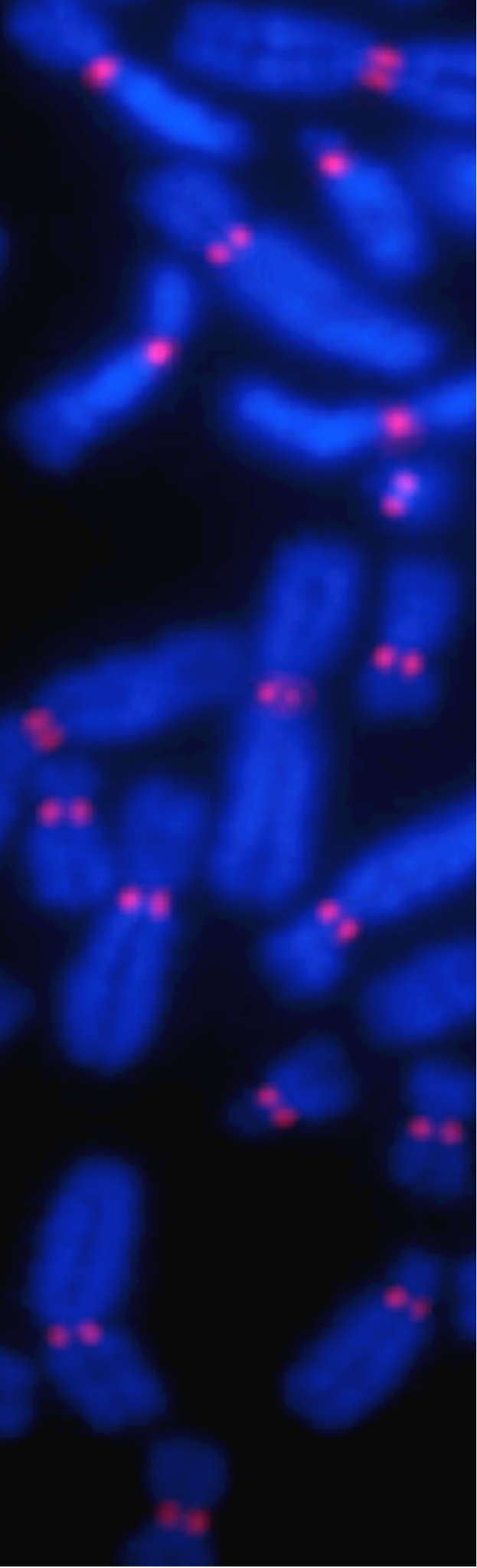
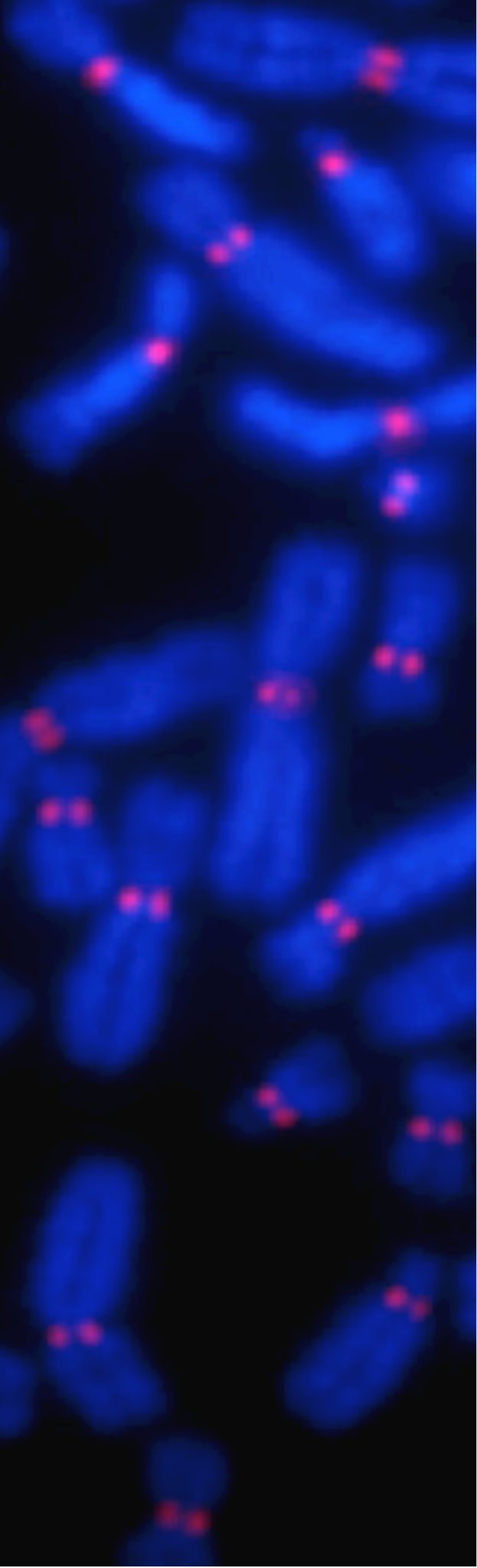


GRCh38 Centromere Reference Models

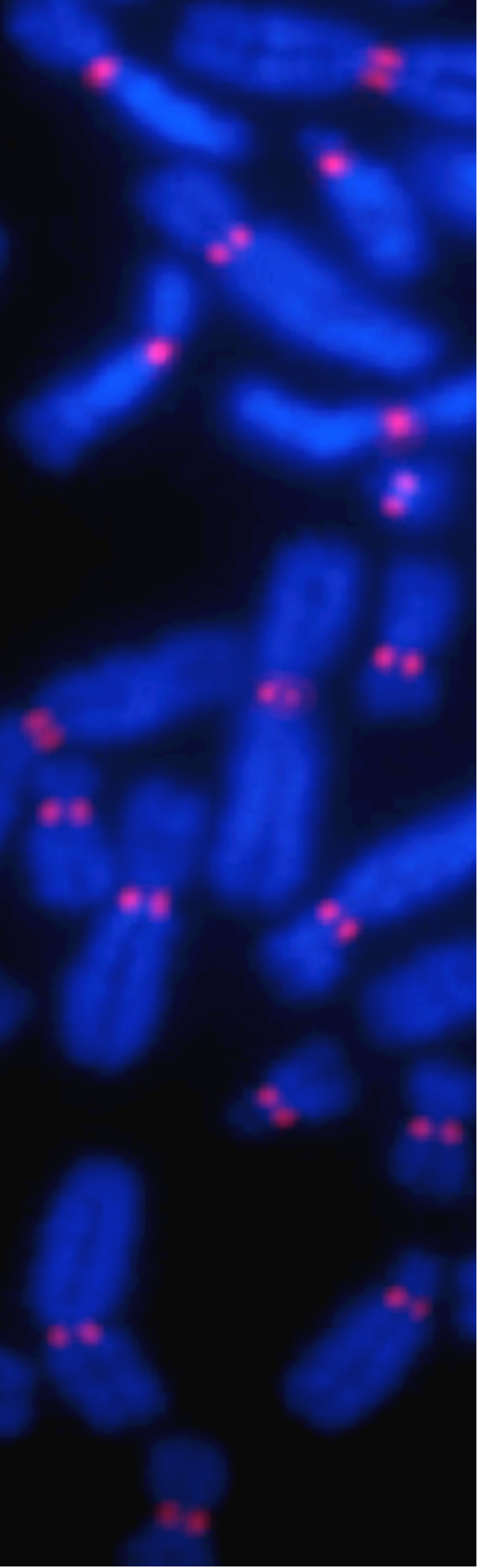
Karen H. Miga
Genome Browser Team Meeting
5/19/14



I. 'Centromere' vs. 'Centromere Gap'

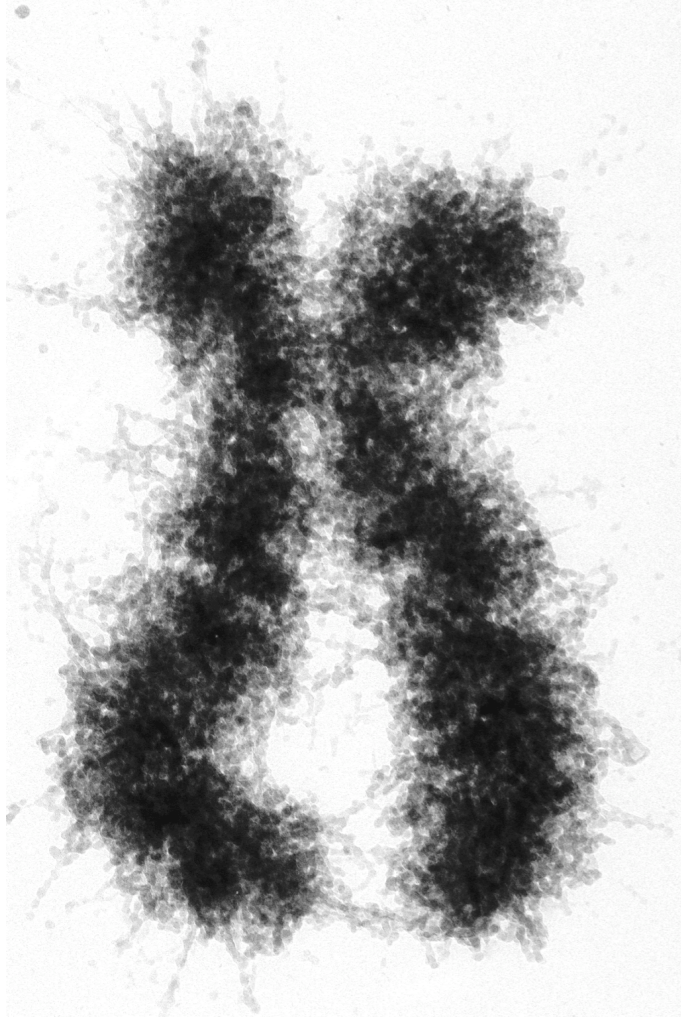


1. 'Centromere' vs. 'Centromere Gap'
2. LinearSat: Satellite reference models vs. Standard Assembly



1. 'Centromere' vs. 'Centromere Gap'
2. LinearSat: Satellite reference models vs. Standard Assembly
3. Sequence annotations that may be useful to include to guide analysis in these regions

'Centromere' vs. 'Centromere Gap'



Package DNA

Segregate DNA

'Centromere' vs. 'Centromere Gap'

Chromatids

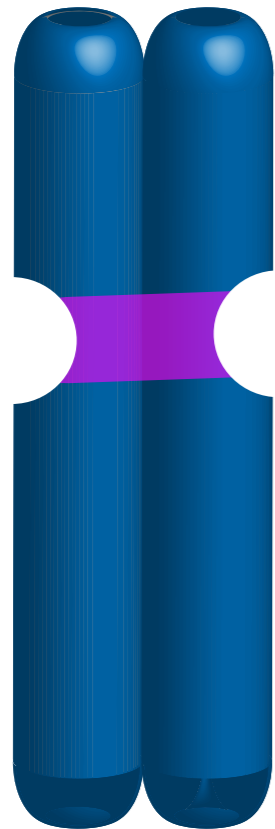


Package DNA

Segregate DNA

'Centromere' vs. 'Centromere Gap'

Chromatids

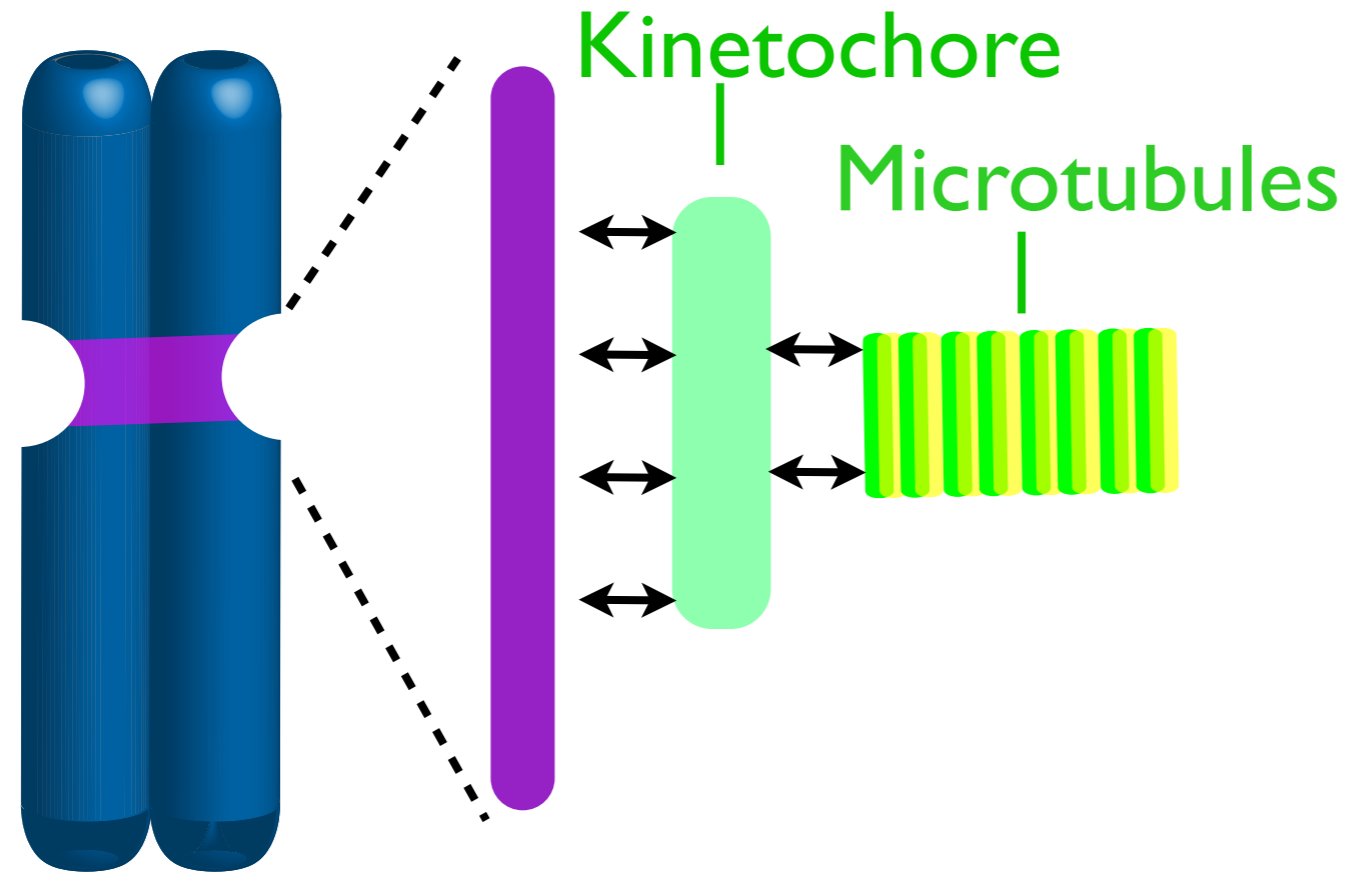


— Centromere

Package DNA

Segregate DNA

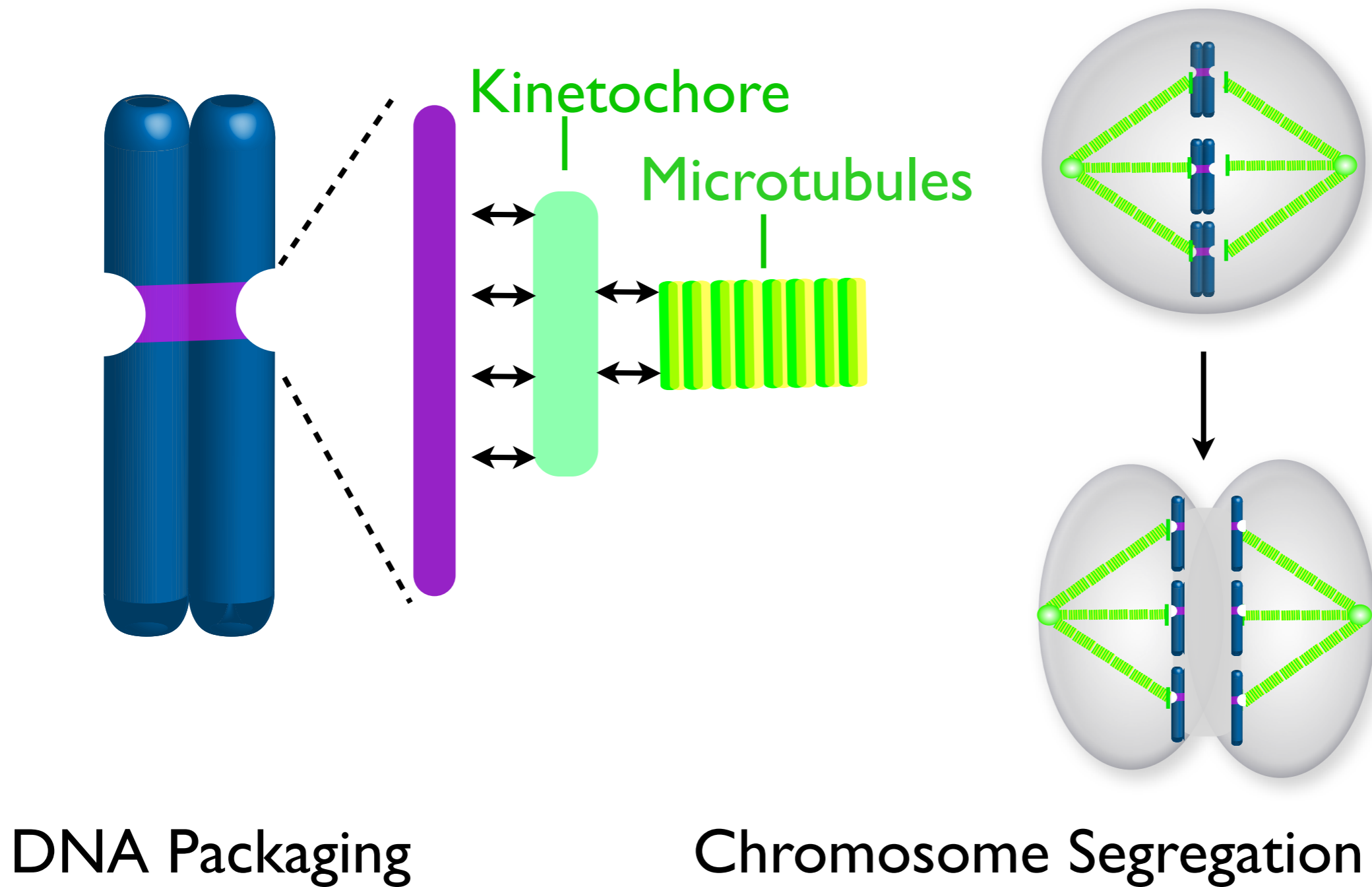
Stable Genome Inheritance



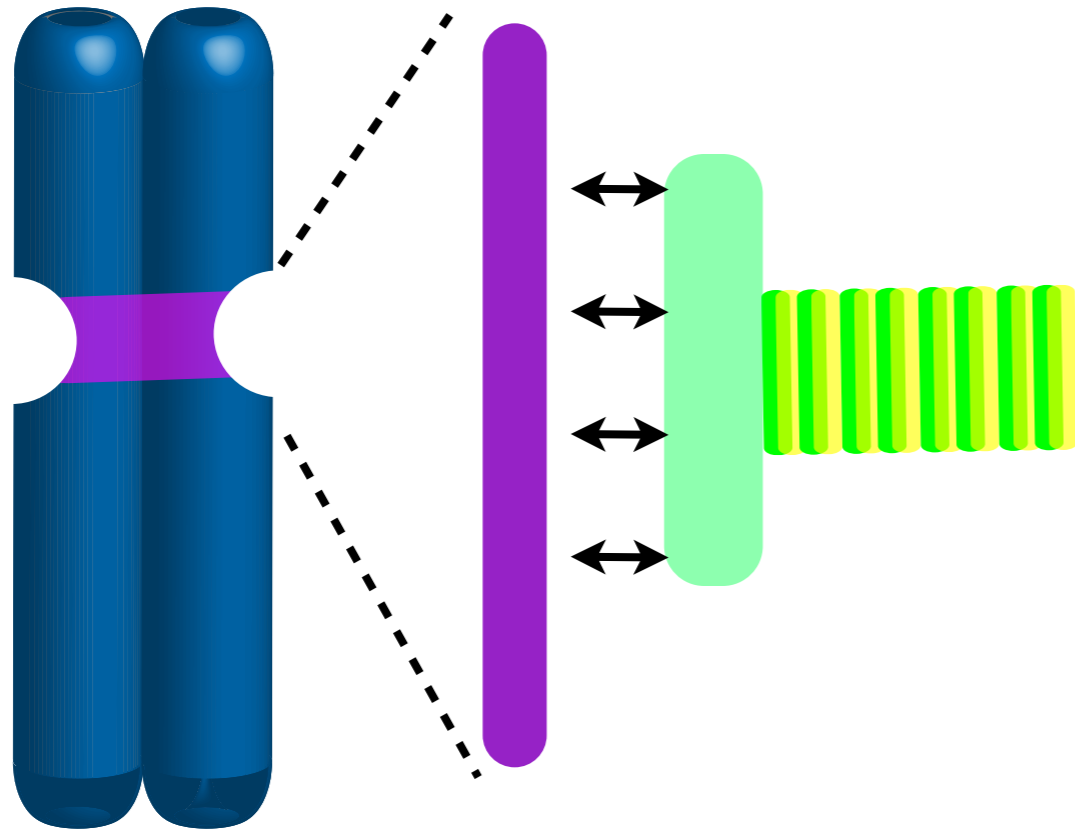
DNA Packaging

Chromosome Segregation

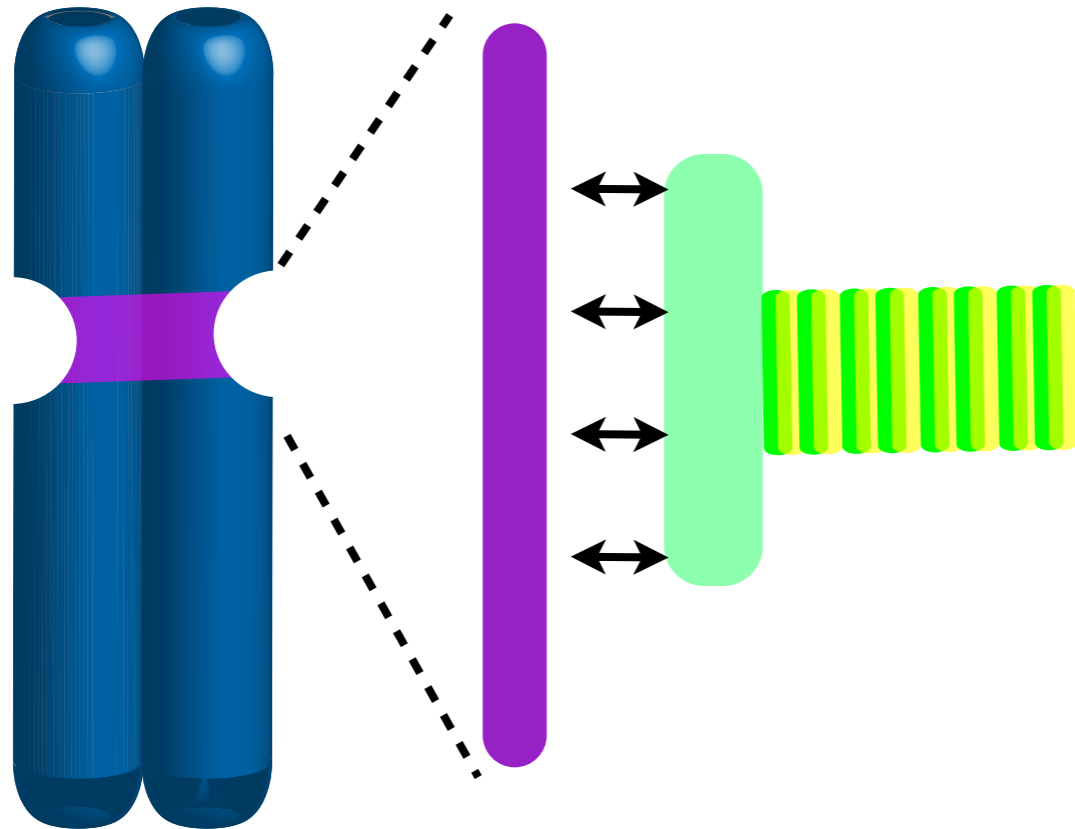
Stable Genome Inheritance



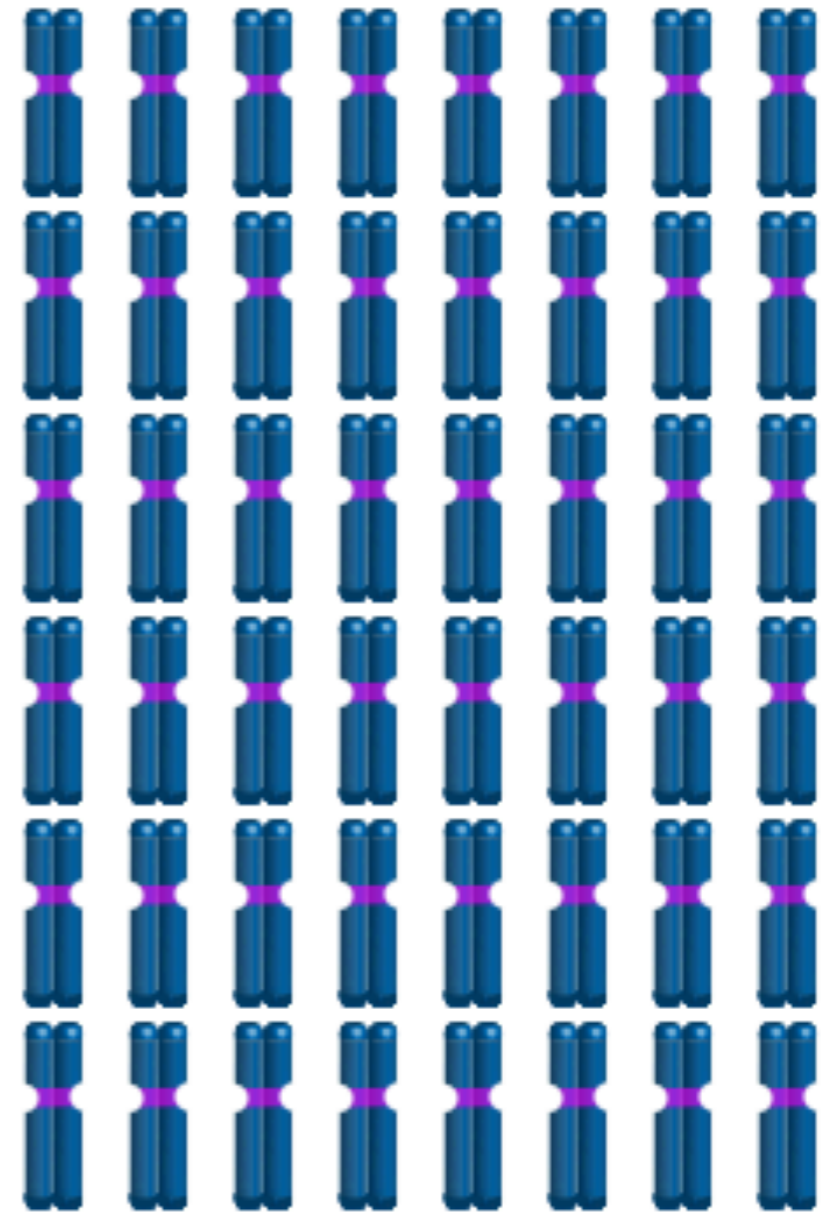
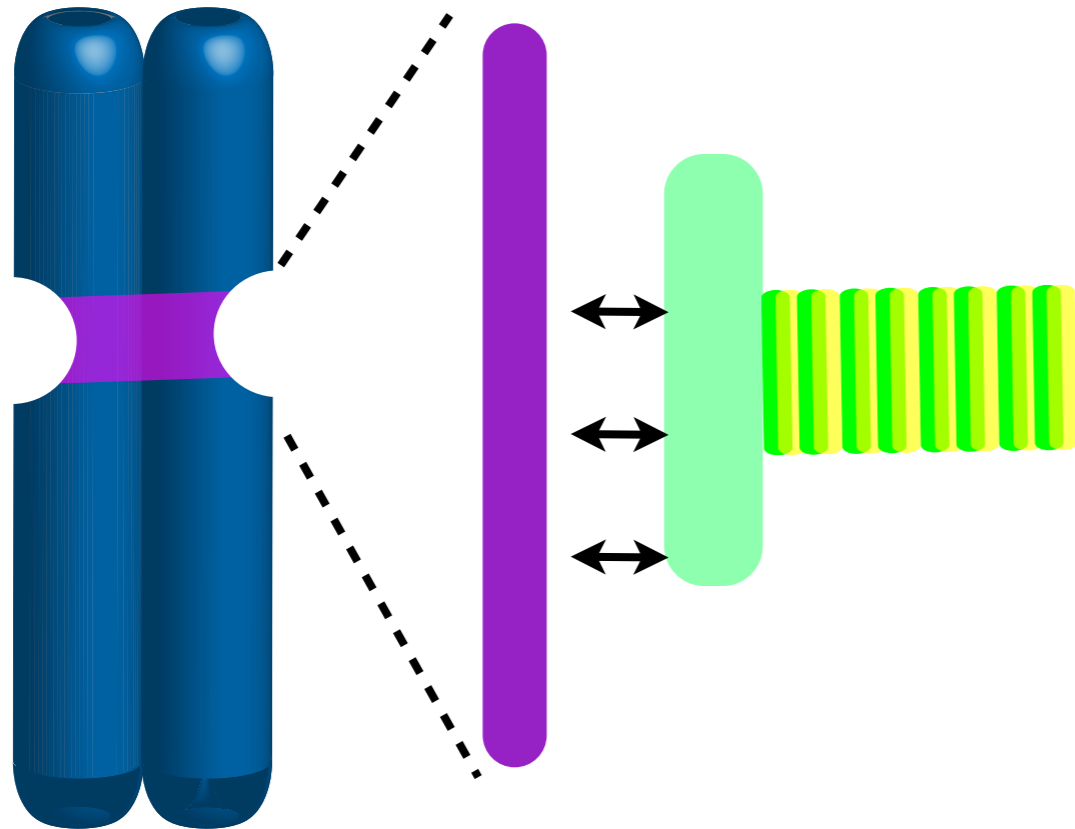
Stable Centromere Position



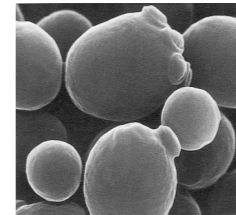
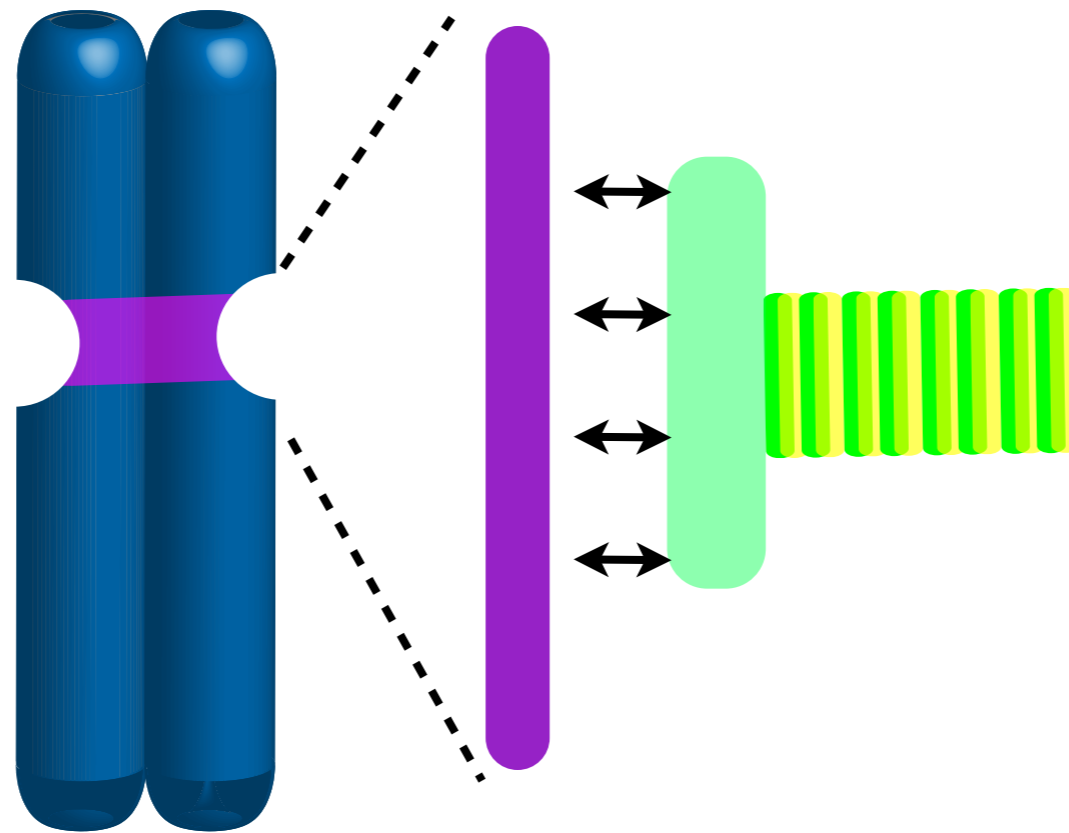
Stable Centromere Position



Stable Centromere Position



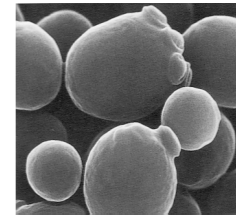
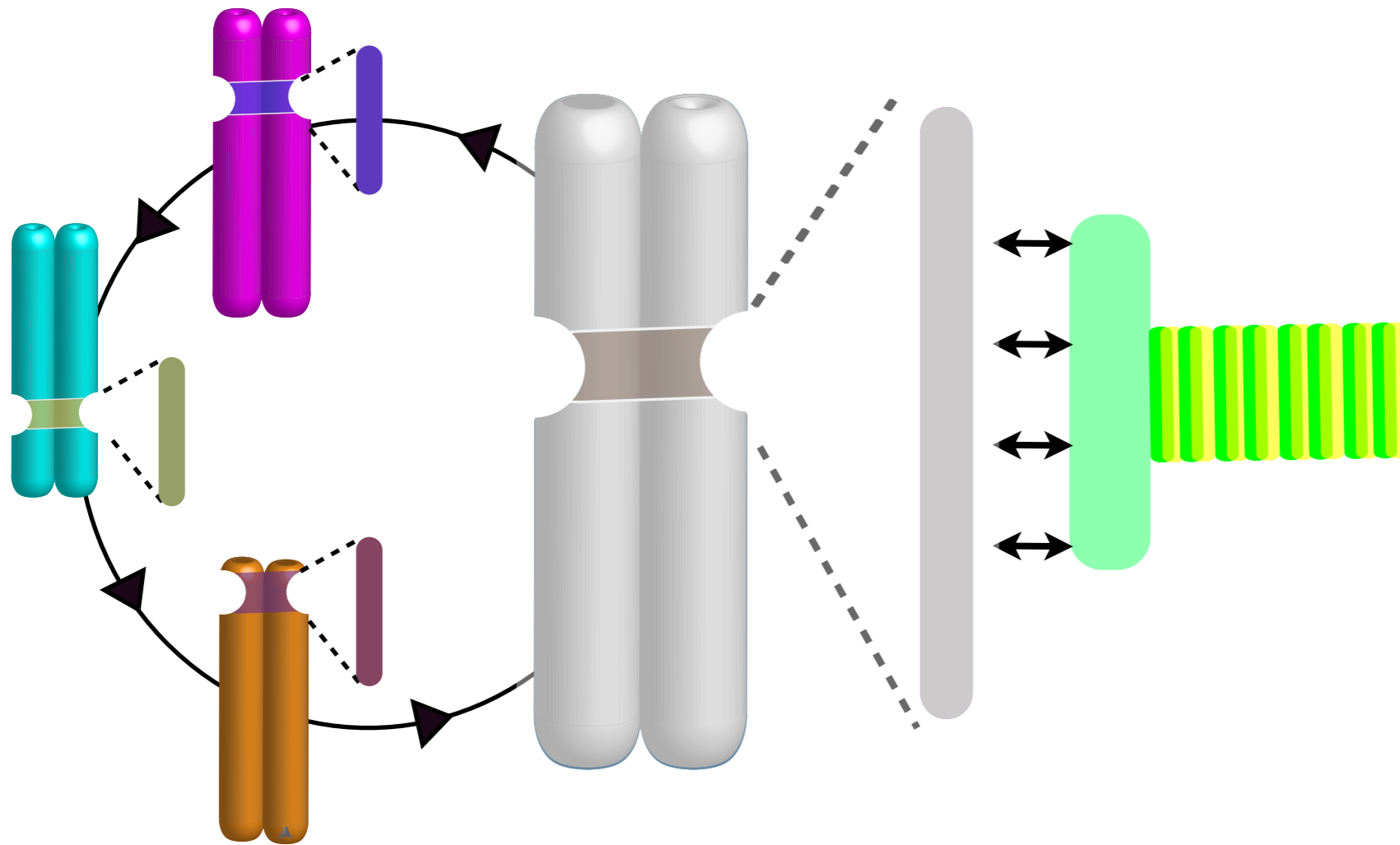
Centromeres Over Vast Evolutionary Time



GENOME

PROTEINS

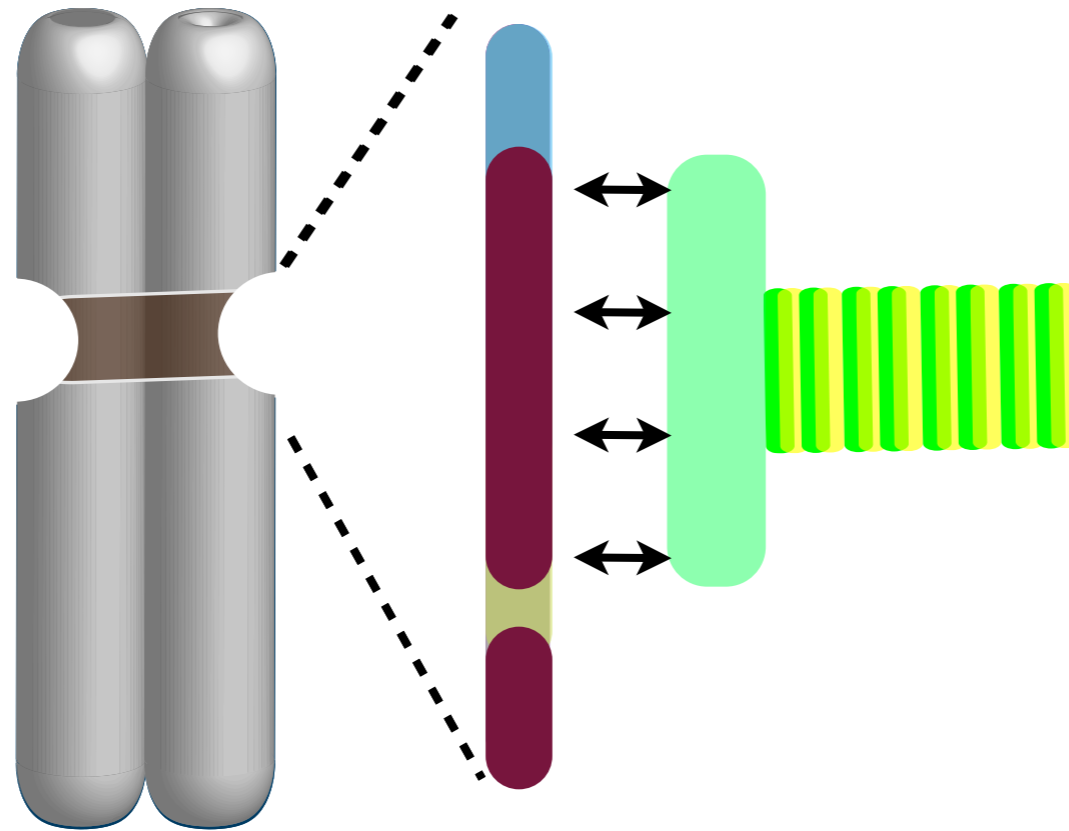
Centromeres Over Vast Evolutionary Time






GENOME

PROTEINS

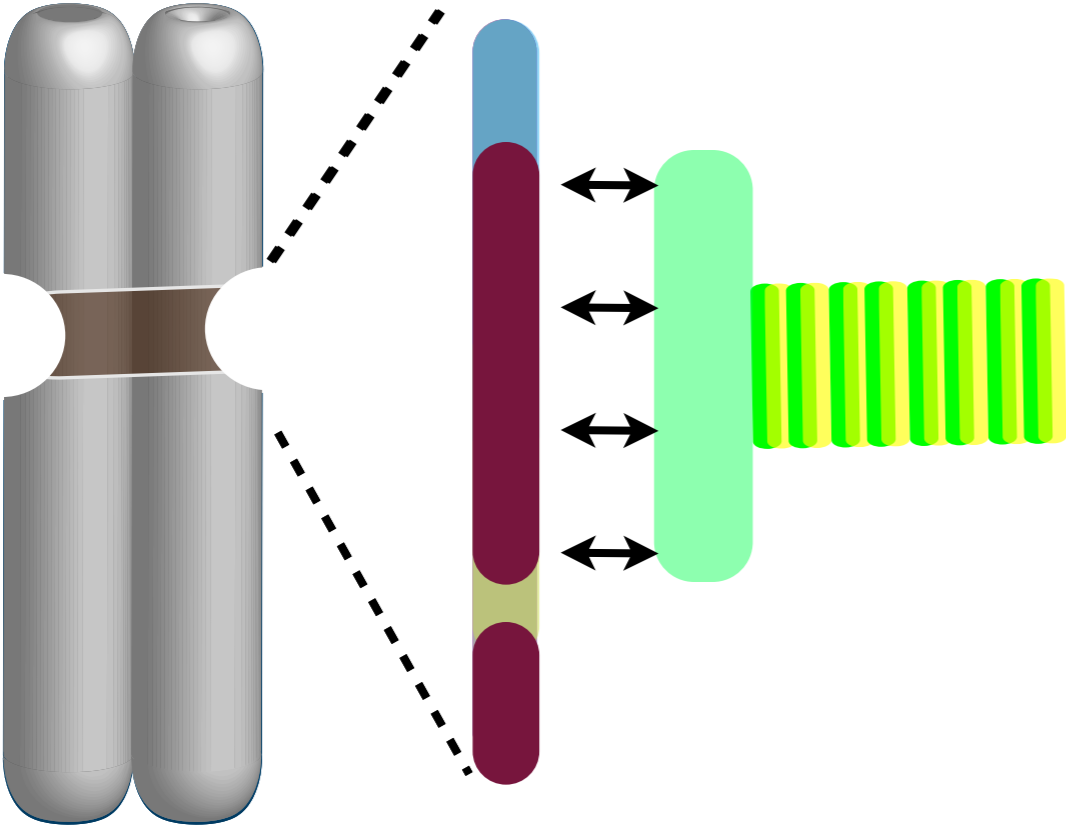
Centromeric regions can be defined by a collection of diverse sequences



