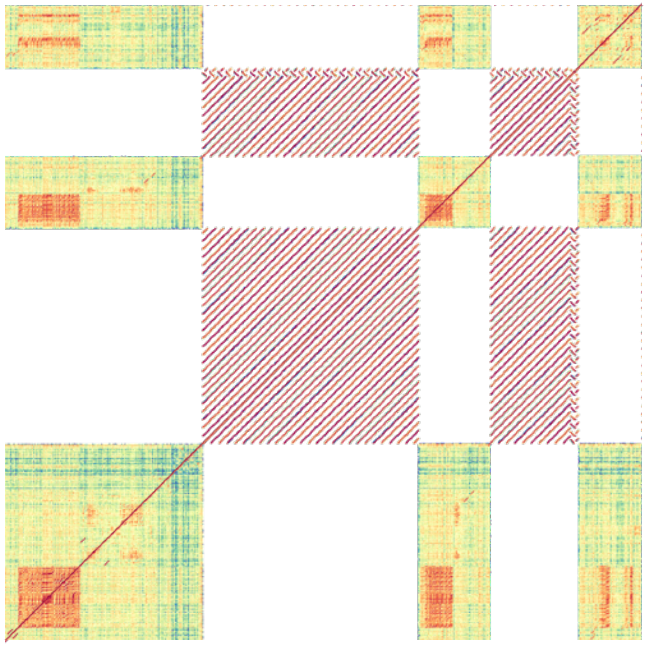
How do I interpret a dotplot?

NIH
NHGRI

Interpreting Self-Identity Plots

- ModDotPlot outputs 3 main files:

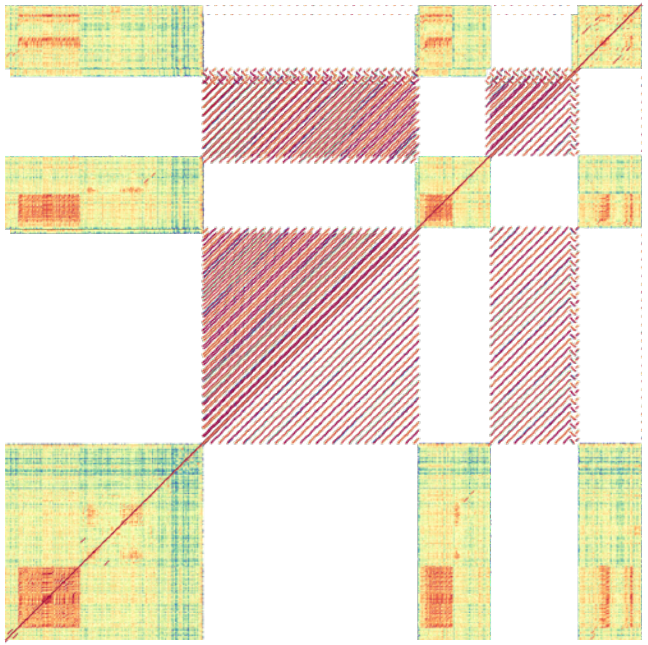
_FULL



Interpreting Self-Identity Plots

- ModDotPlot outputs 3 main files:

_FULL

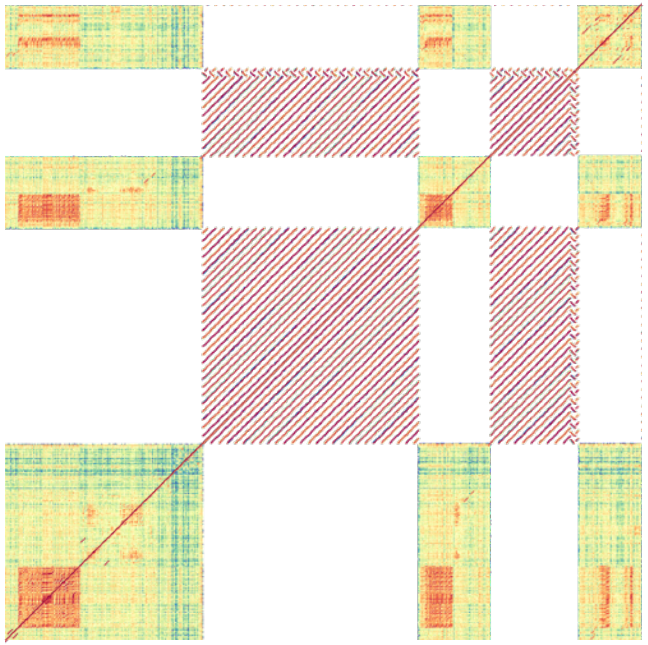


_TRI

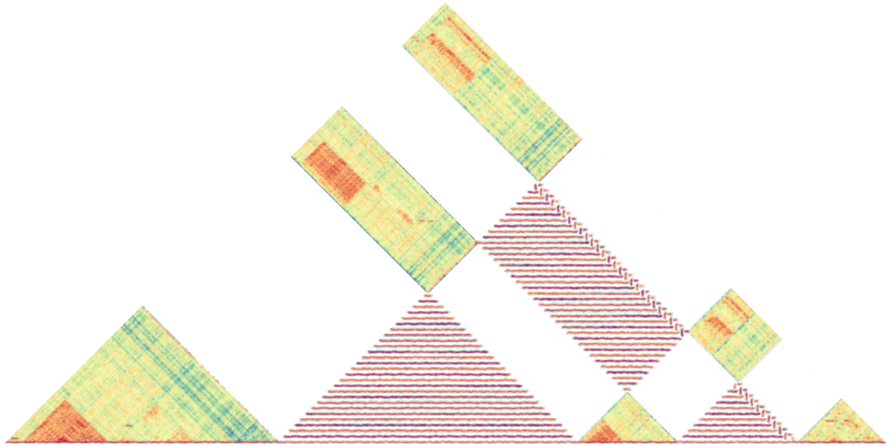
Interpreting Self-Identity Plots

- ModDotPlot outputs 3 main files:

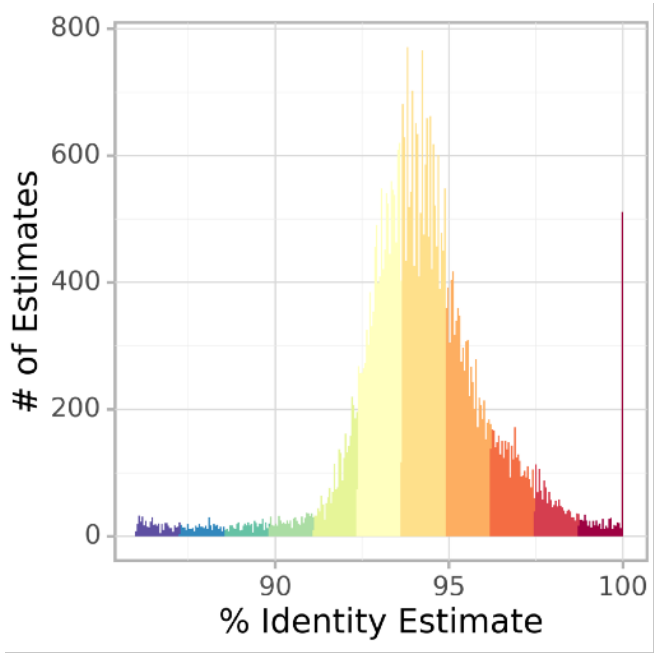
_FULL



_TRI



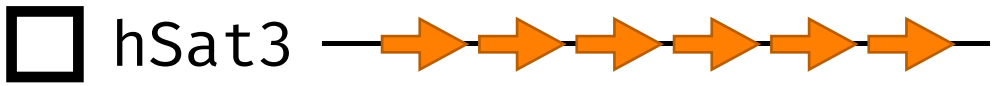
_HIST

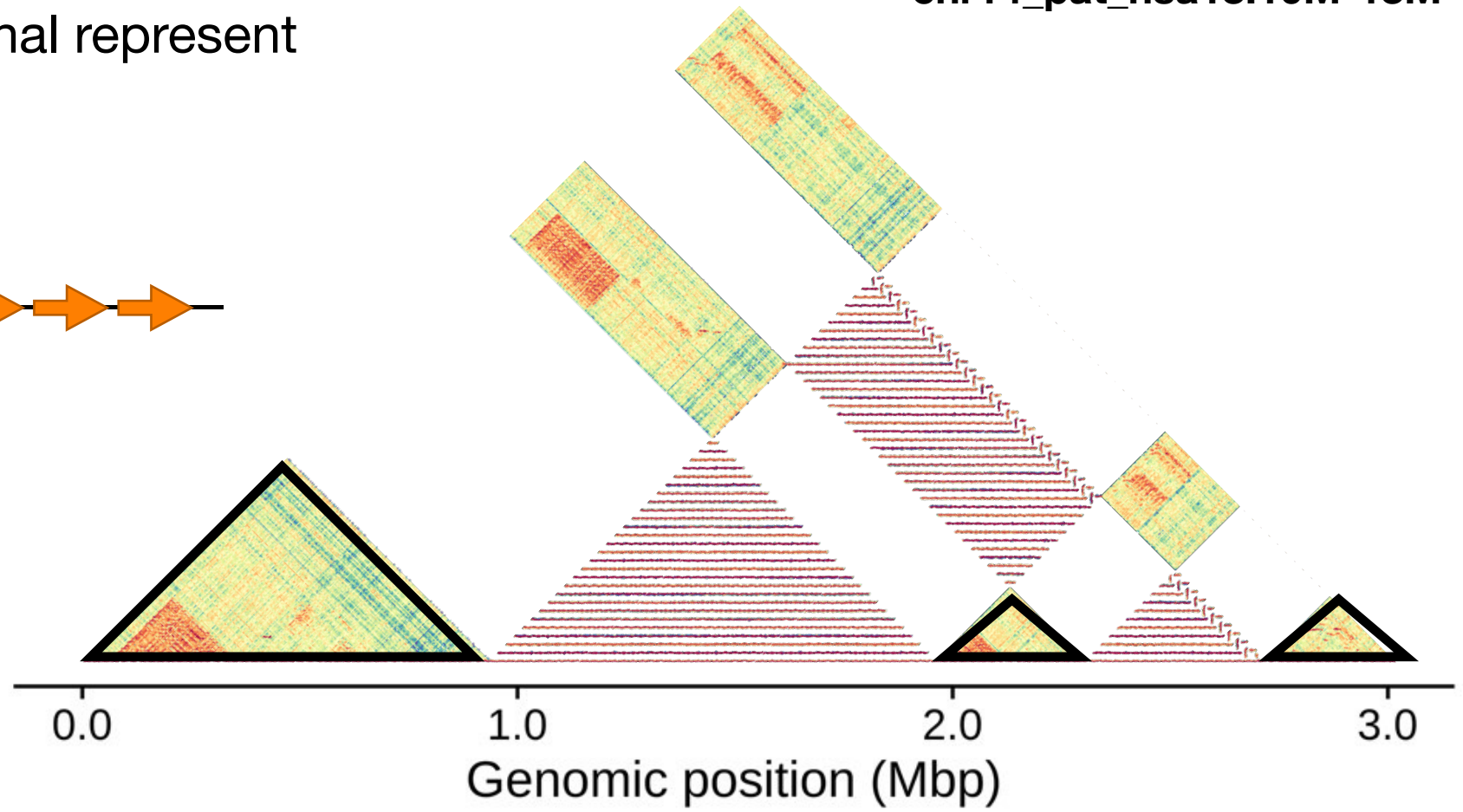


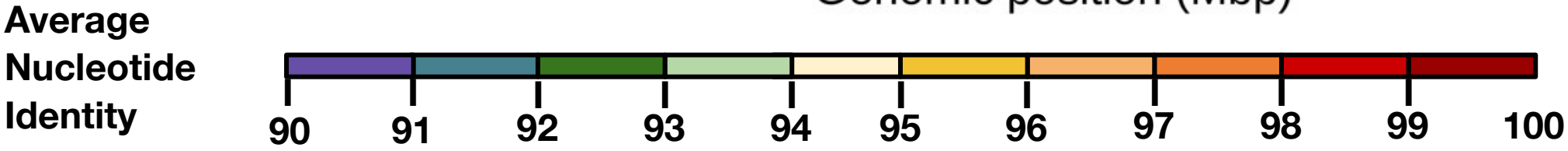
NIH National Human Genome Research Institute