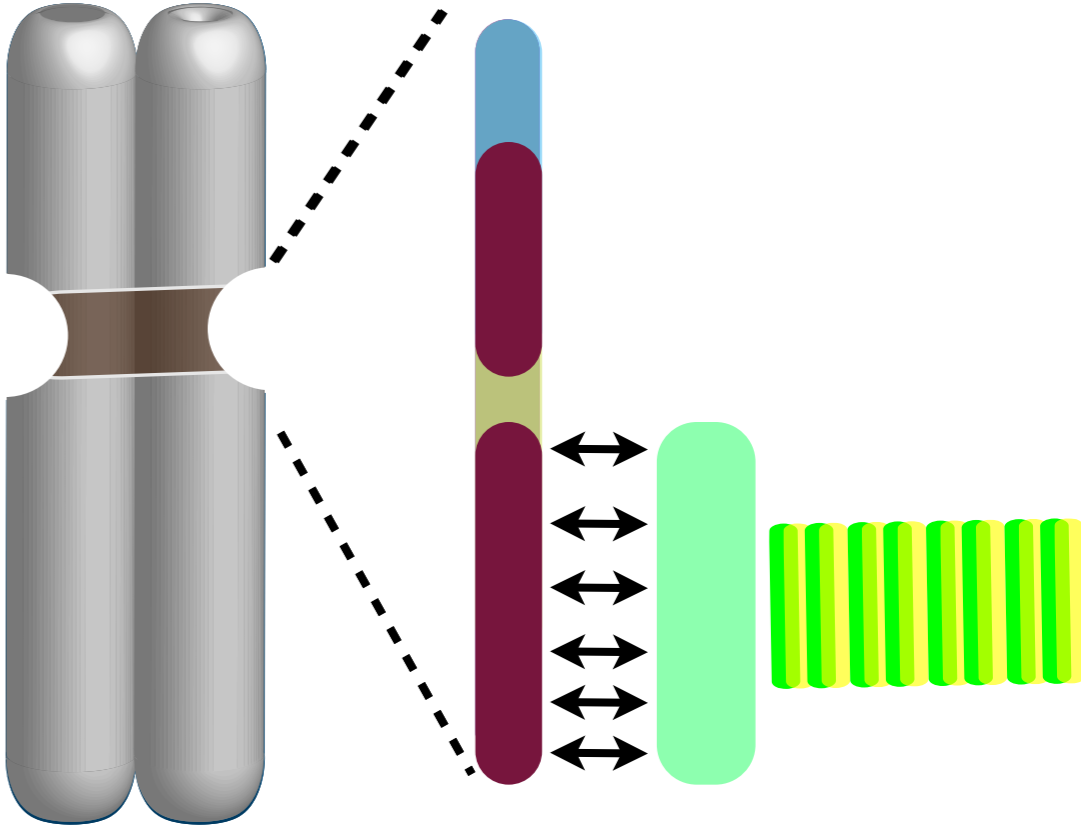
-  Alpha Satellite
-  Human Satellites (2,3)
-  Segmental Duplications

Centromeres are epigenetic

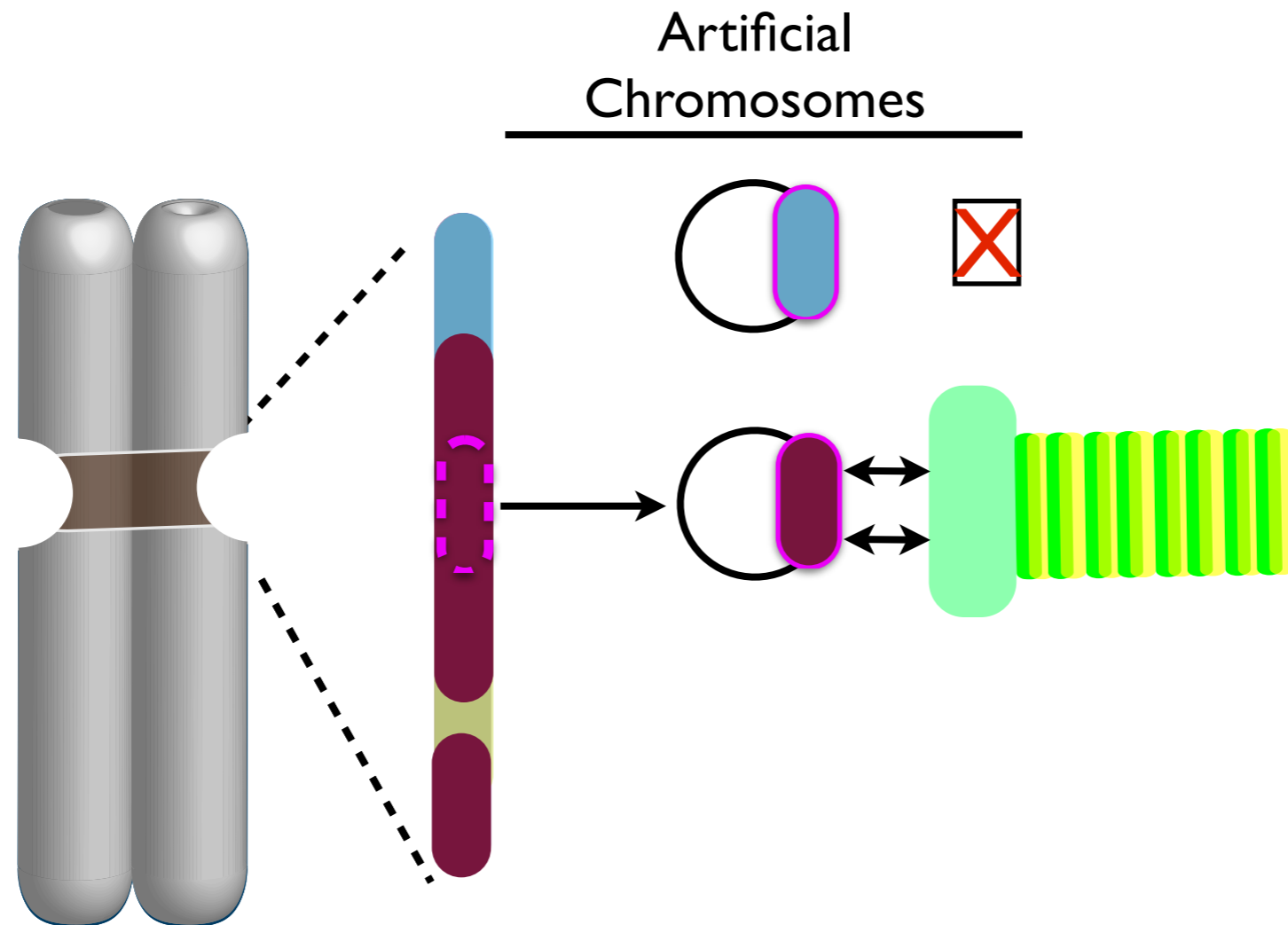
Person #1



Person #2



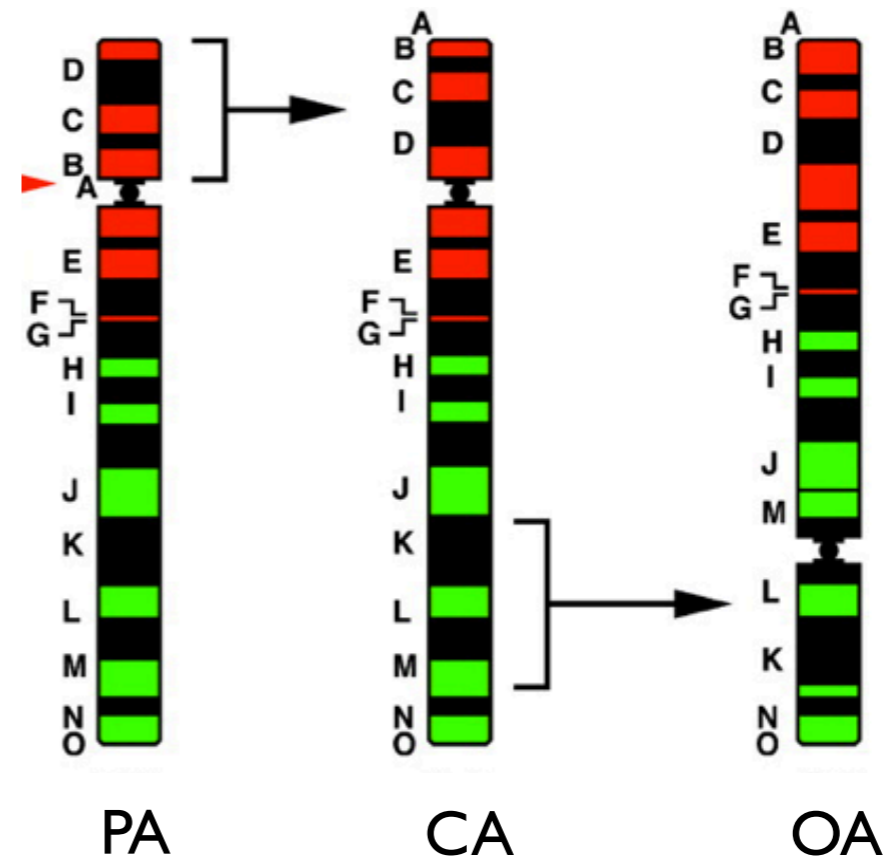
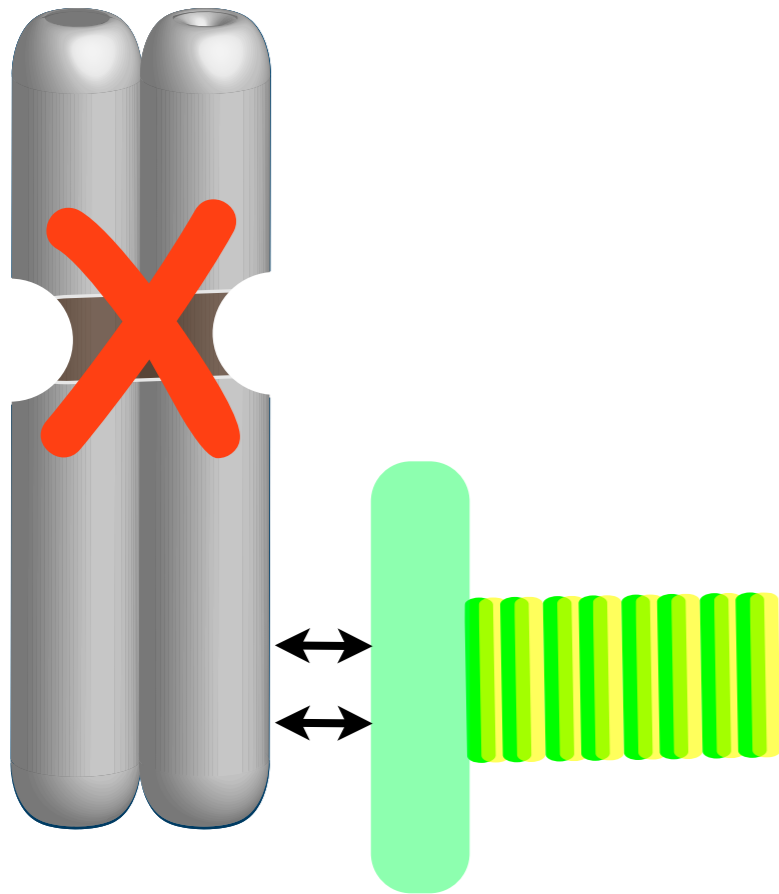
Centromeric Sequences Recruit Epigenetic Marks



- Co-localization of Kinetochore Proteins to Centromere Sequence
- Structural Analysis of Abnormal Chromosomes
- *De Novo* Centromere Formation

Centromeres can form without centromeric regions



Reconstruction of the chromosome 6 phylogeny in primates



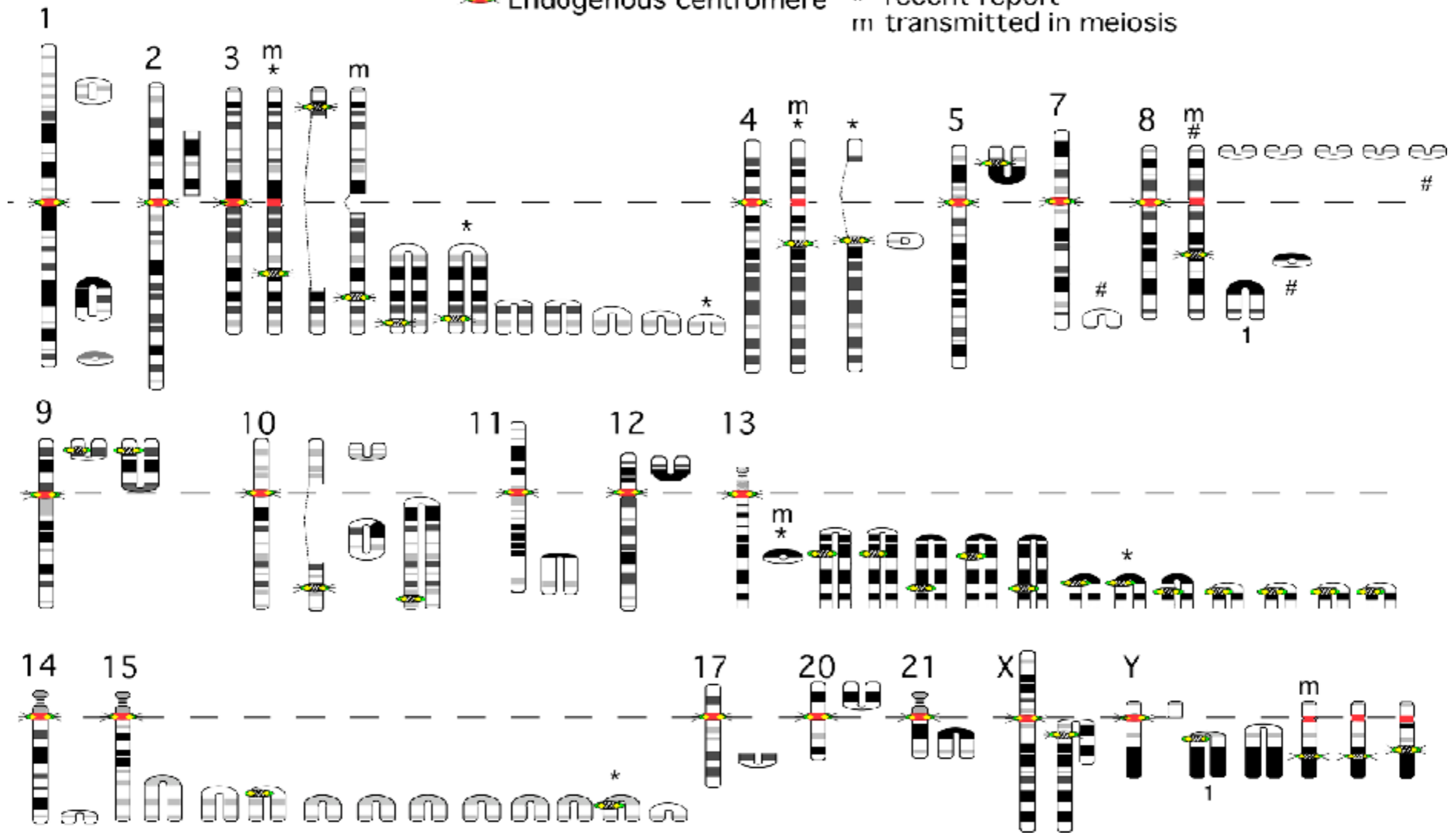
PA=primate ancestor; CA = Catarrhini ancestor; OA = OWM ancestor
modified from Eder V, et al MBE 2003

- Neocentromeres are thought to be rare
- Potential drivers of chromosome evolution

Total: 70, chromosomes: 19

 Neocentromere
 Endogenous centromere

unpublished
* recent report
m transmitted in meiosis



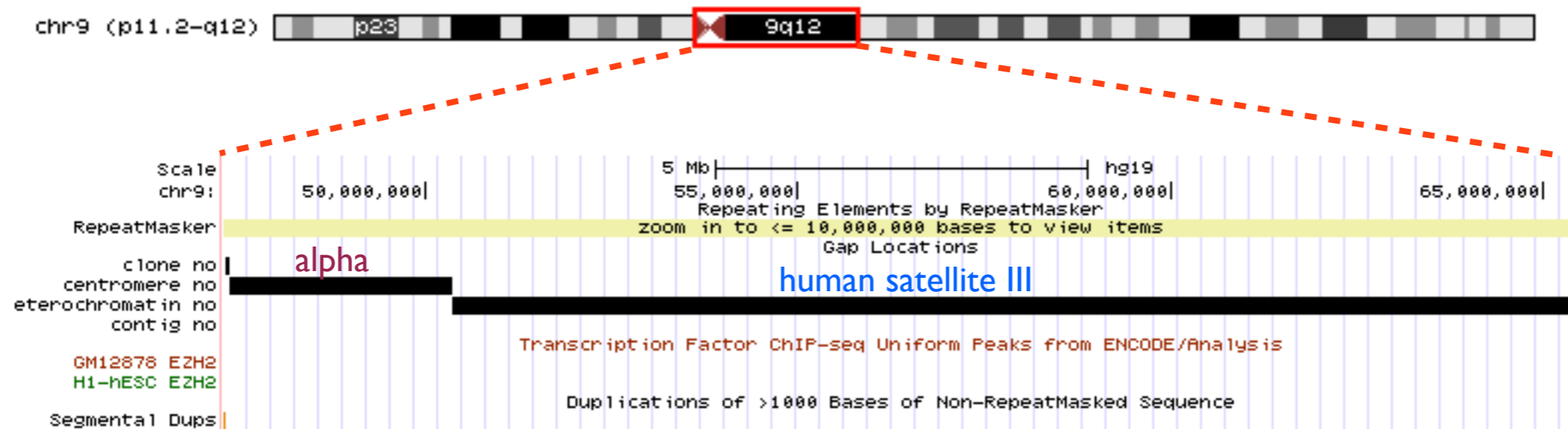
Summary

Human centromeres are currently defined by regions enriched with homogenized alpha satellite.

A single 'centromeric' region can contain more than one region of homogenized alpha satellite.

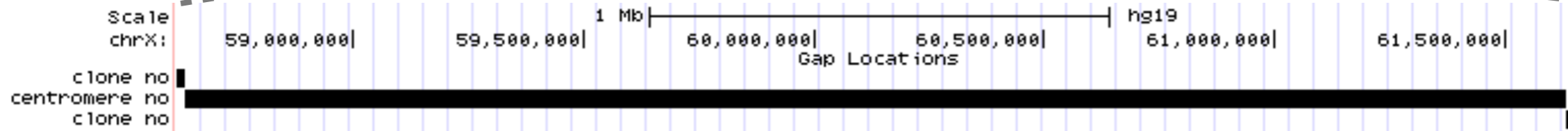
Centromeres are defined epigenetically. They can form over regions of non-centromeric sequence in the p or q arm.

In the reference assembly human “centromeres” are currently defined by 3Mb gaps



- All Centromeric Gaps are designated for alpha satellite DNA
- 3Mb was an educated guess for the size of the gap/placeholder. Centromeric regions can vary within and between individuals
- Adjacent ‘heterochromatin’ gaps typically represent regions enriched for Human Satellites 2,3 (w/ exception of het gap on chromosome 7 = alpha)

CEN

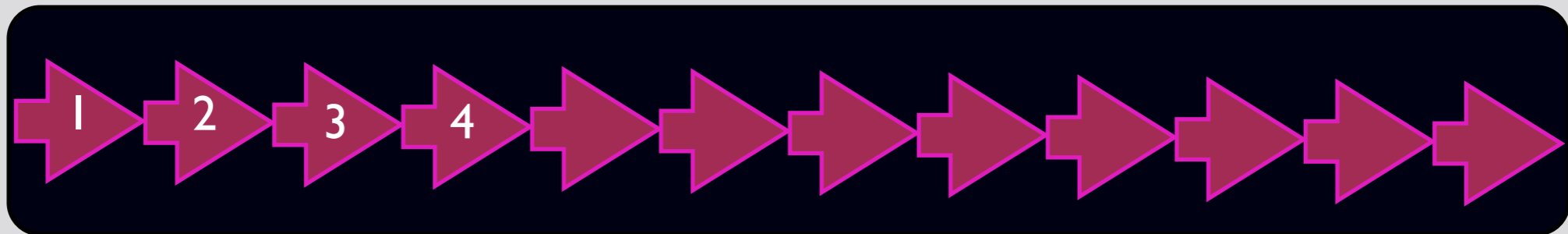


ALPHA SATELLITE

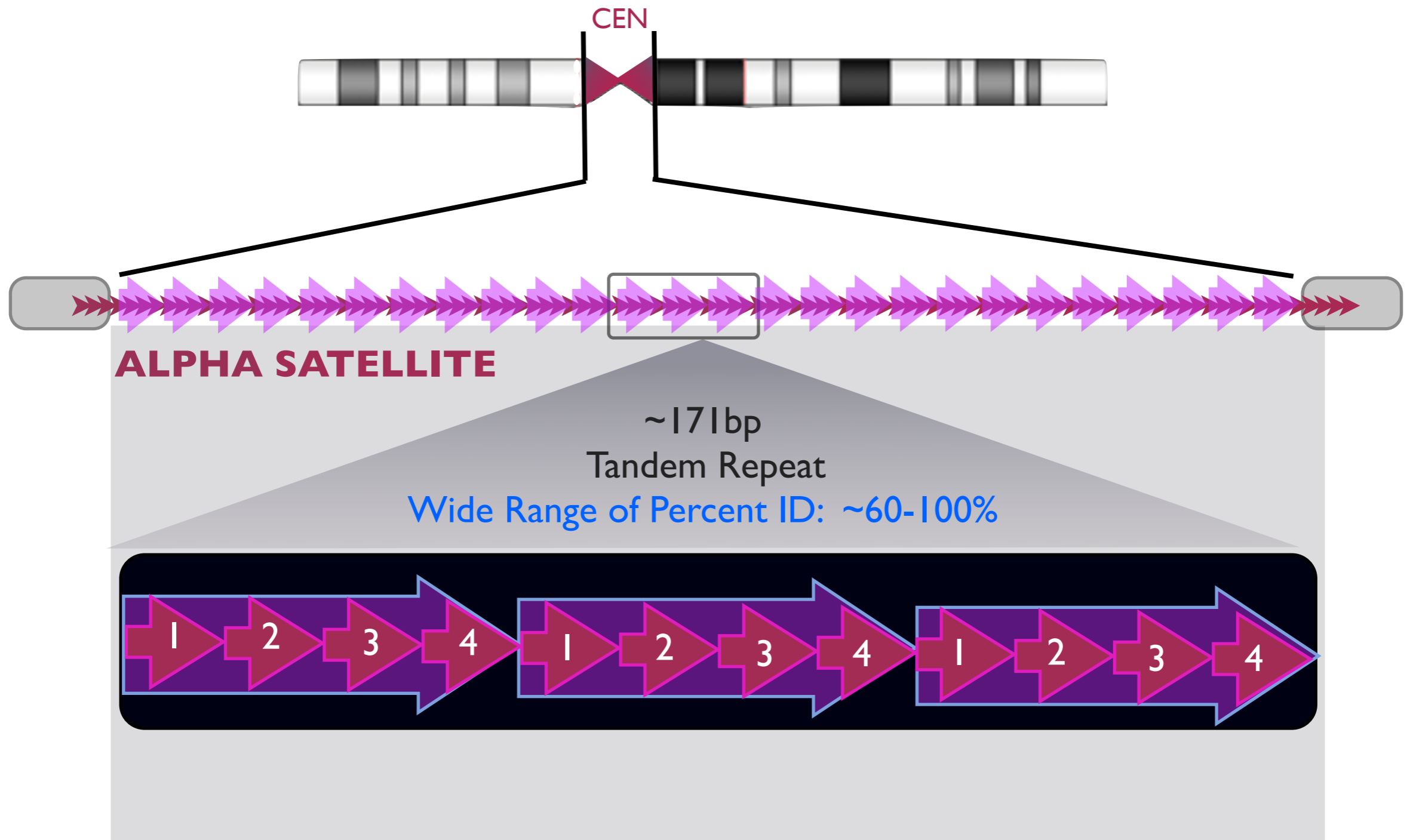
~171bp

Tandem Repeat

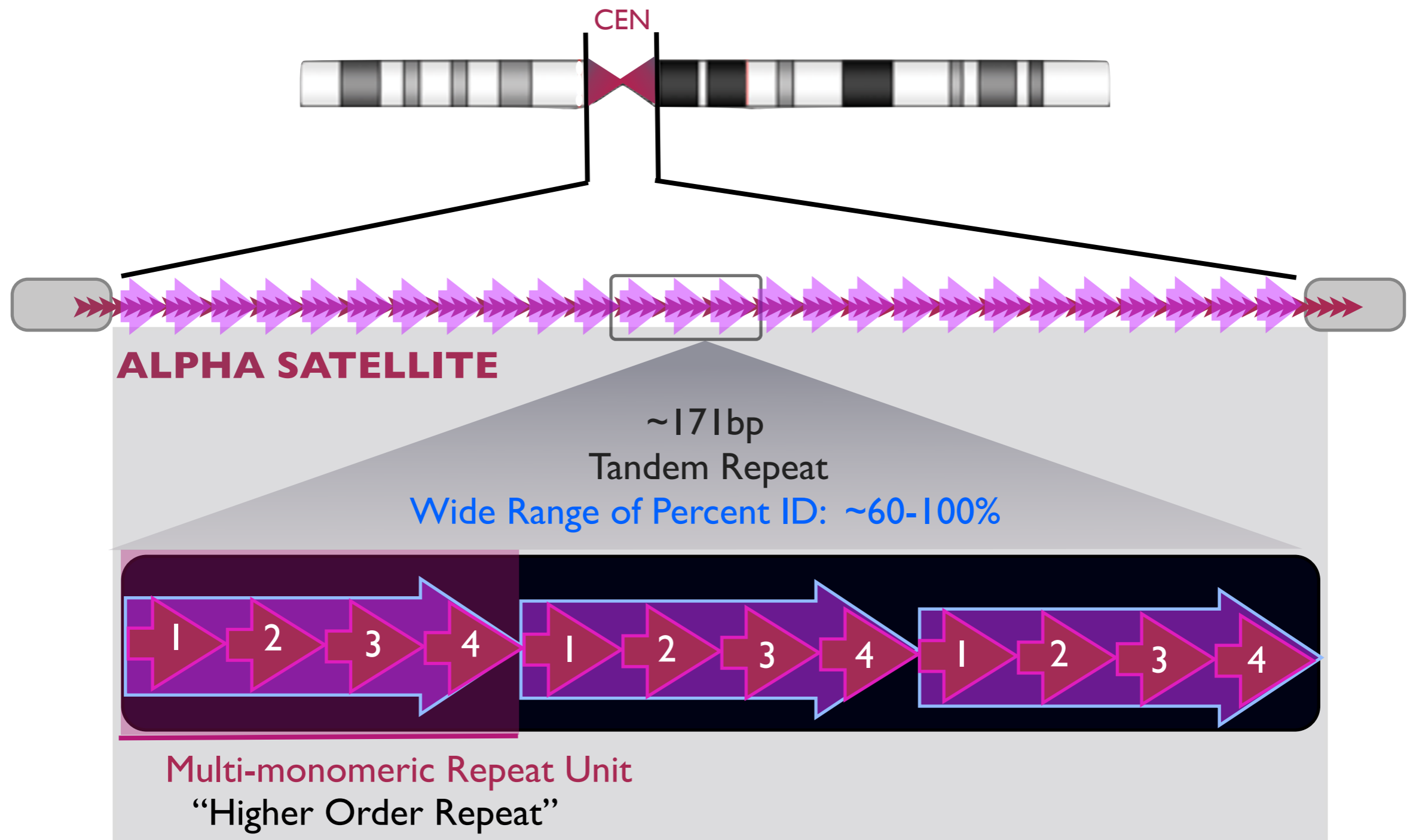
Wide Range of Percent ID: ~60-100%



Alpha Satellite define all normal human centromeres

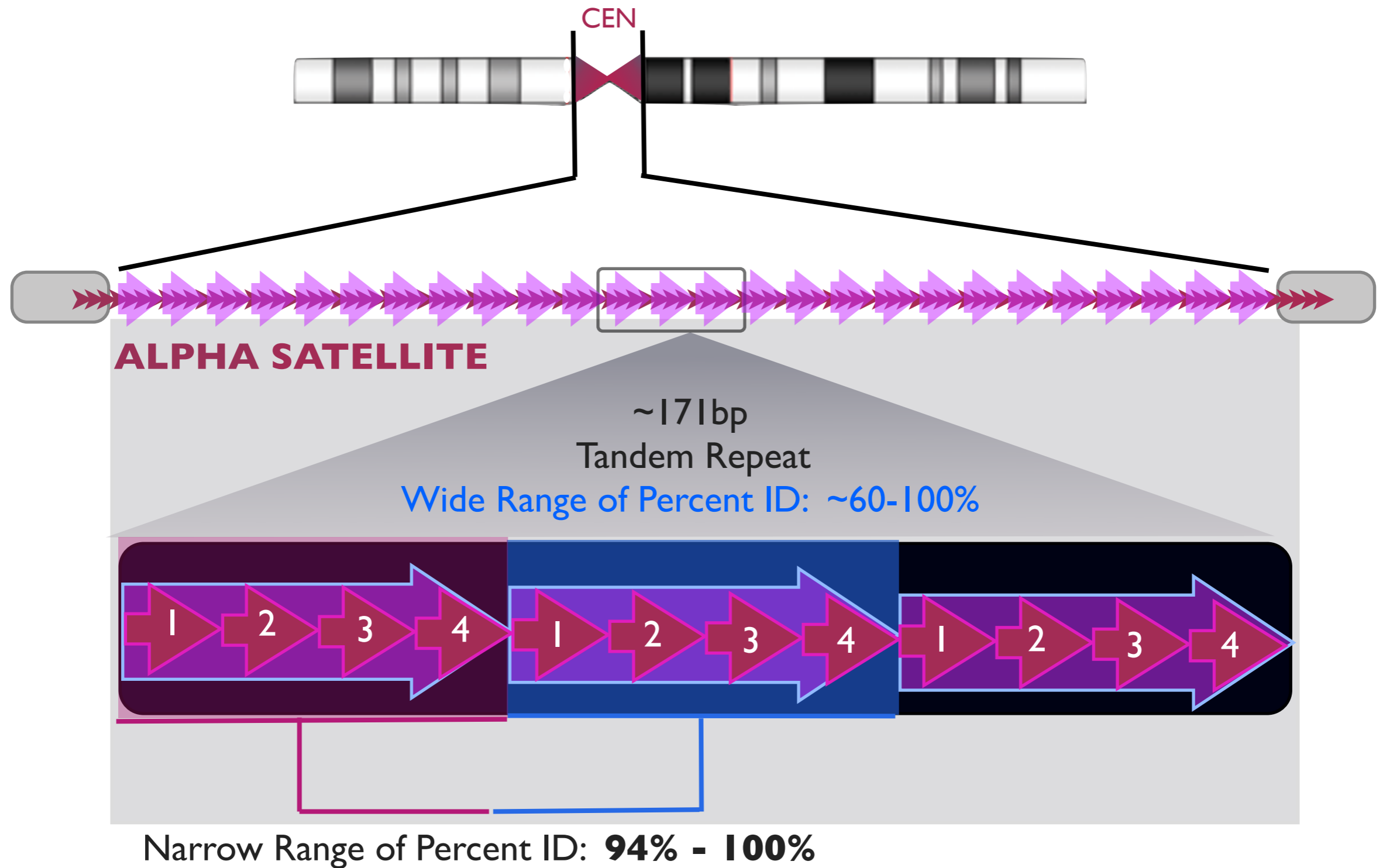


Alpha Satellite repeats (or monomers) are commonly found in long arrays of near-identical higher order repeats



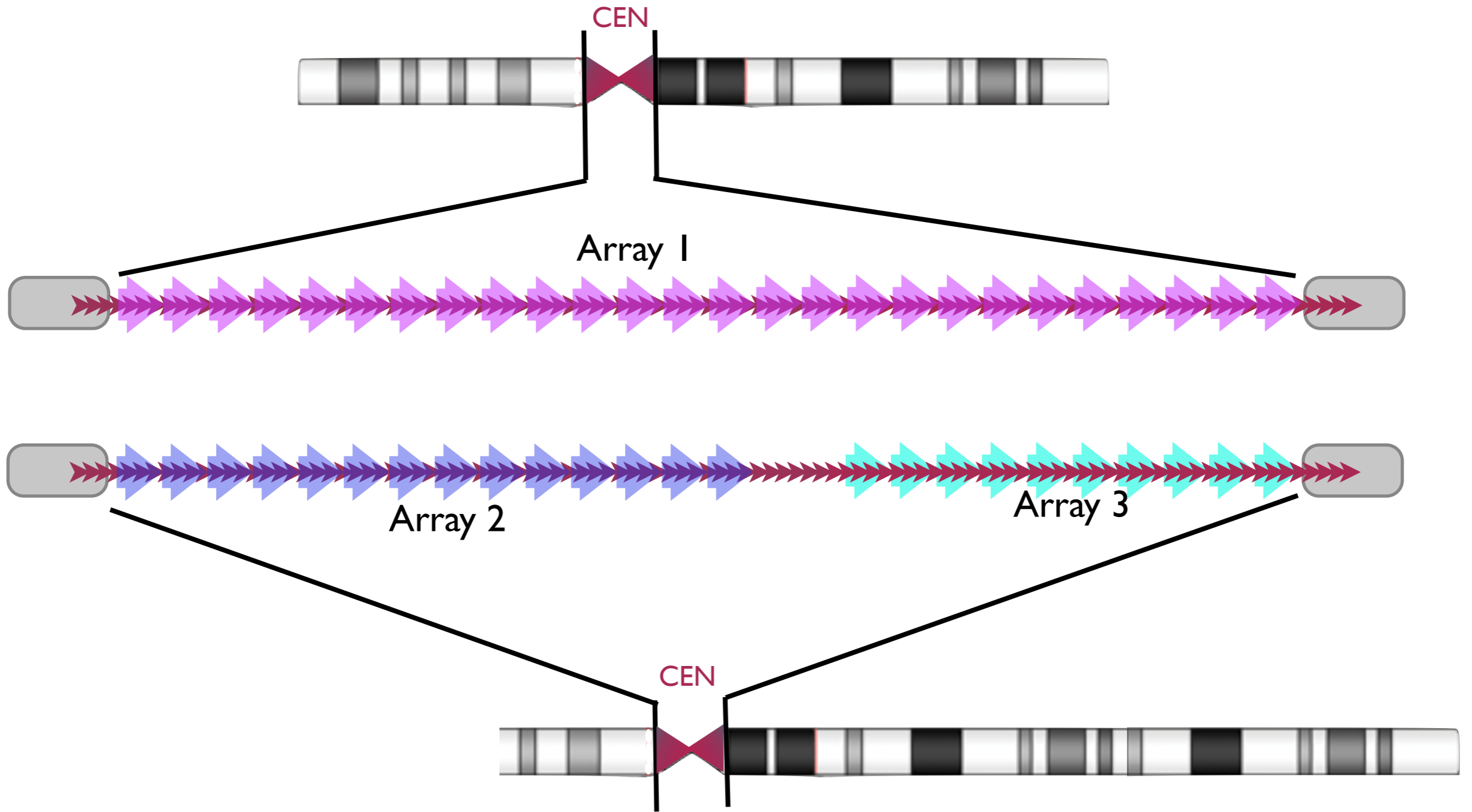
Alpha Satellite repeats (or monomers) are commonly found in long arrays of near-identical higher order repeats

Satellite DNA are the primary sequence in each gap

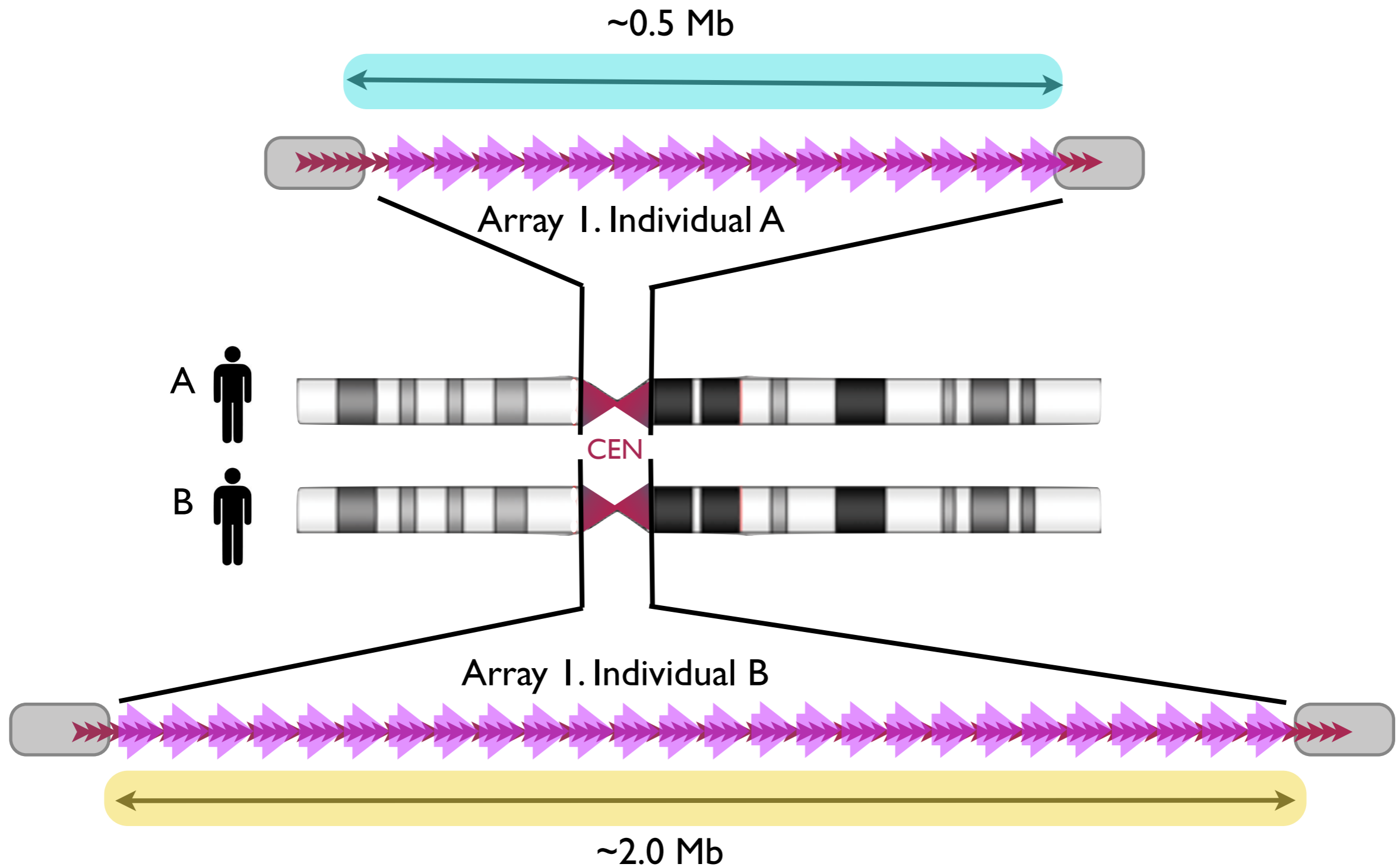


Alpha Satellite repeats (or monomers) are commonly found in long arrays of near-identical higher order repeats