Interpreting Self-Identity Plots

- Triangles on the diagonal represent satellites



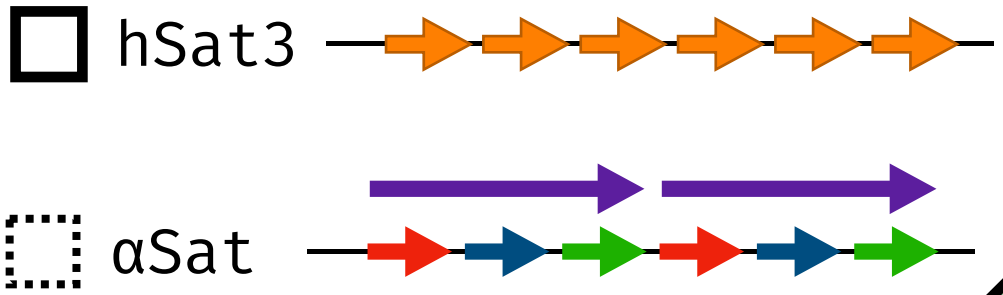


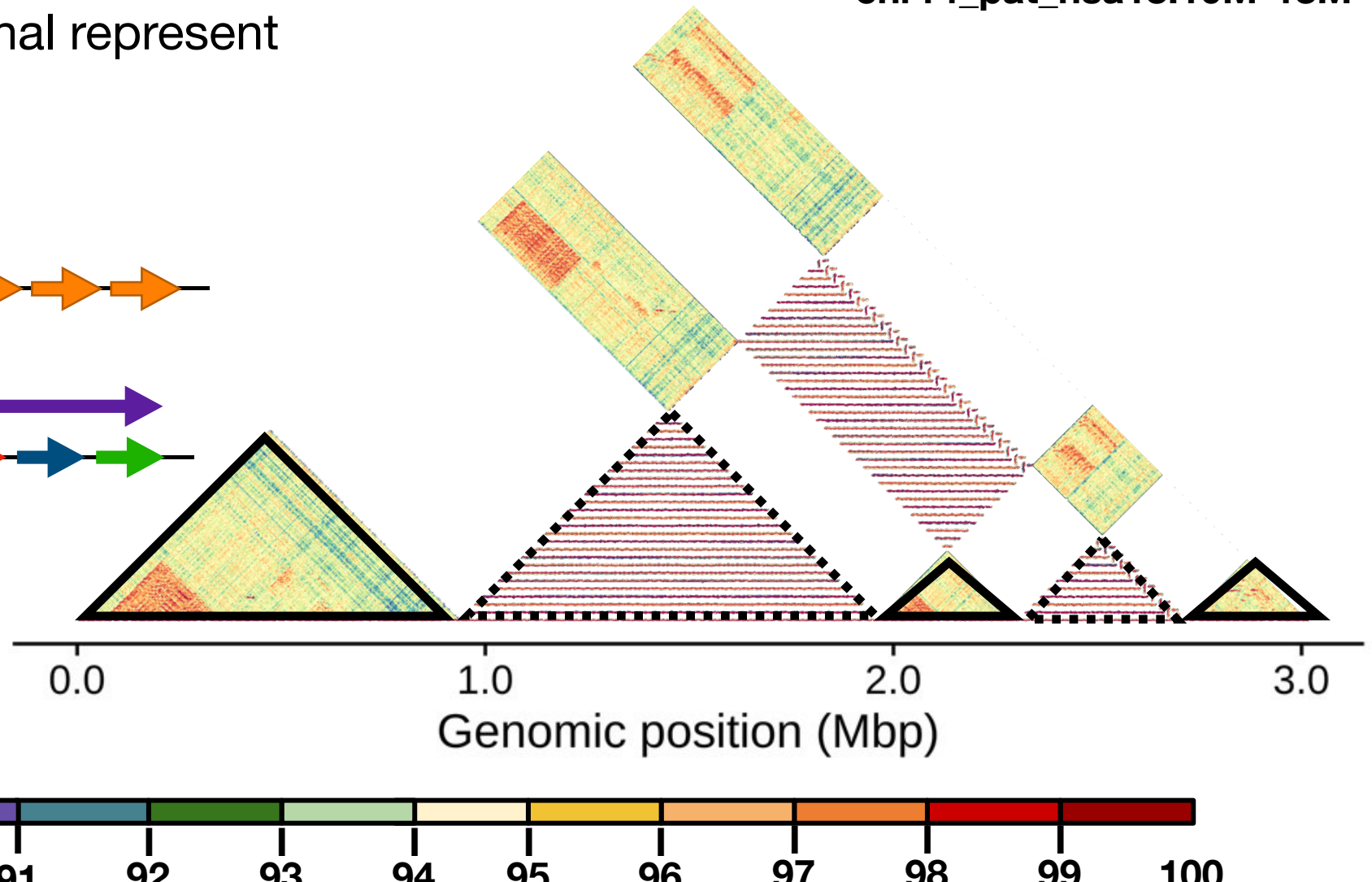


NIH NHGRI

Interpreting Self-Identity Plots

- Triangles on the diagonal represent satellites

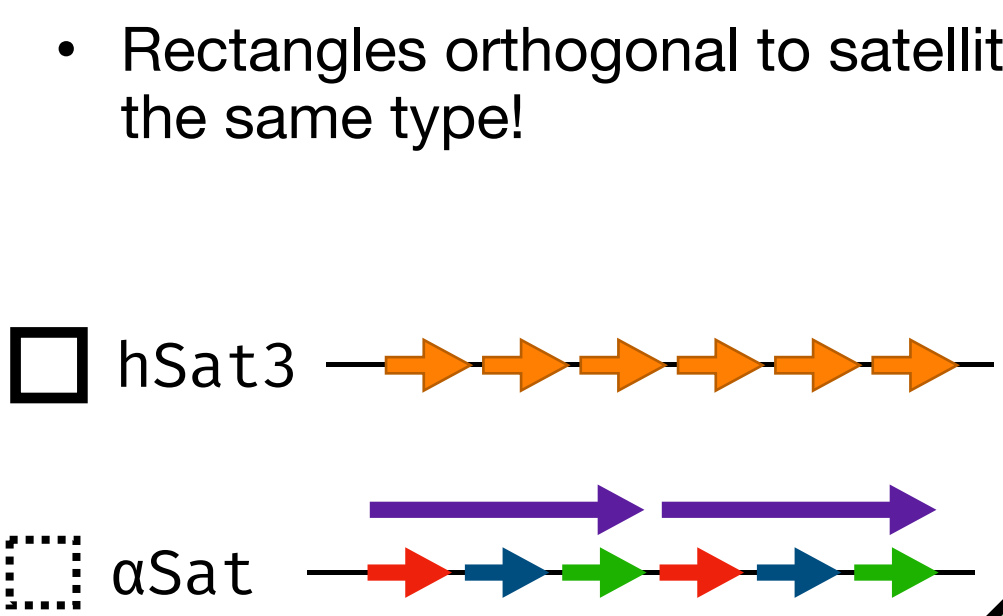


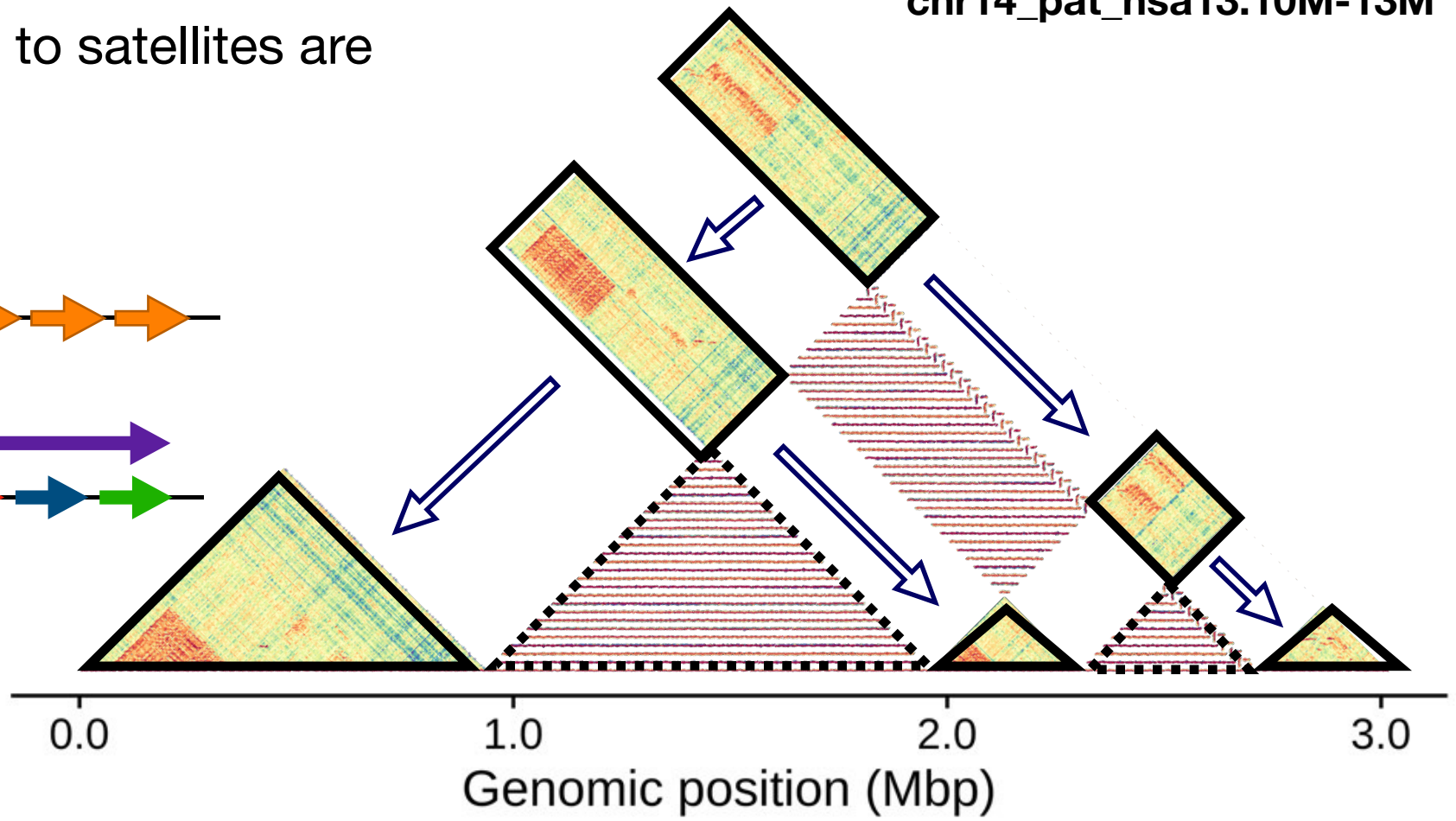


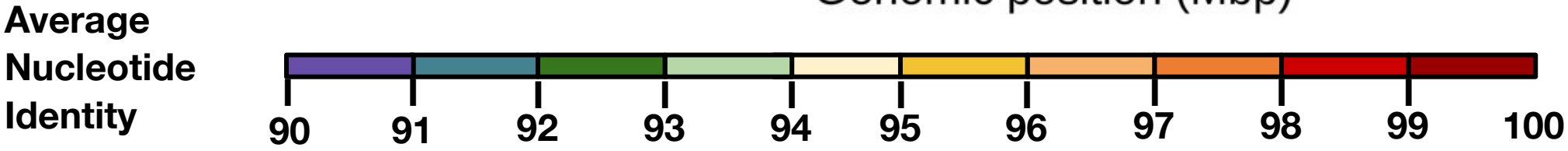
NIH NHGRI

Interpreting Self-Identity Plots

- Rectangles orthogonal to satellites are the same type!



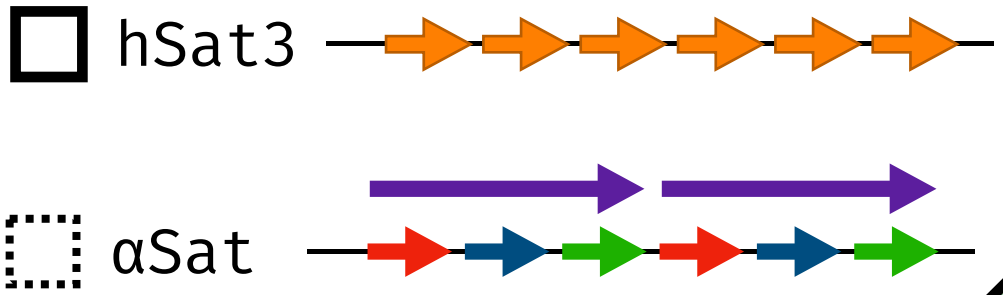


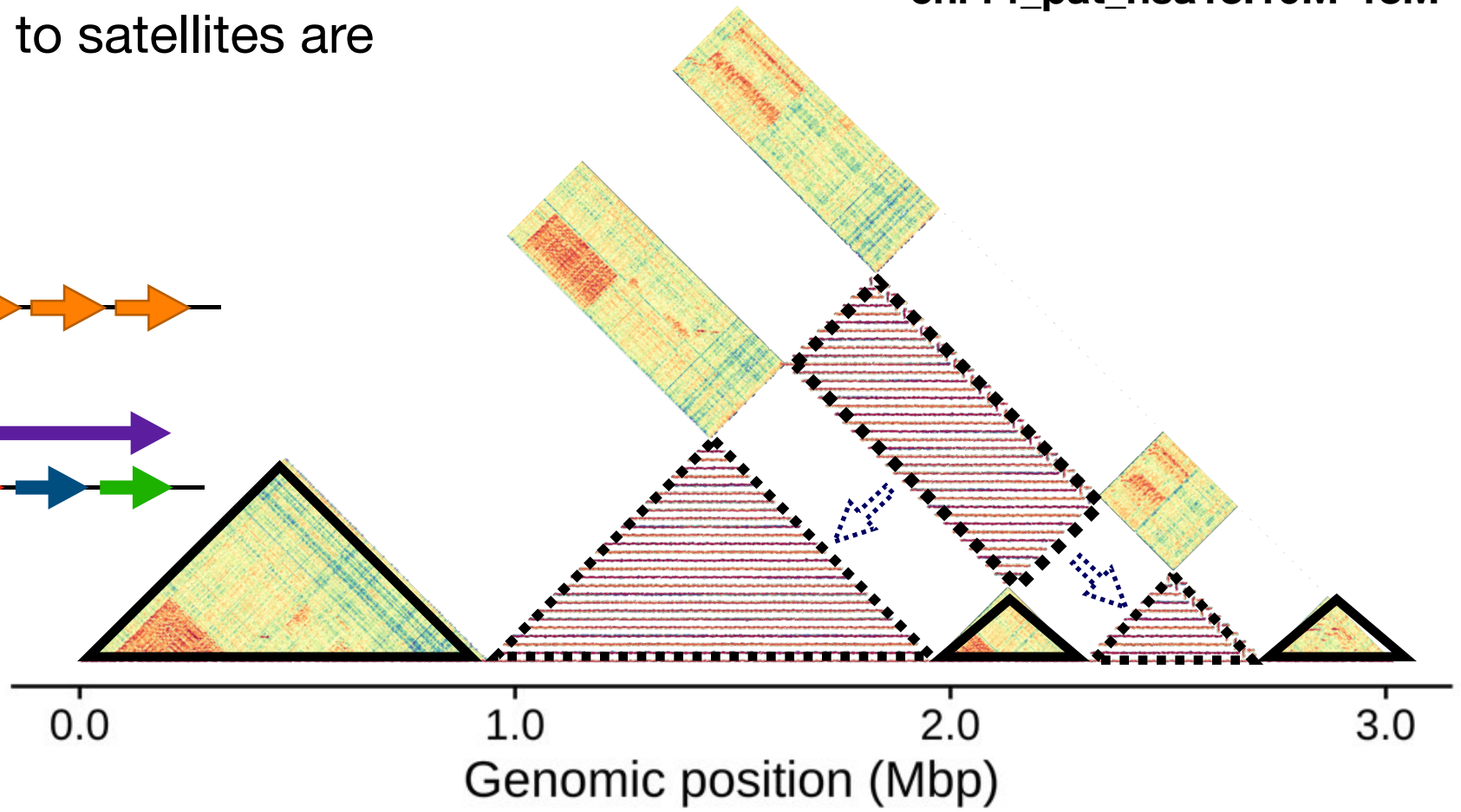


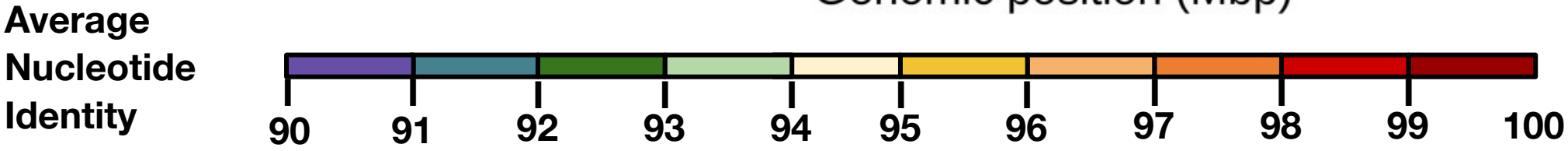
NIH NHGRI

Interpreting Self-Identity Plots

- Rectangles orthogonal to satellites are the same type!