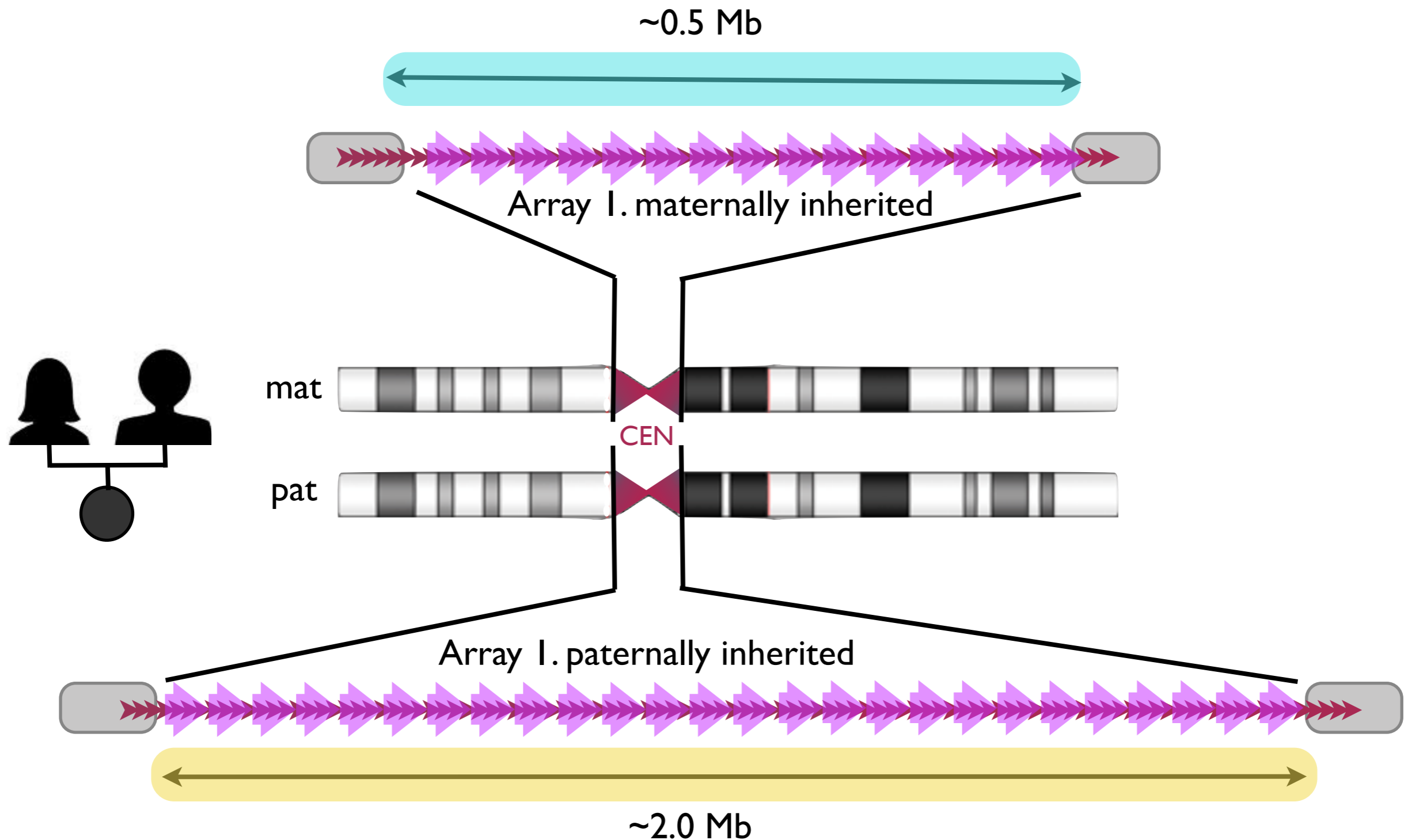
Each chromosome has a different centromeric sequences



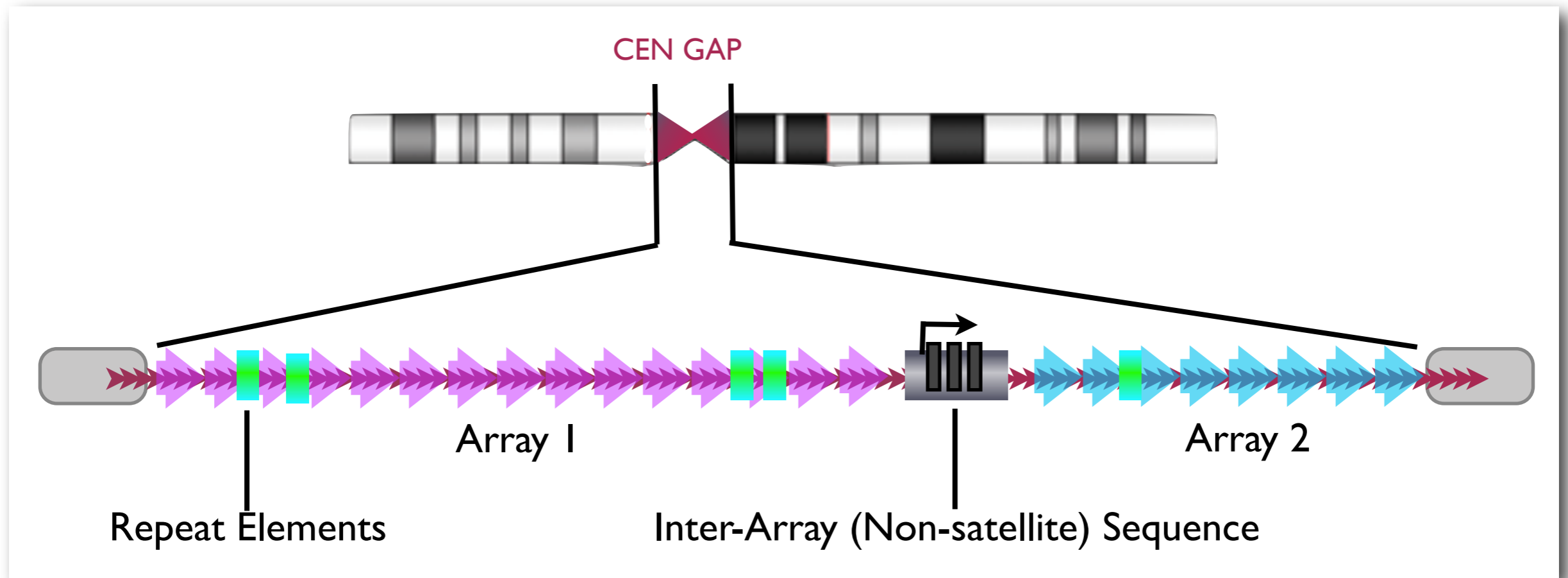
Higher-order arrays vary between individuals



Higher-order arrays can vary between homologous chromosomes in the same individual

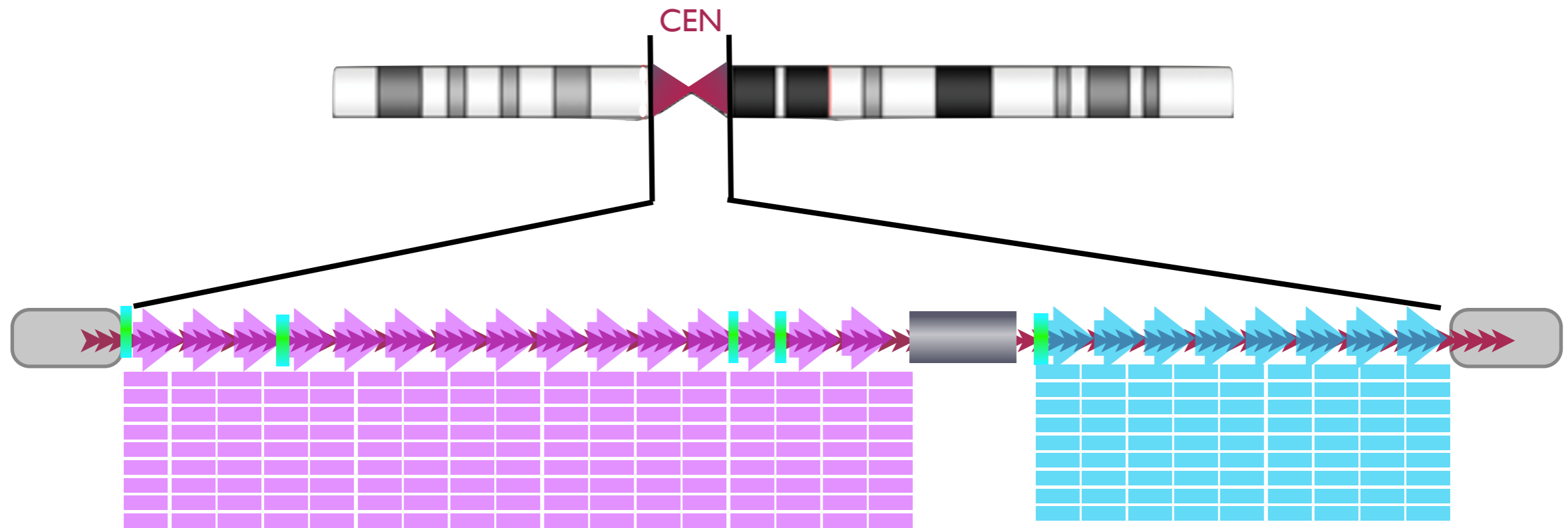


Model Centromere Sequence Organization



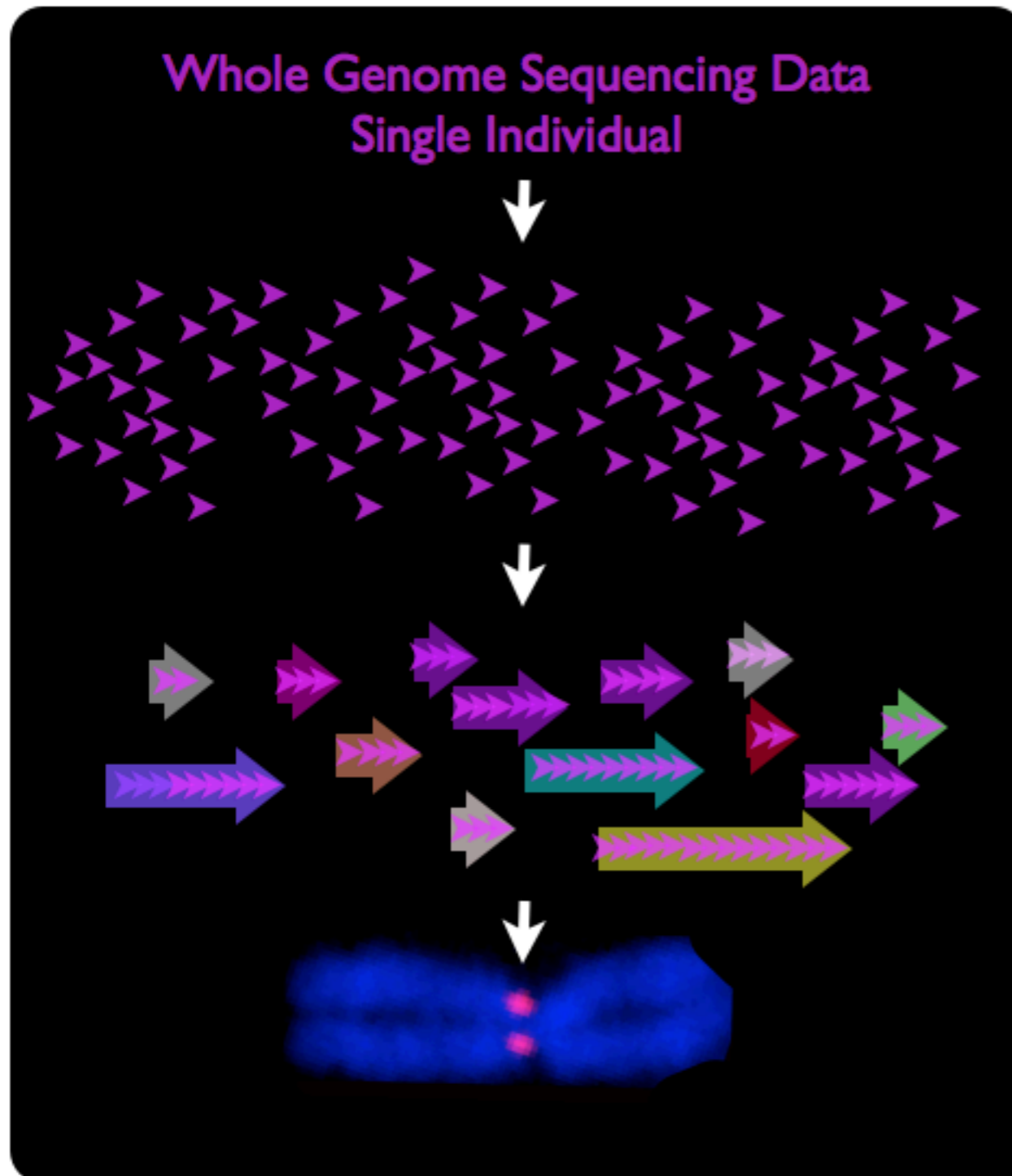
- Standard sequence assembly algorithms fail in these regions.
- Difficult to display the diploid organization: Further, no one haploid representation is expected to provide a true representation for the human population
- However, it is possible to study these regions without a perfect ordering and haploid representation: **provide mapping targets of 'centromere components'**

Reformat sequences observed in each read library into linear reference model



- 1 Constructing Read Libraries for each HOR array:**
Build database of all components in each centromeric region
- 2 LinearSat Software to Convert Reads to Linear Reference Models**
Generate a linear representation of observed components
- 3 Scaffold Reference Models and HuRef assembled contigs using mate pairs**
Order components in each centromeric gap

I Constructing Read Libraries for each HOR array



HuRef Genome

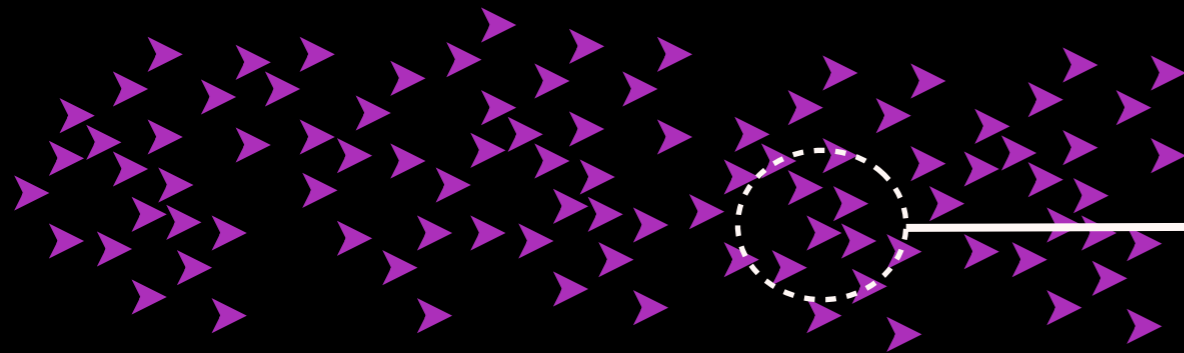
**Centromeric database construction
from reads containing alpha
satellite repeats.
(2.6% of the human genome)**

**Determine chromosome-specific
organization of alpha variants into
higher order repeats.**

**Build statistical models to generate faux
centromere sequence that
will serve as a target for mapping
centromeric reads.**

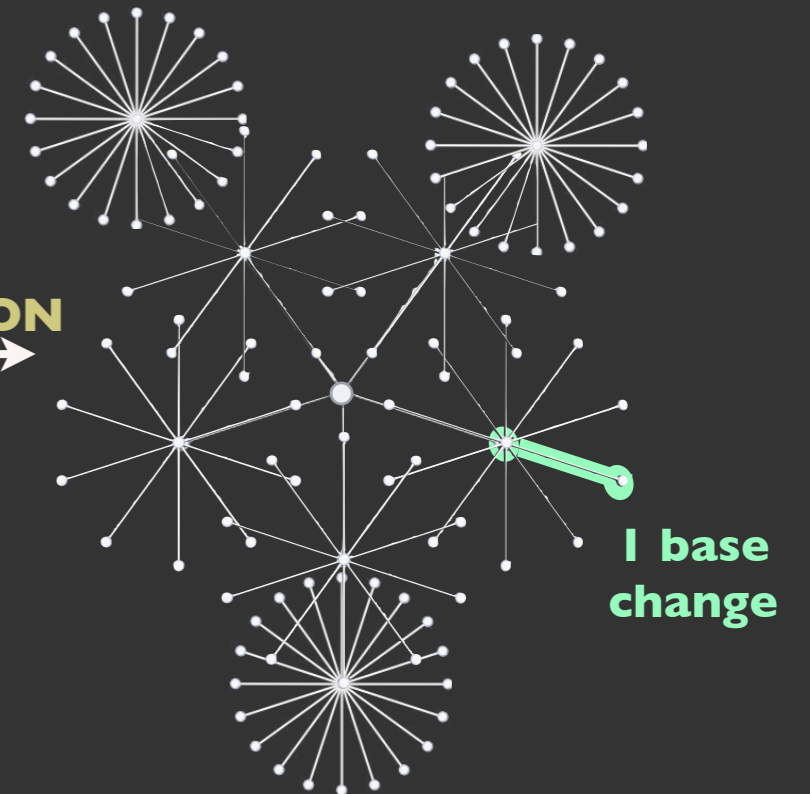
Higher Order Repeat Prediction

HuRef Genome (8x Coverage)



DATA COMPRESSION

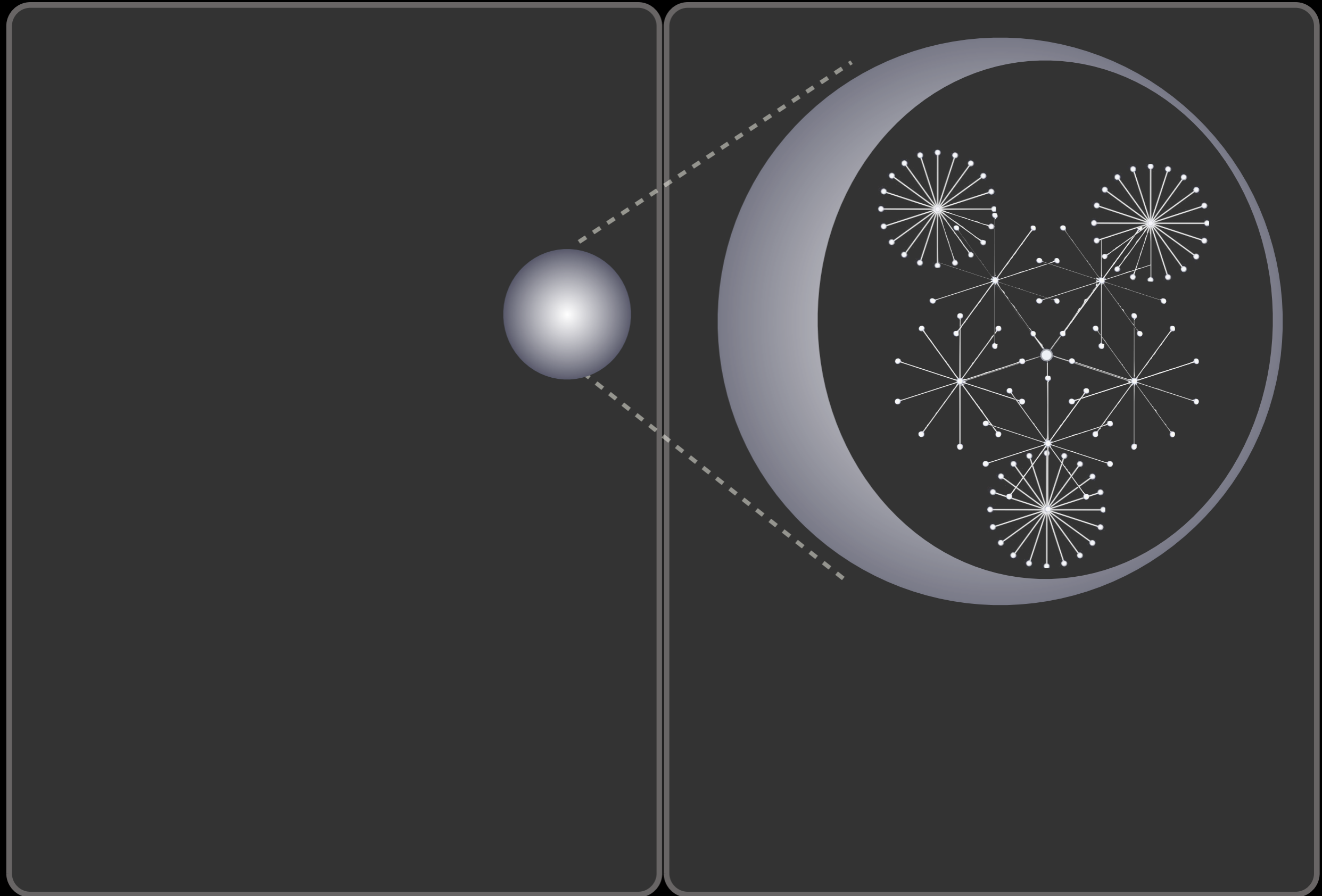
IDENTICAL
MONOMERS



Similarity Clustering:
Defining Epsilon Neighborhood

SEQUENCE RELATIONSHIP

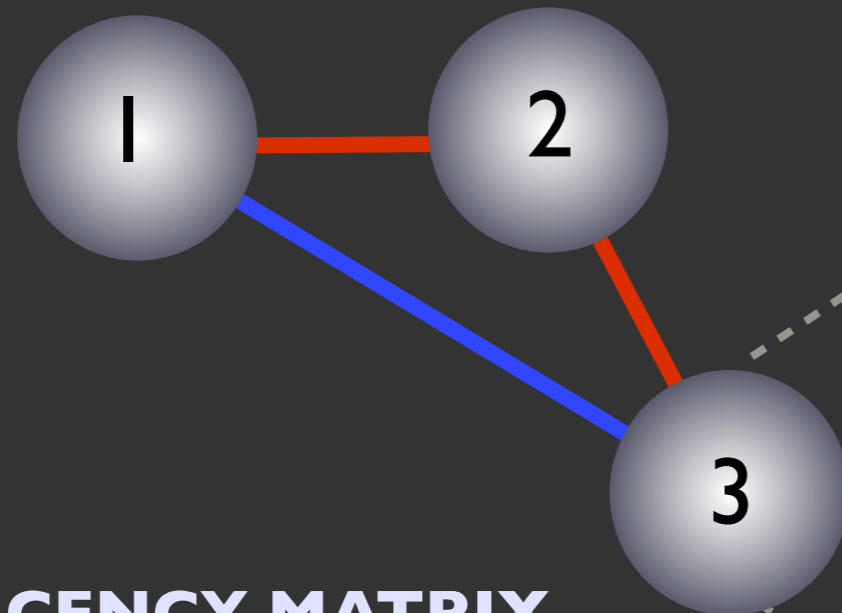
Higher Order Repeat Prediction



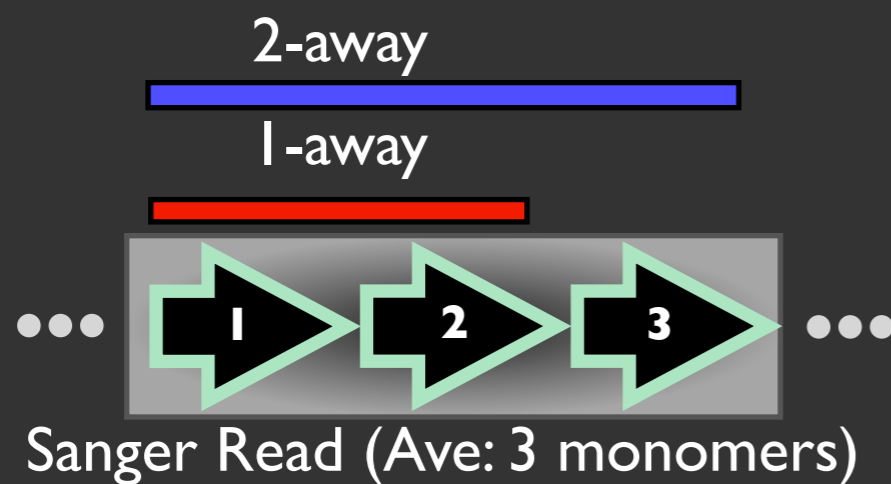
PHYSICAL RELATIONSHIP

SEQUENCE RELATIONSHIP

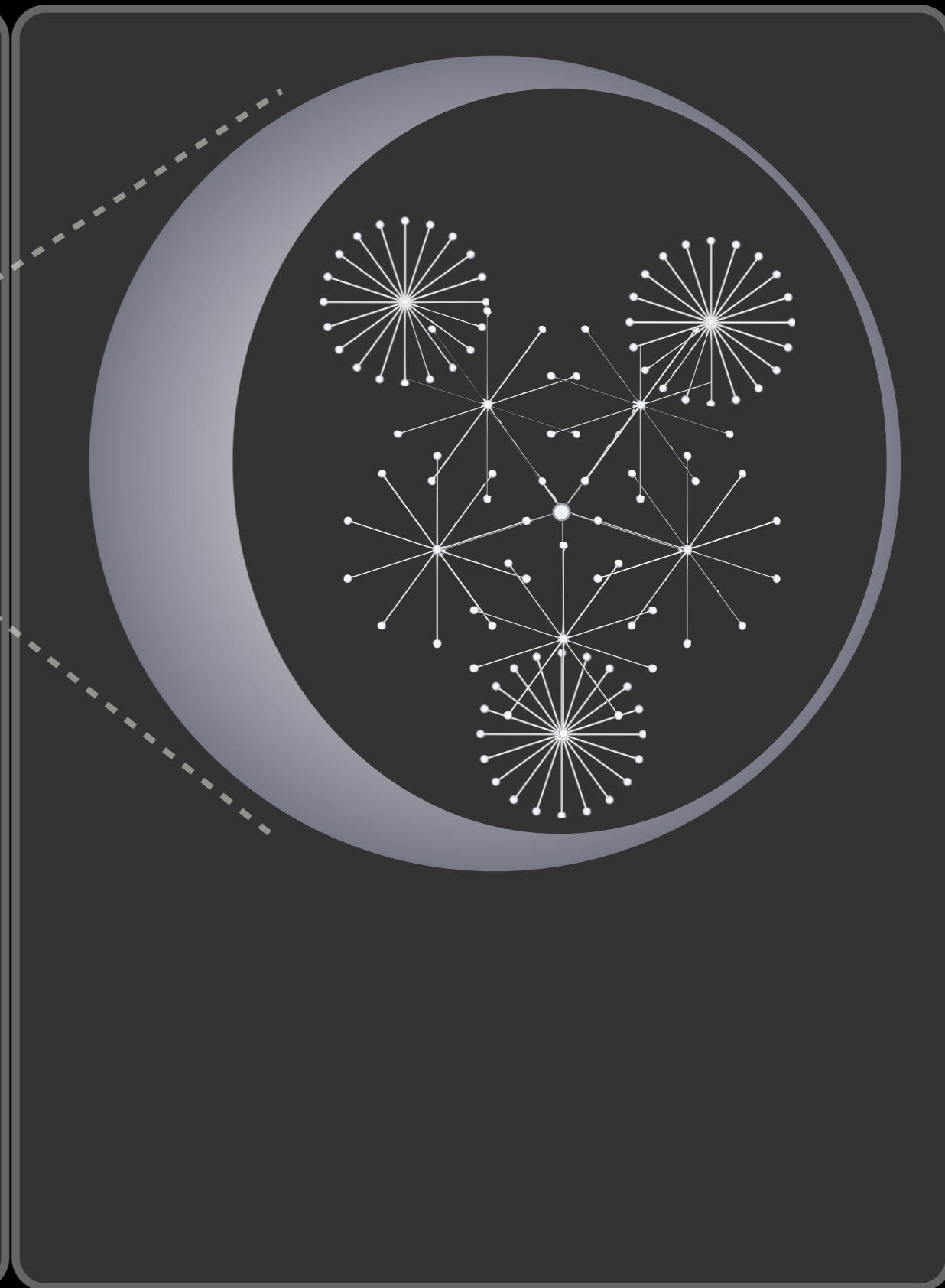
Higher Order Repeat Prediction



ADJACENCY MATRIX
Alpha Satellite Monomers

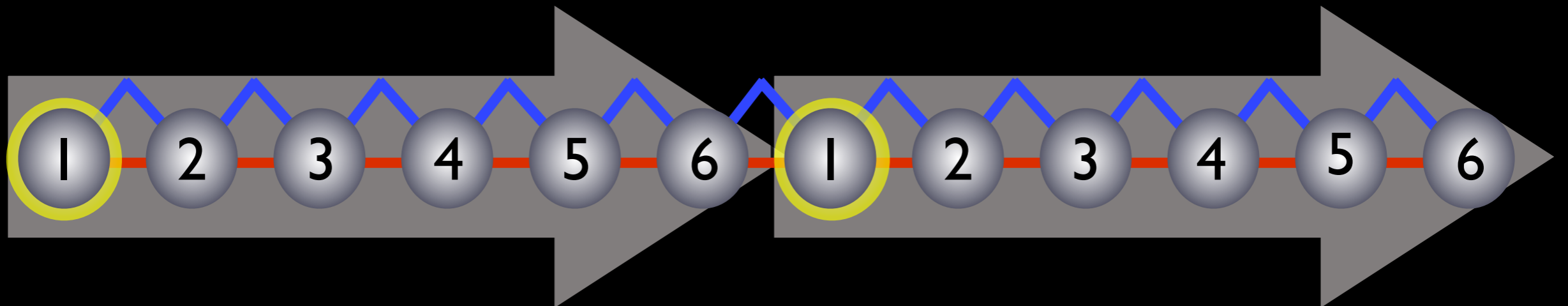


PHYSICAL RELATIONSHIP

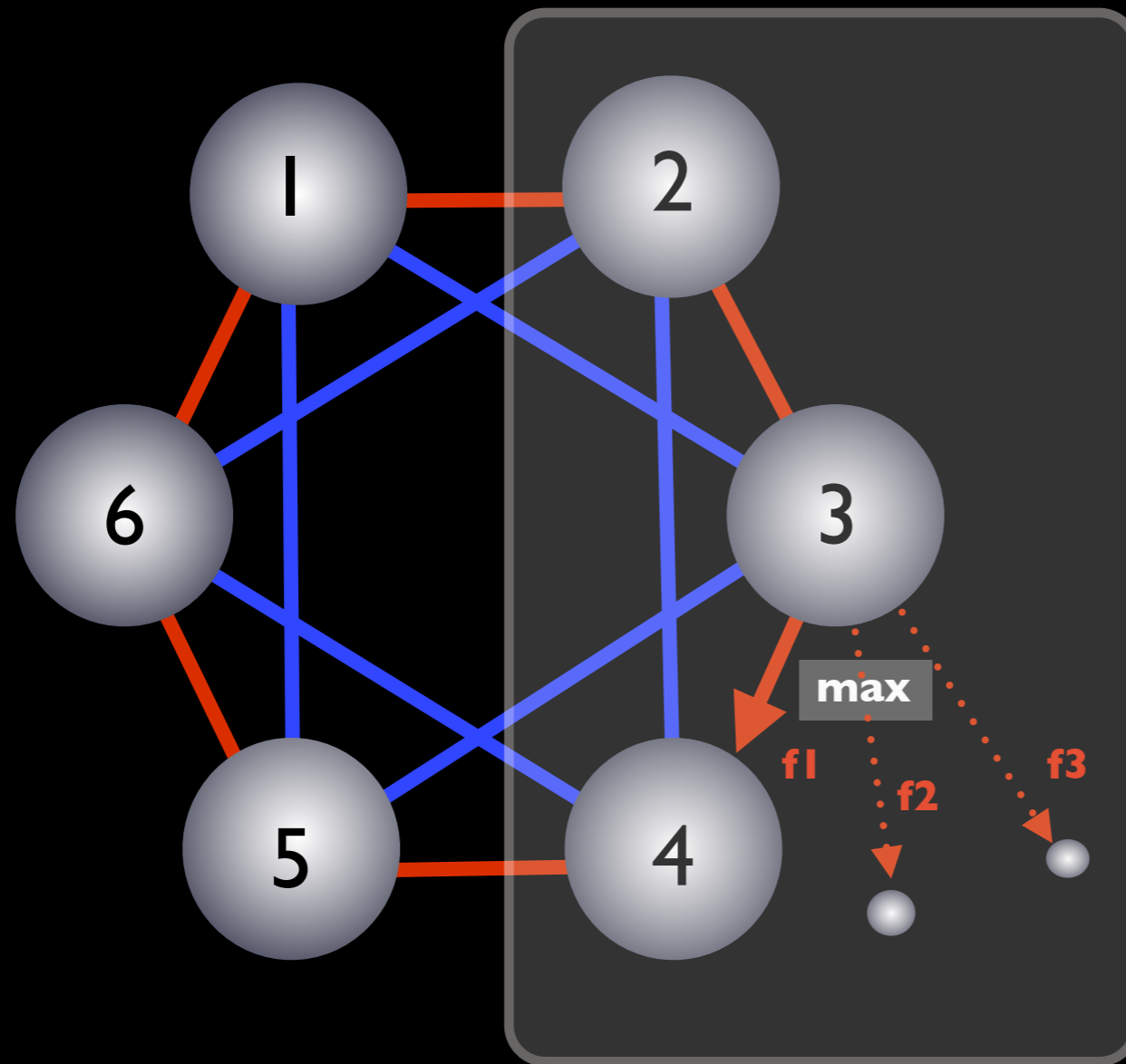


SEQUENCE RELATIONSHIP

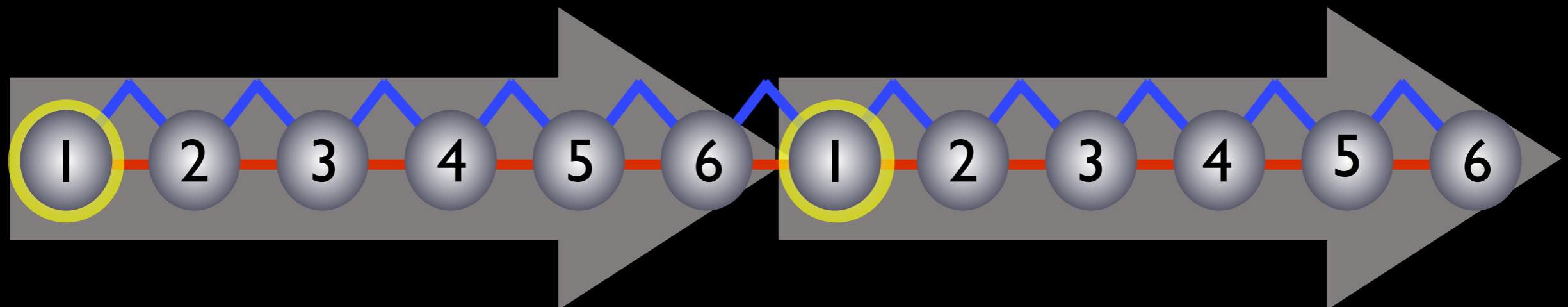
Higher Order Repeat Prediction



Higher Order Repeat Prediction

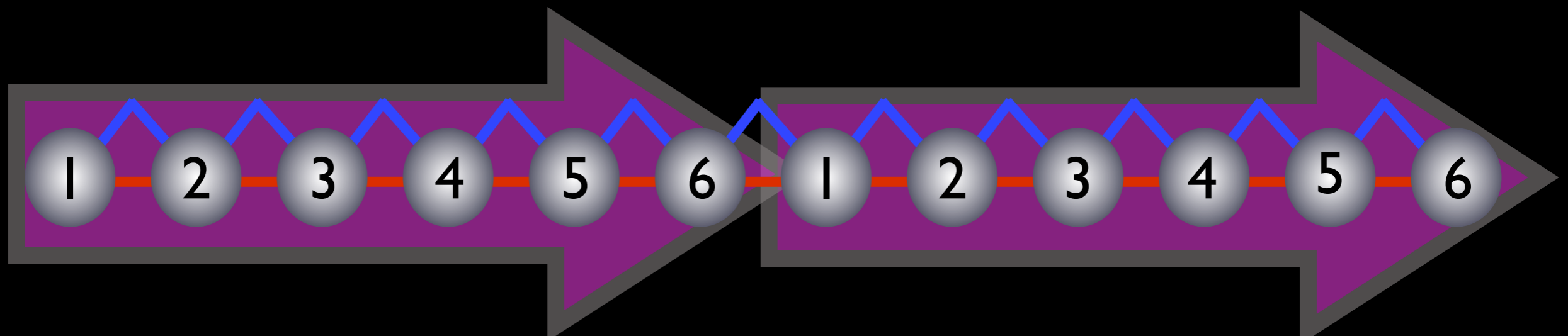
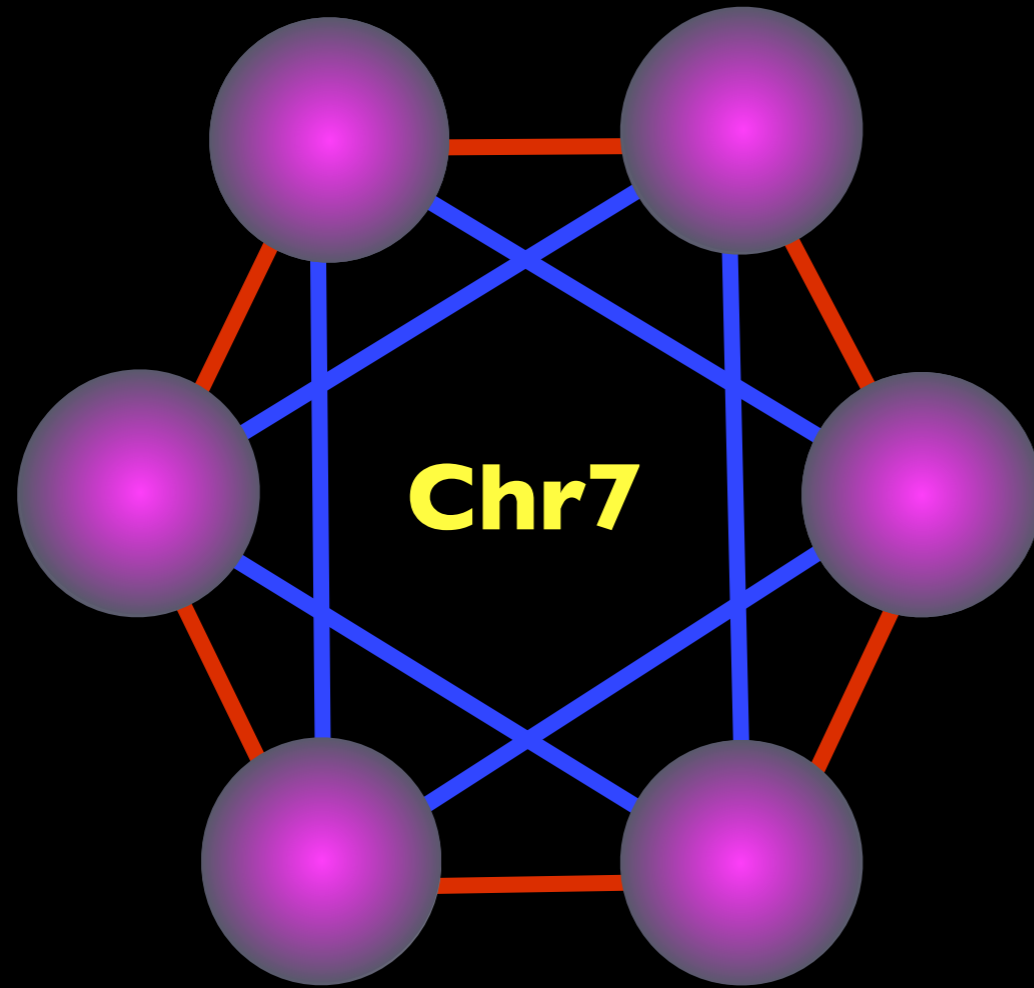


**Viterbi greedy-algorithm
second order markov model**



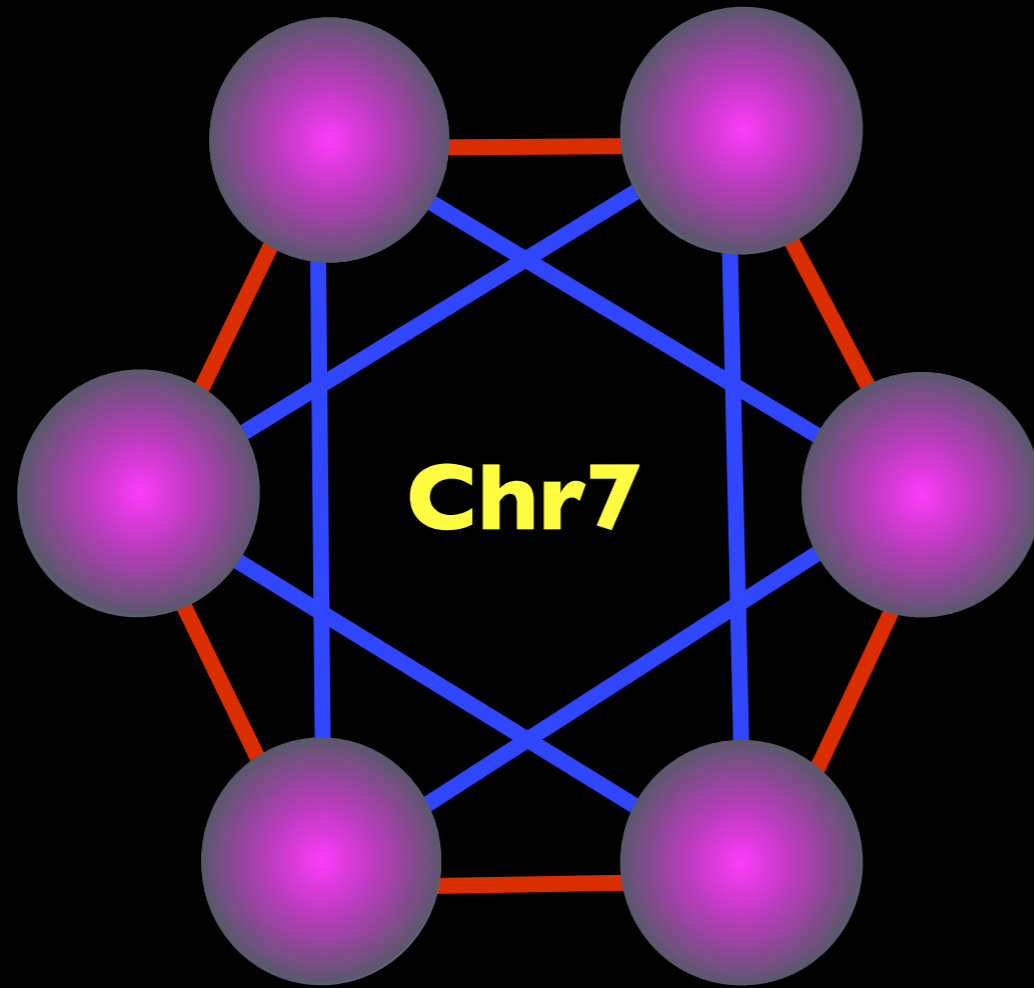
Higher Order Repeat Prediction

Determine Chromosome Specificity:



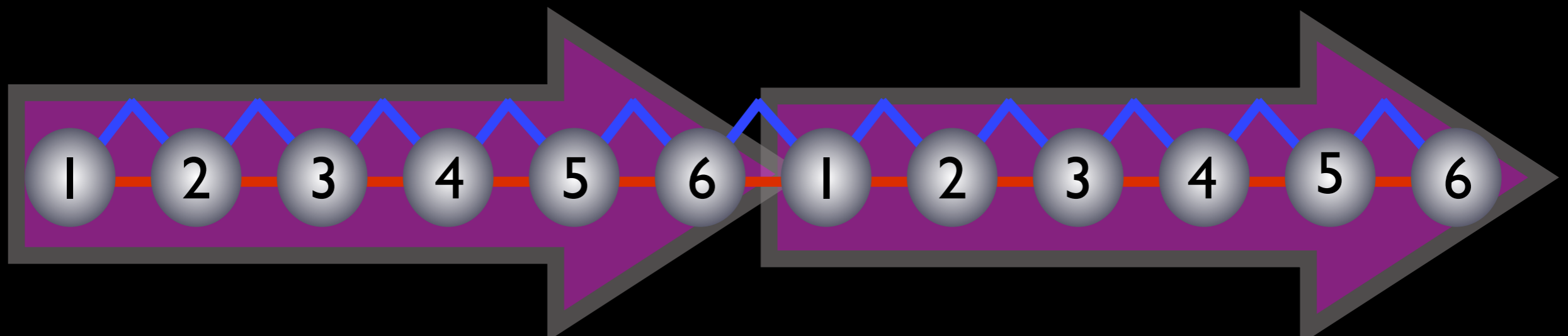
Higher Order Repeat Prediction

Determine Chromosome Specificity:



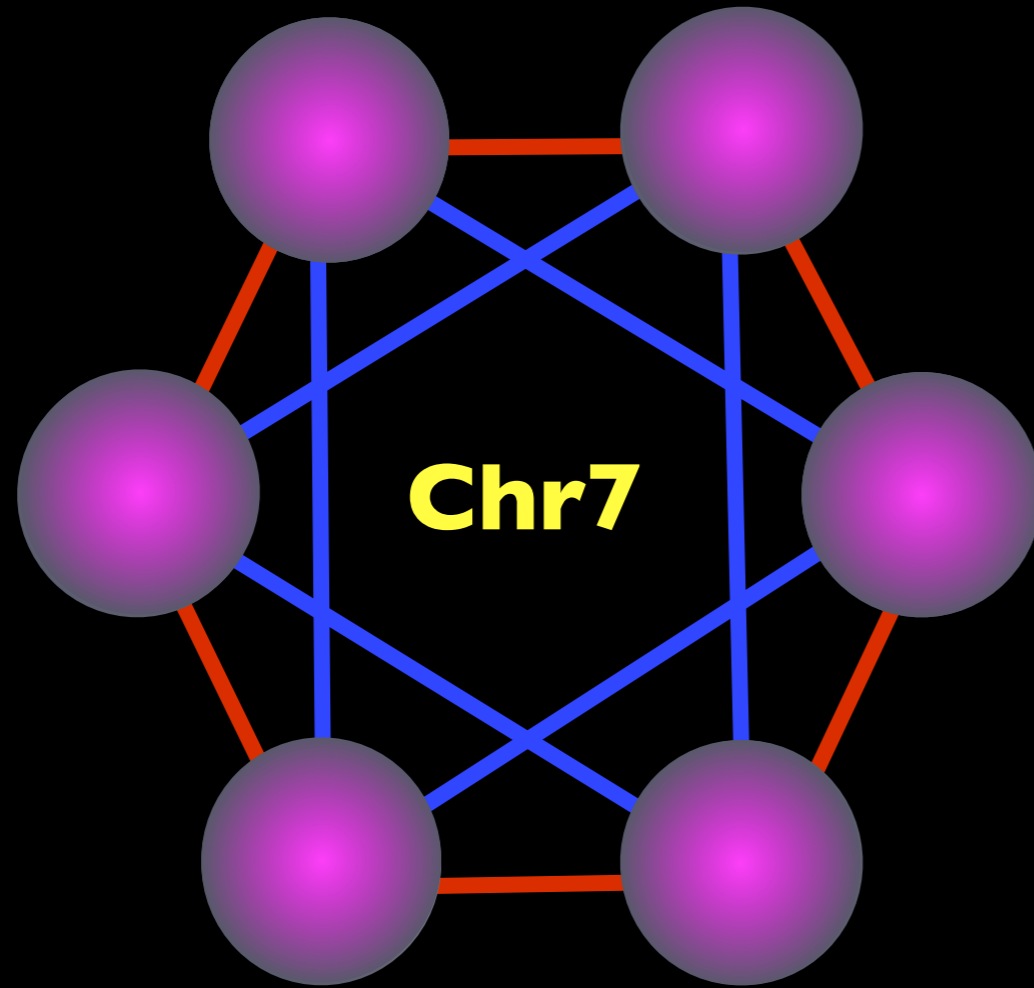
Flow Sorted Chromosome
Alignment/Enrichment

344 Mb of Alpha Satellite from 15 Chromosomes



Higher Order Repeat Prediction

Determine Chromosome Specificity:

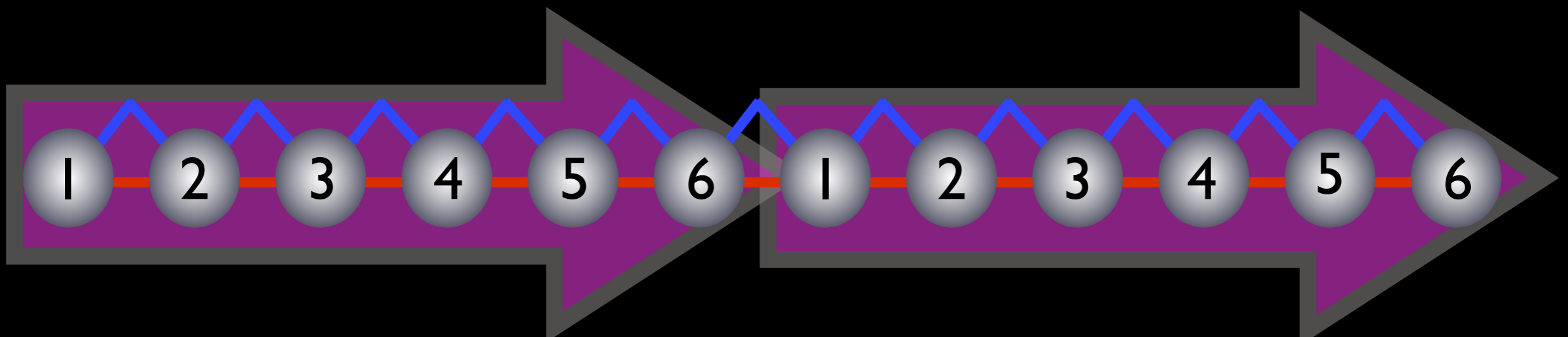


Flow Sorted Chromosome
Alignment/Enrichment

344 Mb of Alpha Satellite from 15 Chromosomes

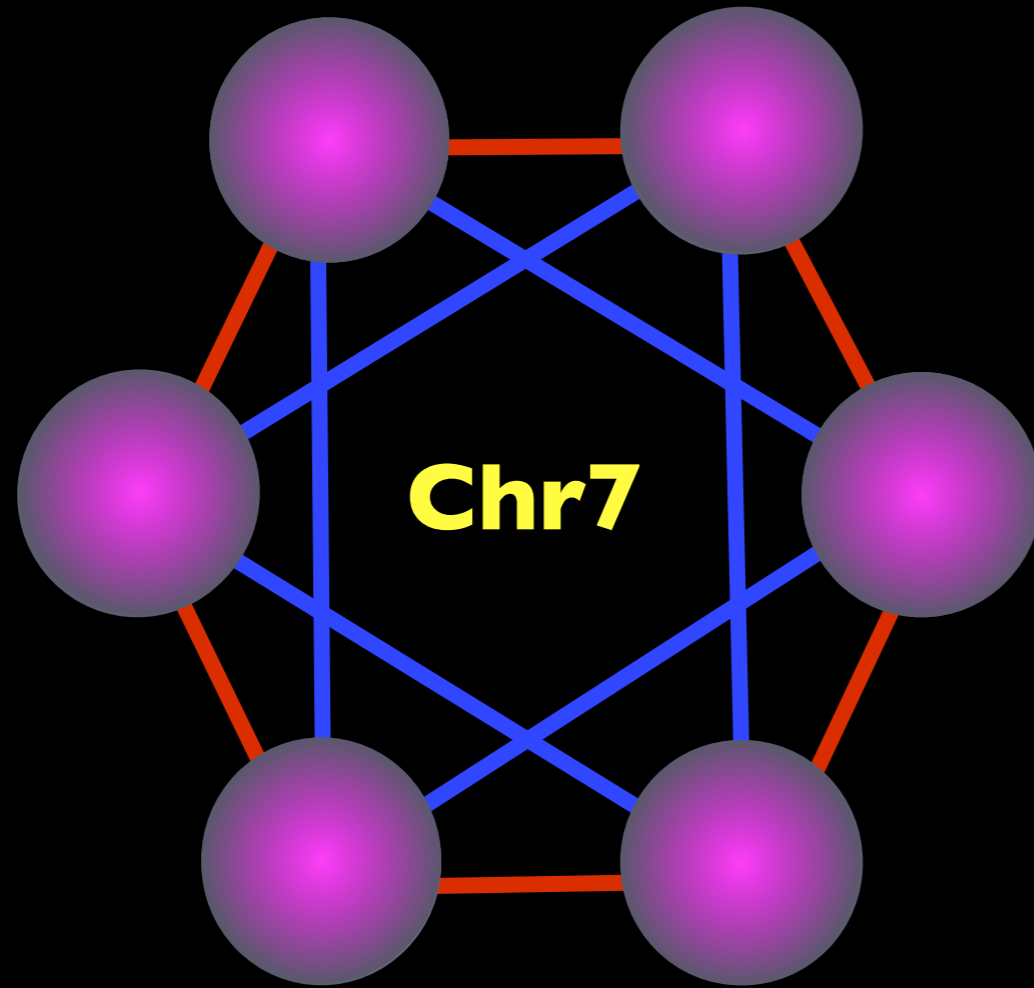
Experimental Evidence

FISH Hybridization and Screening Somatic Cell
Hybrid Panel



Higher Order Repeat Prediction

Determine Chromosome Specificity:



Flow Sorted Chromosome
Alignment/Enrichment

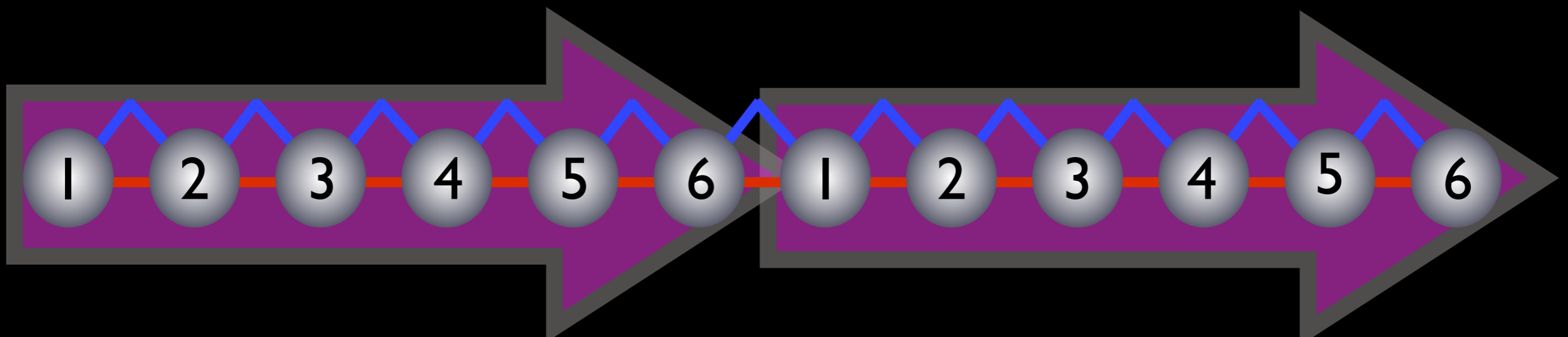
344 Mb of Alpha Satellite from 15 Chromosomes

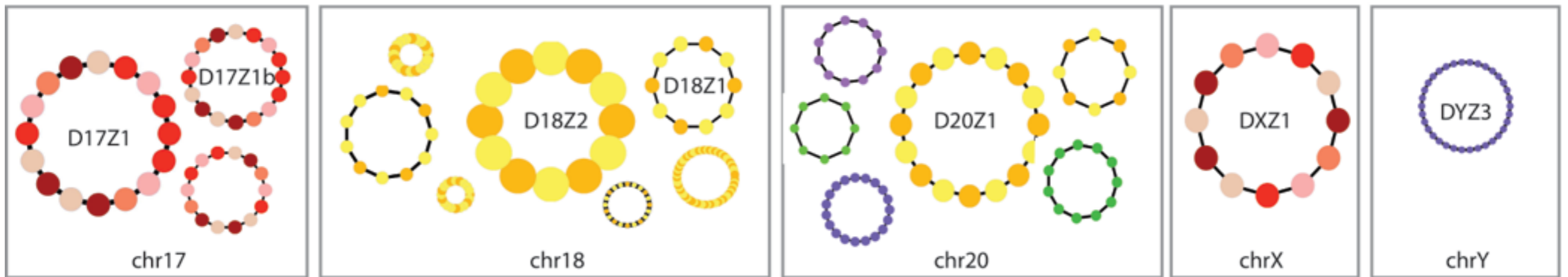
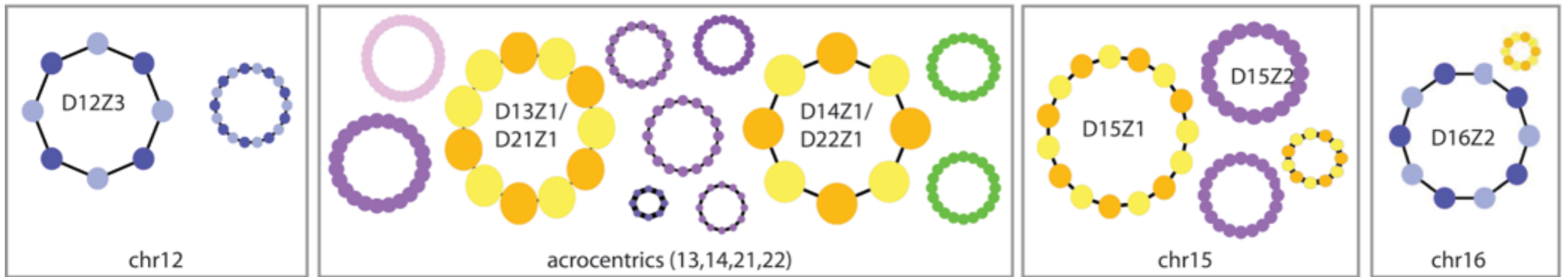
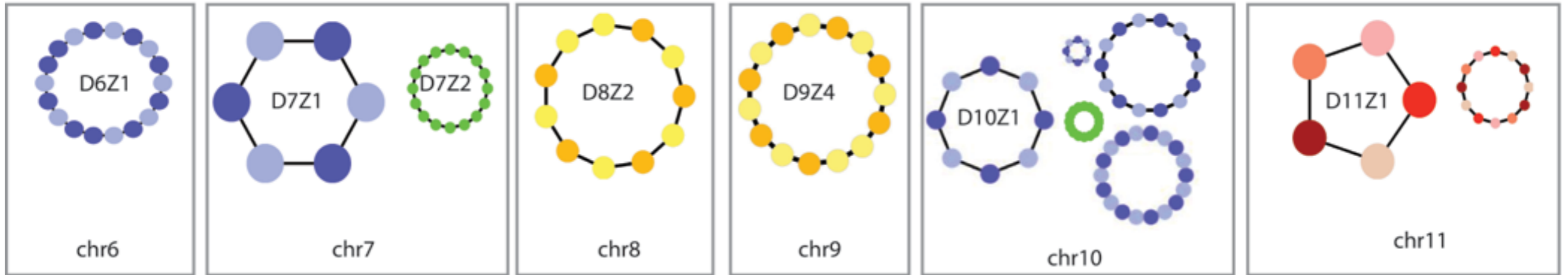
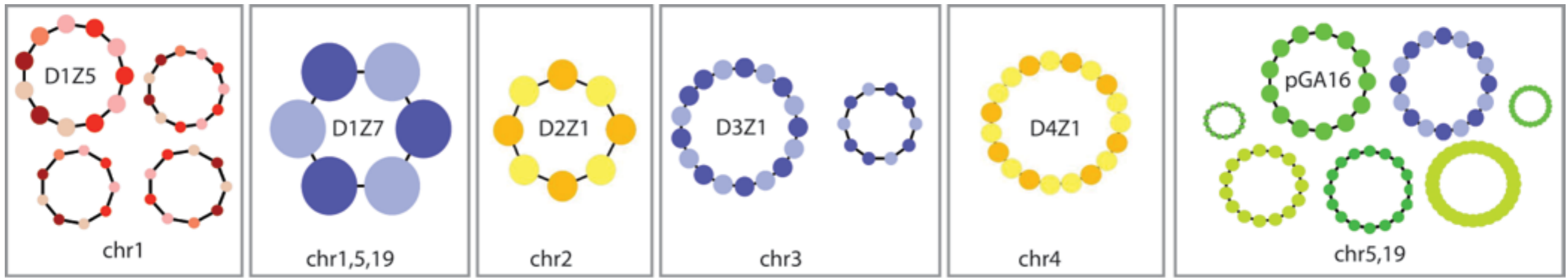
Experimental Evidence

FISH Hybridization and Screening Somatic Cell
Hybrid Panel

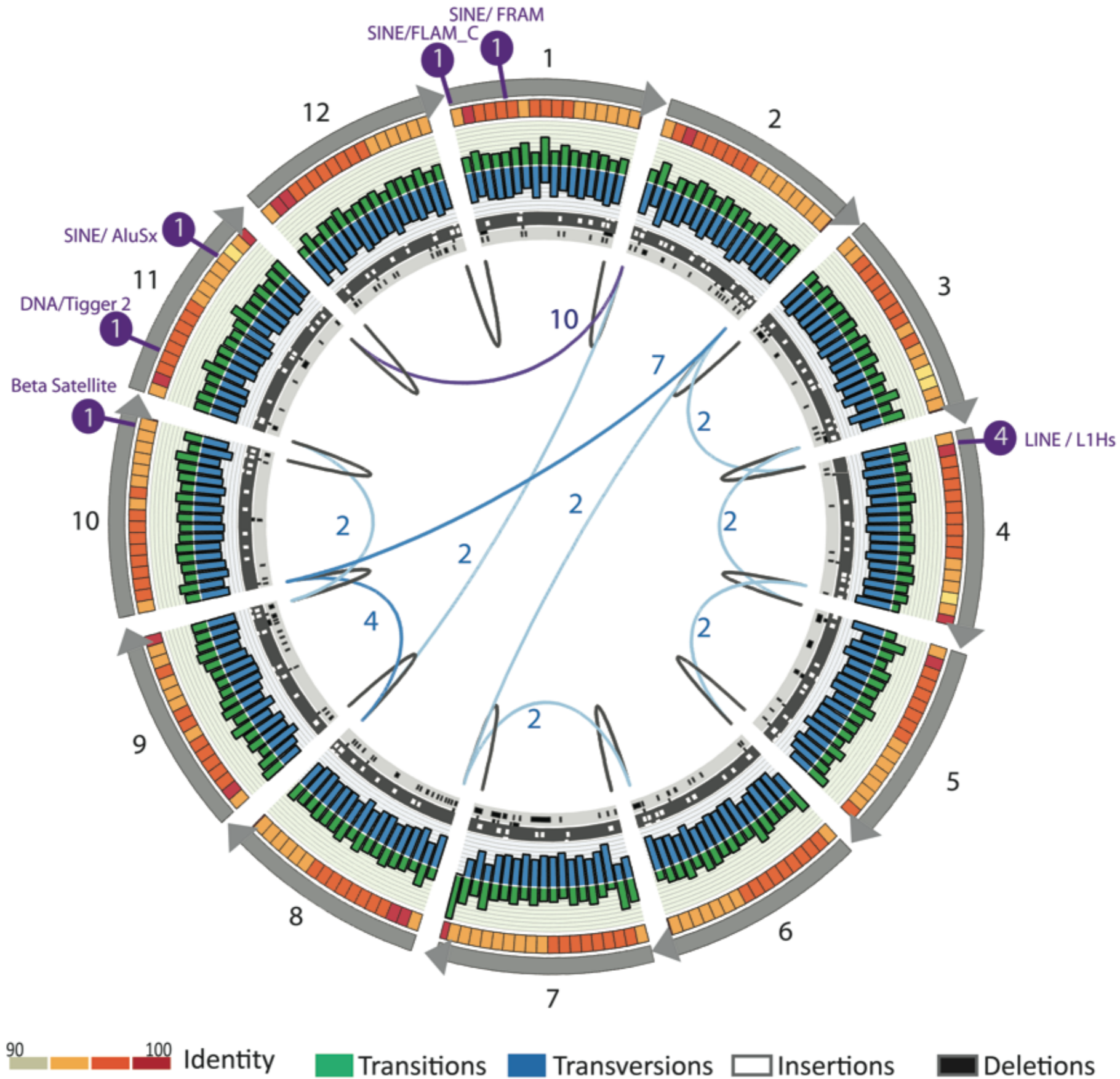
Paired Reads

“Anchor” to adjacent mapped HuRef contigs

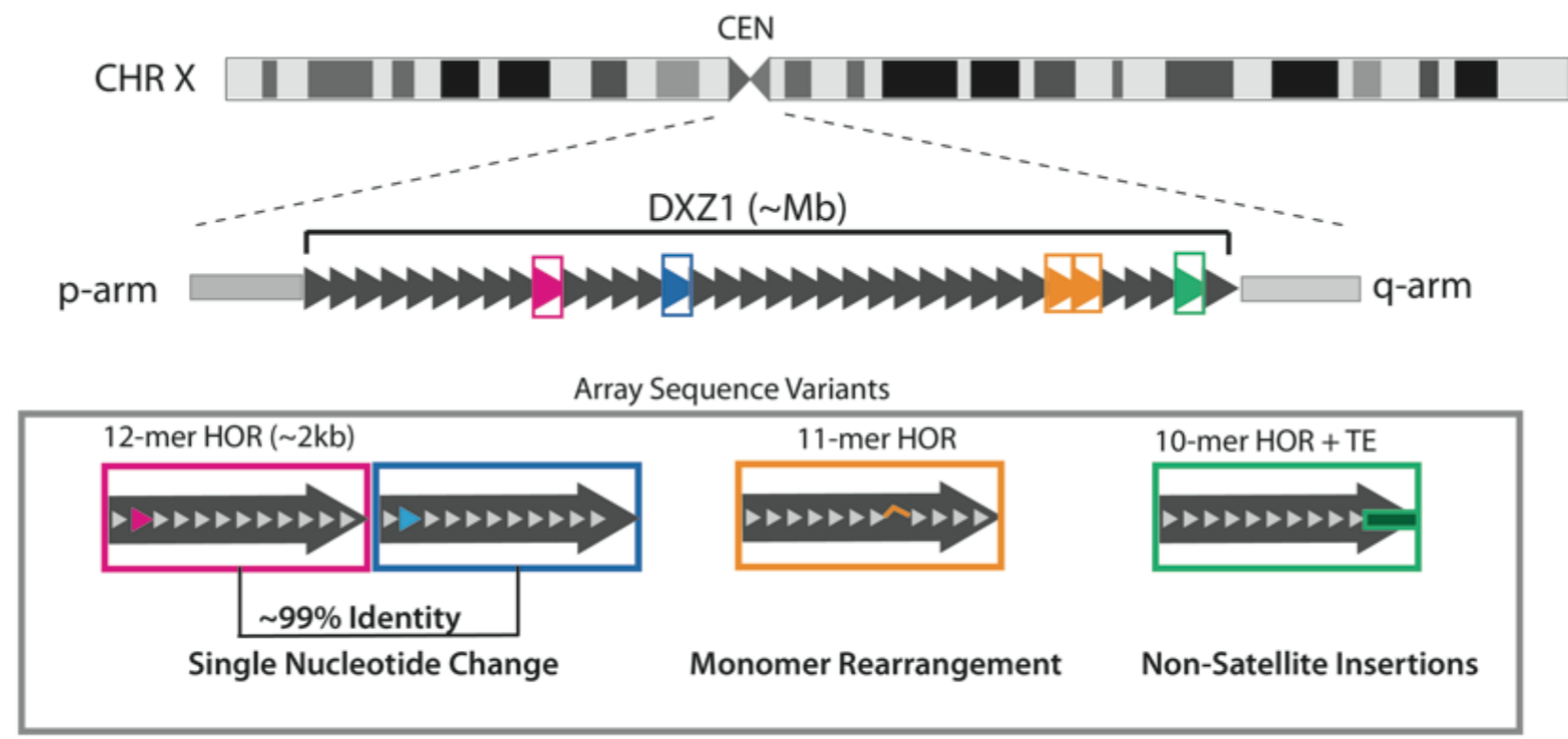


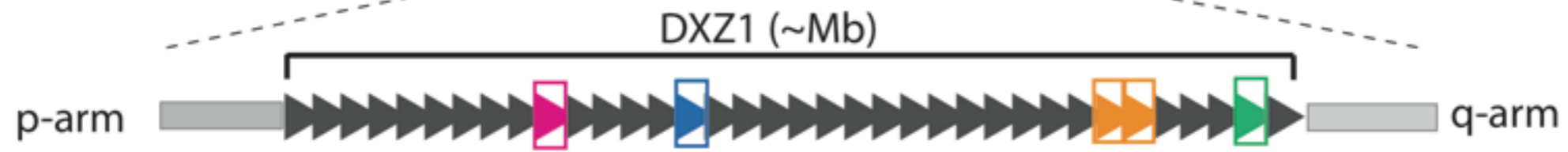


Alpha Satellite Array (DXZI) on Chromosome X

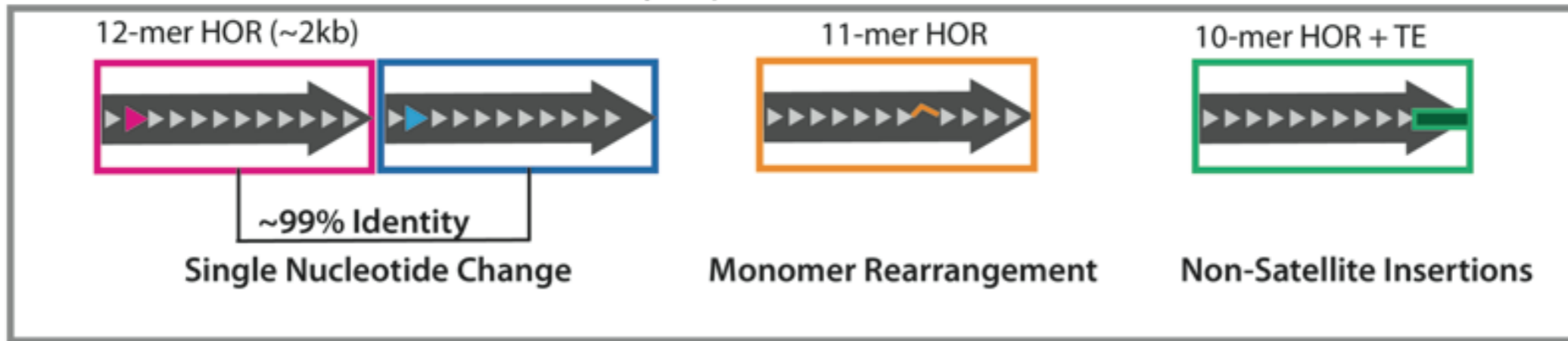


2 LinearSat Software to Convert Reads to Linear Reference Models



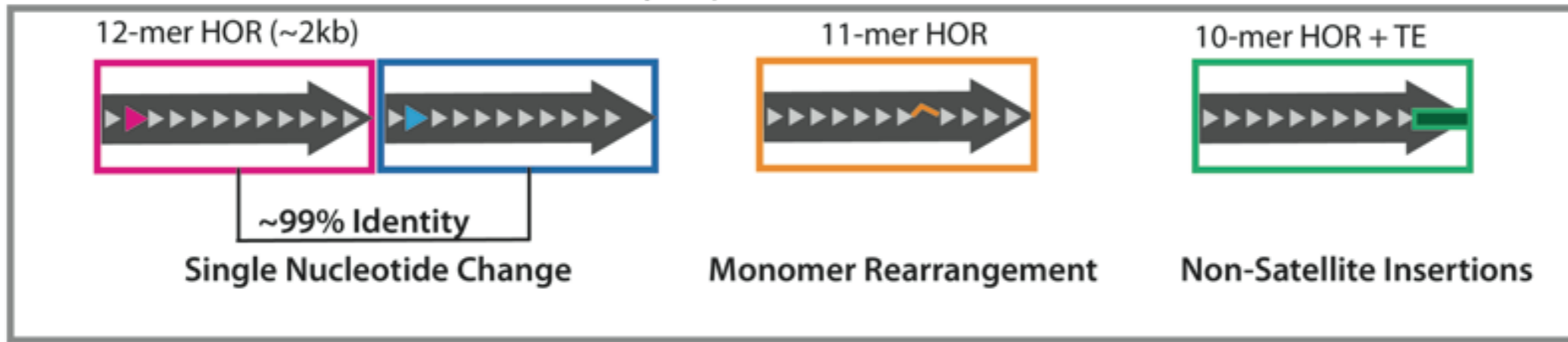


Array Sequence Variants

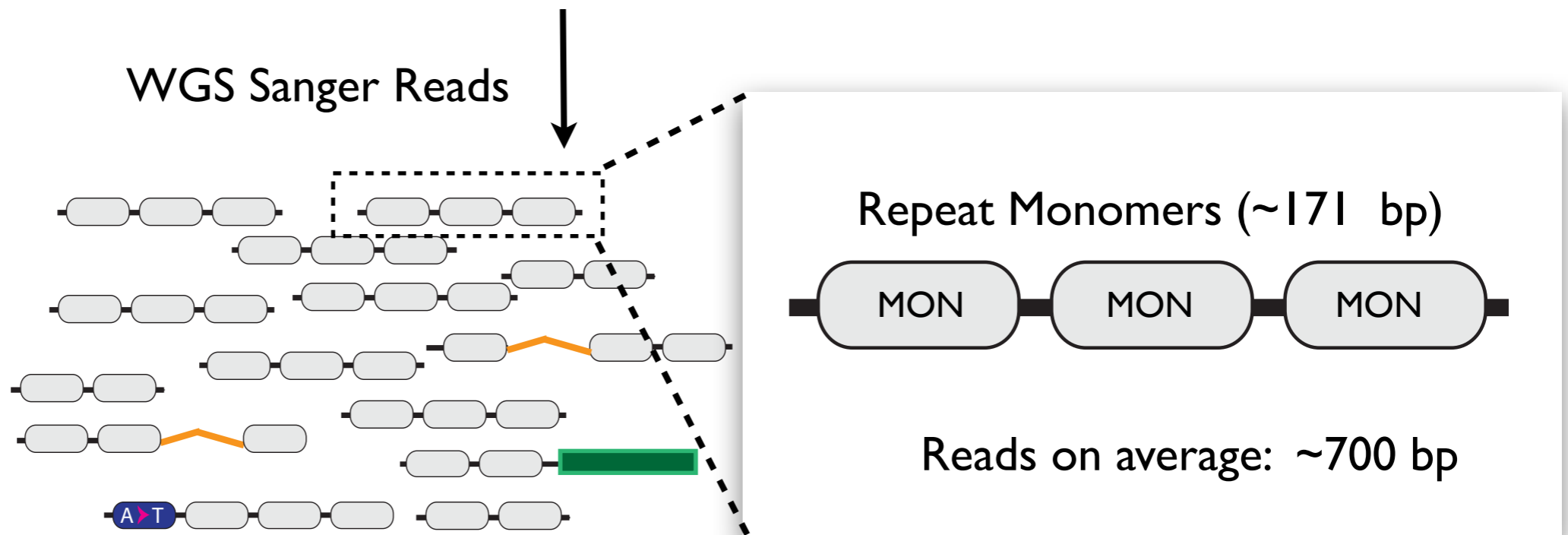




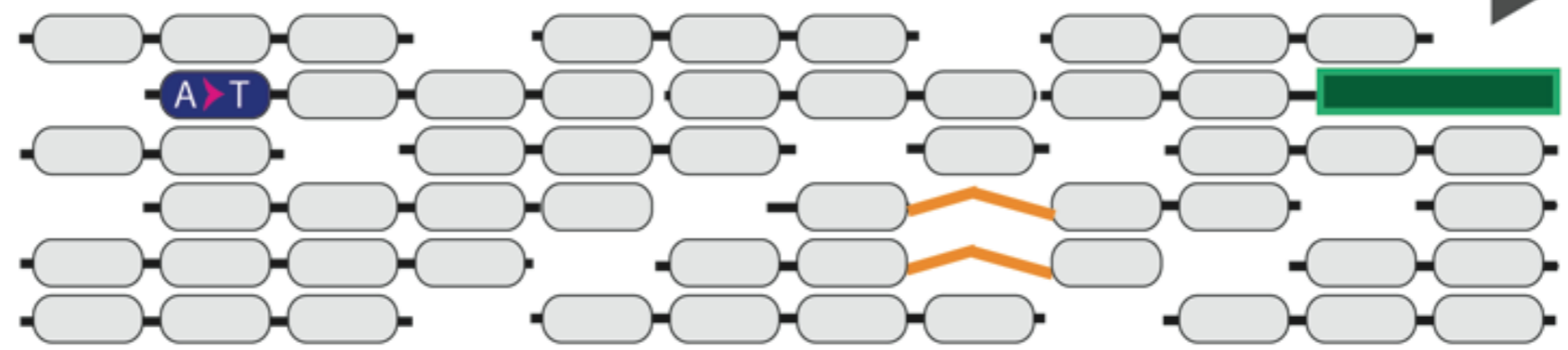
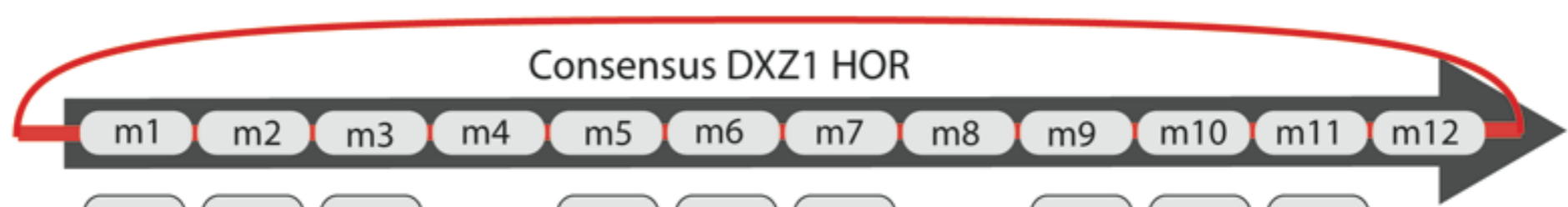
Array Sequence Variants



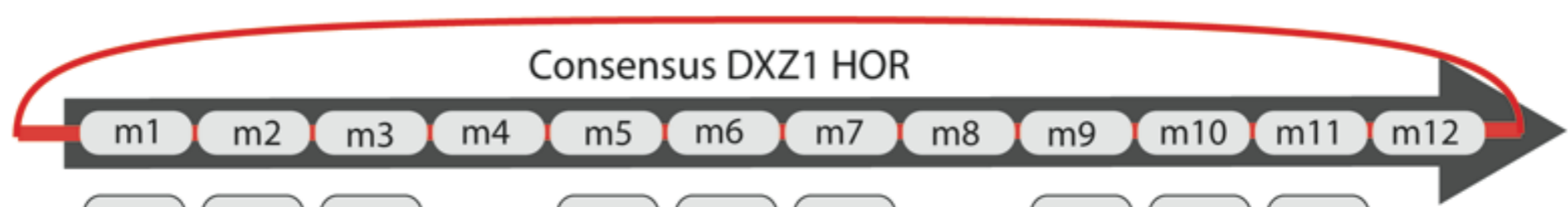
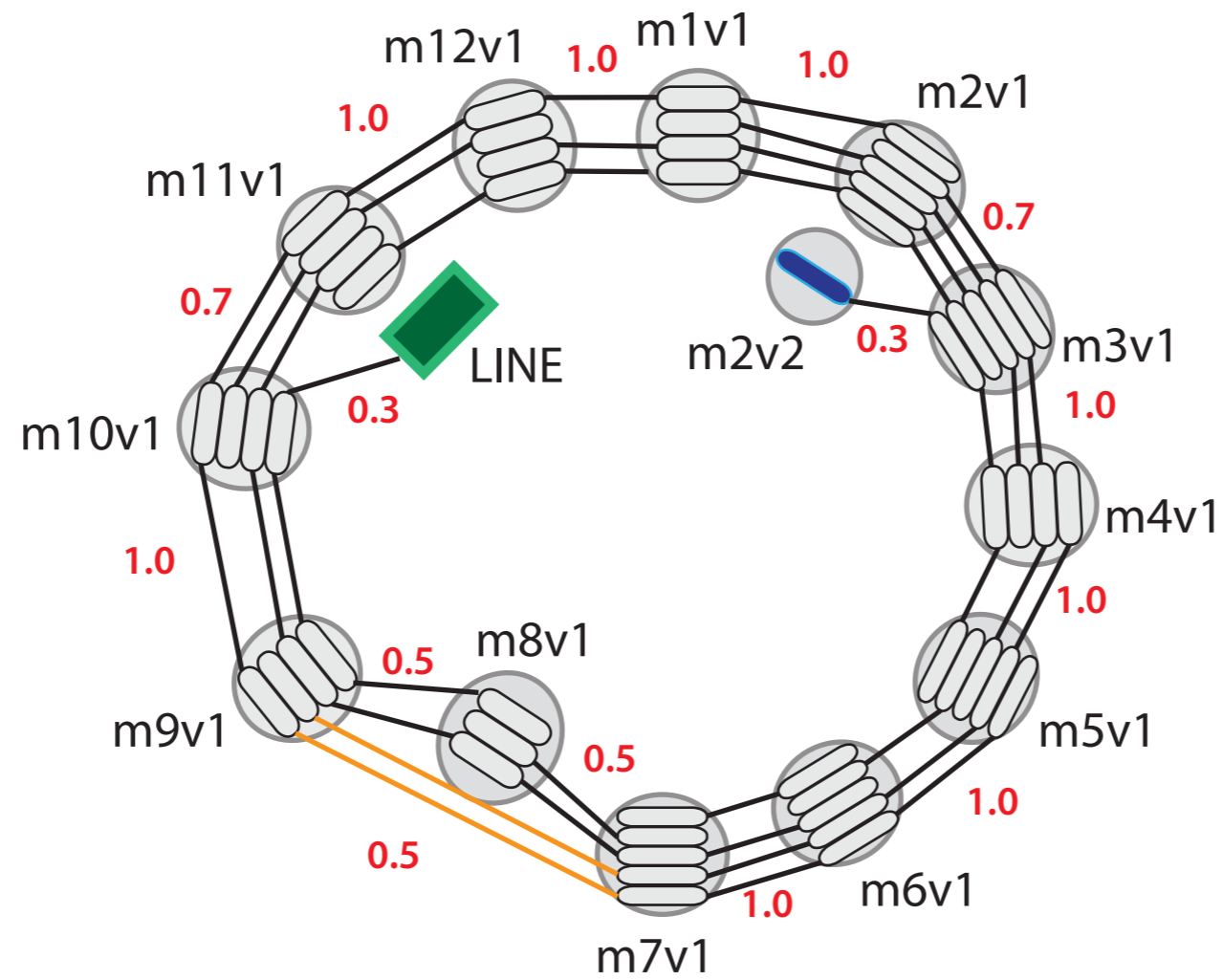
WGS Sanger Reads