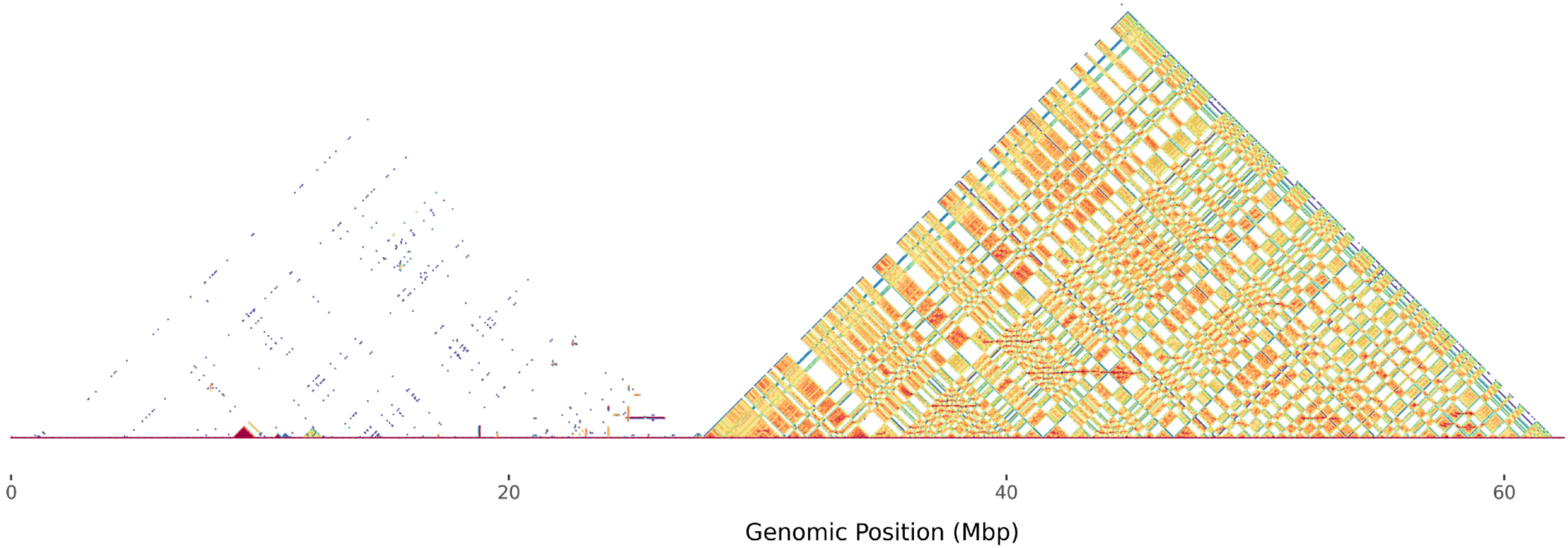




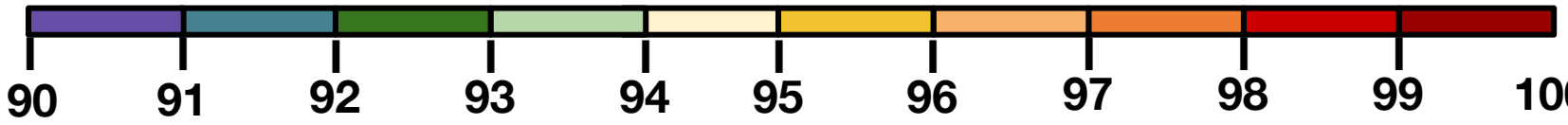


NIH NHGRI

Interpreting Self-Identity Plots

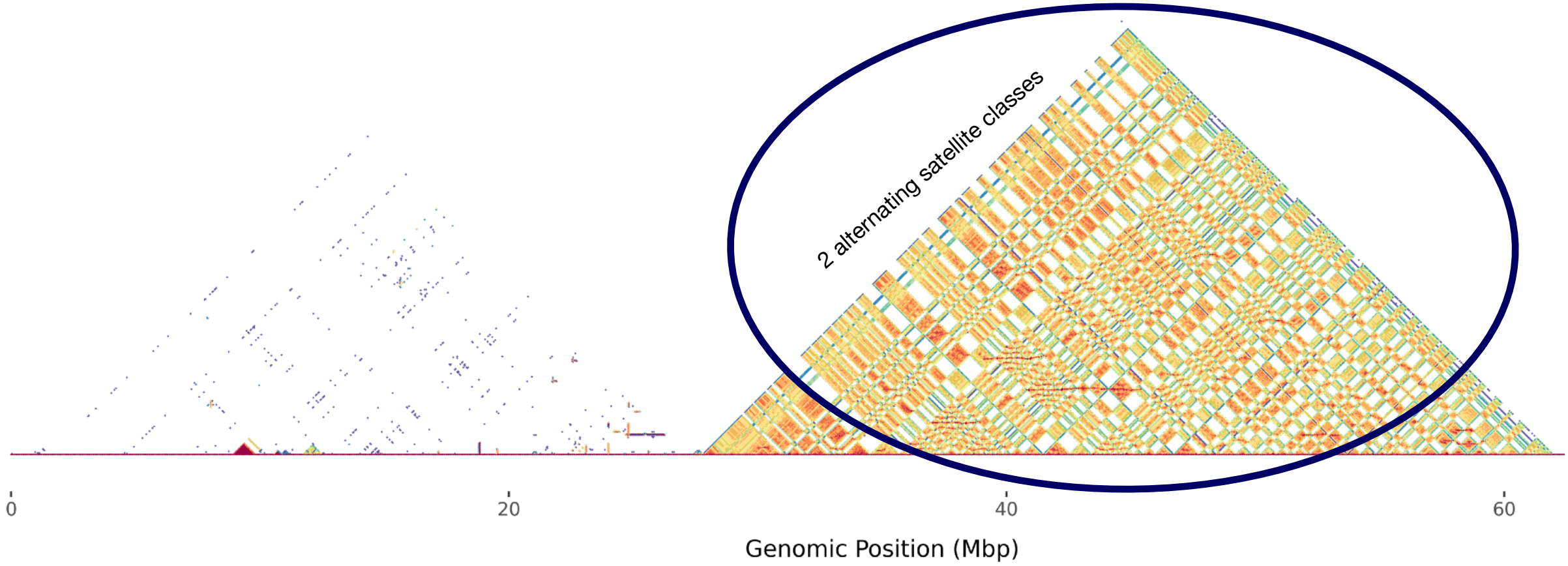


Average Nucleotide Identity

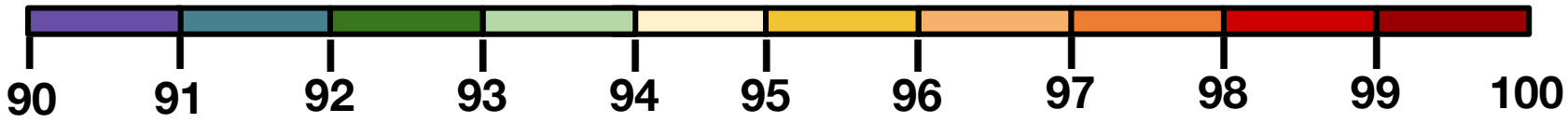


NIH NHGRI

Interpreting Self-Identity Plots

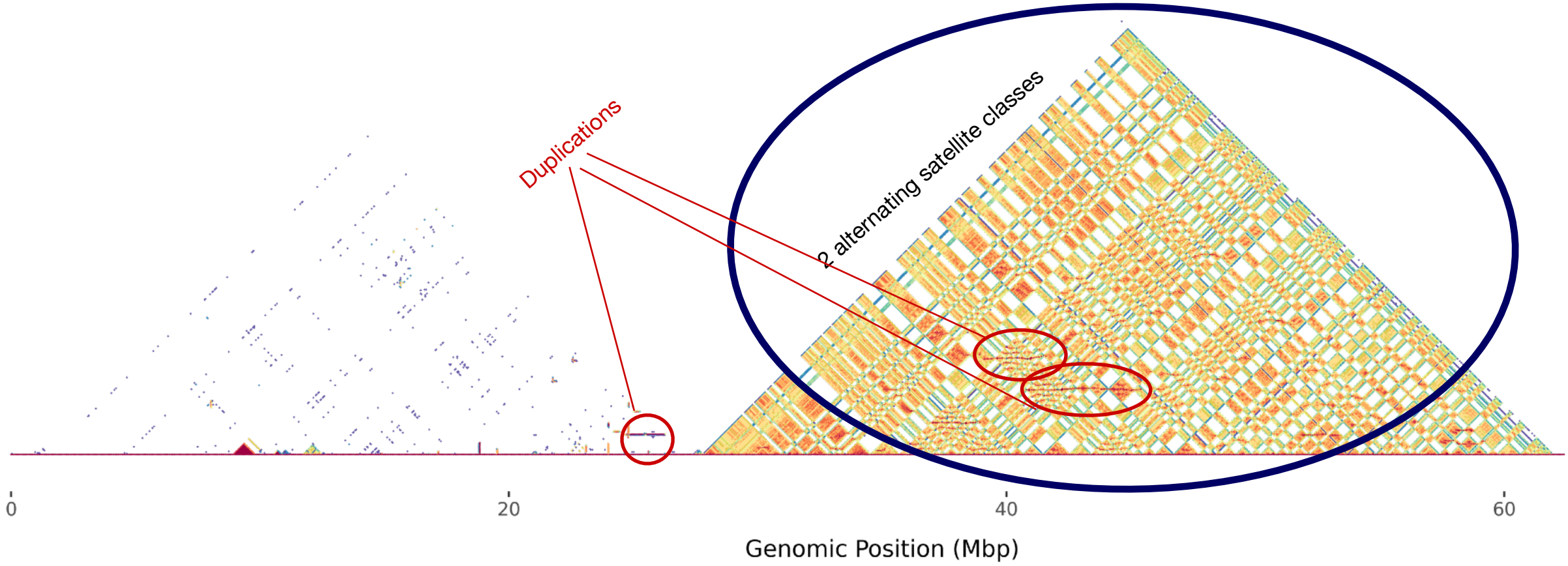


Average Nucleotide Identity

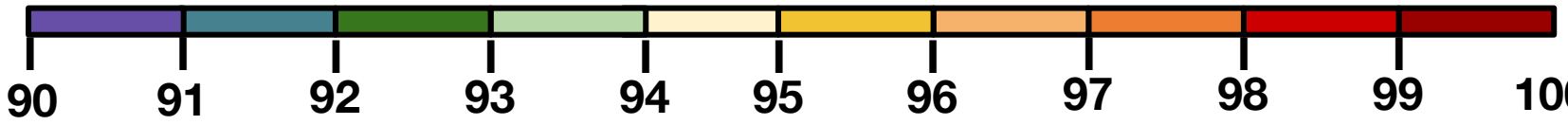


NIH NHGRI

Interpreting Self-Identity Plots

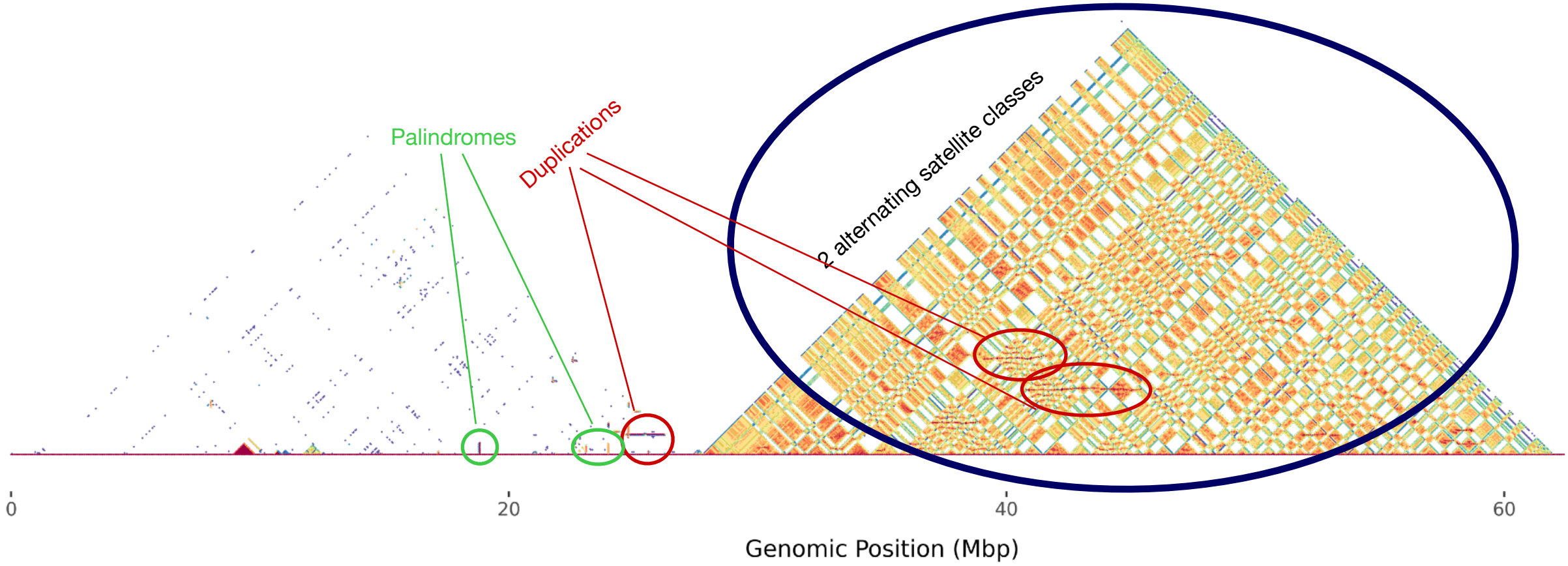


Average Nucleotide Identity

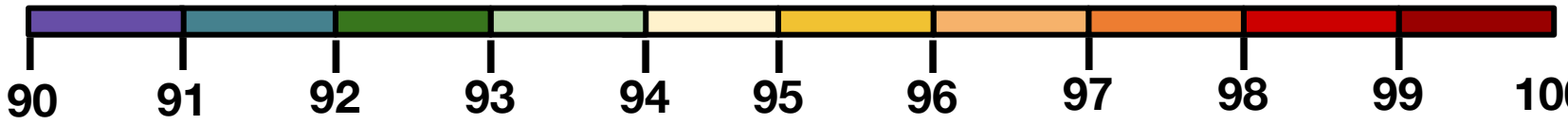


NIH NHGRI

Interpreting Self-Identity Plots

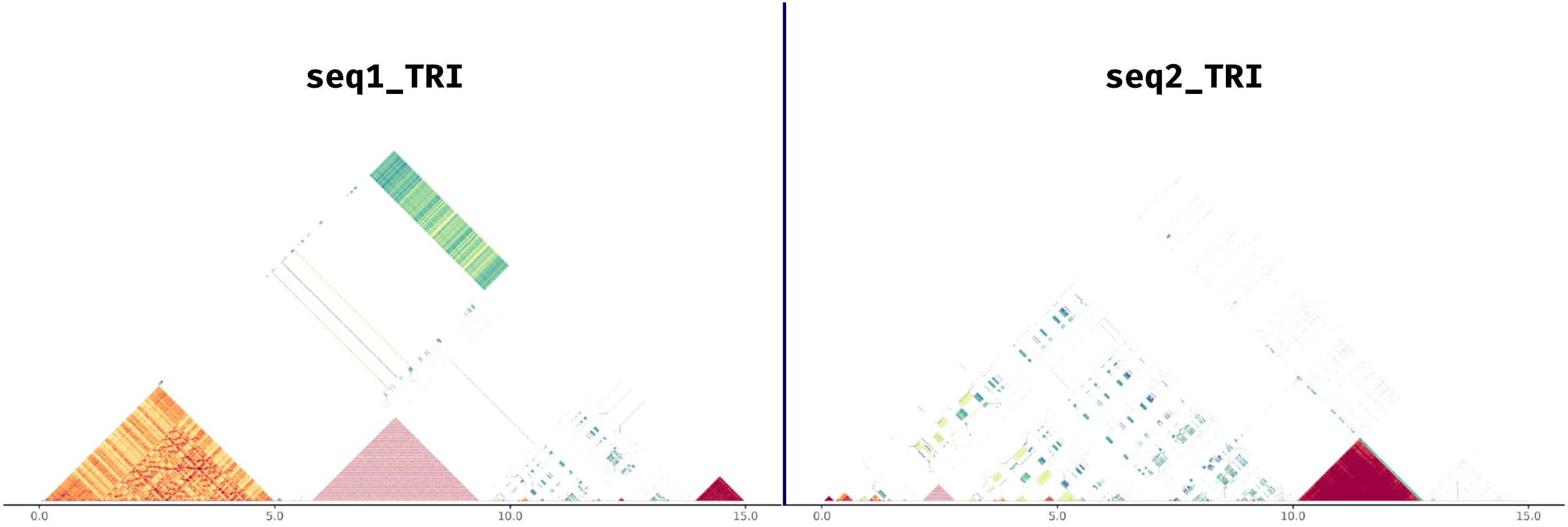


Average Nucleotide Identity



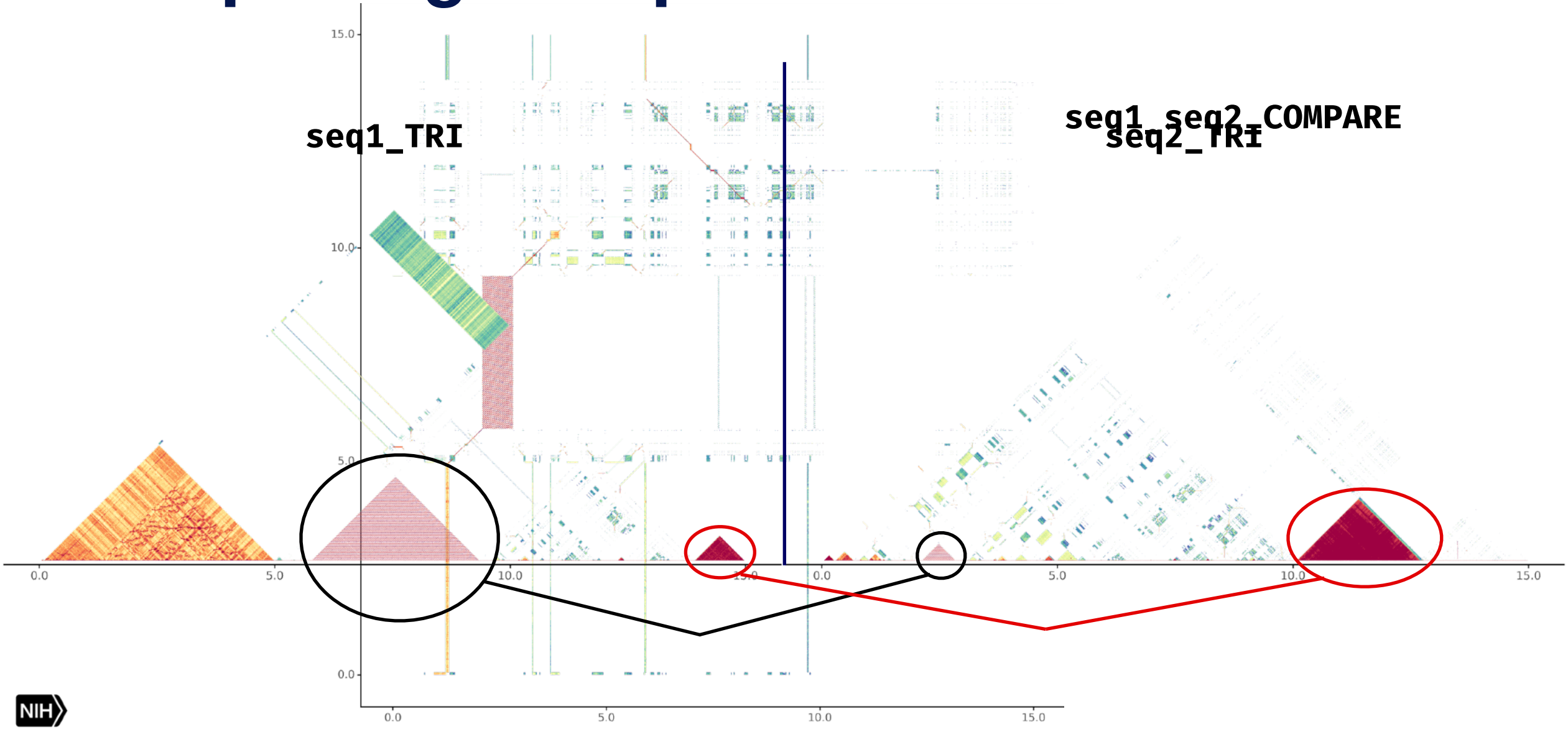
NIH NHGRI

Interpreting Comparative Plots



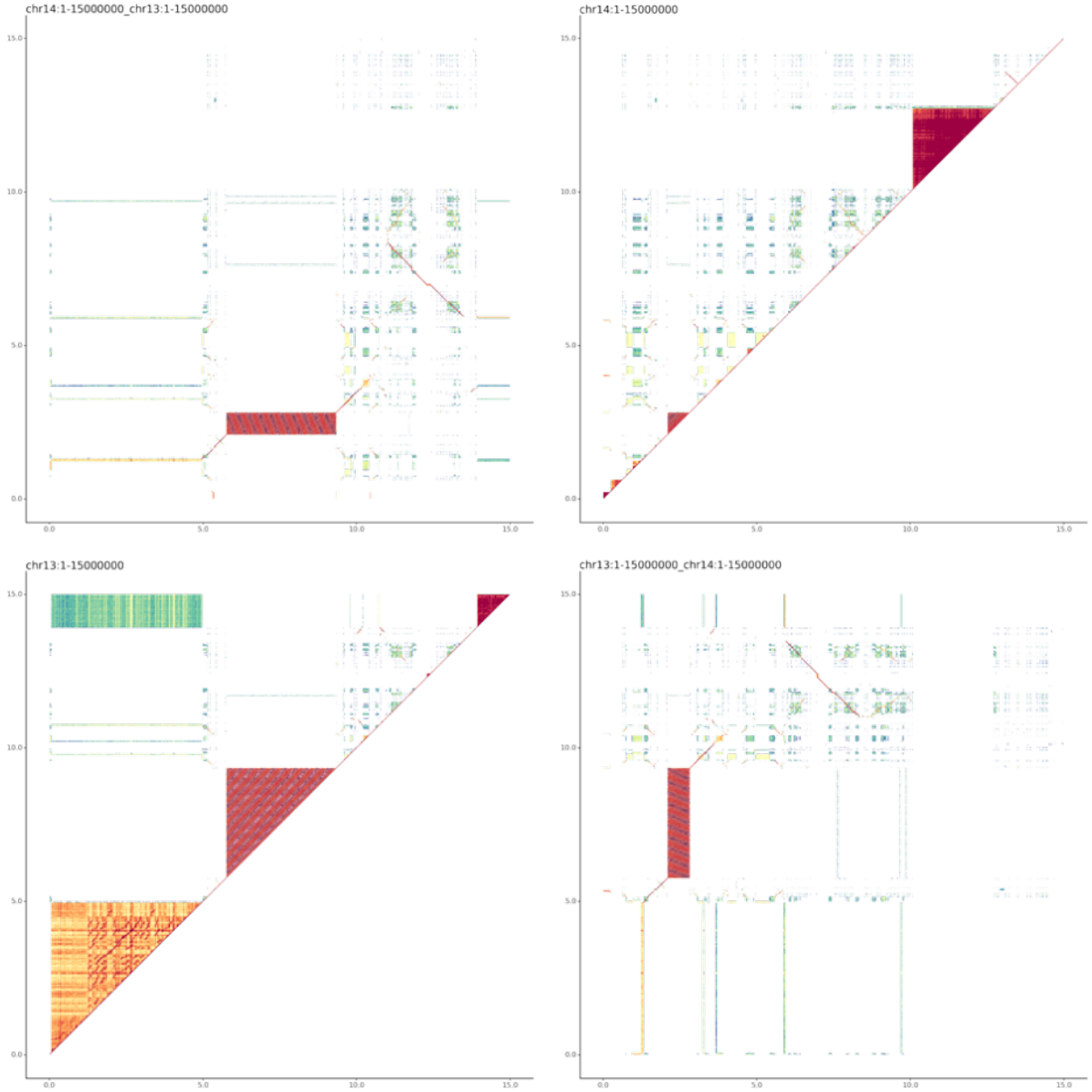
NIH
NHGRI

Interpreting Comparative Plots



NIH
NHGRI

Interpreting Comparative Plots

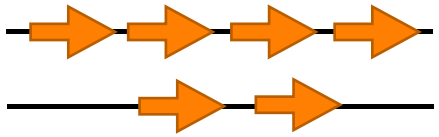


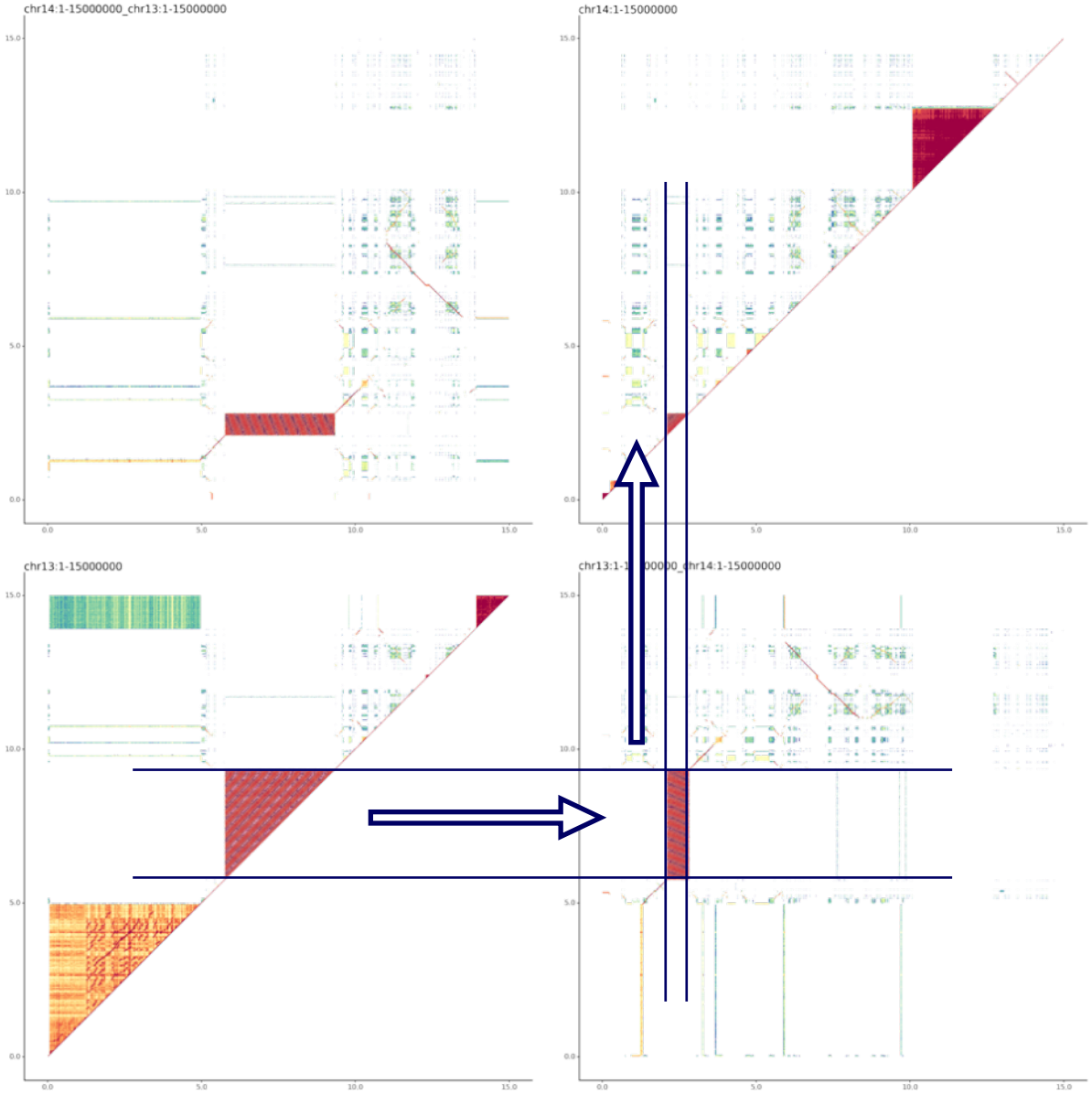
`2×2_GRID`

NIH NHGRI

Interpreting Comparative Plots

rDNA





2×2_GRID

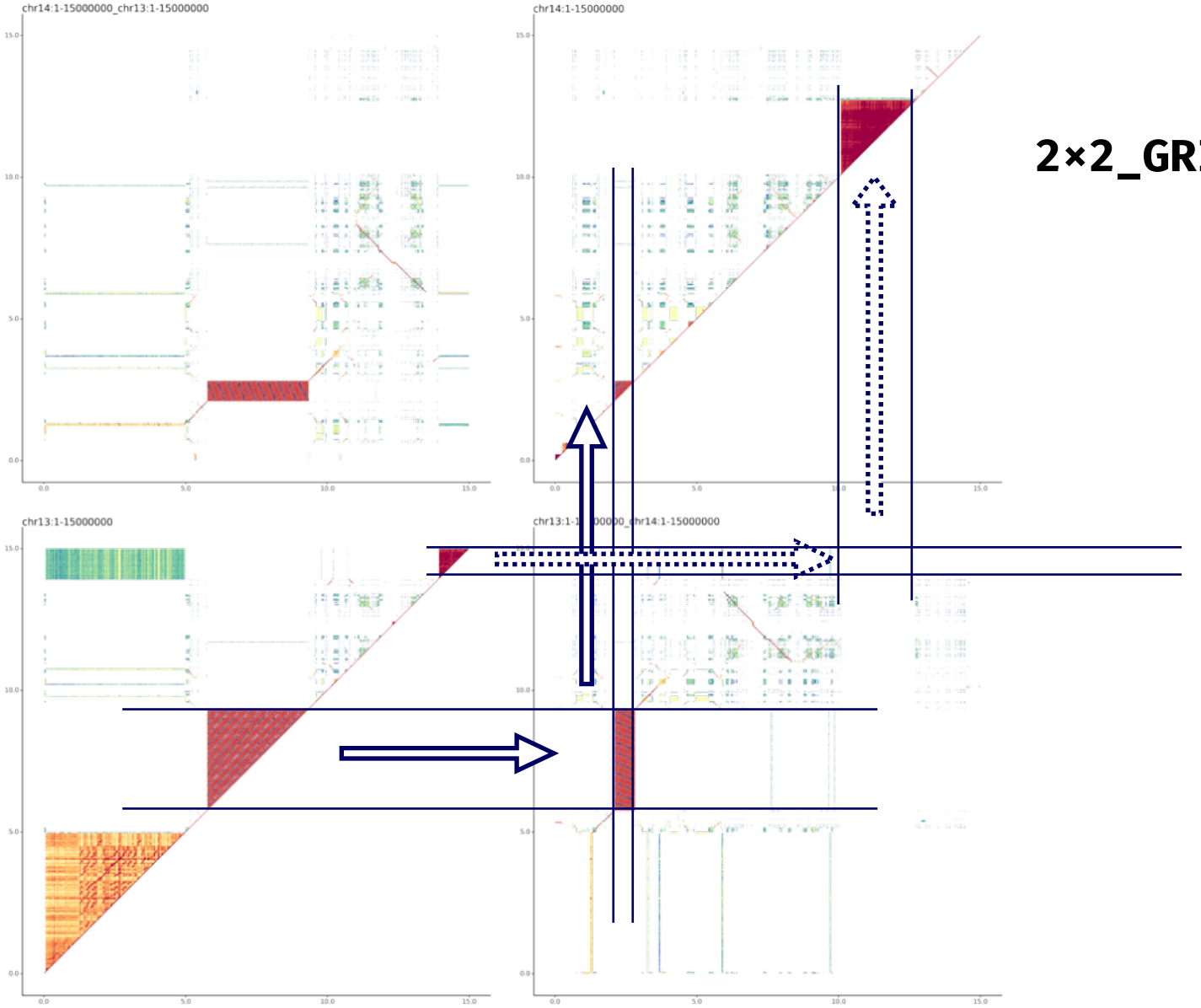
NIH NHGRI

Interpreting Comparative Plots

rDNA

hSat1

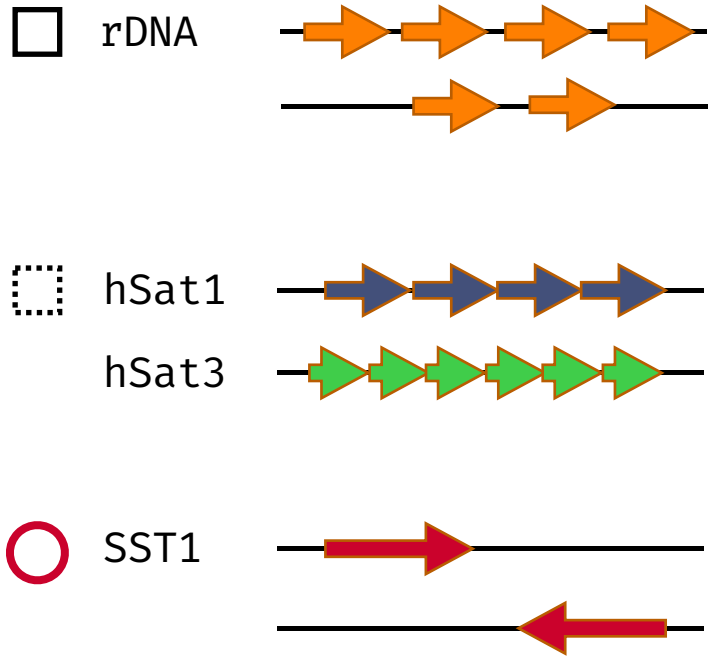
hSat3

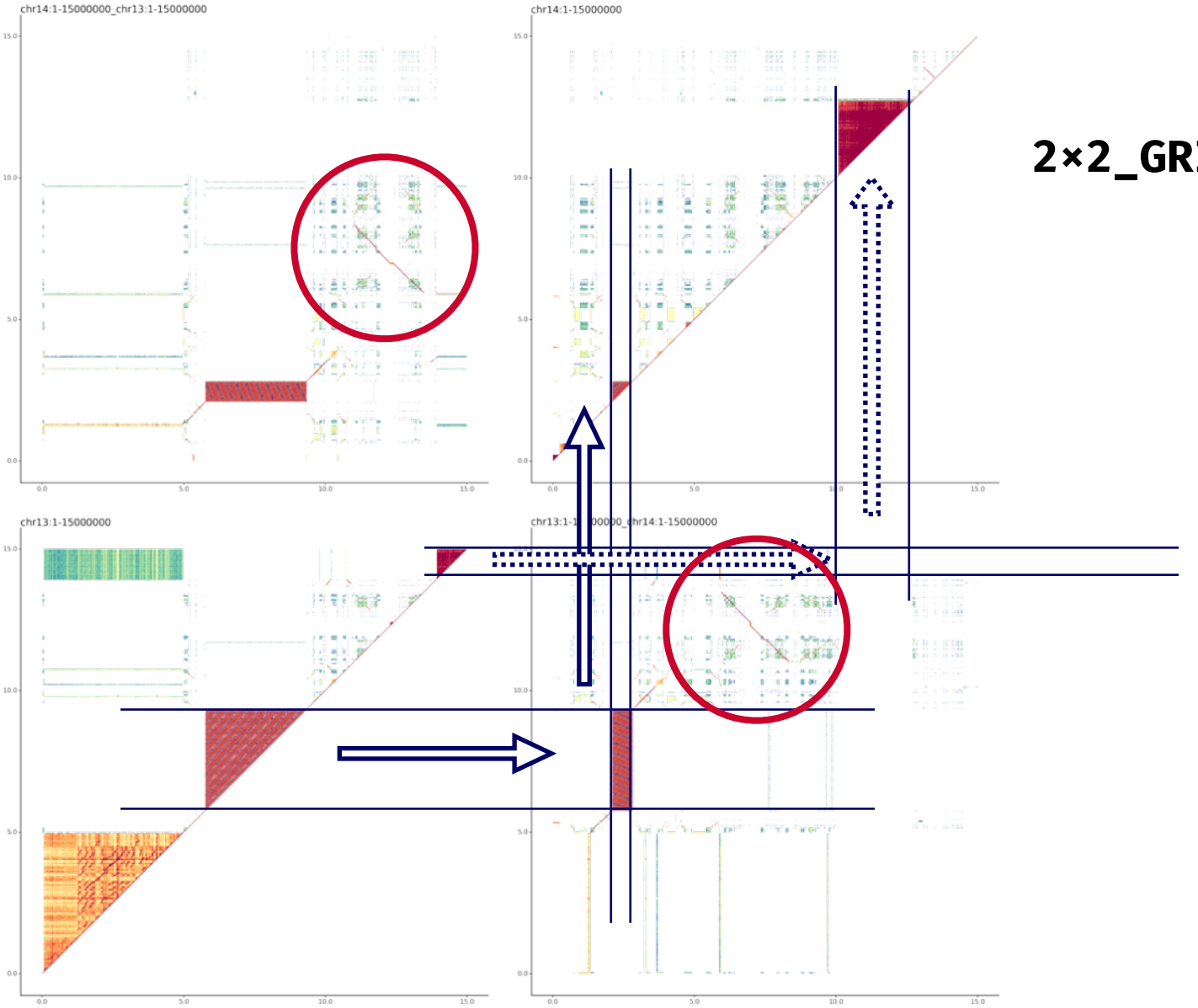


2×2_GRID

NIH NHGRI

Interpreting Comparative Plots





2×2_GRID

NIH NHGRI

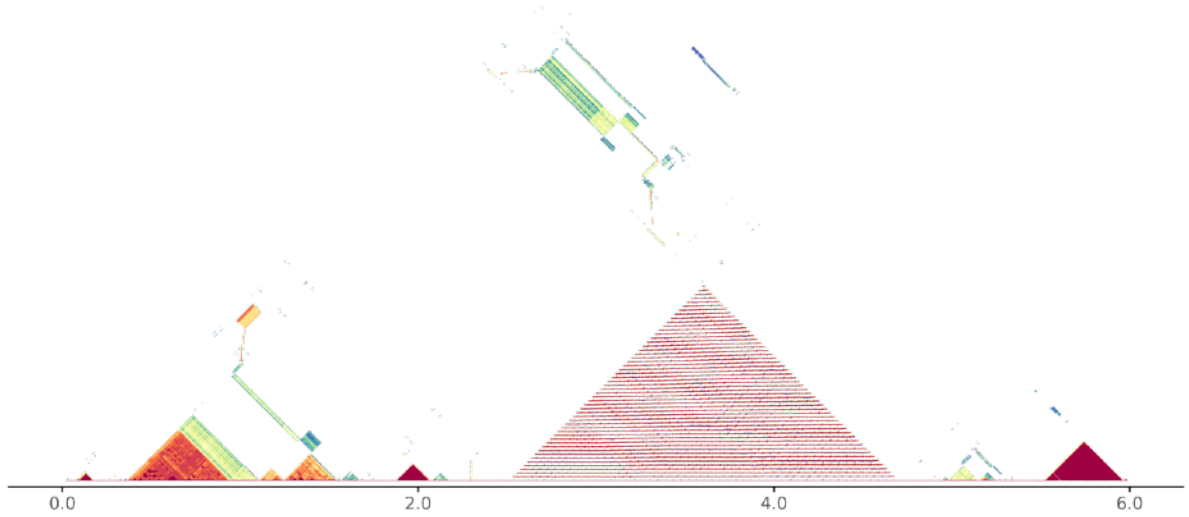
Running ModDotPlot

NIH
NHGRI

Running ModDotPlot

- ModDotPlot can be run in one of 2 modes:
- Static Mode:
  - Produces static image files
  - Good for publication quality figures
  - Fast!

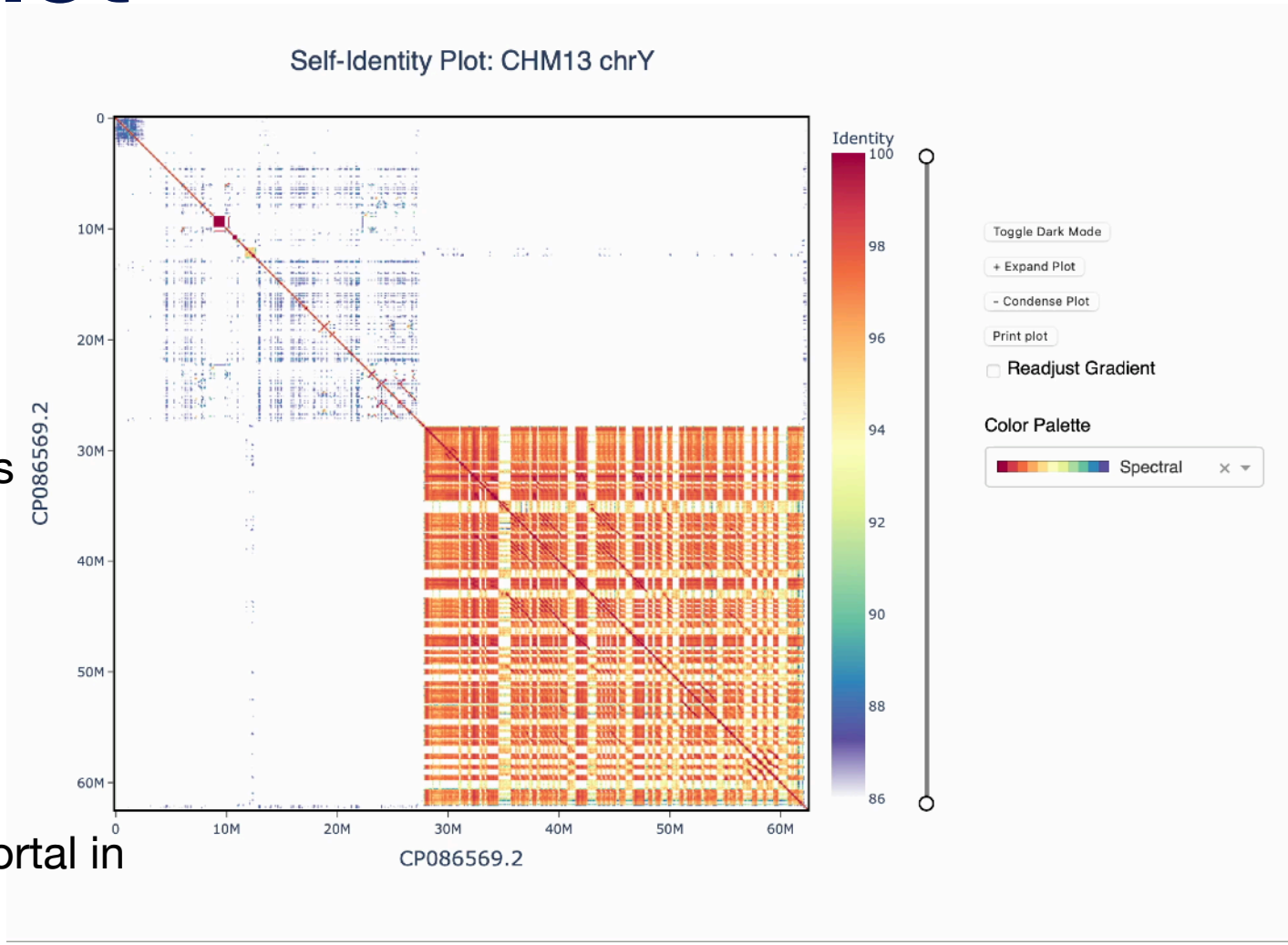
```
 __  __           _ ____        _   ____  _       _
|  \/  | ___   __| |  _ \  ___ | |_|  _ \| | ___ | |_
| |\/| |/ _ \ / _` | | | |/ _ \| __| |_) | |/ _ \| __|
| |  | | (_) | (_| | |_| | (_) | |_|  __/| | (_) | |_
|_|  |_|\___/ \__,_|____/ \___/ \__|_|   |_|\___/ \__|

v0.8.7

Running ModDotPlot in static mode

Retrieving k-mers from Chr15....

Progress: |████████████████████████████████████████| 100.0% Completed

Chr15 k-mers retrieved!
```



NIH NHGRI

Running ModDotPlot

- ModDotPlot can be run in one of 2 modes:
- Static Mode:
    - Produces static image files
    - Good for publication quality figures
    - Fast!
- Interactive Mode:
    - Builds a plot hierarchy.
    - Launches an interactive viewing portal in your web-browser.