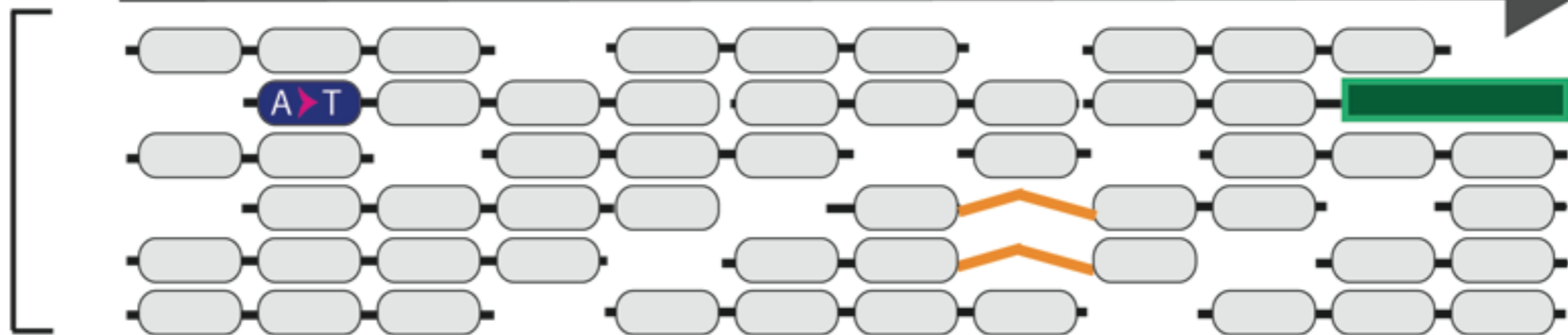
Consensus DXZ1 HOR

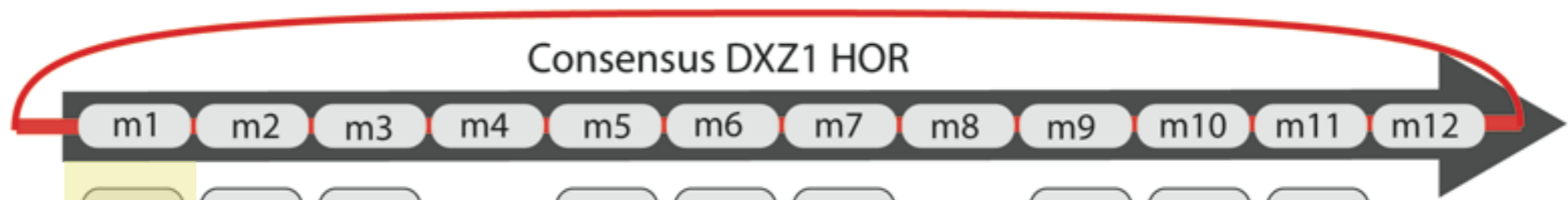
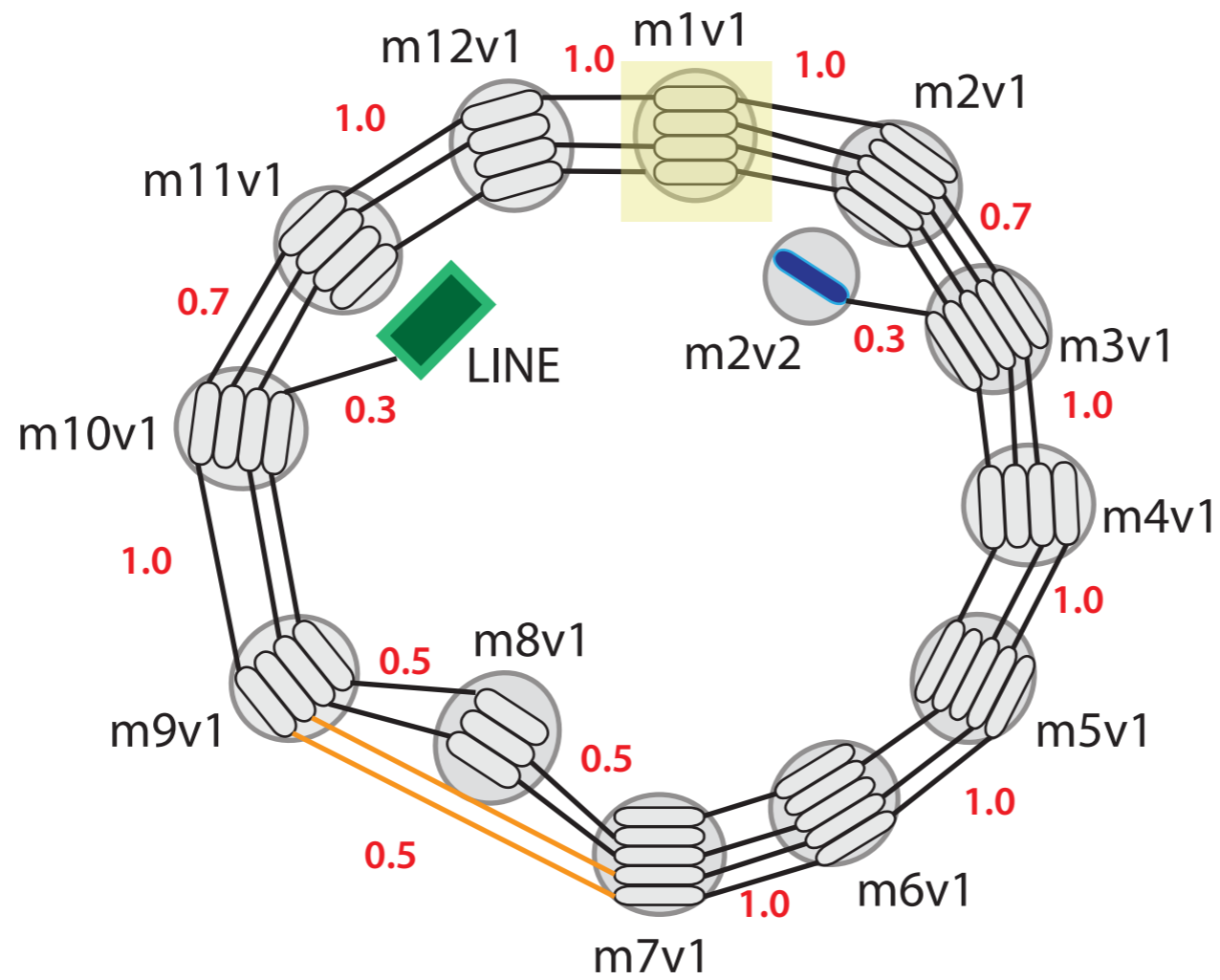


WGS Read Database:

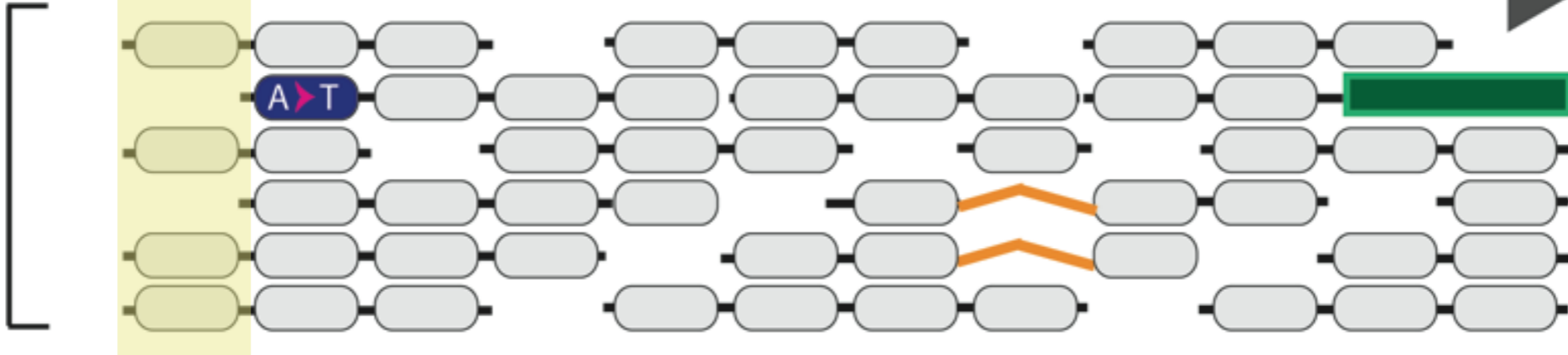


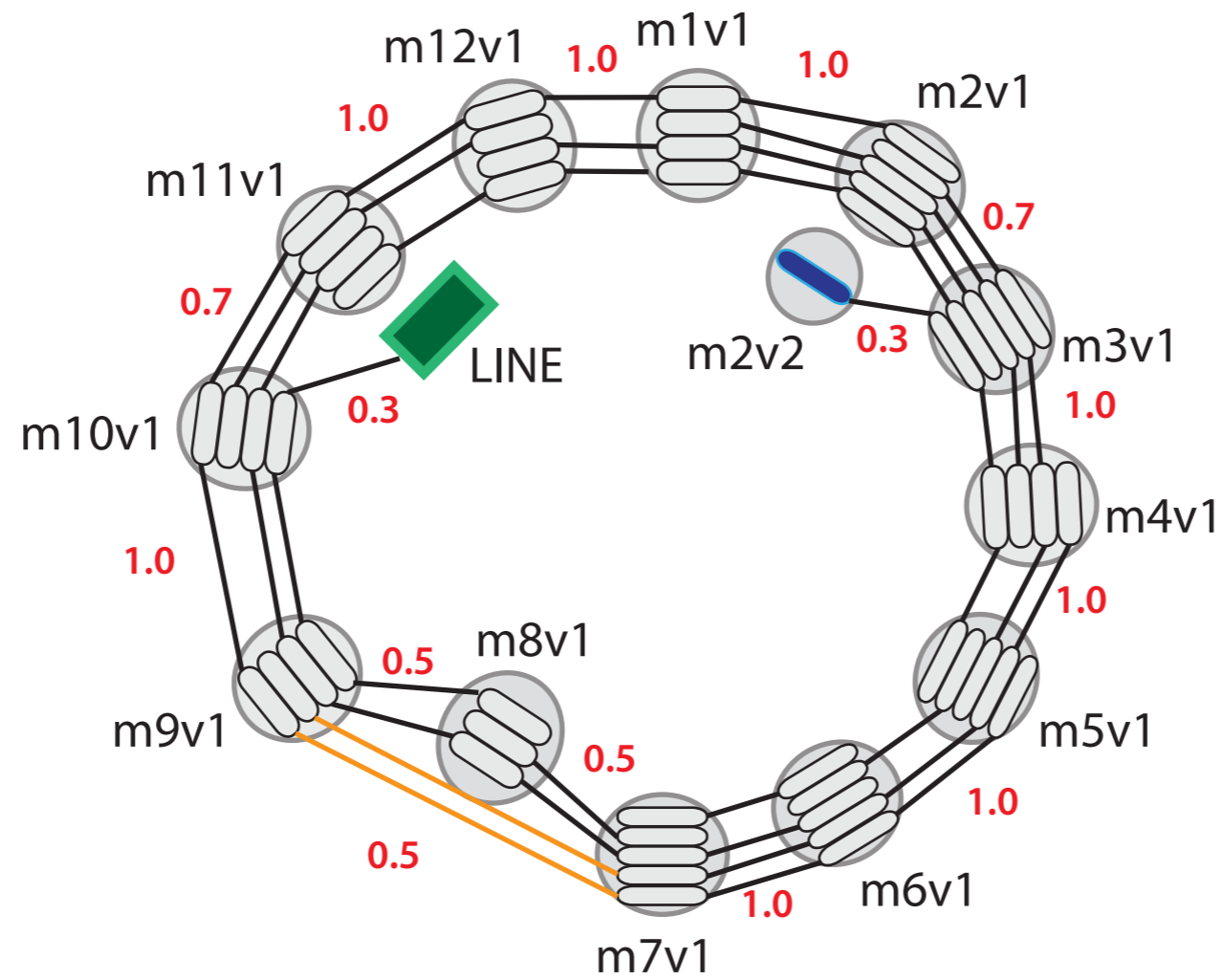
WGS Read Database:





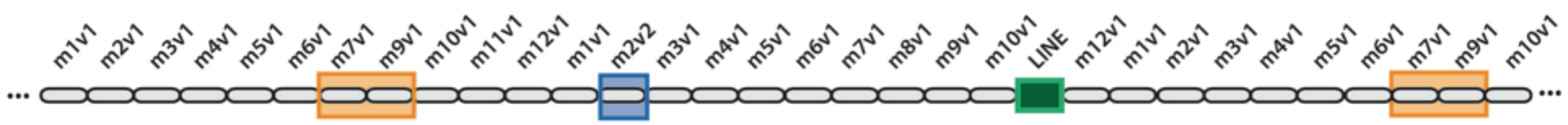
WGS Read Database:



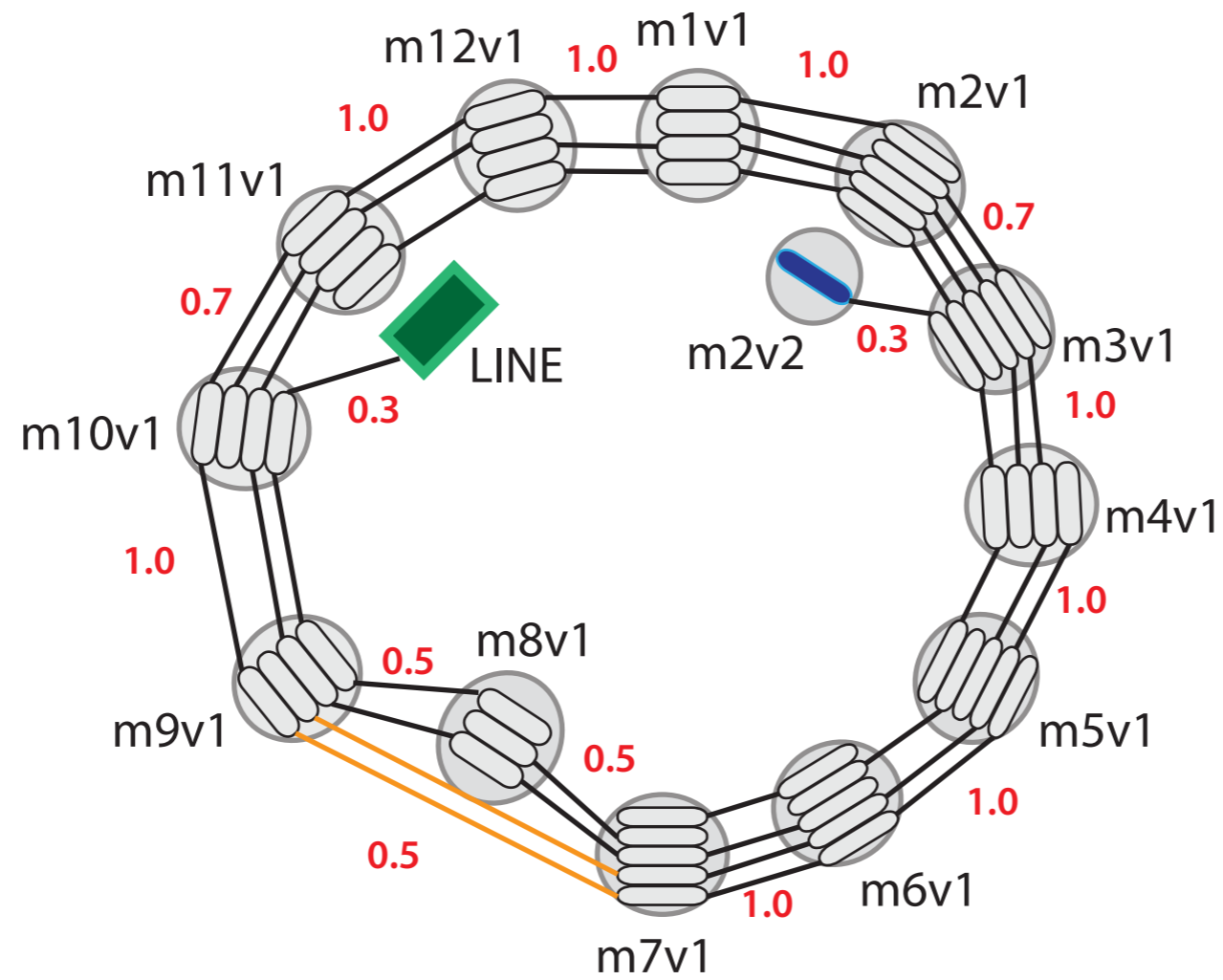


LinearSat

- 2nd Order Markov Chain
- Length determined by normalized array length estimates



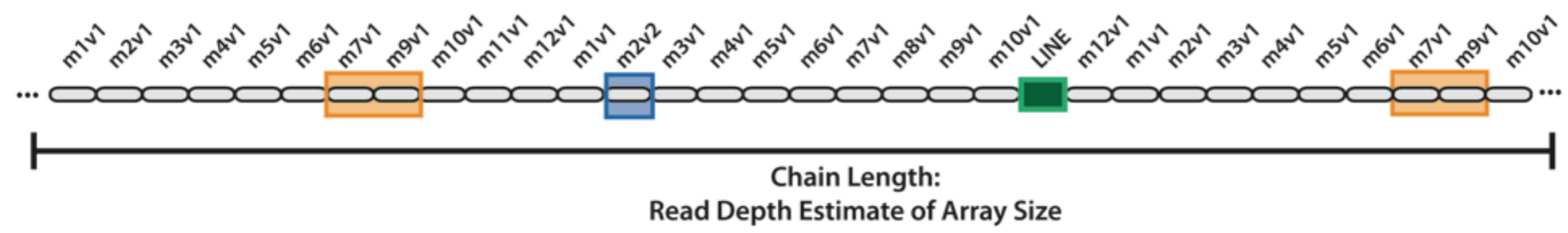
Chain Length:
Read Depth Estimate of Array Size



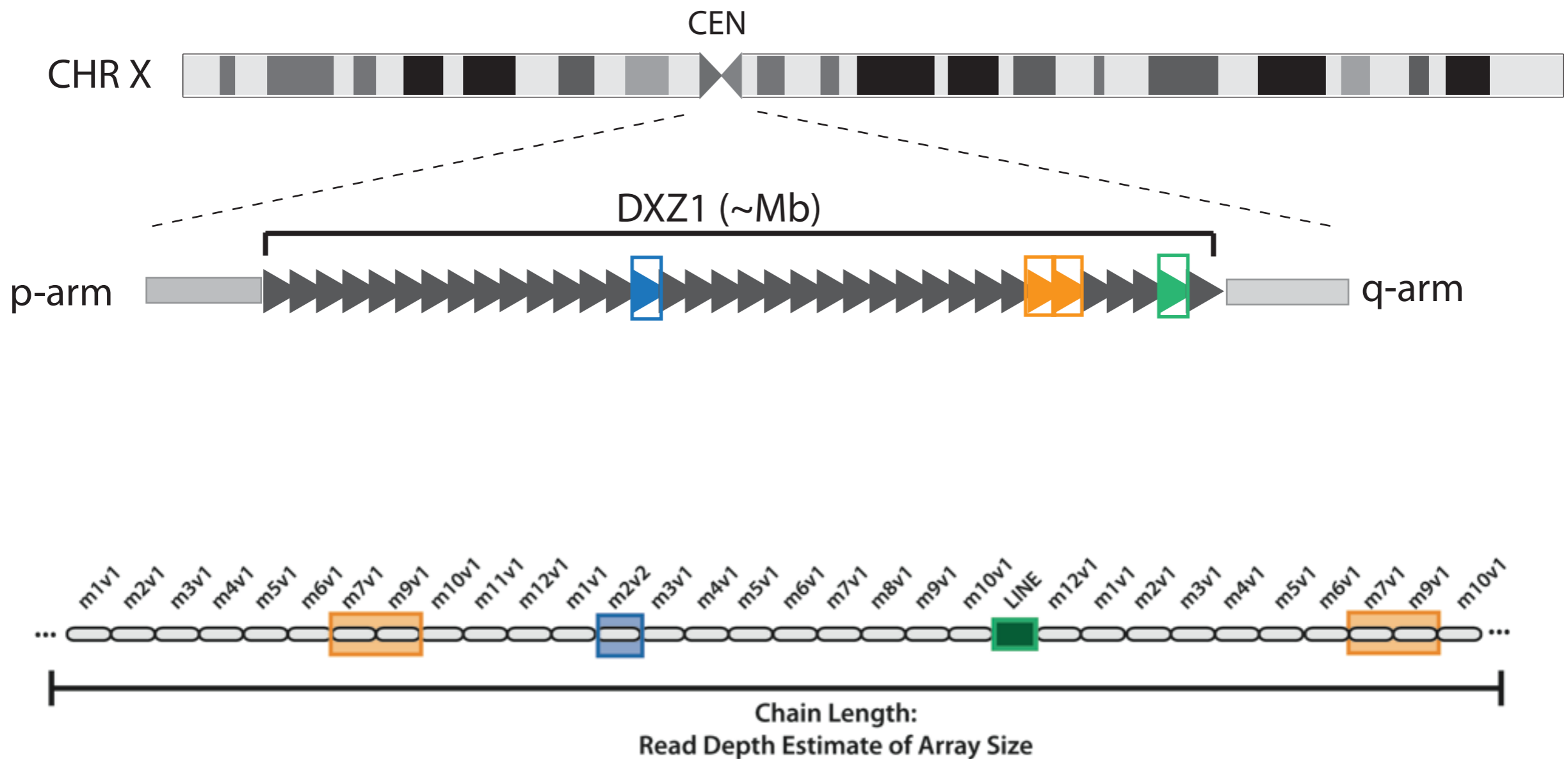
Not the “true” long-range organization, yet adequately represents the alpha satellite array sequence

LinearSat

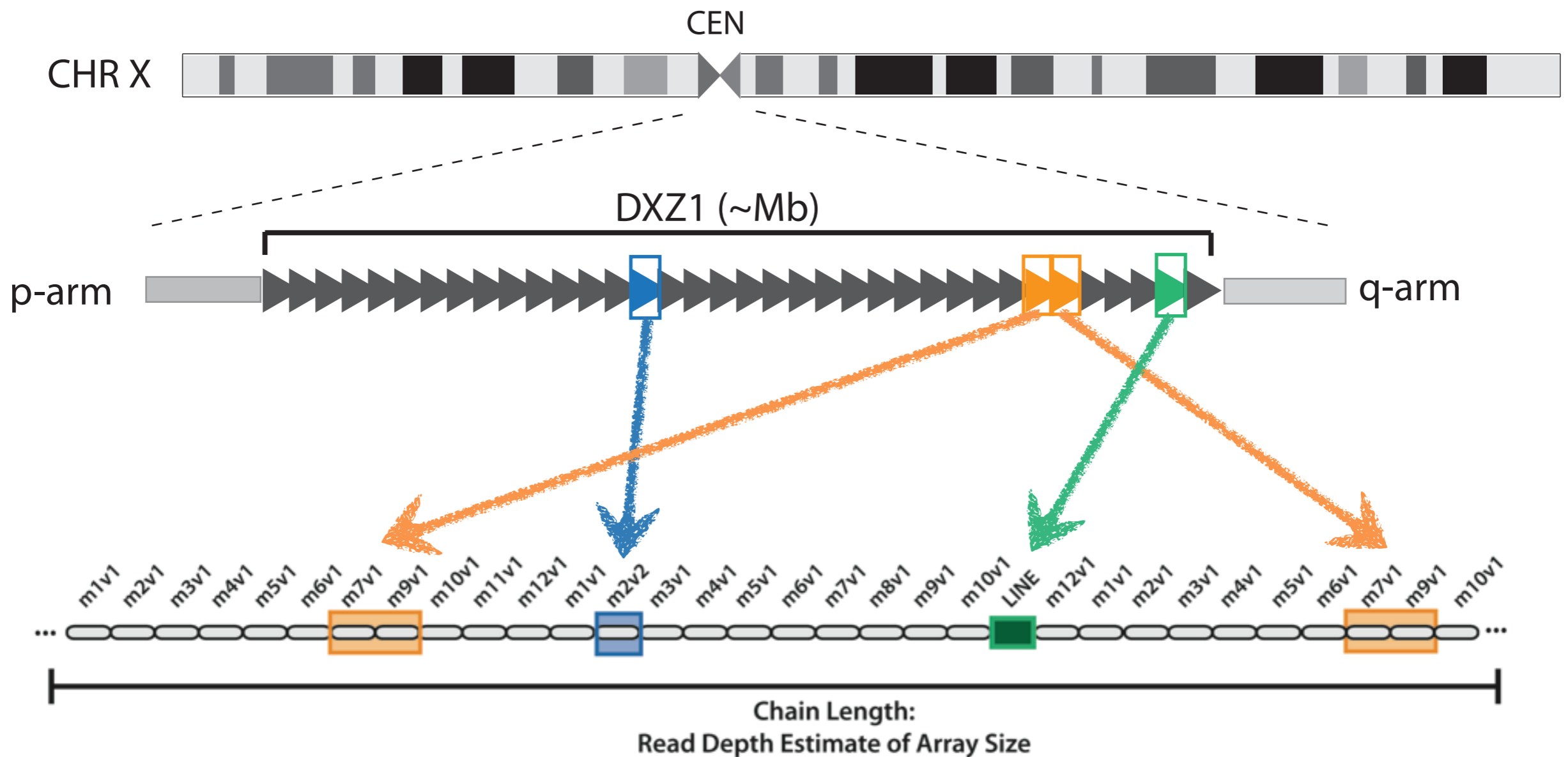
- 2nd Order Markov Chain
- Length determined by normalized array length estimates



Provide a linear array “model” of satellite sequence variation, proportional to that observed in the original array.



Provide a linear array “model” of satellite sequence variation, proportional to that observed in the original array.



Foundation Data Structure:

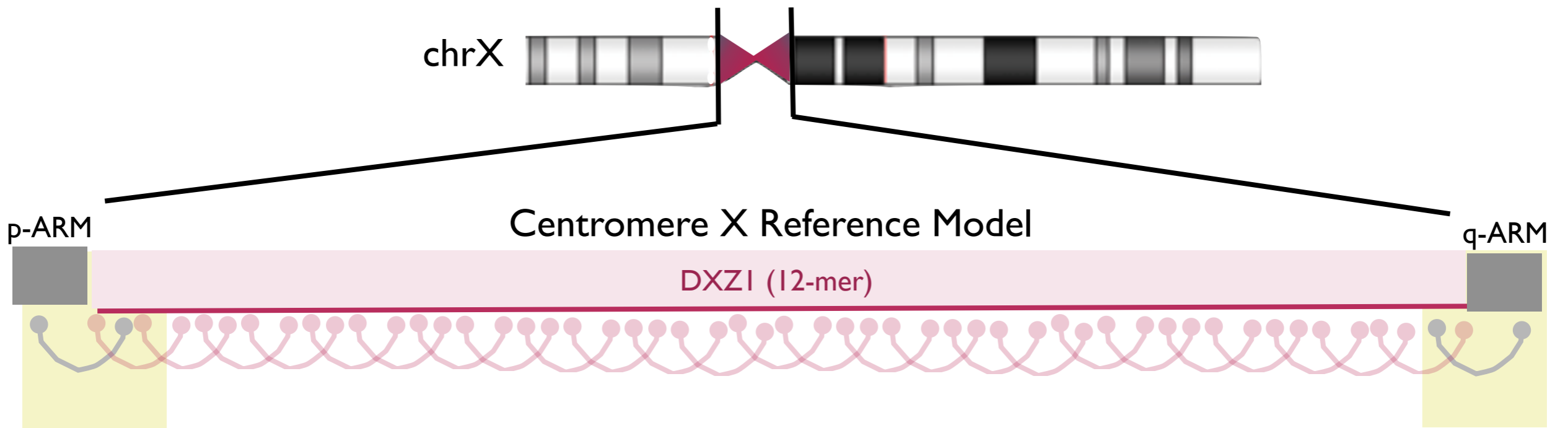
We could assemble these sequence using the *in house* alpha satellite probabilistic model

Positive:

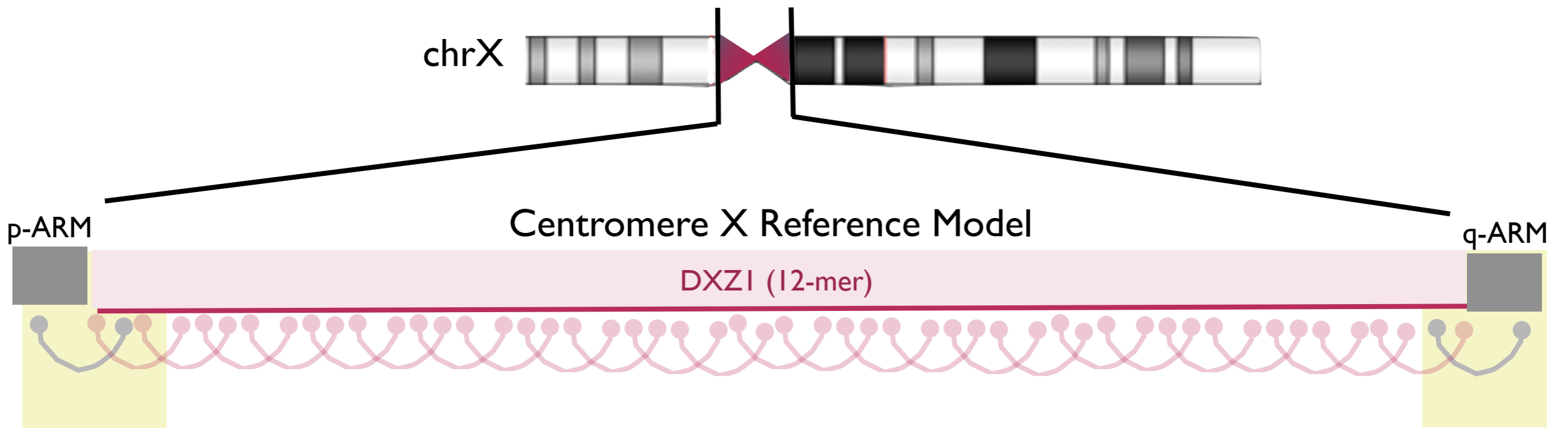
- Full coverage, high quality genome
- Read depth is meaningful (copy number est)
- All monomers (& sat adj seq) have a genome location

Negative:

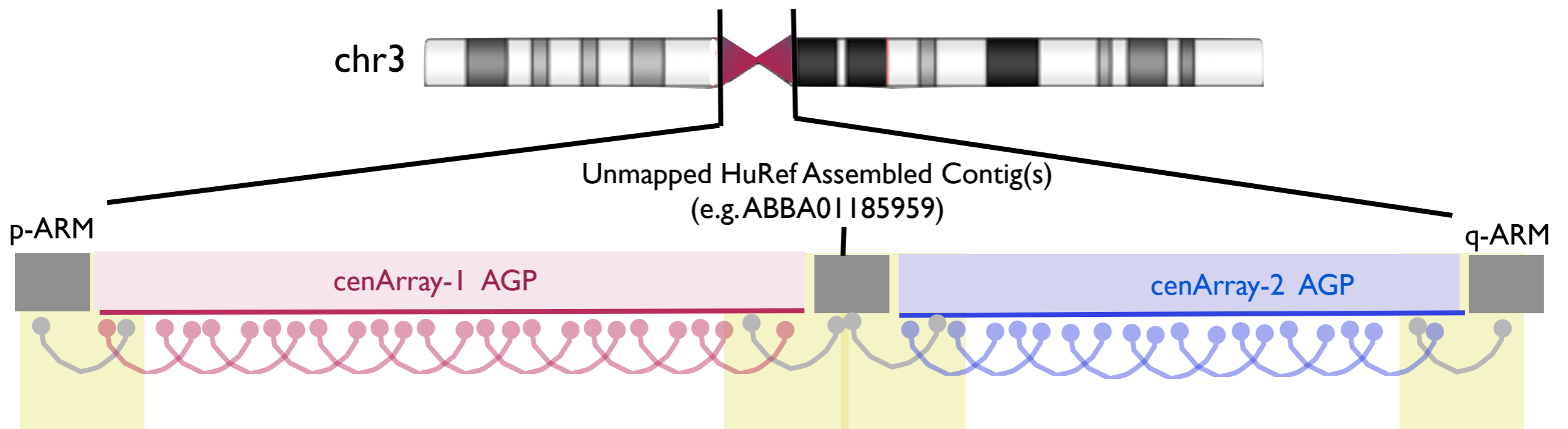
- Forced to make a number of assumptions in linear order

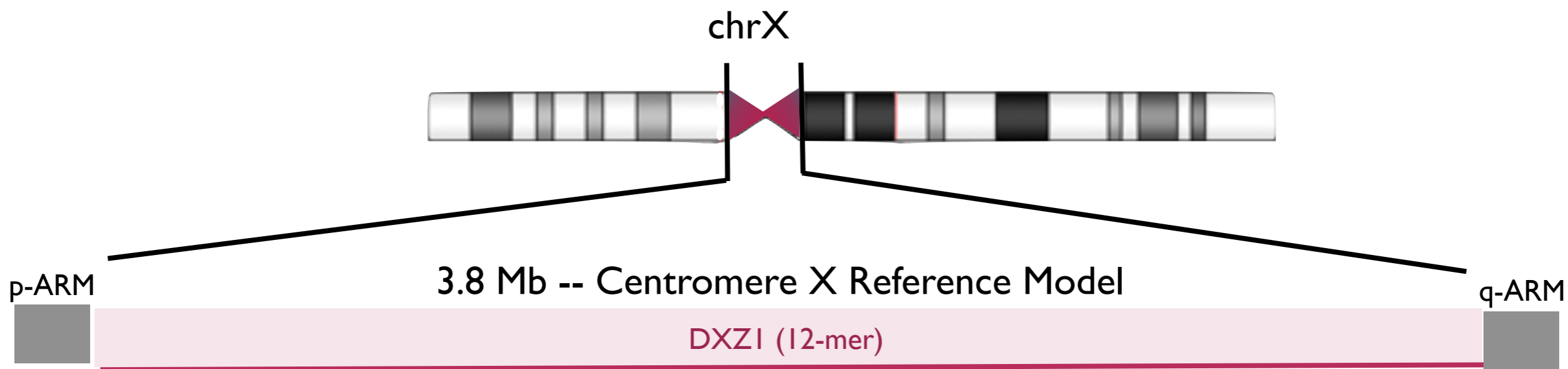


Scaffold models and assembled contigs using mate pairs



Scaffold models and assembled contigs using mate pairs





NCBI Resources How To

Nucleotide Limits Advanced

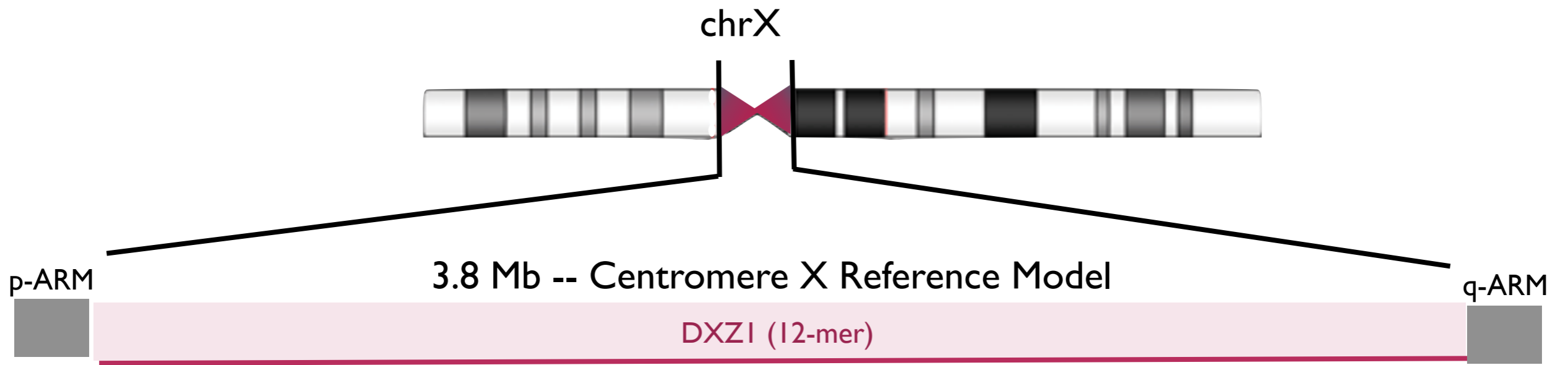
Display Settings: FASTA Send to:

TPA_asm: Homo sapiens chromosome X map Xcen assembly LinearCen1.0, whole genome shotgun sequence

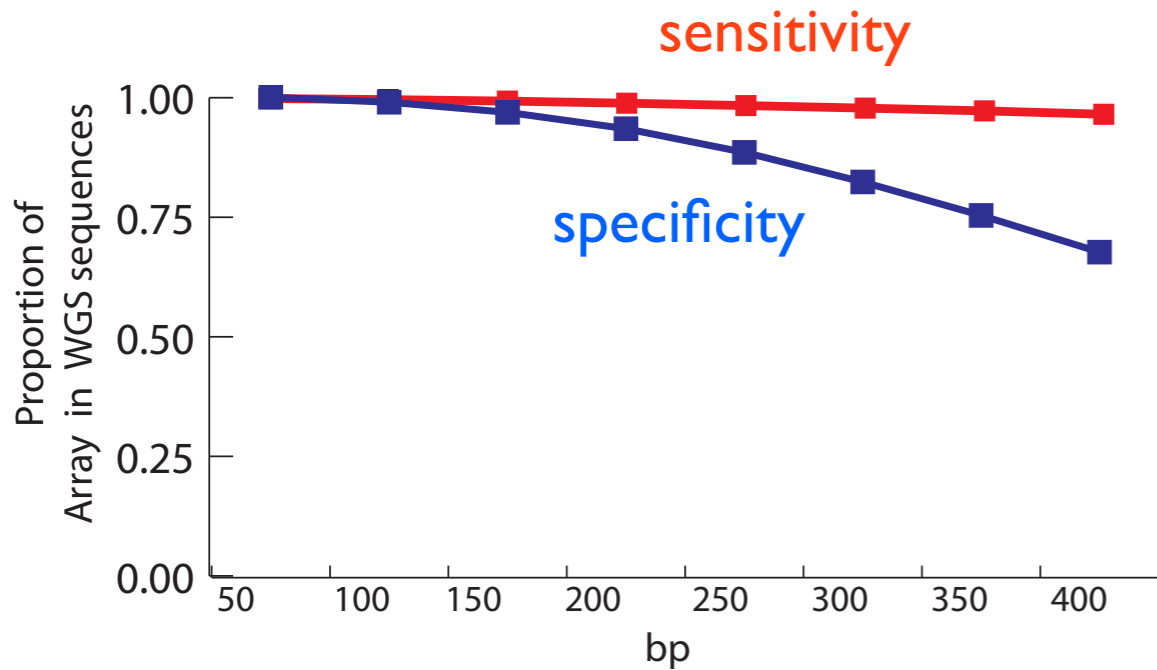
GenBank: GK000058.1

[GenBank](#) [Graphics](#)

```
>gi|529053718|tpg|GK000058.1| TPA_asm: Homo sapiens chromosome X map Xcen
assembly LinearCen1.0, whole genome shotgun sequence
AGCATTCTCAGAAACGACTTTGTGAGGATGGCATTCAACTCATGGAGTTGAACAATCCCATTGATAGAGC
AGATTGGAATCAGCTCTTTTGTAGAACTGCAAAATGGAGATTTGGACTGCTTTGGGGCCACGGTAGTAT
AGGAAGGAACTTCATATAAAAGGCCAAACGGGAAGCATTCTCAGAAATATCTTTGTGATGATGGAGTTTAC
TCACAGAGCTGAACGTGCCTTTTGTAGGAGCAGTTTCCAAATACACTTTTGGTAGAATCTGCAGGTGGAT
ATTTGGAGCTCTCGGAAGATTTGTTGAAACGGGAATAATTTCCATAACTAAACACAAACACGCTGAG
AAAGTTCTCATGATGAATGCATTTAACTCGCAGAGATGAACCTGCCTTTGAGAGTTTCAAGTTGAAACA
CTCTTTCTGTAGAACTCTGCAAGTGGATATTTGGACCCTGGCTGGCTTCGTTTCGAAACGGGTATATGTT
CACGTAAAAACTAAAGAGAAGCGTTCTCAGAAACTTCTGAGTGATGATTGCATTCAAGTCACACAGTTGA
ACCTTCTTTTGTGATTGAGCAGTTTGAACCTGCTTTTTGTAGAACTGTAAAGTGTATGCGTGGACCTCT
TTGAAGATTTCTTTGAAACGGGAATAATTTCCACAGAAAACTAAACTGAAGCATTCTCAGAAACCGCTT
TGTGATGTTTGTGTTTCGAGCGACAGAGTTTAACTTGCCTTTTCATAGAGCAGTTTGAATATCTTTTGG
GCAGAACTGCAAGTGGACATTTGGAGCGCTTTTCAGGCCTGTGGTGGCAAAGGCCTGAAAGCCTTTTCT
TTATCTTACAGAAAGACGAGAGAGAAGCATTGTCAGAAACTTCTTTGTGATGATGCAATCAACTCACA
GAGTTGAAGATTTCTTTGAAACAGCAGTTTTCGAAACACTCTTTCTGTGGGATCCGCAAGGGGATATTTG
GACCTCTTTGAAAGTTTCTGTTGAAACGGGATAATCTTCACTAAAGCTAAACGGGAAGCATTCTCAGAA
ACTTCTTTGGGATGTTTGCATTACACTCAGAGTTGAACTTTCCCTTTGATAGCGCAGCTTTGACACAC
TTTTTCTACAATGTGCAAGTGGCTATTTAGCGGGCTTGGAGGACTGTGTTGAAAGGAAATACTCTTCTA
AAAACGACATAGAAGCATTCTCAGAAACTGCTCTGTGATGATTGCATTCAACTCCAGAGTTGAACATTC
CTTTGATAGAGCAGTTTGCAGAACTCTTTTGTAGAACTCTGCAAGTGGAGATTTGGACCGCTTTGAGG
CCTGTGGTAGTGAAGGAAAGAACTTCATATAAAAACCAGAGCGTAGCACTCTCAGAAAAATCTTTGTGAC
GATGGAGTTTAACTCAGGGAGCTGAACATTCGTTATGATGGAGCAGTTTCCAAACACATGTTTGTAGAA
TCTGCGAGGGGATATTTGGACCTCTGAGGATTTCTGTTGAAACGGGATCAACTTCCATAACTGAACG
GAAGCAACTCAGAACATTTCTTTGTGATGTTTGTATTCATTCACAGAGTTGAACCTTCTTTGATAGTT
CAGGTTTGCAGAACCTTTGTAGTAGAATCTGCAAGTGTATATTTGACCACTTTGTAGCCTTCTGTTGAA
ACGCTTATATCTTACATCAAACTAGACAGAAGCATTCTCAGAAAGTTTCTGCGATGACTGCATTCAA
CTCACAGAGTTGAACAATCTTCTGATGGAGCAGTTTGTATACCTCTTTCTTTGGAATCTGCAAGGGGA
```

K-mer Comparison with Original DXZI read dataset



NCBI Resources How To

Nucleotide Limits Advanced

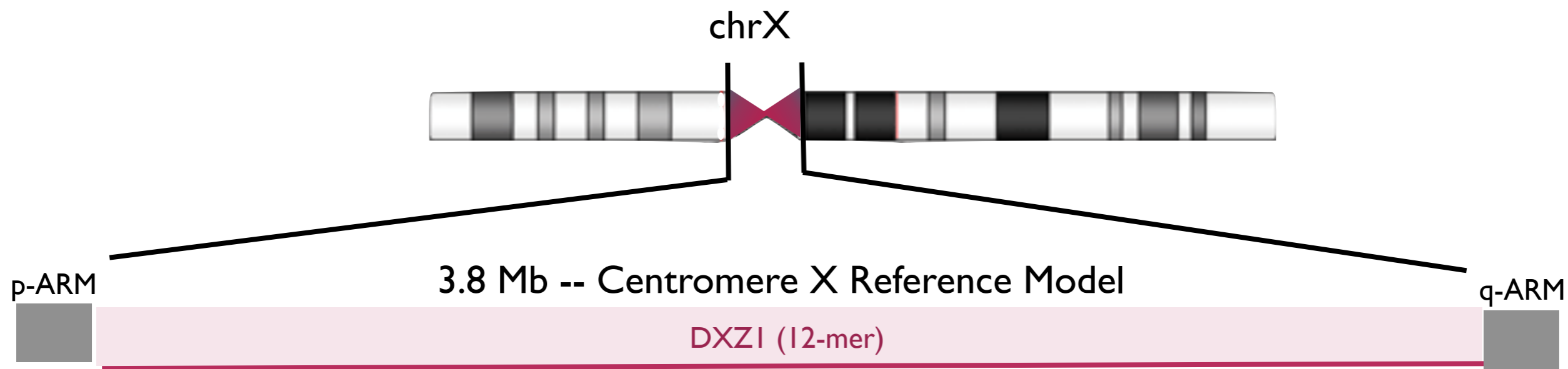
Display Settings: FASTA Send to: [v]

TPA_asm: Homo sapiens chromosome X map Xcen assembly LinearCen1.0, whole genome shotgun sequence

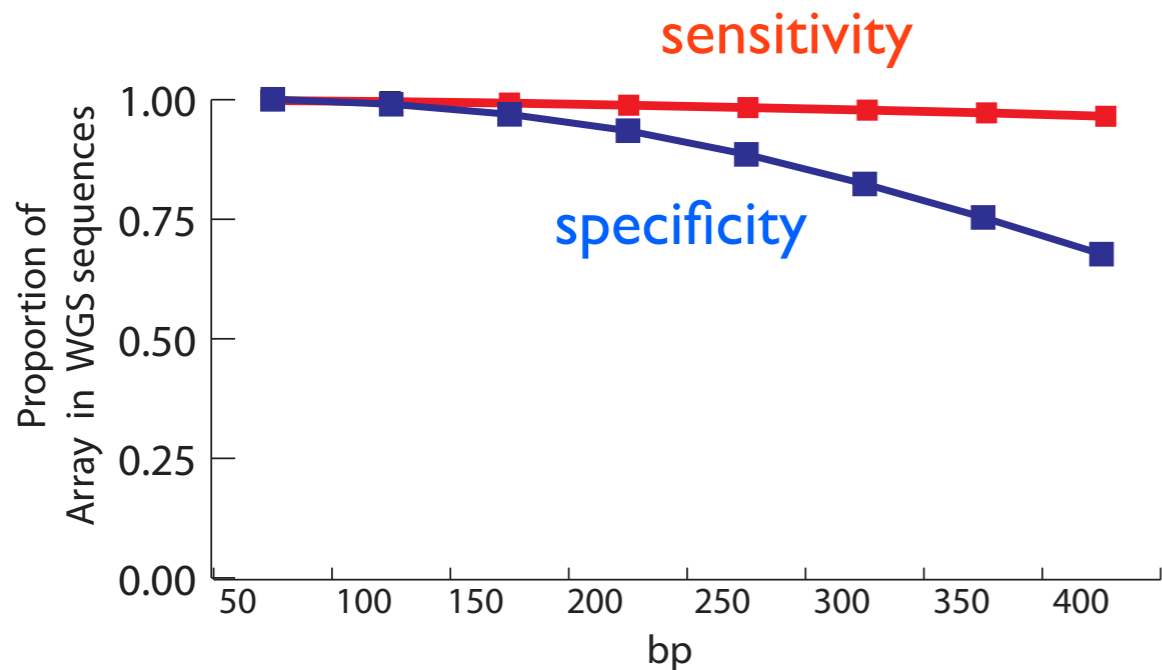
GenBank: GK000058.1

[GenBank](#) [Graphics](#)

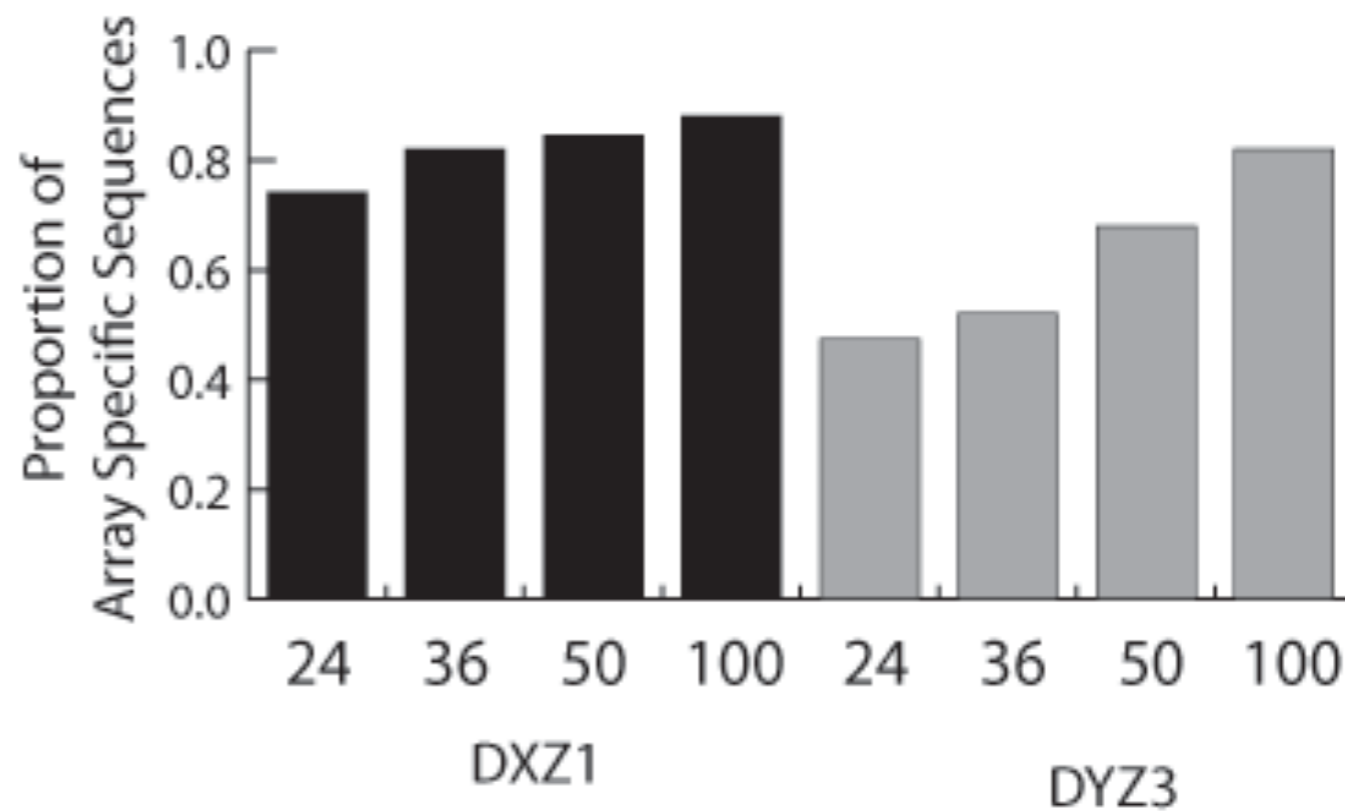
```
>gi|529053718|tpg|GK000058.1| TPA_asm: Homo sapiens chromosome X map Xcen
assembly LinearCen1.0, whole genome shotgun sequence
AGCATTCTCAGAAACGACTTTGTGAGGATGGCATTCAACTCATGGAGTTGAACAATCCCATTGATAGAGC
AGATTGGAATCAGCTCTTTTGTAGAAATCTGCAAAATGGAGATTTGGACTGCTTTGGGGCCACGGTAGTAT
AGGAAGGAACTTCATATAAAAGGCCAAACGGAAGCATTCTCAGAAATATCTTTGTGATGATGGAGTTTAC
TCACAGAGCTGAACGTGCCTTTTGTAGGAGCAGTTTCCAAATACACTTTTGGTAGAATCTGCAGGTGGAT
ATTTGGAGCTCTCGGAAGATTTCTGTTGAAACGGGAATAATTTCCATAACTAAACACAAACACGCTGAG
AAAGTTCTTCATGATGAATGCATTTAACTCGCAGAGATGAACCTGCCTTTGAGAGTTTCAGGTTGAAACA
CTCTTTCTGTAGAAATCTGCAAGTGGATATTTGGACCCTGGCTGGCTTCGTTTCGAAACGGGTATATGTT
CACGTAAAAACTAAAGAGAAGCGTTCTCAGAAACTTCTGAGTGATGATTGCATTCAAGTCACACAGTTGA
ACCTCTCTTTGATTGAGCAGTTTGAACCTGCTTTTTGTAGAAATCTGTAAGTGTATGCGTGGACCTCT
TTGAAGATTTCTTTGAAACGGGAATAATTTCCACAGAAAACTAAACTGAAGCATTCTCAGAAACCGCTT
TGTGATGTTTGTGTTTCGAGCGACAGAGTTTAAACATTGCTTTTCATAGAGCAGTTTGAATATCTTTTG
GCAGAACTGCAAGTGGACATTTGGAGCGCTTTTCAGGCCTGTGGTGGCAAAGGCCTGAAAGCCTTTTCT
TTATCTTCACAGAAAGACGAGAGAGAAGCATTGTCAGAAACTTCTTTGTGATGATGCAATCAACTCACA
GAGTTGAAGATTTCTTTTGAACAGCAGTTTTCGAAACACTCTTTCTGTGGGATCCGCAAGGGGATATTTG
GACCTCTTTGAAGTTTCTGTTGAAACGGGATAATCTTCACCTAAAAGCTAAACGGAAGCATTCTCAGAA
ACTTCTTTGGGATGTTTGCATTTCACCTCAGAGTTGAACTTTCCCTTTGATAGCGCAGCTTTGACACAC
TTTTCTACAATGTGCAAGTGGCTATTTAGCGGGCTTGGAGGACTGTGTTGGAAAAGGAAATATCTTTTA
AAAACGACATAGAAGCATTCTCAGAAACTGCTCTGTGATGATTGCATTCAACTCCAGAGTTGAACATTC
CTTTGATAGAGCAGTTTGCAAACACTCTTTTGTAGAAATCTGCAAGTGGAGATTTGGACCGCTTTGAGG
CCTGTGGTGTGAAAGGAAAGAACTTCATATAAAAACAGAGCGGTAGCACTCTCAGAAAAATCTTTGTGAC
GATGGAGTTTAACTCAGGGAGCTGAACATTCGTTATGATGGAGCAGTTTCCAAACACATGTTTGTAGAA
TCTGCGAGGGGATATTTGGACCTCTCTGAGGATTTCTGTTGAAACGGGATCAACTTCCATAACTGAACG
GAAGCAACTCAGAACATTTCTTTGTGATGTTTGTATTCATTCACAGAGTTGAACCTTCTTTGATAGTT
CAGGTTTGAACACCTTGTAGTAGAATCTGCAAGTGTATATTTGACCCTTTGTAGCCTTCTGTTTGA
ACGCTATATCTTCACATCAAACTAGACAGAAGCATTCTCAGAAAGTTTCTGCGATGACTGCATTCAA
CTCACAGAGTTGAACAATCTCTGTAGTGGAGCAGTTTGTATACCTCTTTCTTTGGAATCTGCAAGGGGA
```



K-mer Comparison with Original DXZI read dataset



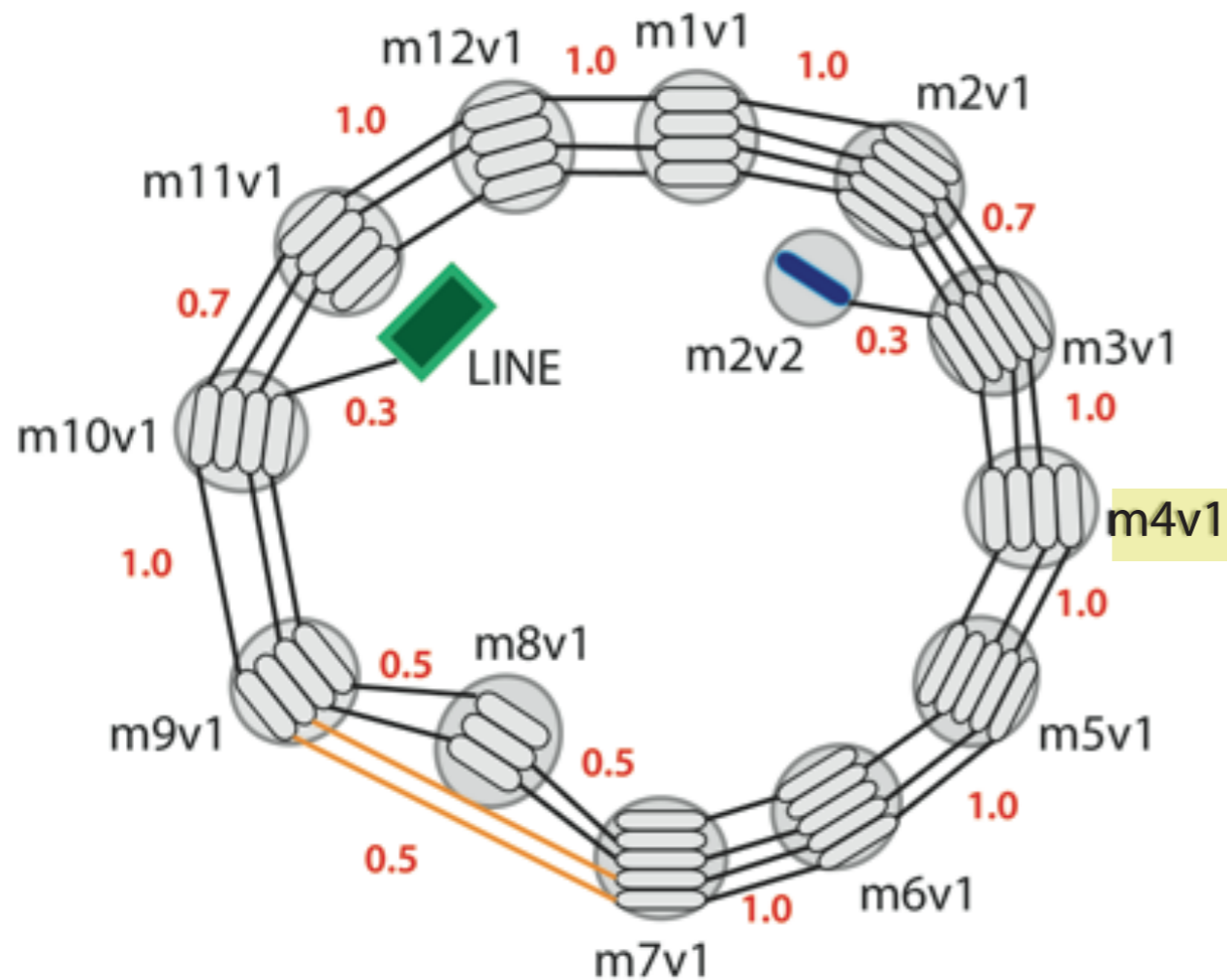
Mappability



Evaluate as a Potential Read Mapping Target

GRCh38 Data Structure

Level 1: Repeat Components



Database all unique sequence in each array graph

>m4v1 4 identical monomers

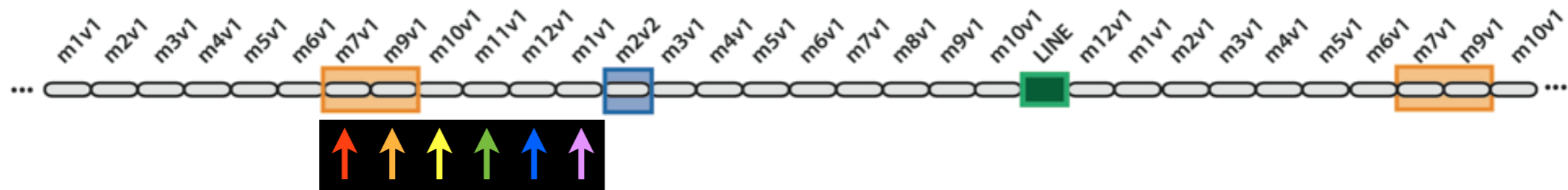
```
CACTTGCAGATTCTACAAAAGAGTGCTTCAAAC
TGCTCTGTCAAAGGAAGGTTCAACTCTGTTACTT
GAGTACACACATCACAAGGAAGTTTCTGAGAATGC
TTCTGTCTGGTTTTTAGGAGAAGATATTTCTTTT
TCAACATAGGCCTCAAAGCGCTGCAAATGTCCACT
TCC
```







Deposit (NCBI, TPA) individual component fasta sequence of each centromere reference model

GRCh38 Data Structure

Level 2: AGP describing the order of sequence components

Level 2: Centromere Reference Model "cenArray" AGP



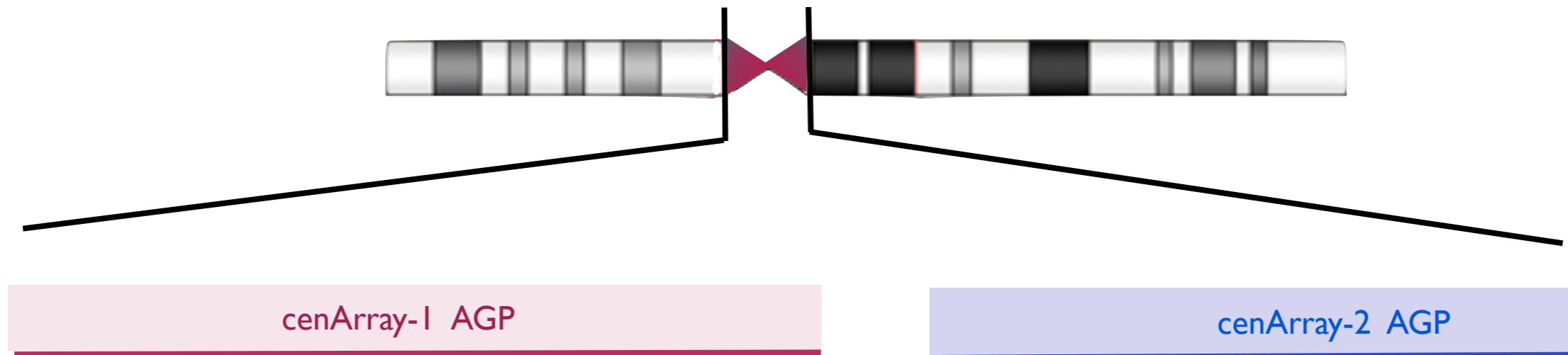
Array Name	Array Start	Array End	UID	UID	Level I Entry	LI Start	LI End	Level I Ori
 cenArray	129970	130138	759	O	m7v1	1	169	+
 cenArray	130139	130309	760	O	m9v1	1	171	+
 cenArray	130310	130608	761	O	m10v1	1	170	+
 cenArray	130609	130708	762	N	m11v1	1	171	+
 cenArray	130709	130878	763	O	m12v1	1	170	+
 cenArray	130879	131049	764	O	m1v1	1	171	+

Array Coordinates

Level I Sequence

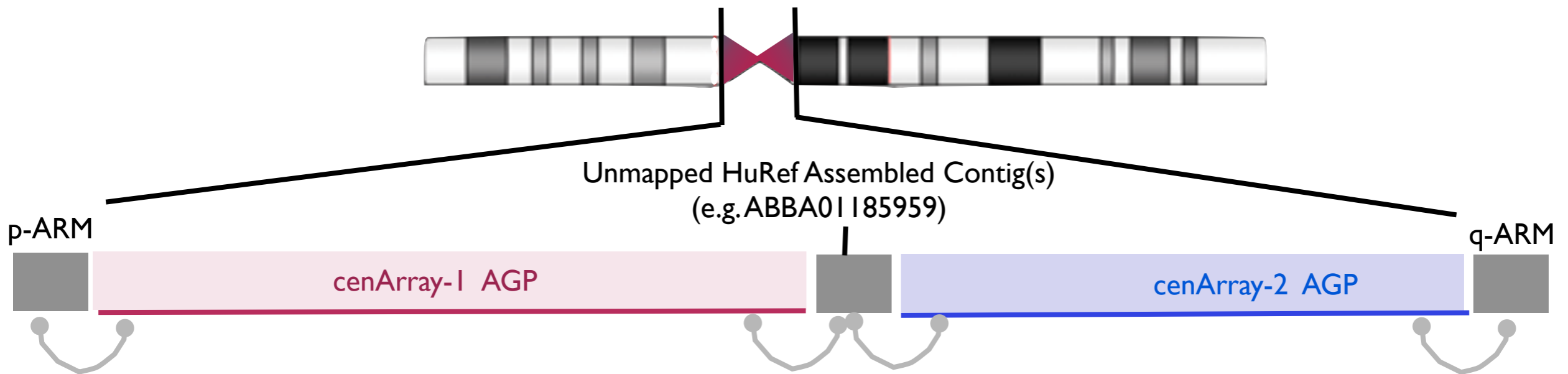
GRCh38 Data Structure

Level 3: AGP describing the order of Array components



Single centromeric gap can contain more than one array

3 Scaffold Reference Models and HuRef assembled contigs using mate pairs



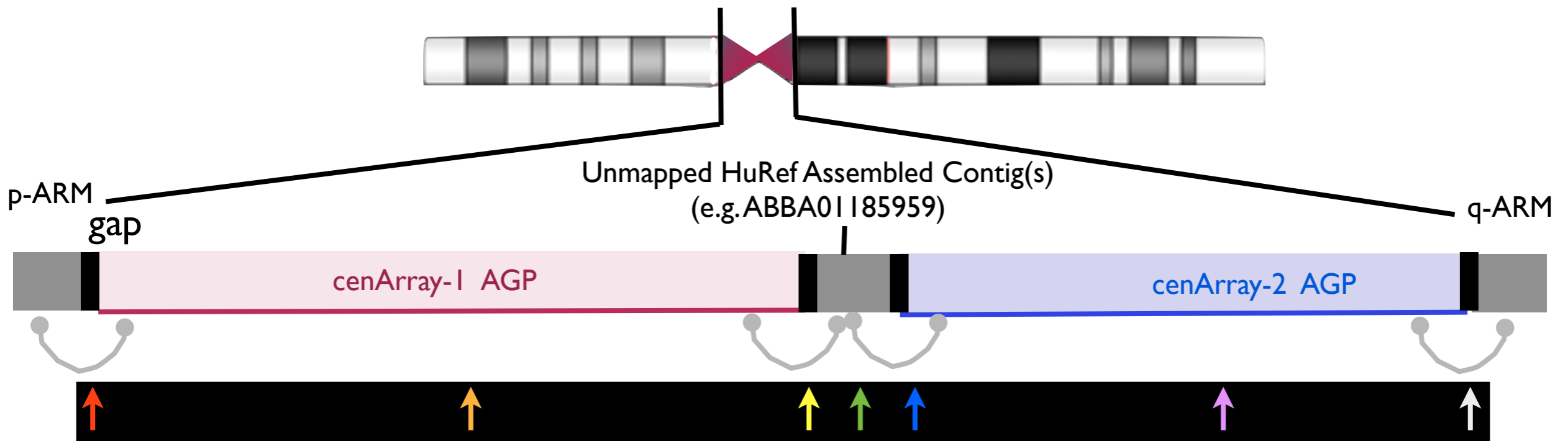
Single centromeric gap can contain more than one array








Scaffolding Order: Weighted by Mate Pairs

-- Bundled paired read information informs array component order

GRCh38 Data Structure

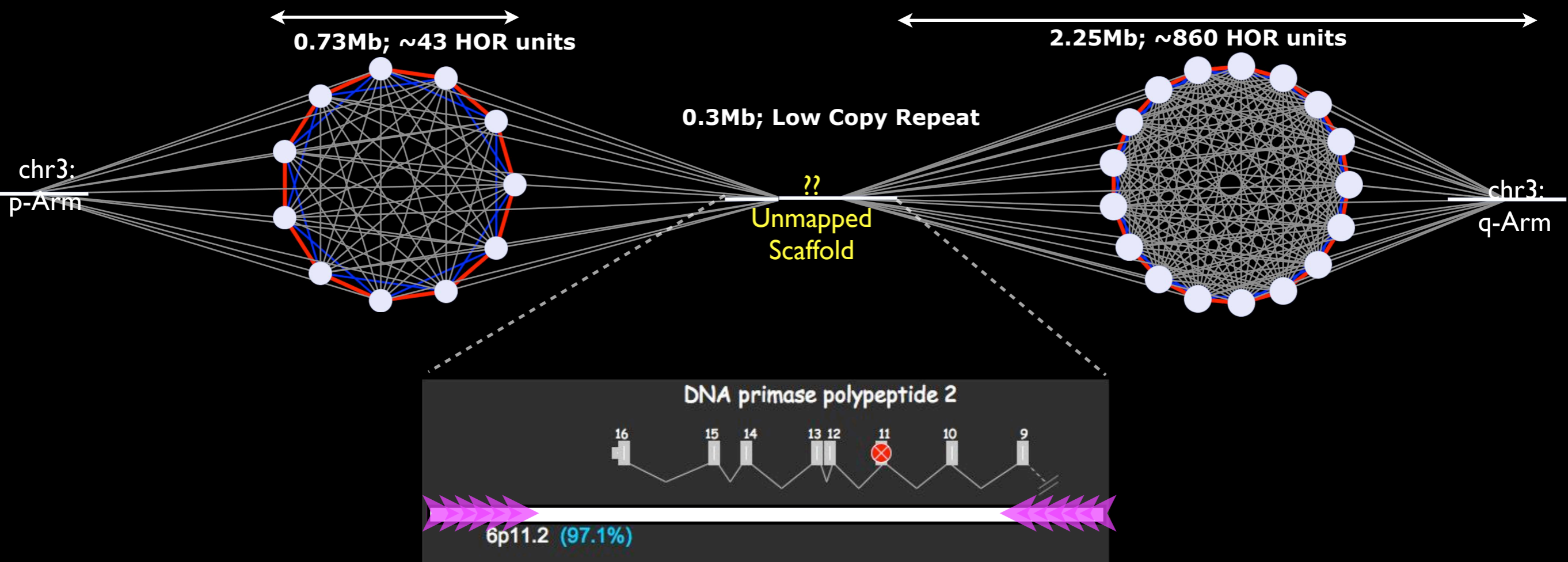
Level 3: AGP describing the order of Array components



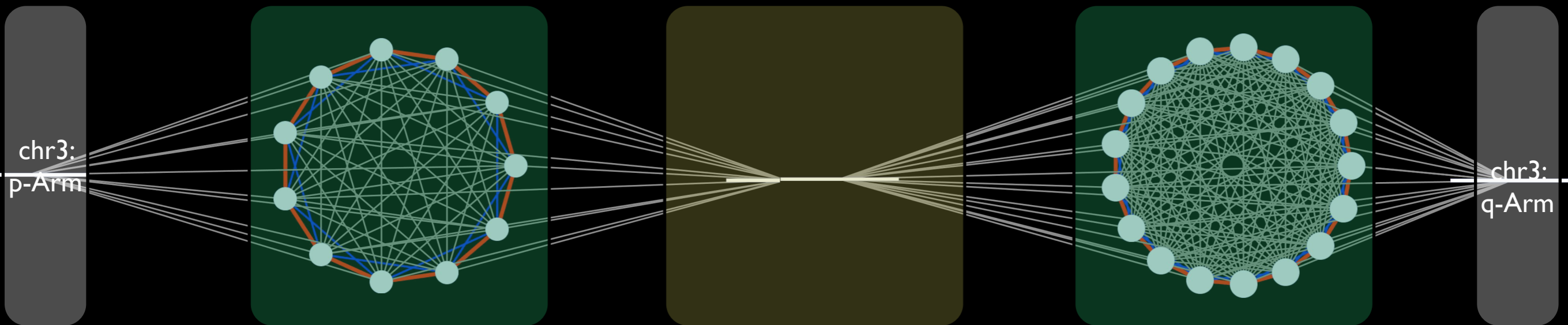
Array Name	Array Start	Array End	UID	UID	Level 1 Entry	LI Start	LI End	Level 1 Ori
 cenArray	0	100	1	N	p-ARM gap	1	100	paired-read
 cenArray	101	2836899	2	O	cenArray-1	1	2836798	+
 cenArray	2836899	2837000	3	N	gap	1	100	paired-read
 cenArray	2837000	2842055	4	O	ABBA01185959	1	5055	paired-read
 cenArray	2836899	2837000	5	N	gap	1	100	paired-read
 cenArray	2837001	4369982	6	O	cenArray-2	1	1532981	+
 cenArray	4369983	4370083	7	N	gap	1	100	paired-read

CEN Coordinates

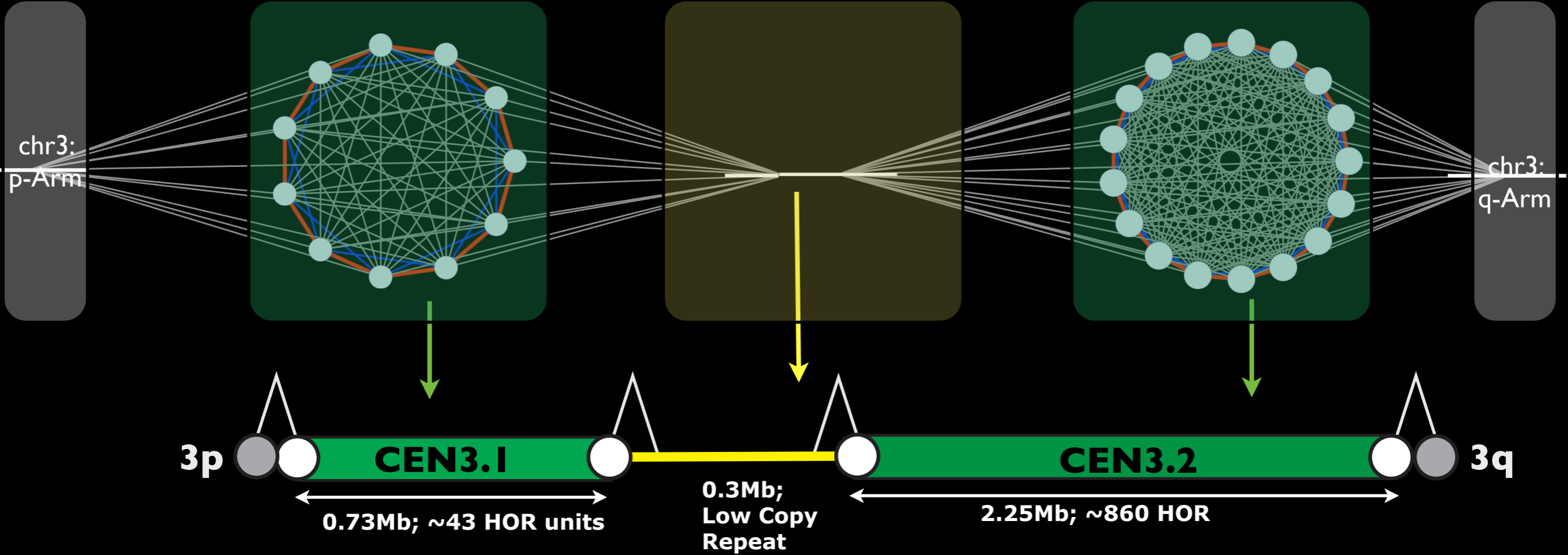
Level 2 Array Sequences

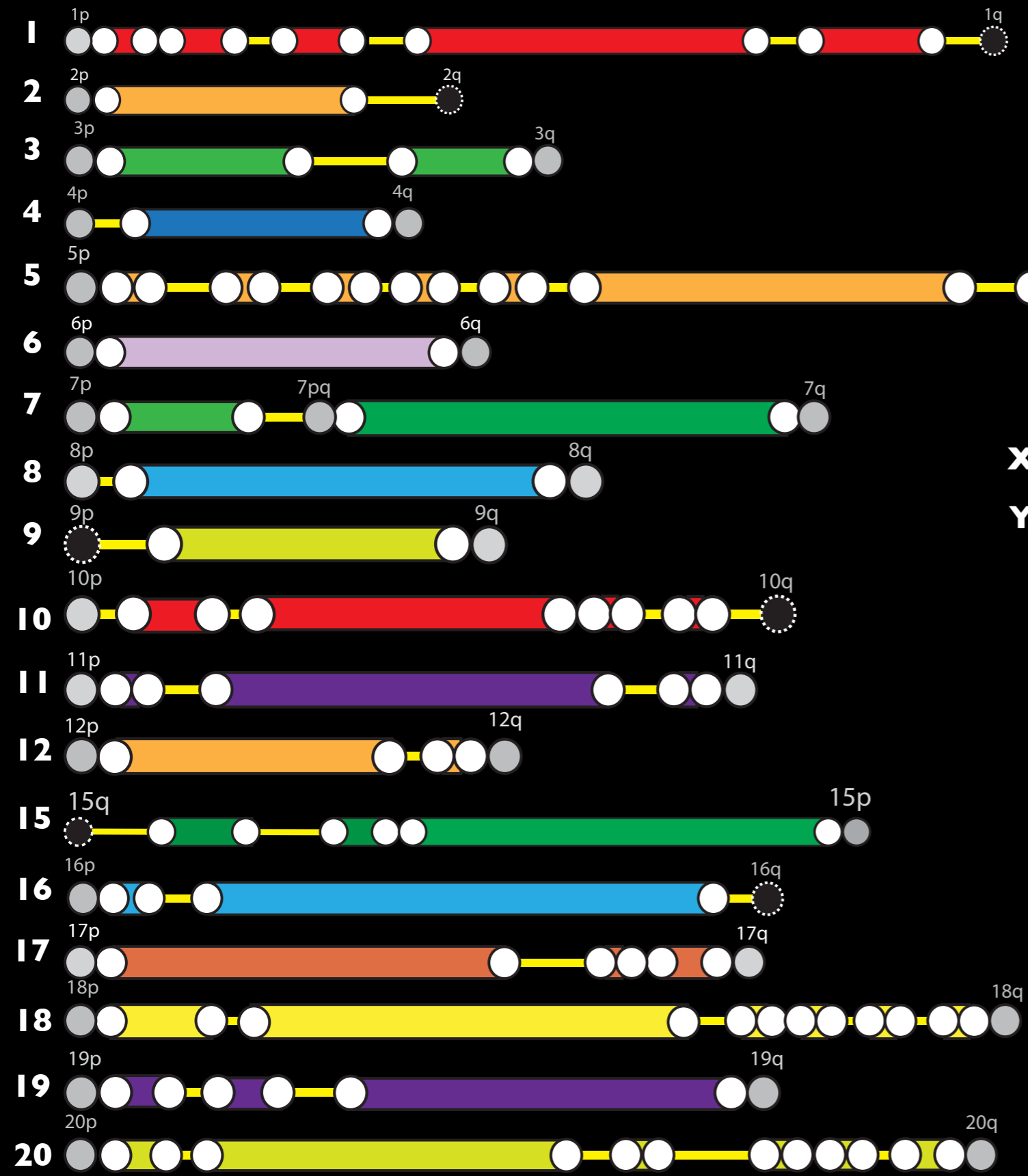


Scaffolding Problem: Order Elements by Paired Reads



Scaffolding Problem: Order Elements by Paired Reads



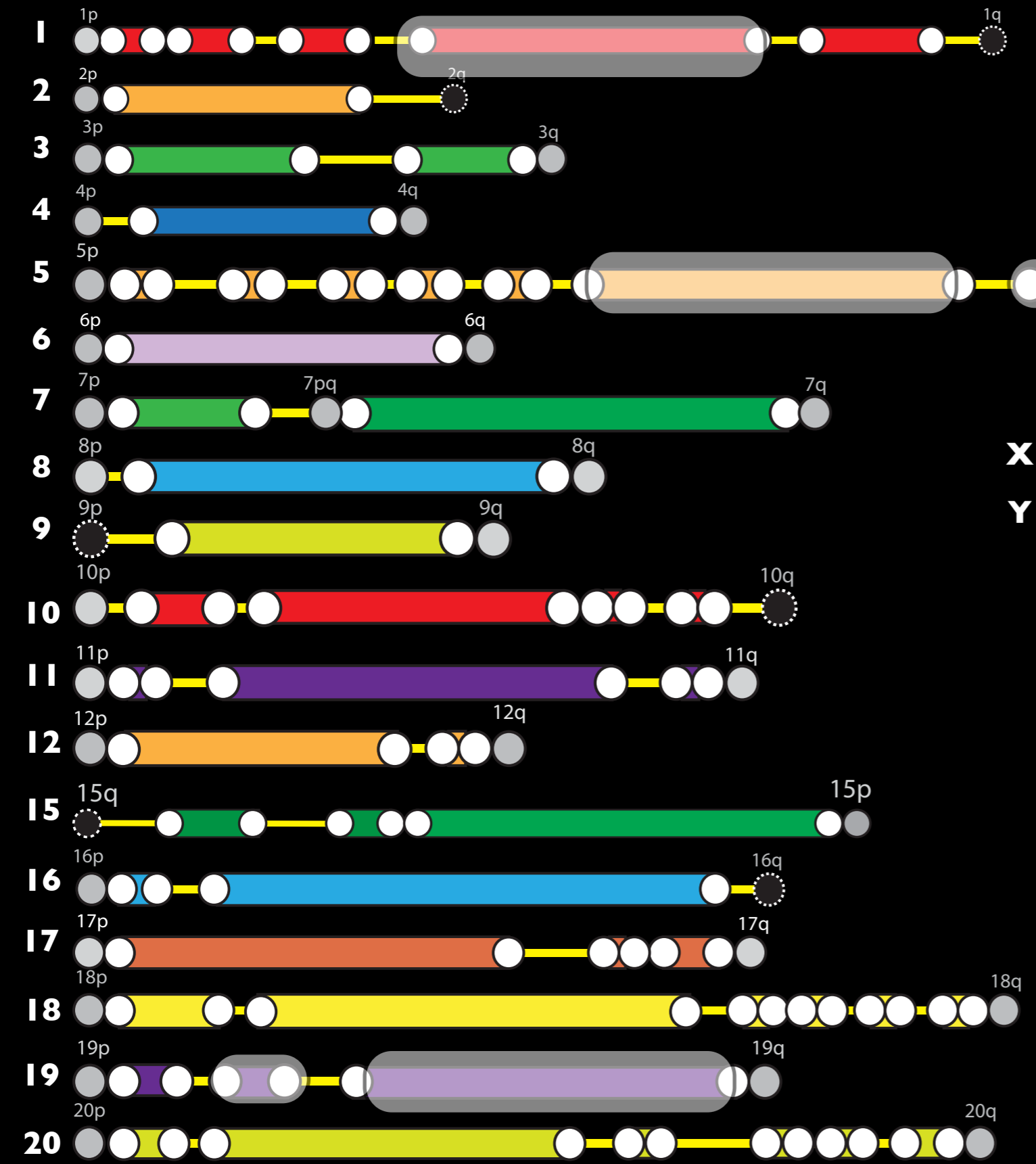


An Initial Draft of Human Centromere Sequence Composition

Alpha Satellite Reference Models:
~60 Mb (59571670 bp)

Acrocentric Chr
(13,14,21,22)





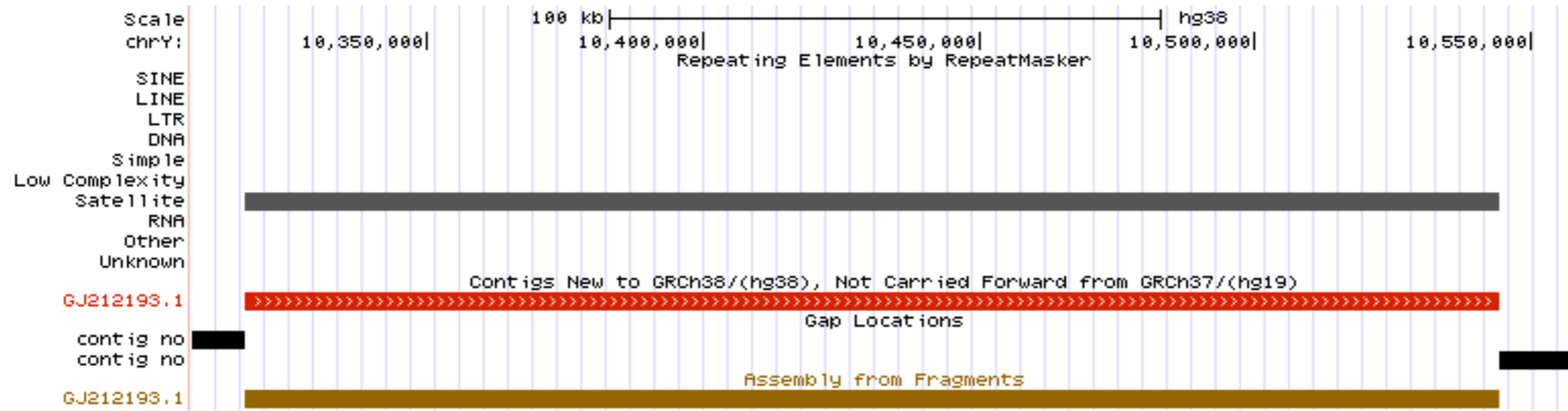
An Initial Draft of Human Centromere Sequence Composition

Redundant Arrays: Cannot assign to a specific chromosome that is normalized appropriately