    - Slower, but allows for resolution adjustment & exact coordinates.



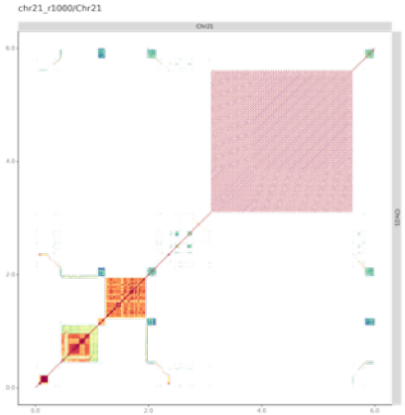
NIH NHGRI

Running ModDotPlot - Static Mode

- Parameters to be familiar with:
  - `-f/--fasta`: Input sequence. *Fasta format
  - `-b/--bed`: Input sequence. *Paired end bed format
  - `-o/--output-dir`: Output directory. Default: `./`

```
>Chr21_segment
AAACCCGTAACCCTAACCCGTAACCCTAACCGTCTAACCCTAACCTAACCCTAACCCTAACCCTAAACCC
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
CCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA
```

```
#query_name query_start query_end   reference_name  reference_start reference_end
Chr21   1   6000    Chr21   1   6000    100.0
Chr21   1   6000    Chr21   6001    12000   99.74286819318638
Chr21   1   6000    Chr21   372001  378000  86.91744899102561
Chr21   1   6000    Chr21   378001  384000  86.91744899102561
Chr21   1   6000    Chr21   2196001 2202000 89.83420588379086
Chr21   1   6000    Chr21   2202001 2208000 89.83420588379086
Chr21   1   6000    Chr21   2592001 2598000 86.91744899102561
Chr21   1   6000    Chr21   2598001 2604000 86.91744899102561
Chr21   1   6000    Chr21   3066001 3072000 88.0433911418159
Chr21   1   6000    Chr21   3072001 3078000 86.91744899102561
Chr21   1   6000    Chr21   5760001 5766000 86.91744899102561
```



fasta → **bed** → **plots**

NIH NHGRI

Running ModDotPlot - Static Mode

- Parameters to be familiar with:
    - `n`: Sequence size. *Not a parameter you can control through ModDotPlot!
    - `-r/--resolution`: Number of intervals (plot resolution). Default: 1000.
    - `-w/--window`: Interval length. *Note this is inversely proportional to resolution!

Running ModDotPlot - Static Mode

- Parameters to be familiar with:
    - `n`: Sequence size. *Not a parameter you can control through ModDotPlot!
    - `-r/--resolution`: Number of intervals (plot resolution). Default: 1000.
    - `-w/--window`: Interval length. *Note this is inversely proportional to resolution!

Runtime quadratically scales with ***r*** *(*not ***n*****!**)

NIH
NHGRI

Running ModDotPlot - Static Mode

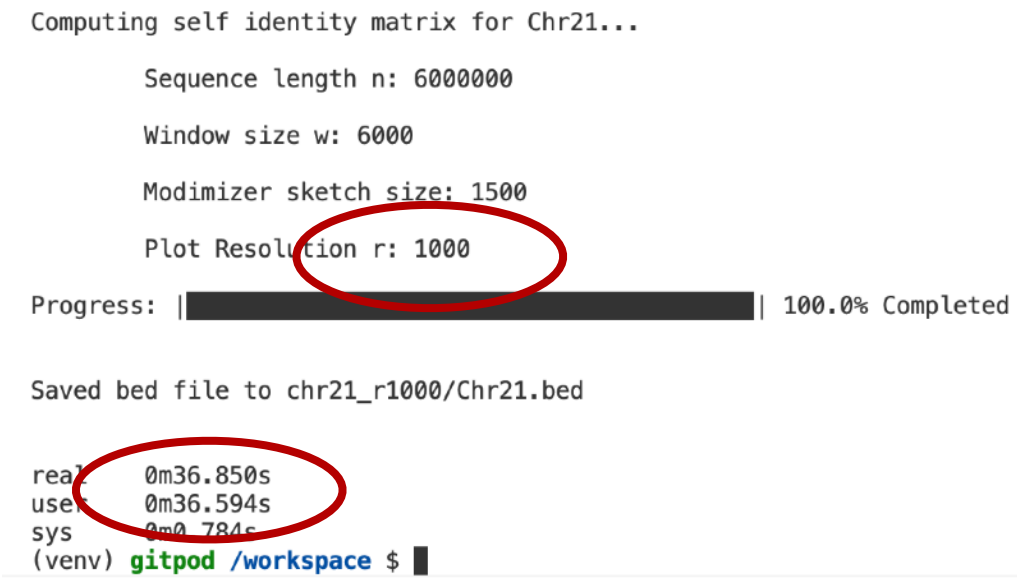
- Parameters to be familiar with:
    - `n`: Sequence size. *Not a parameter you can control through ModDotPlot!
    - `-r/--resolution`: Number of intervals (plot resolution). Default: 1000.
    - `-w/--window`: Interval length. *Note this is inversely proportional to resolution!

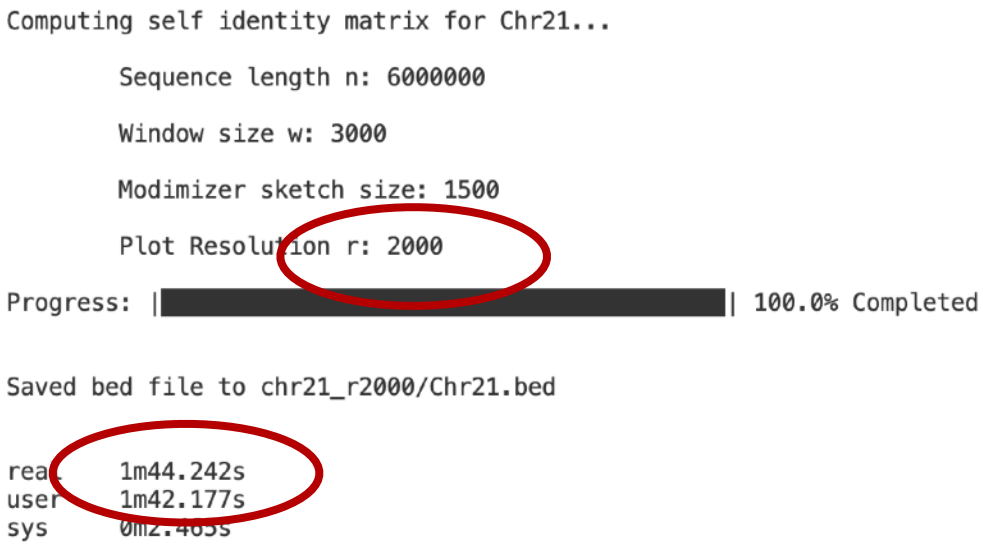
```
Computing self identity matrix for Chr21...
        Sequence length n: 6000000
        Window size w: 6000
        Modimizer sketch size: 1500
        Plot Resolution r: 1000
Progress: |████████████████████████████████████████| 100.0% Completed

Saved bed file to chr21_r1000/Chr21.bed

real    0m36.850s
user    0m36.594s
sys     0m0.784s
(venv) gitpod /workspace $
```

NIH NHGRI

Running ModDotPlot - Static Mode

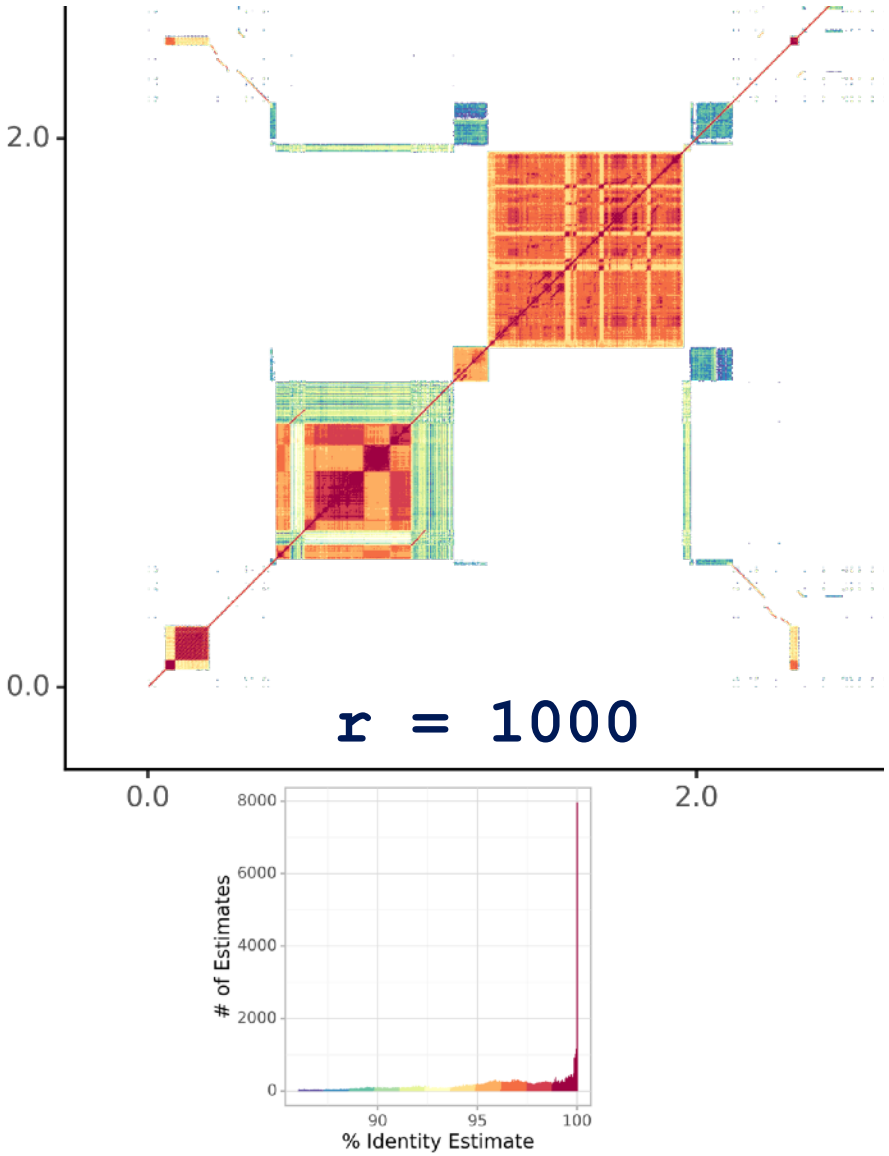
- Parameters to be familiar with:
    - `n`: Sequence size. *Not a parameter you can control through ModDotPlot!
    - `-r/--resolution`: Number of intervals (plot resolution). Default: 1000.
    - `-w/--window`: Interval length. *Note this is inversely proportional to resolution!

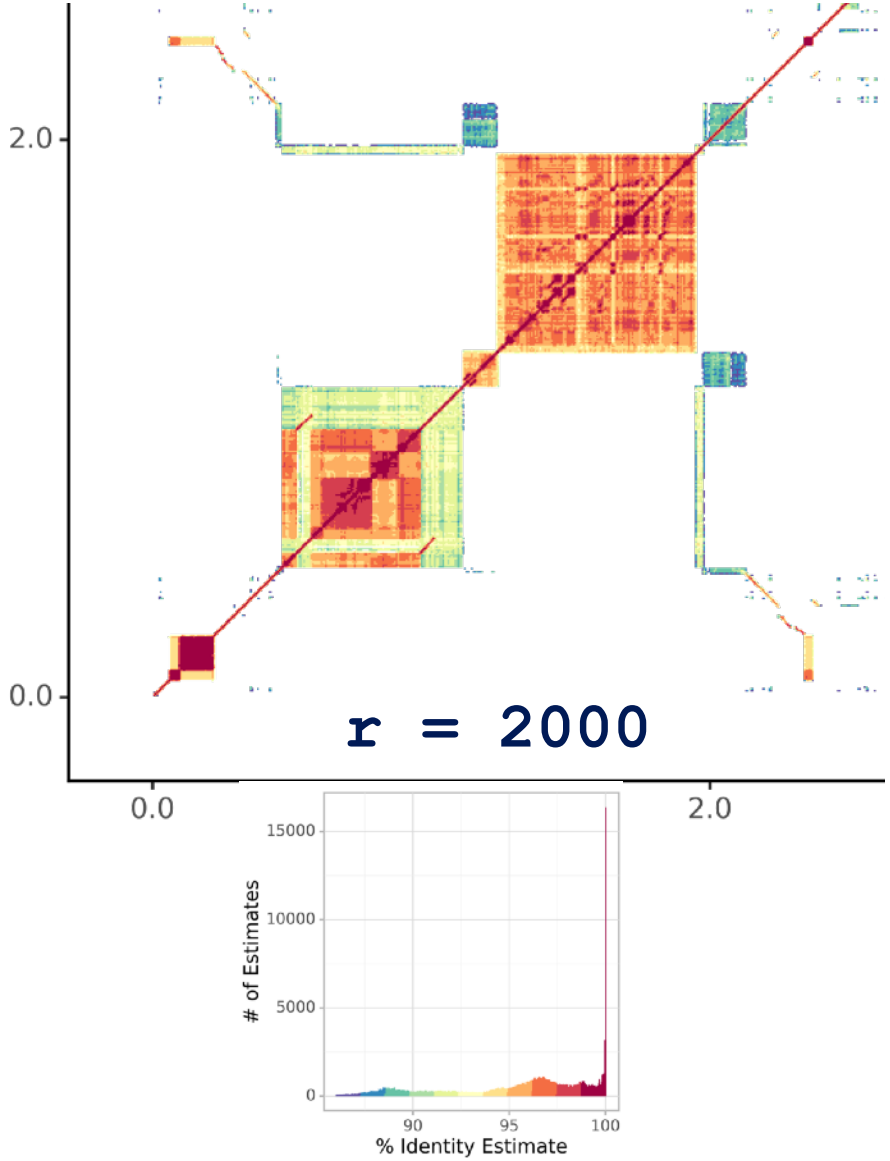




NIH NHGRI

Running ModDotPlot - Static Mode





NIH
NHGRI

Running ModDotPlot - Static Mode

- Other parameters:
    - `-k/--kmer`: *k*-mer length. Default: 21
    - `--compare`: Will produce a comparative plot if 2+ fasta/bed files input
    - `--grid`: Produce an nxn grid, used with compare. *Limited to 6 sequences!

```
Computing pairwise identity matrix for Chr15 and chr1:14M-18M...
        Sequence length Chr15: 6000000
        Sequence length chr1:14M-18M: 4000000
        Window size w: 6000
        Modimizer sketch size: 1500
        Plot Resolution r: 1000
Progress: |████████████████████████████████████████| 100.0% Completed
```

NIH NHGRI

Running ModDotPlot - Interactive Mode

- Most parameters stay the same! Different ones:
    - `--port`: Port number to show Dash on. Default: 8050
    - `--save`: Save matrices to folder (`-o`)
    - `--load`: Load a save matrix hierarchy.

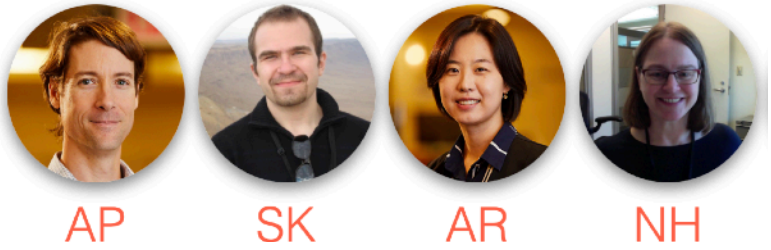
Acknowledgements

NIH NHGRI

JOHNS HOPKINS UNIVERSITY

BGA

Phillippy Lab - NHGRI



Schatz Lab - JHU

BGA24

Thanks Damon!

NIH NHGRI