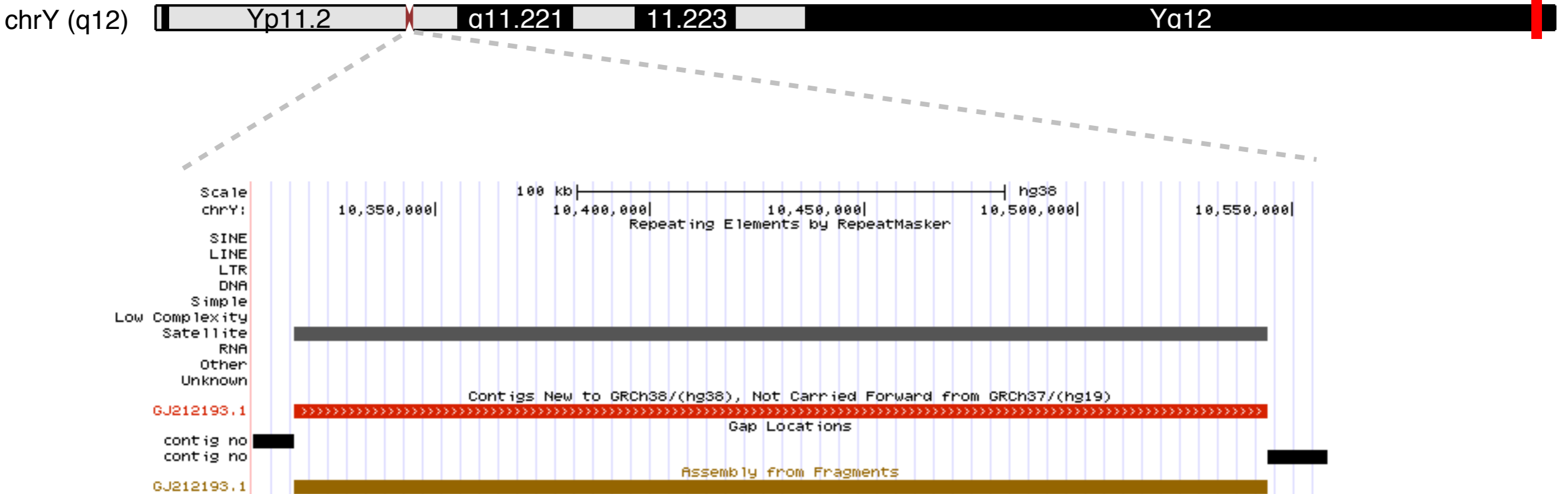
Acrocentric Chr
(13,14,21,22)



Centromeric Sequence Annotation



Centromeric Sequence Annotation



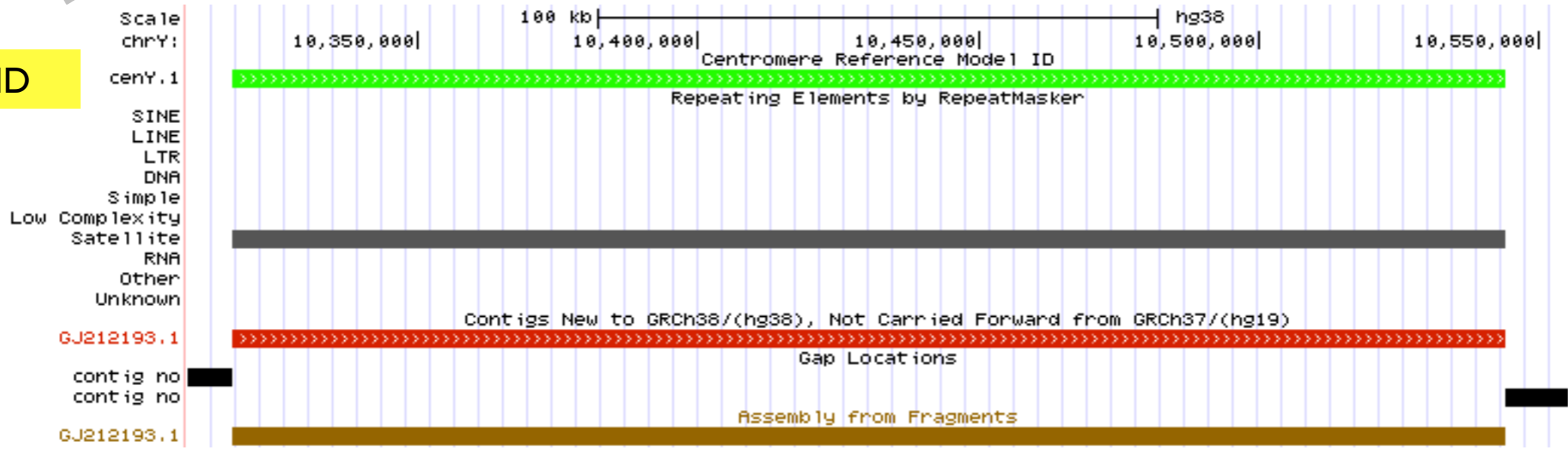
Primary Annotation Goals:
“The Basics”

- Array ID
- HOR patterns
- Monomers
- Confidence of ordering in Linear Sat
- Chromosome assignment
- Known Centromere motifs
- Mappability
- Paired Read Support: Ordering

Centromeric Sequence Annotation



Array ID



Array ID

HOR patterns

Monomers

Confidence of ordering in Linear Sat

Chromosome assignment

Known Centromere motifs

Mappability

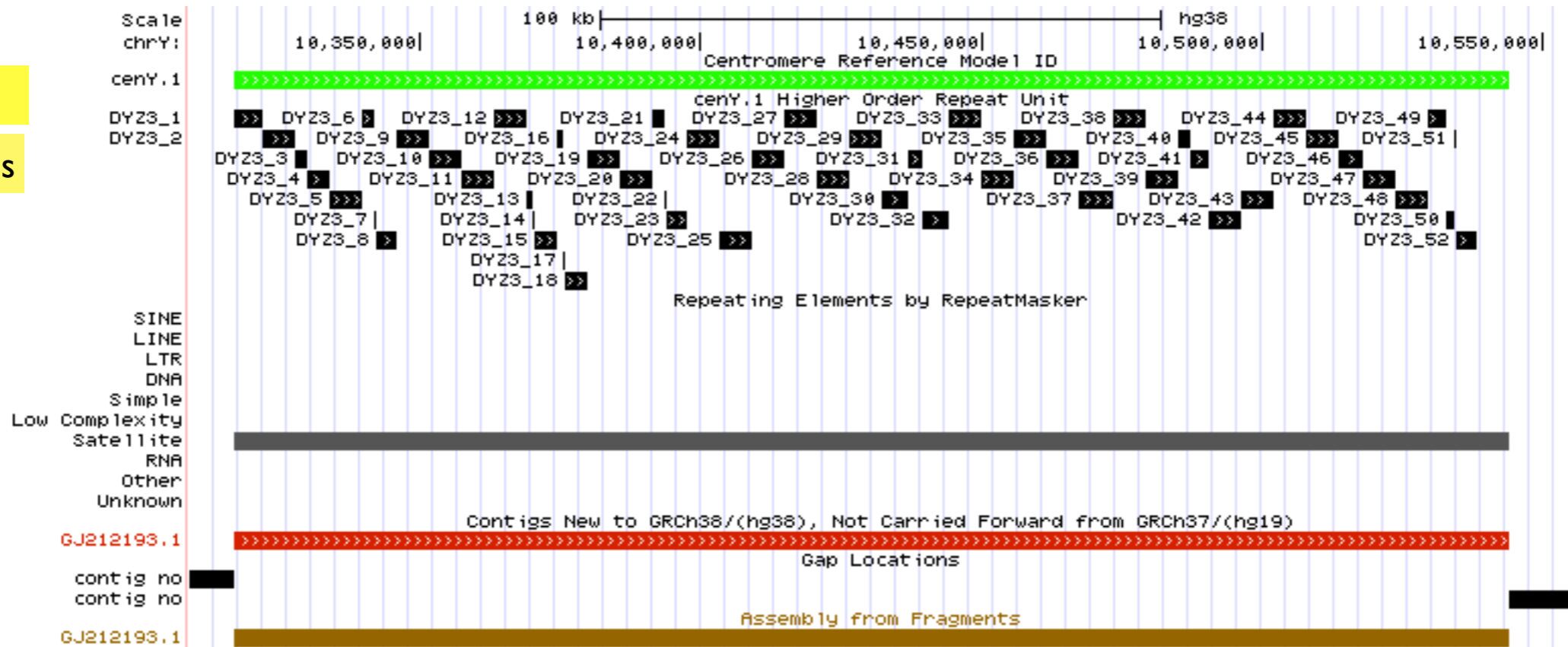
Paired Read Support: Ordering

Primary Annotation Goals:
"The Basics"

Centromeric Sequence Annotation



Array ID
HOR patterns



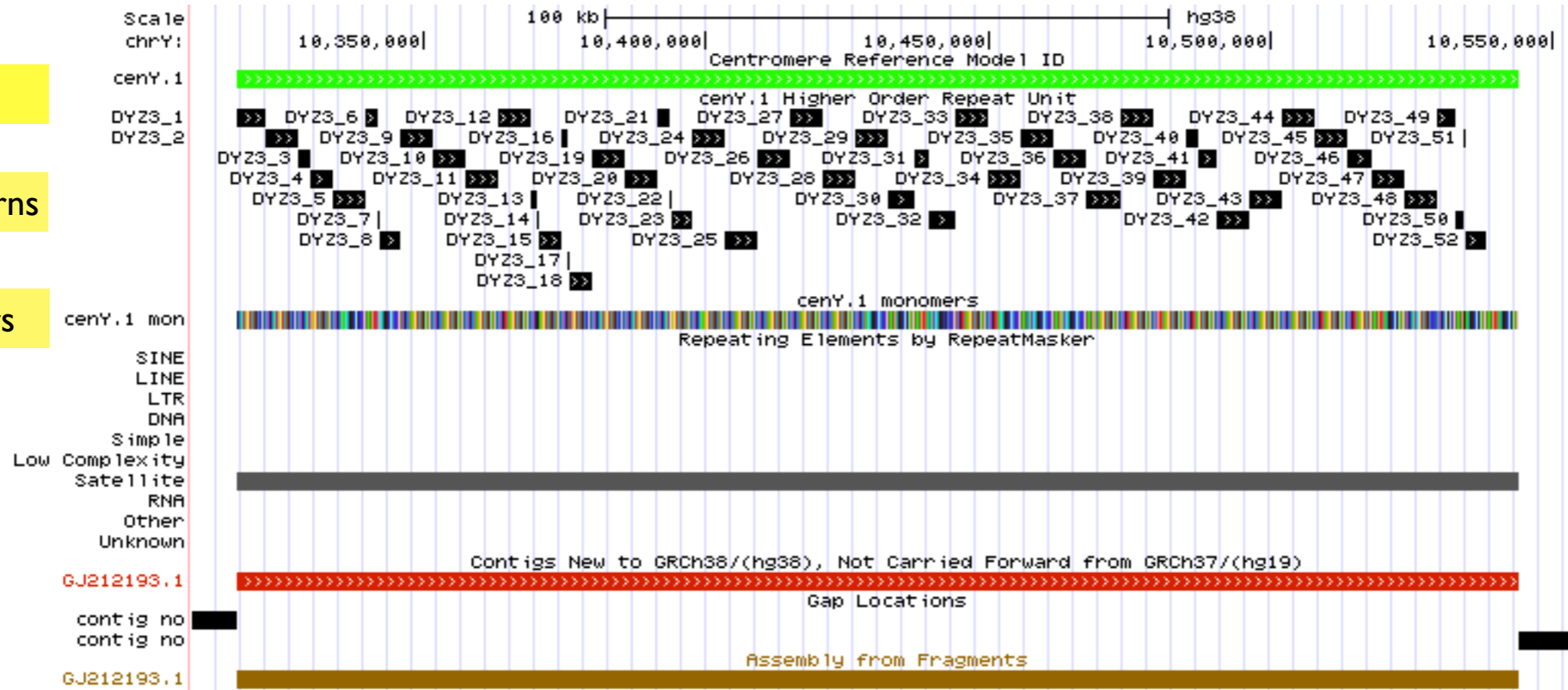
Centromeric Sequence Annotation



Array ID

HOR patterns

Monomers



Centromeric Sequence Annotation

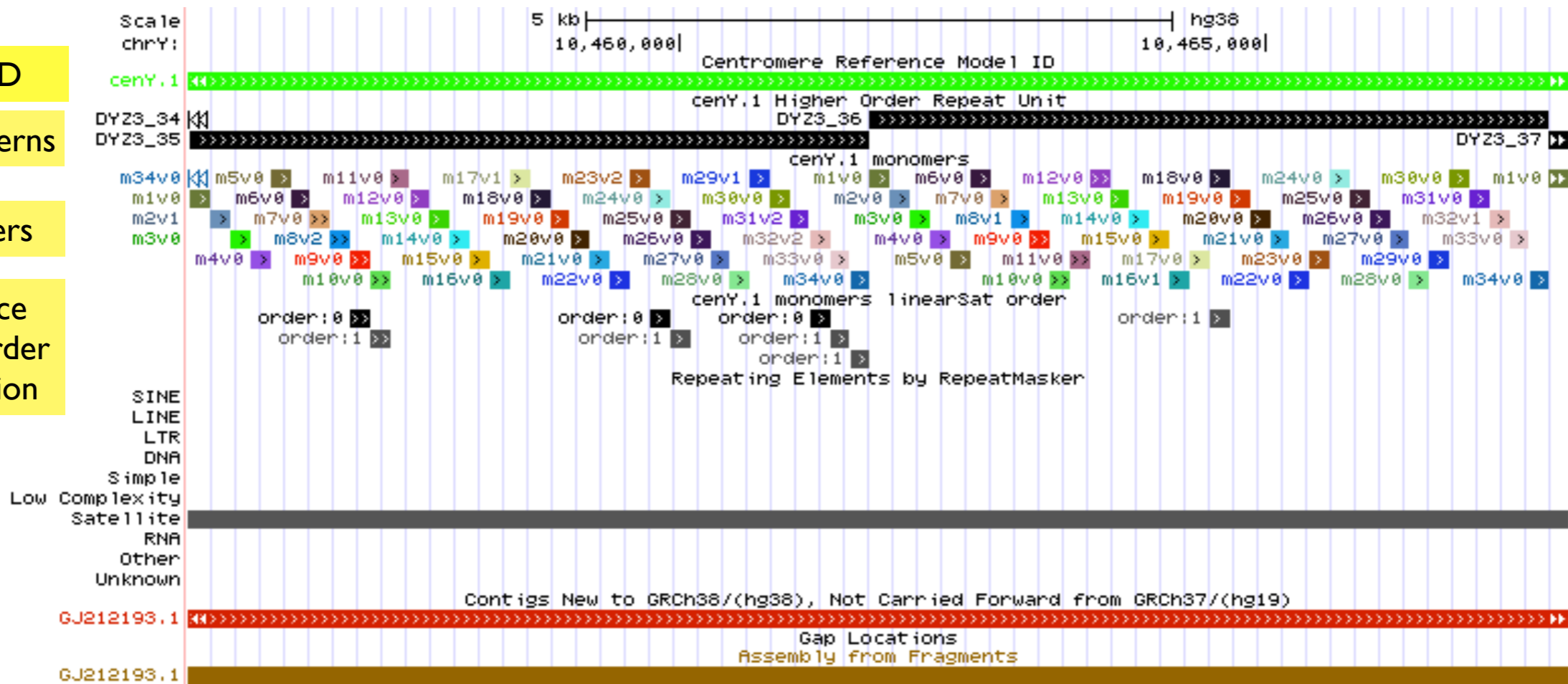


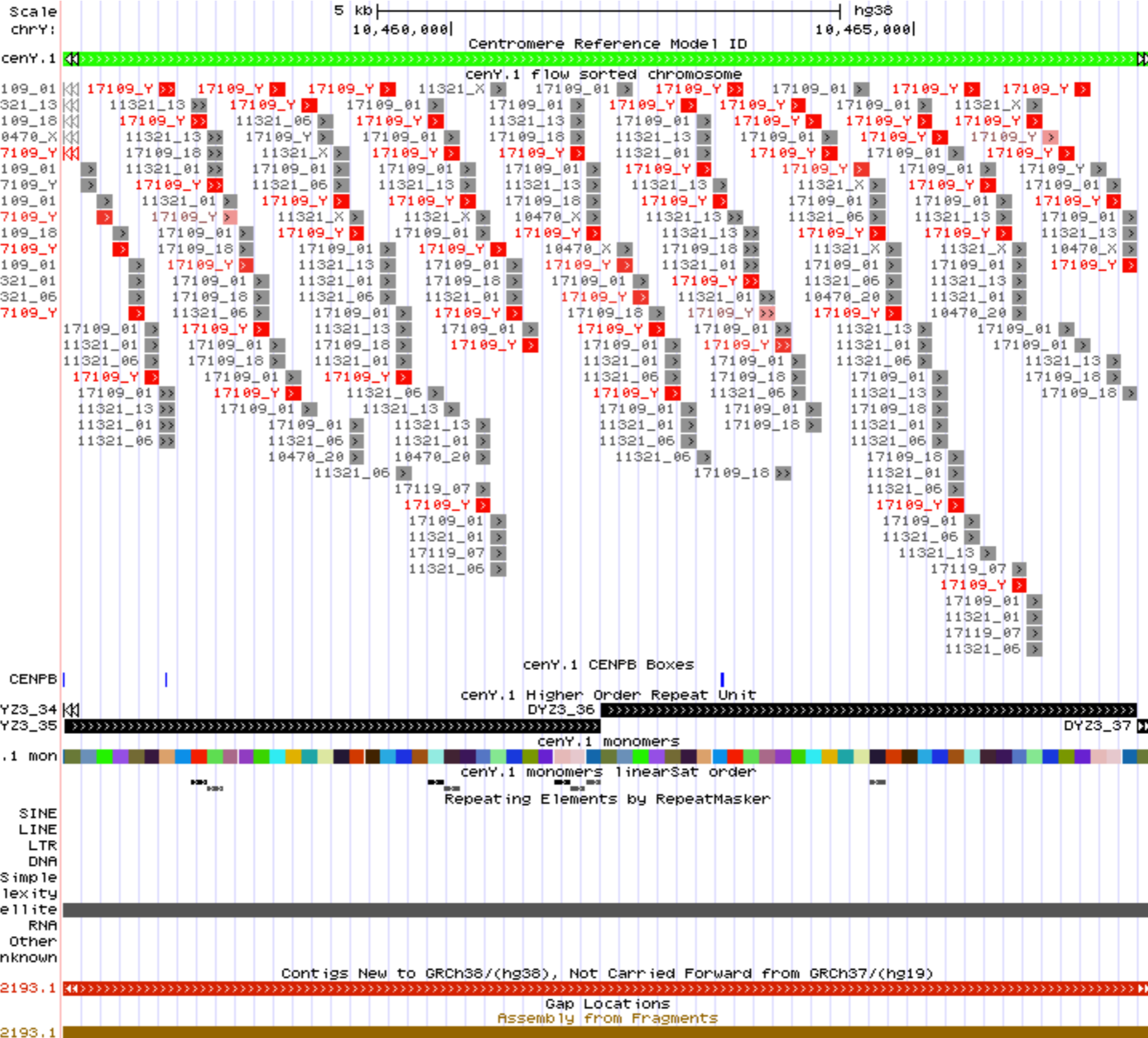
Array ID

HOR patterns

Monomers

Reference Model Order Information





Paired Rd Support

Array ID

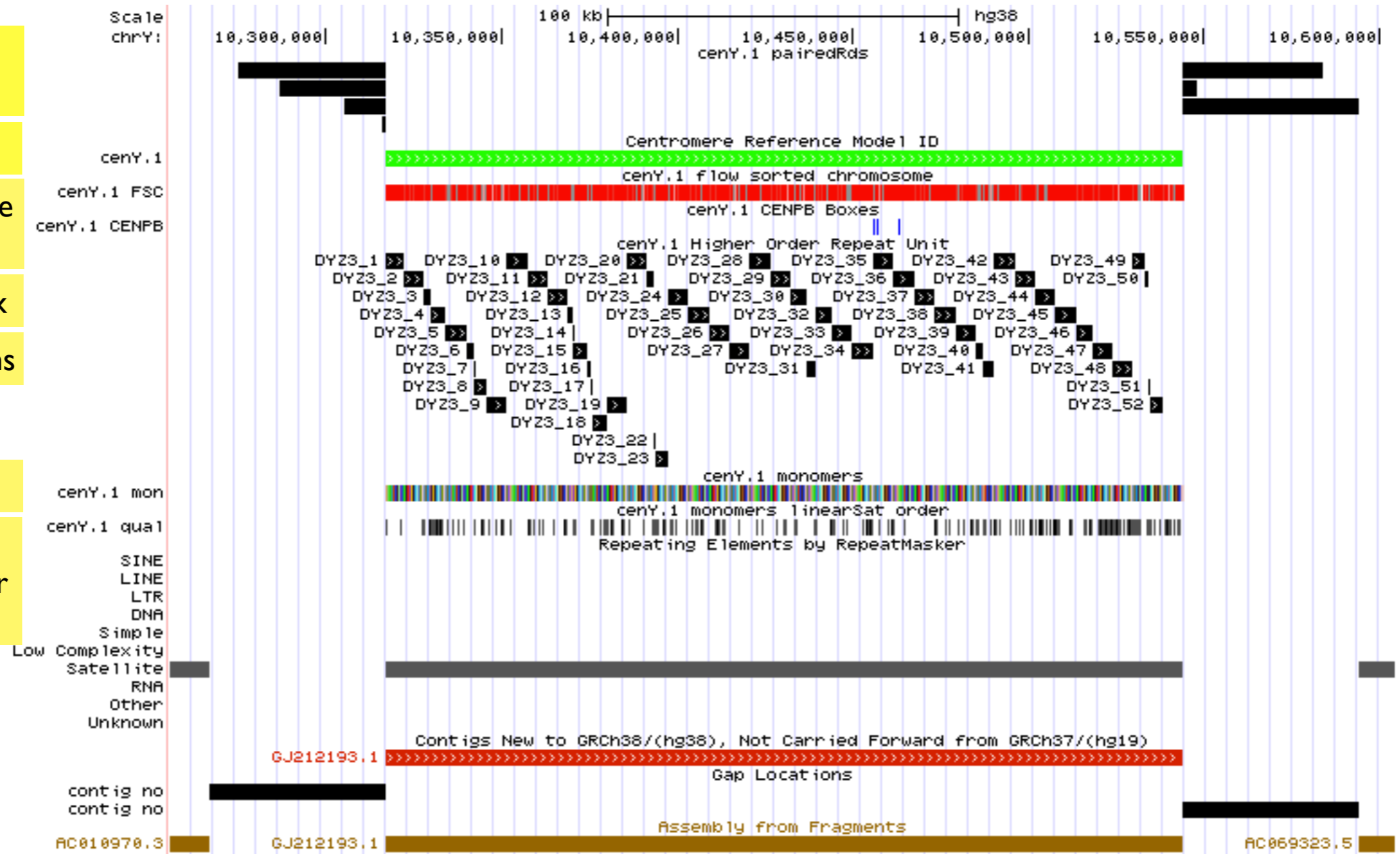
Chromosome Assignment

CENP-B Box

HOR patterns

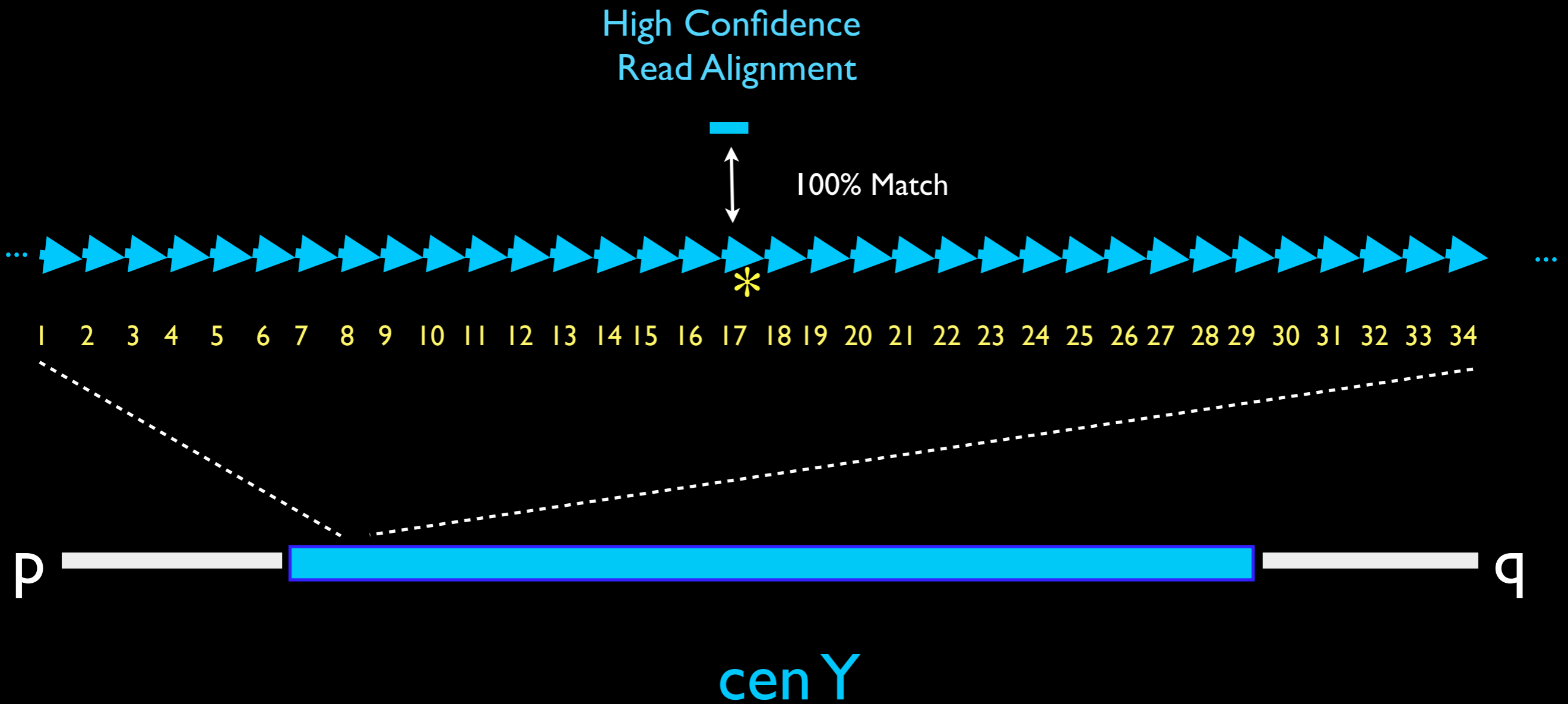
Monomers

Reference Model Order Information



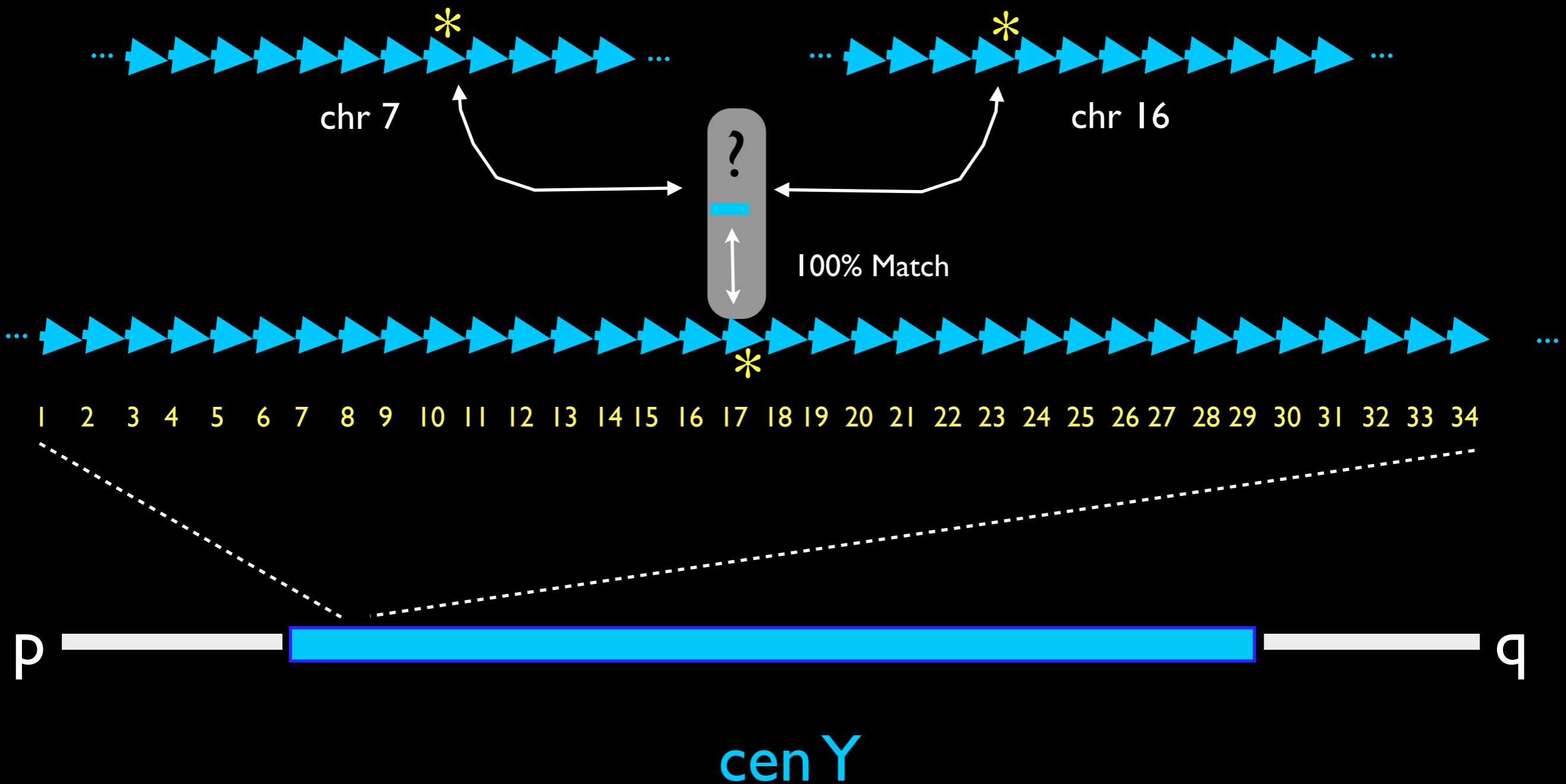
Divergent Satellite

Shared Monomer Homology Challenge Short-read Mapping



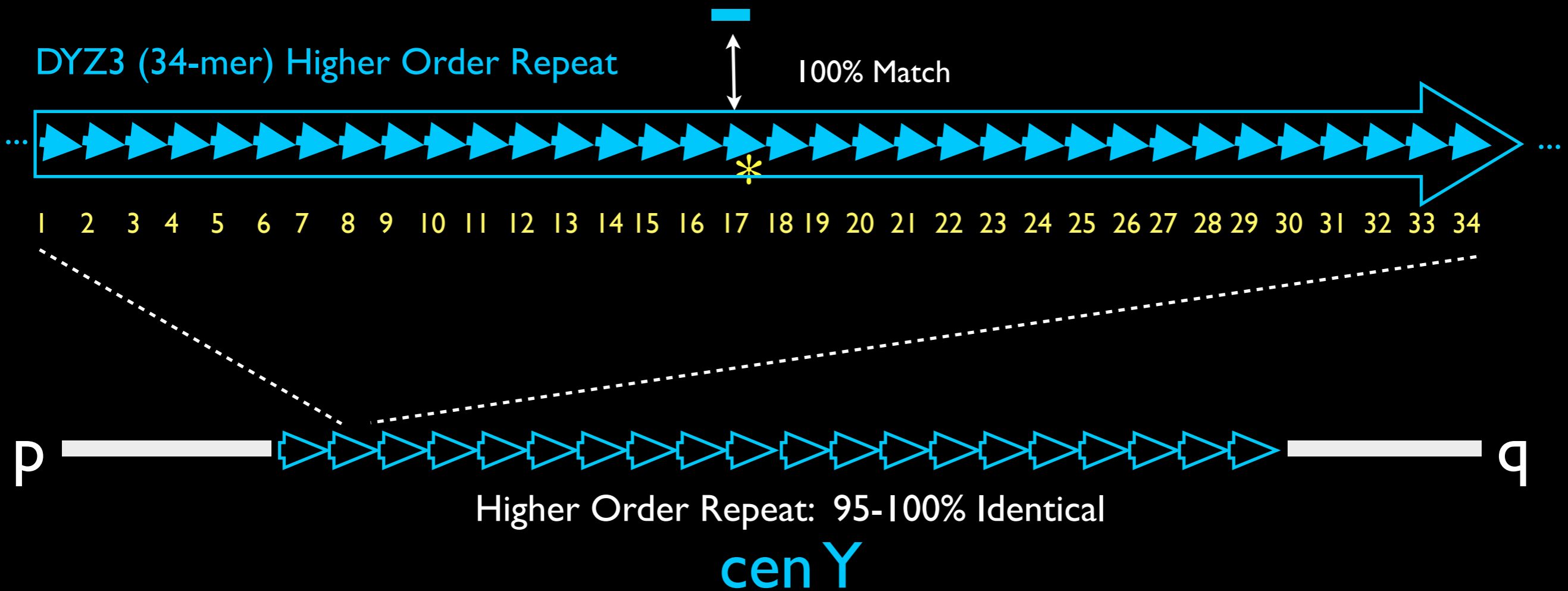
Challenge: Mapping Interpretation

I. Inter-Array Homology



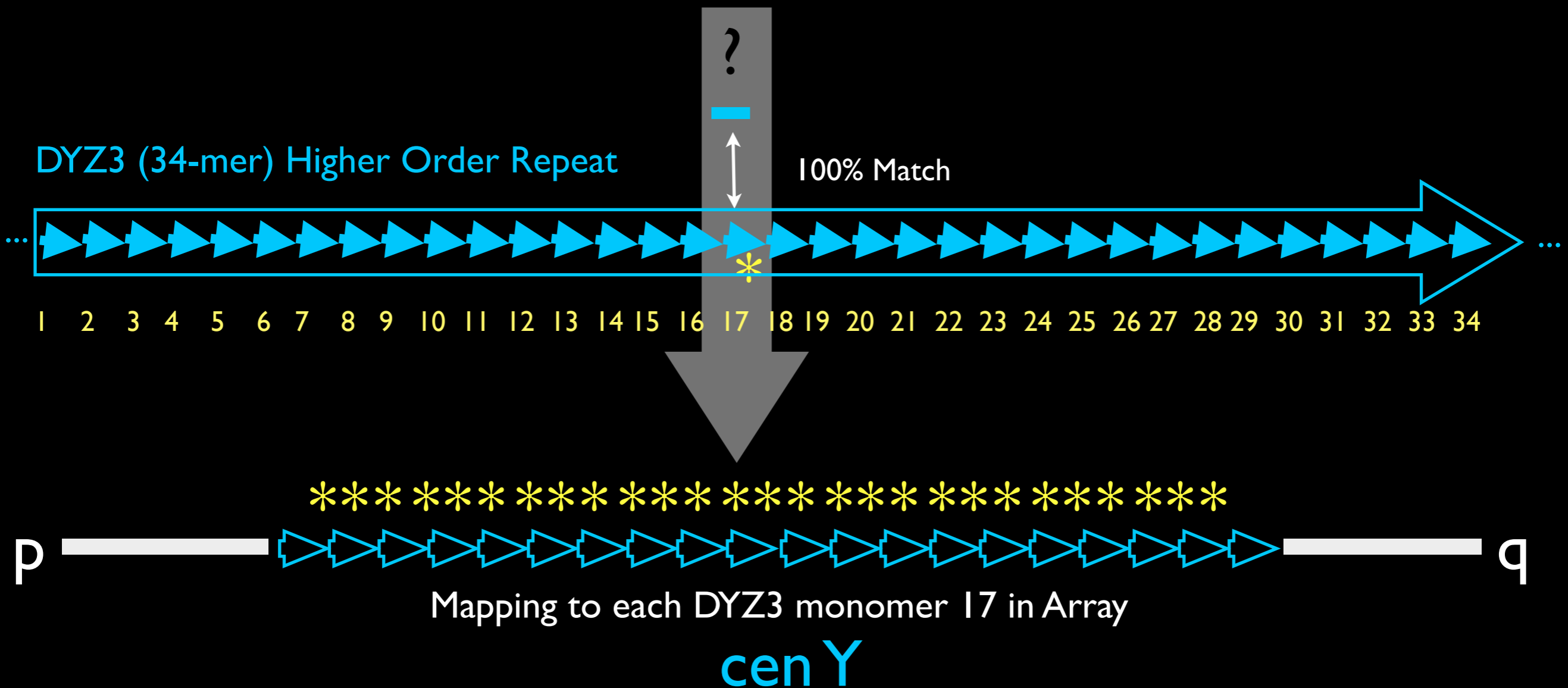
Challenge: Mapping Interpretation

II. Intra-Array Homology



Challenge: Mapping Interpretation

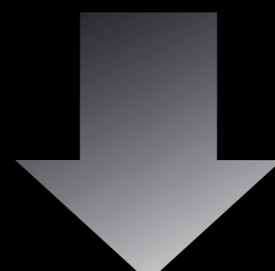
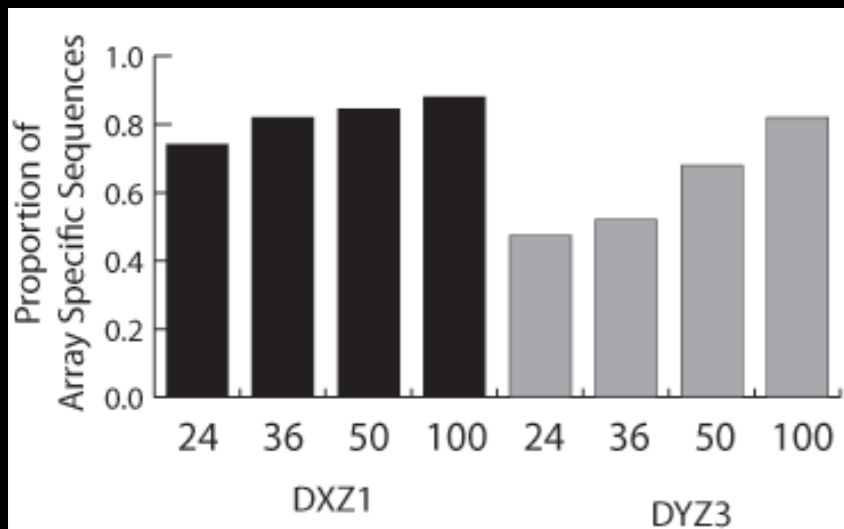
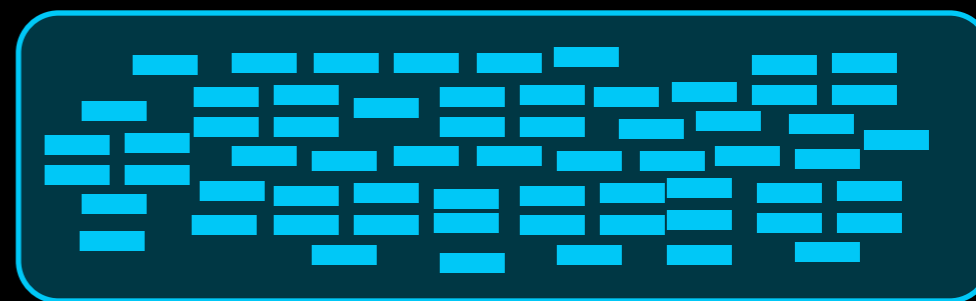
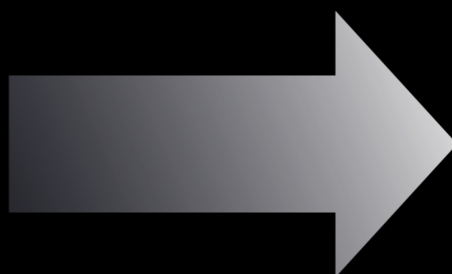
II. Intra-Array Homology



Mappability

Alpha Satellite Reads

Whole Genome Dataset

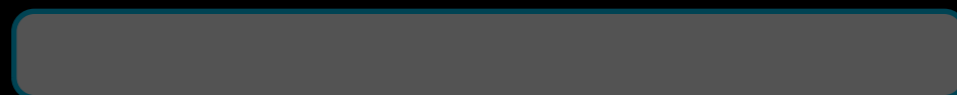


DYZ3 K-mers

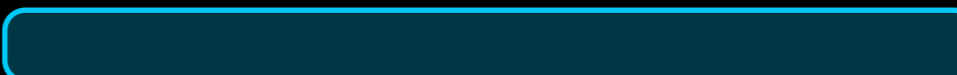
Alpha K-mer Library



DYZ3 Genomic Coordinates



Non-Specific



Specific

Primary Annotation Goals: “The Basics”

Array ID

HOR patterns

Monomers

Confidence of ordering in Linear Sat

Chromosome assignment

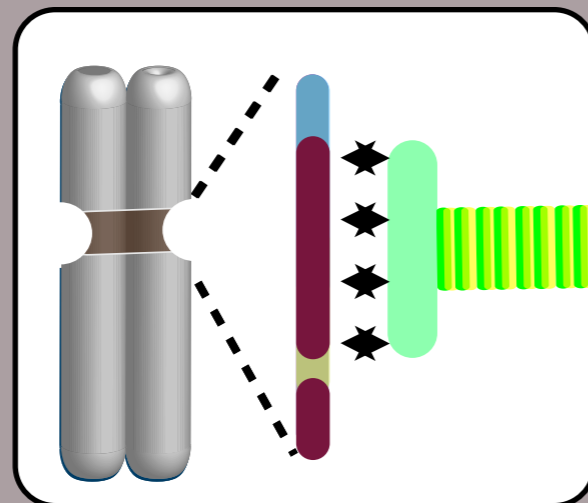
Known Centromere motifs

Mappability

Paired Read Support: Ordering

Up Next:

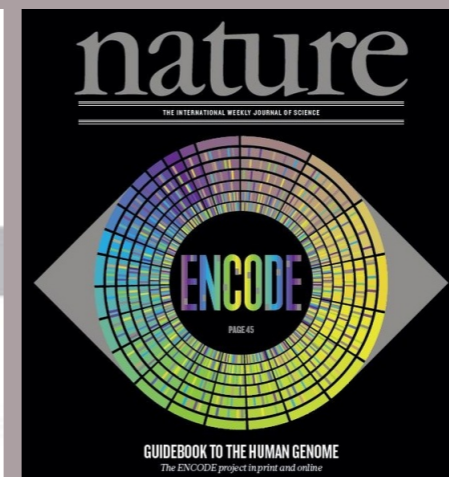
Mapping Epigenetic
Centromere

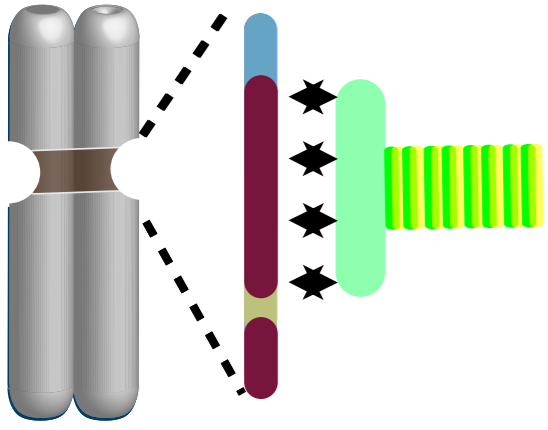


Population Data

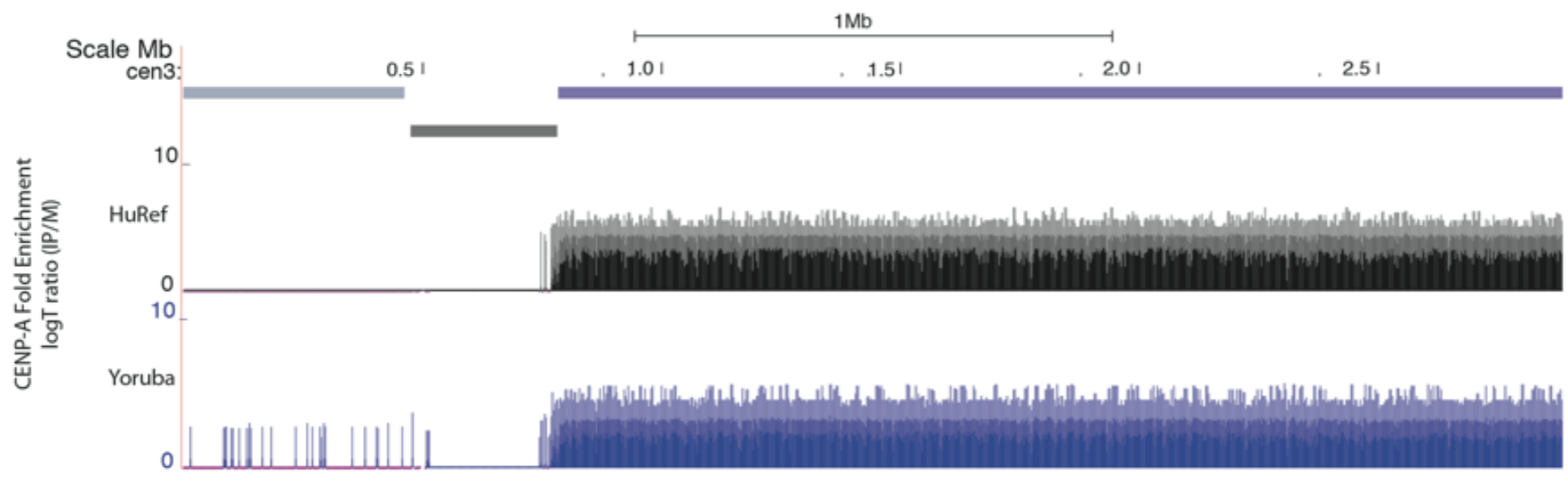
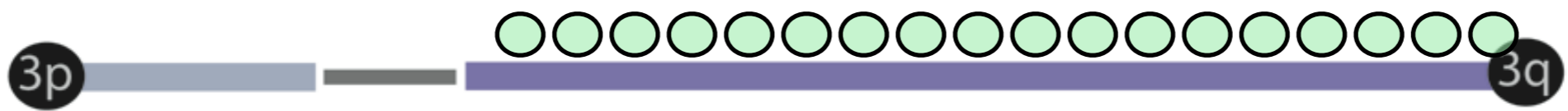
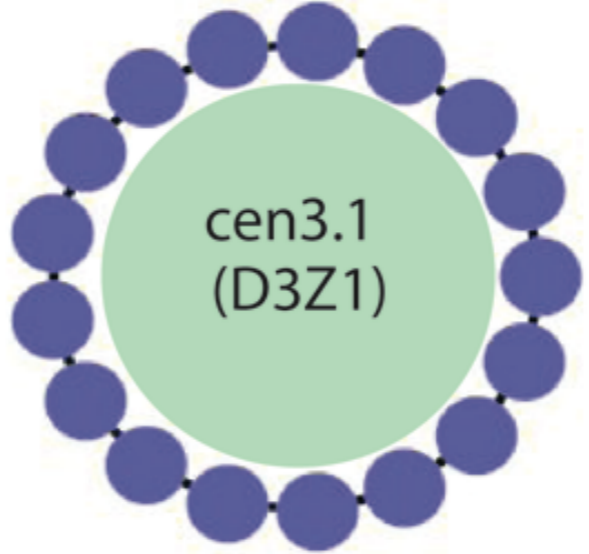
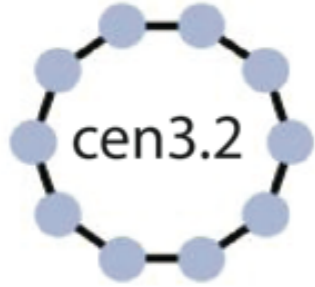


ENCODE Data





Mapping Epigenetic Centromere

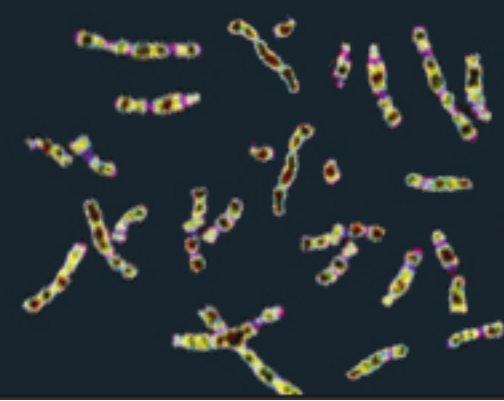


Datasets

- HuRef
- YRI
- NAI2878 *
- HeLa *

1000 Genomes

A Deep Catalog of Human Genetic Variation



GENOMICS 7, 325–330 (1990)

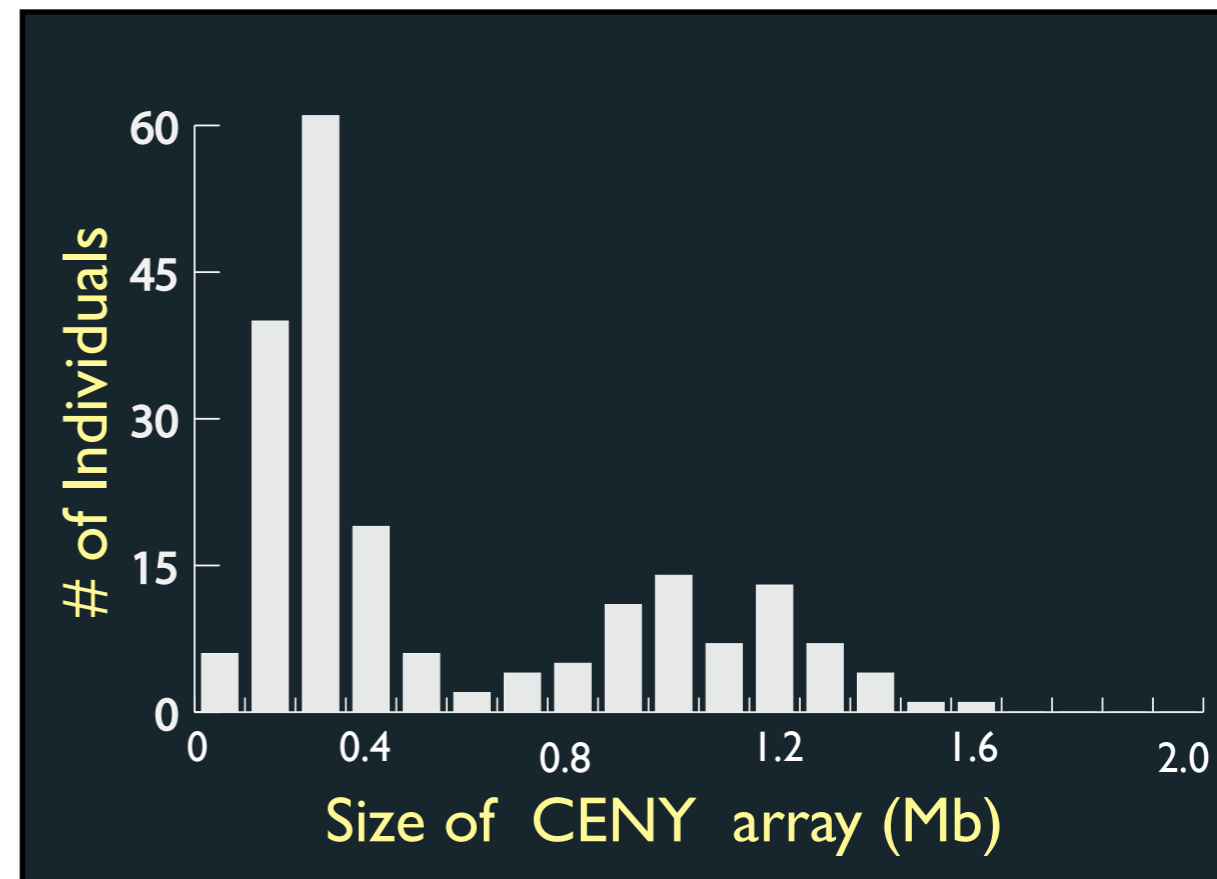
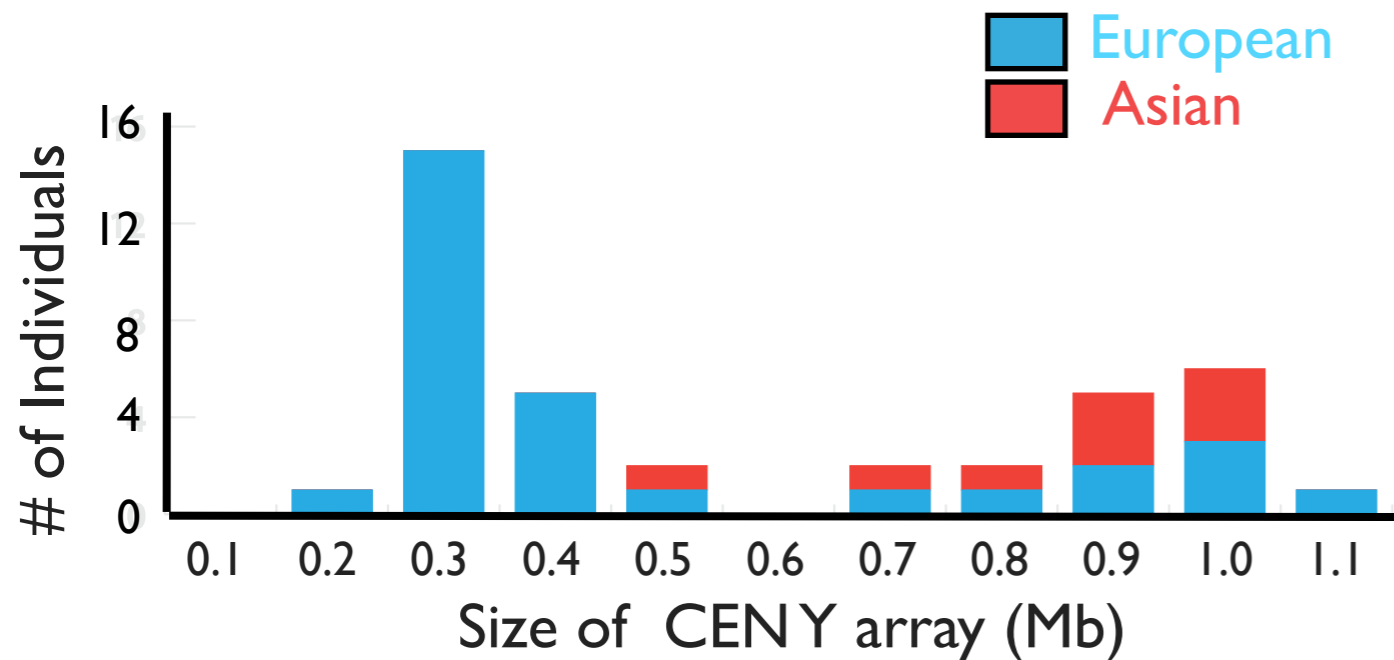
Y Chromosome DNA Haplotyping Suggests That Most European and Asian Men Are Descended from One of Two Males

REBECCA OAKY¹ AND CHRIS TYLER-SMITH²

CRC Chromosome Molecular Biology Group, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, United Kingdom

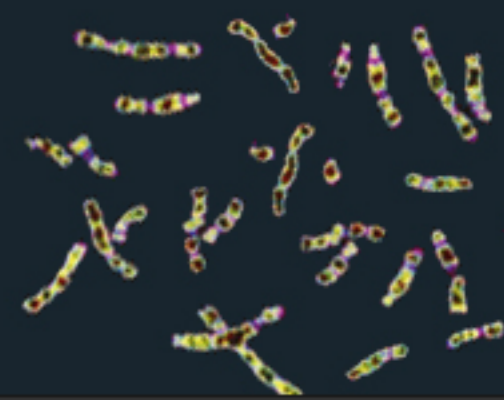
Received November 15, 1989; revised February 23, 1990

HuRef k-mers (24mers) useful in predicting array length across ~400 male individuals



1000 Genomes

A Deep Catalog of Human Genetic Variation



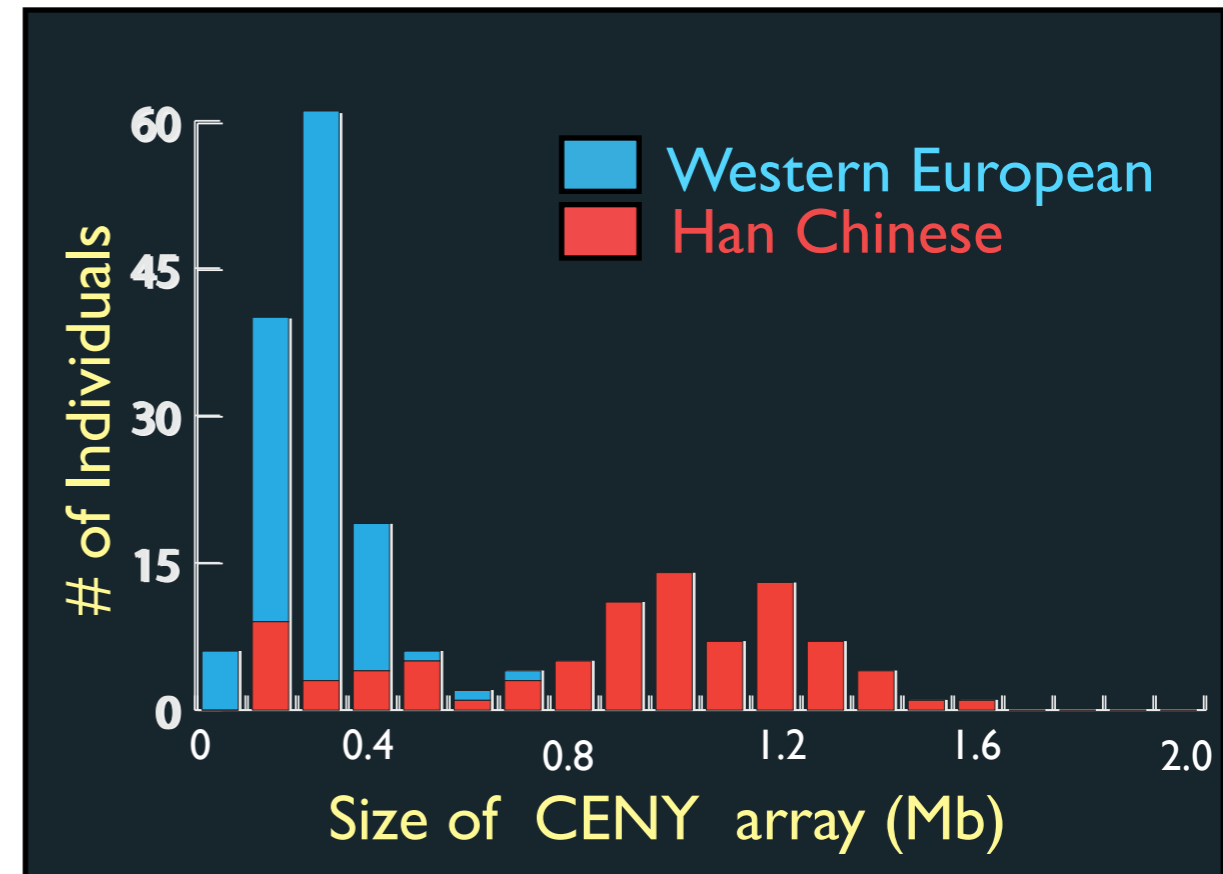
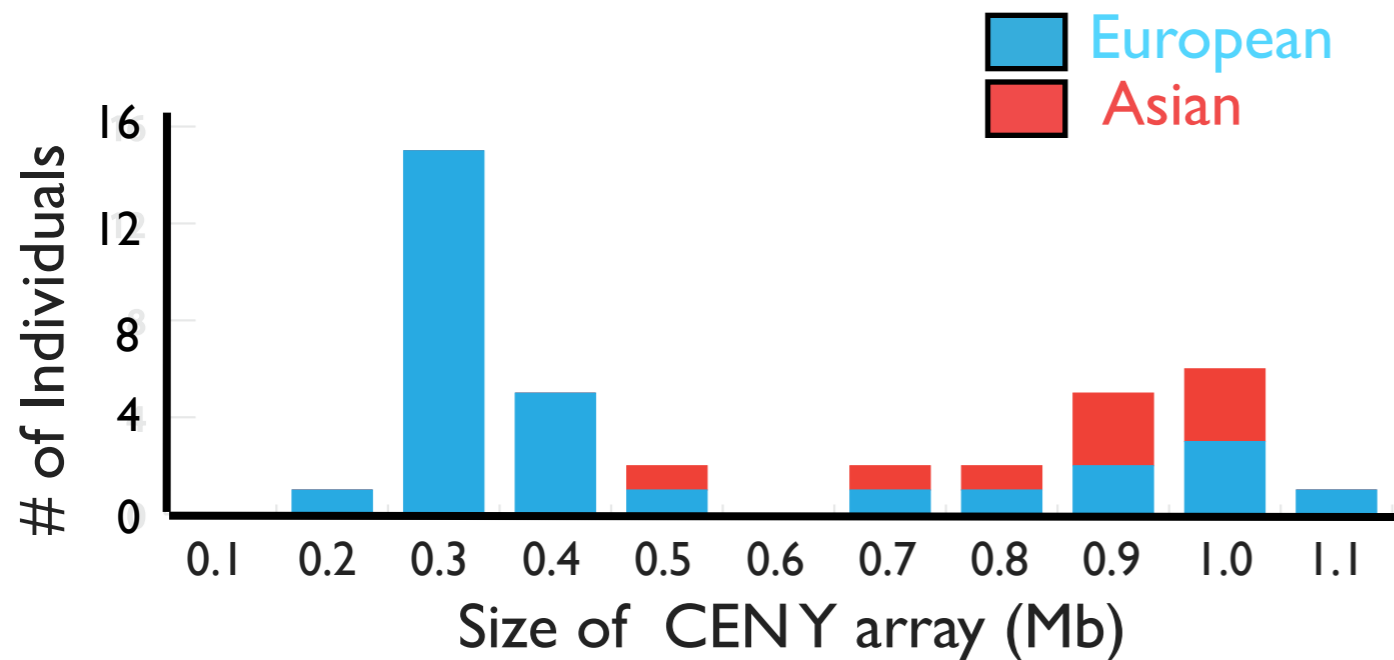
GENOMICS 7, 325–330 (1990)

Y Chromosome DNA Haplotyping Suggests That Most European and Asian Men Are Descended from One of Two Males

REBECCA OAKLEY¹ AND CHRIS TYLER-SMITH²

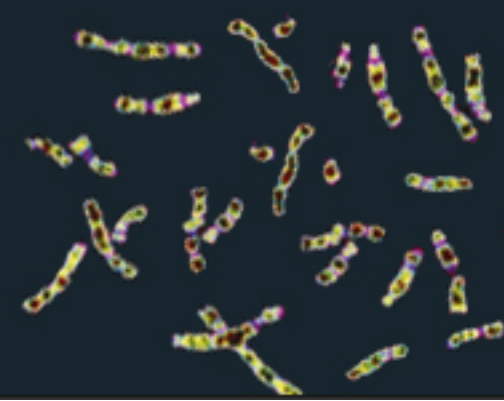
CRC Chromosome Molecular Biology Group, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, United Kingdom

Received November 15, 1989; revised February 23, 1990



1000 Genomes

A Deep Catalog of Human Genetic Variation

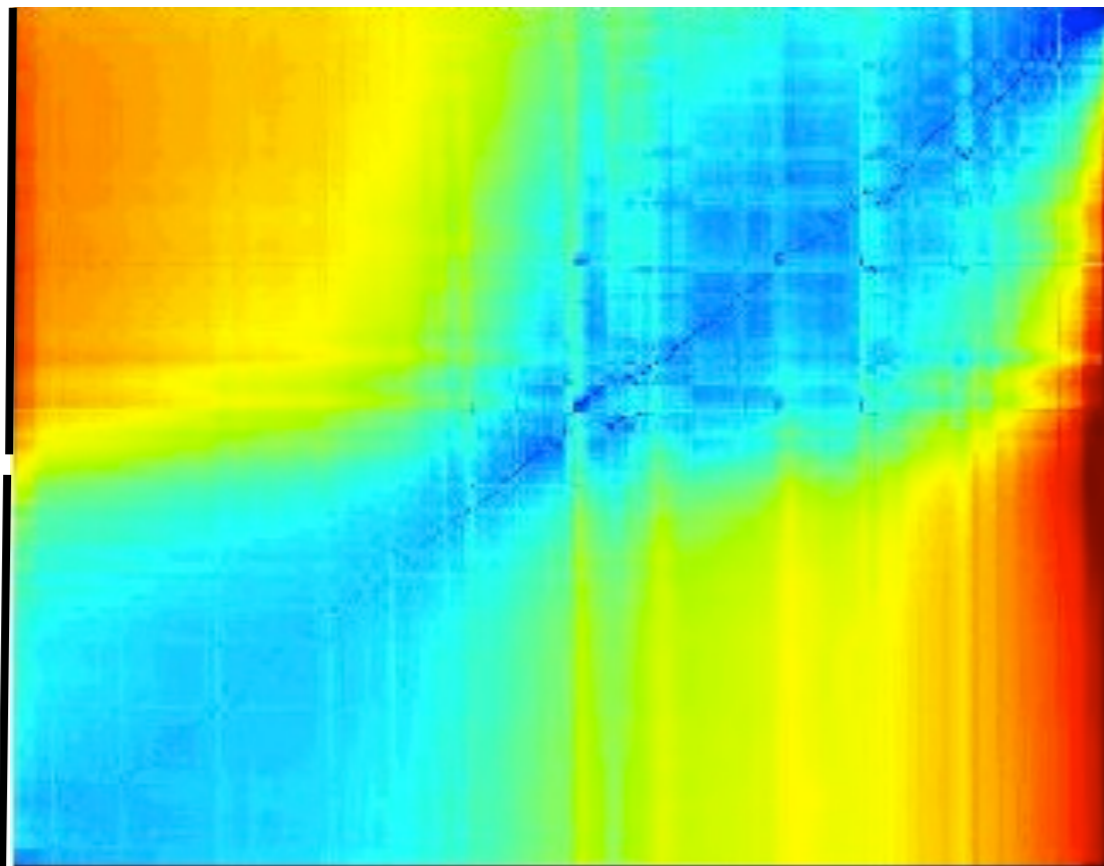


HuRef k-mer profiles are useful in predicting array classification across ~400 male individuals into two distinct groups

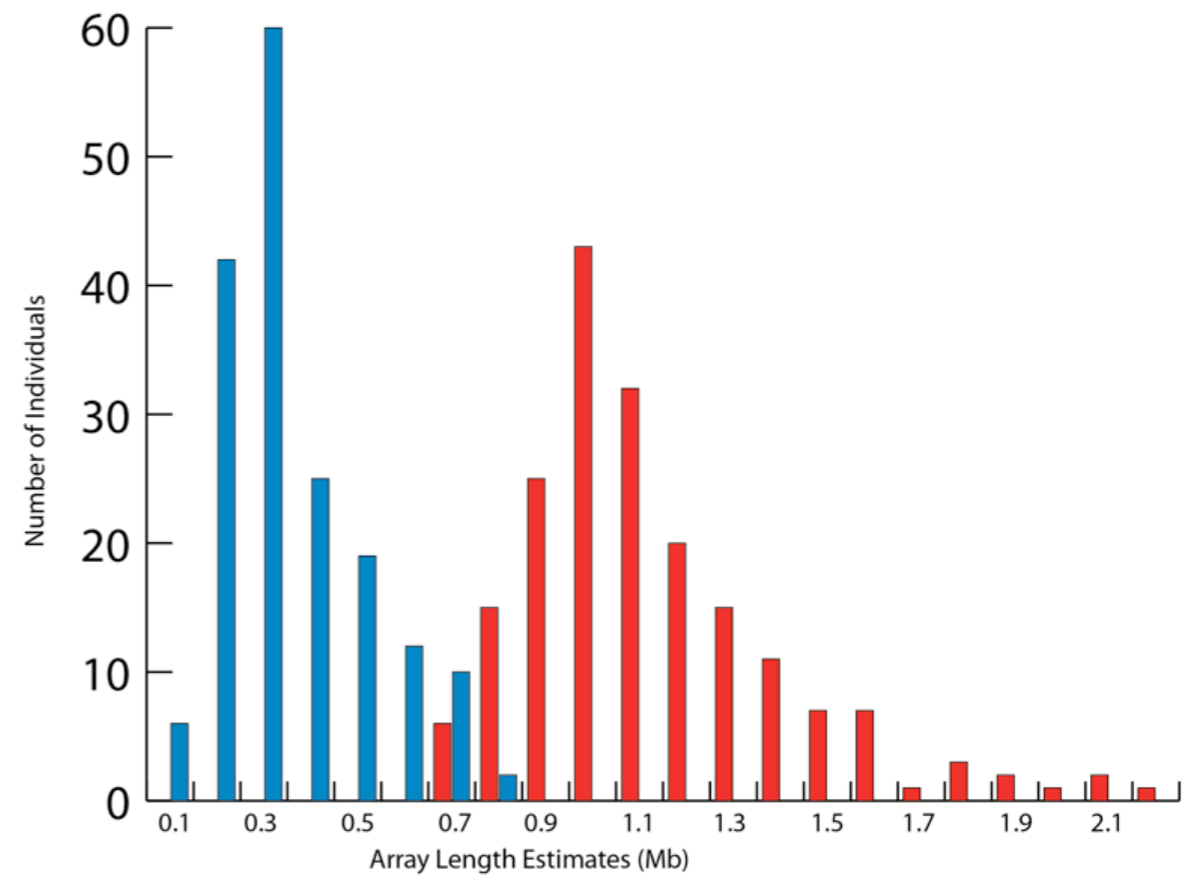
Group 1

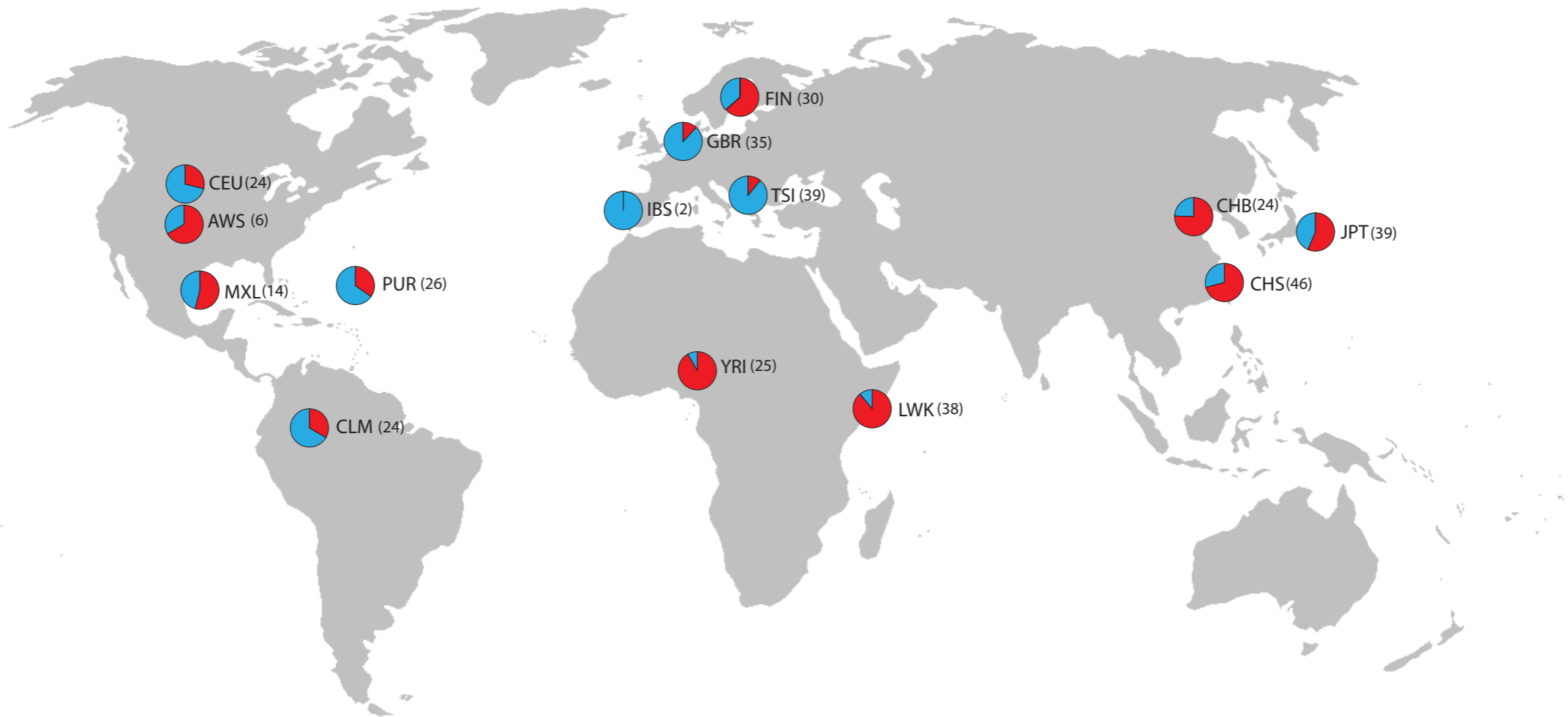


Group 2



Clustergram: K-mer Identity Matrix between Male Individuals

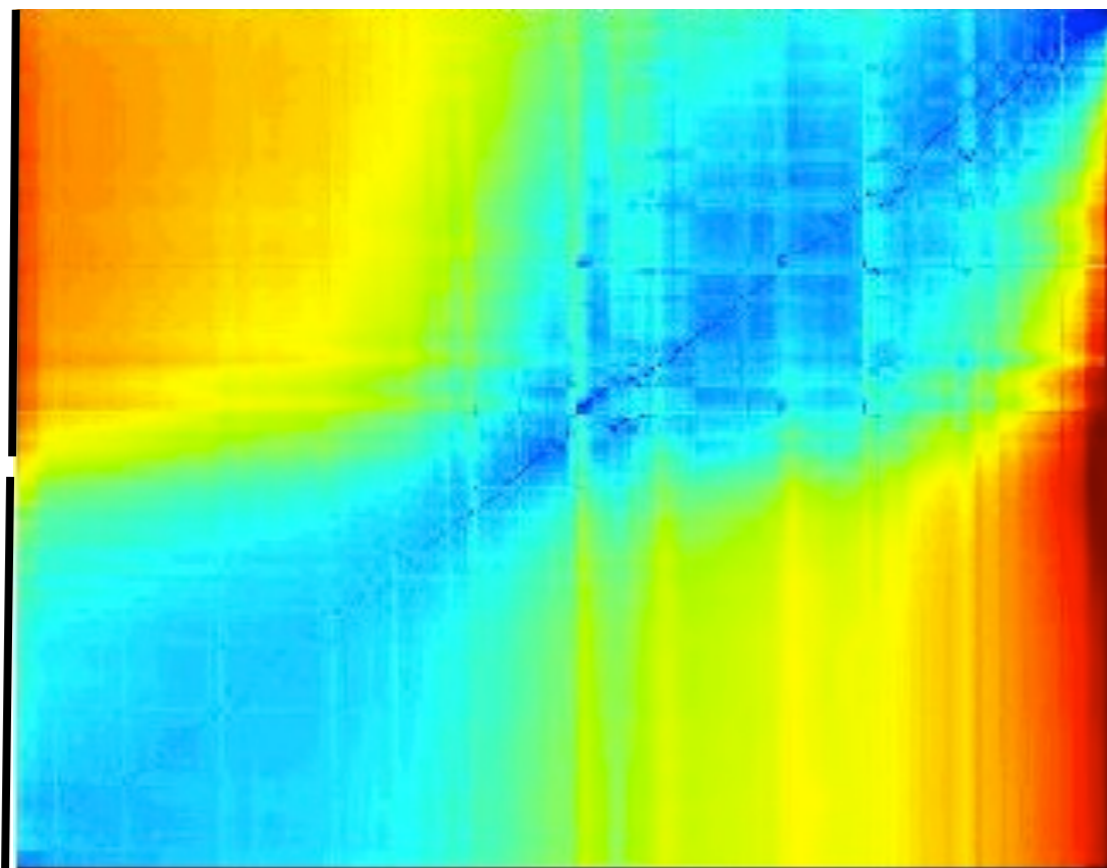




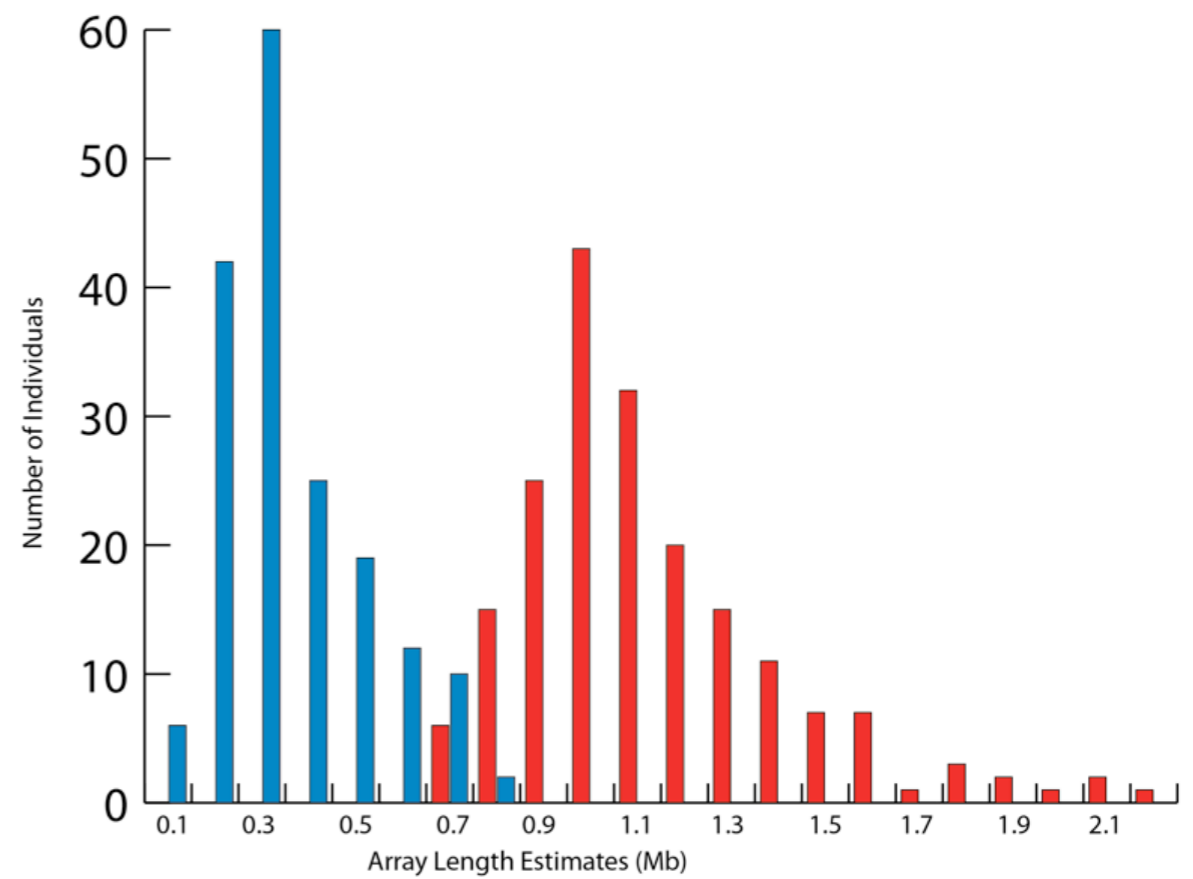
Group 1

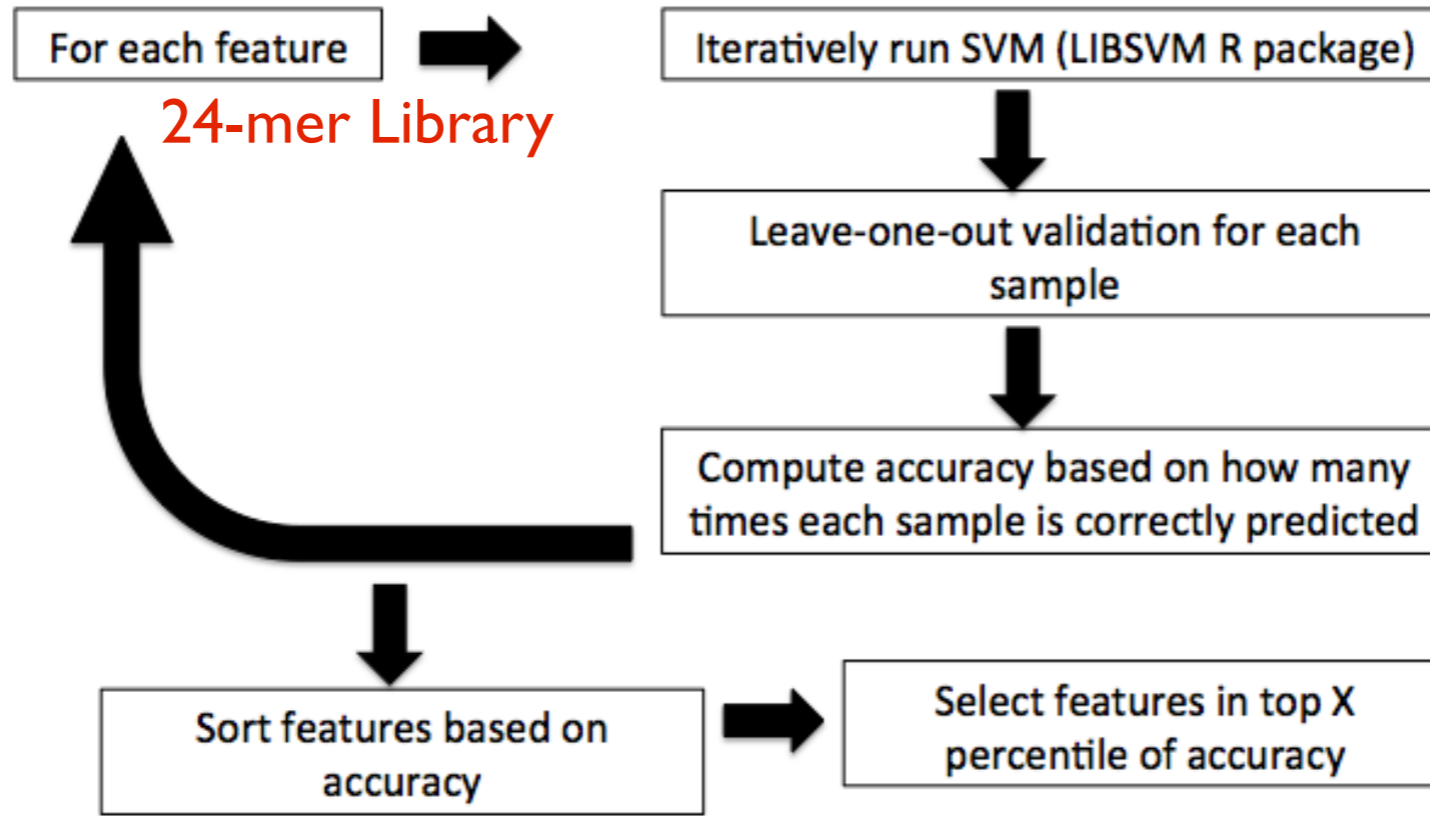


Group 2



Clustergram: K-mer Identity Matrix between Male Individuals

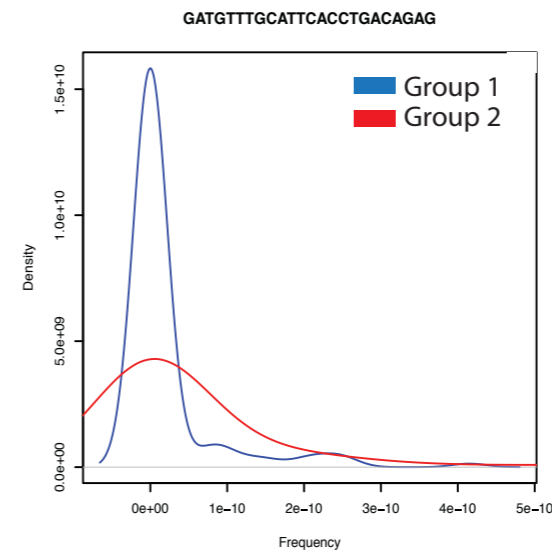
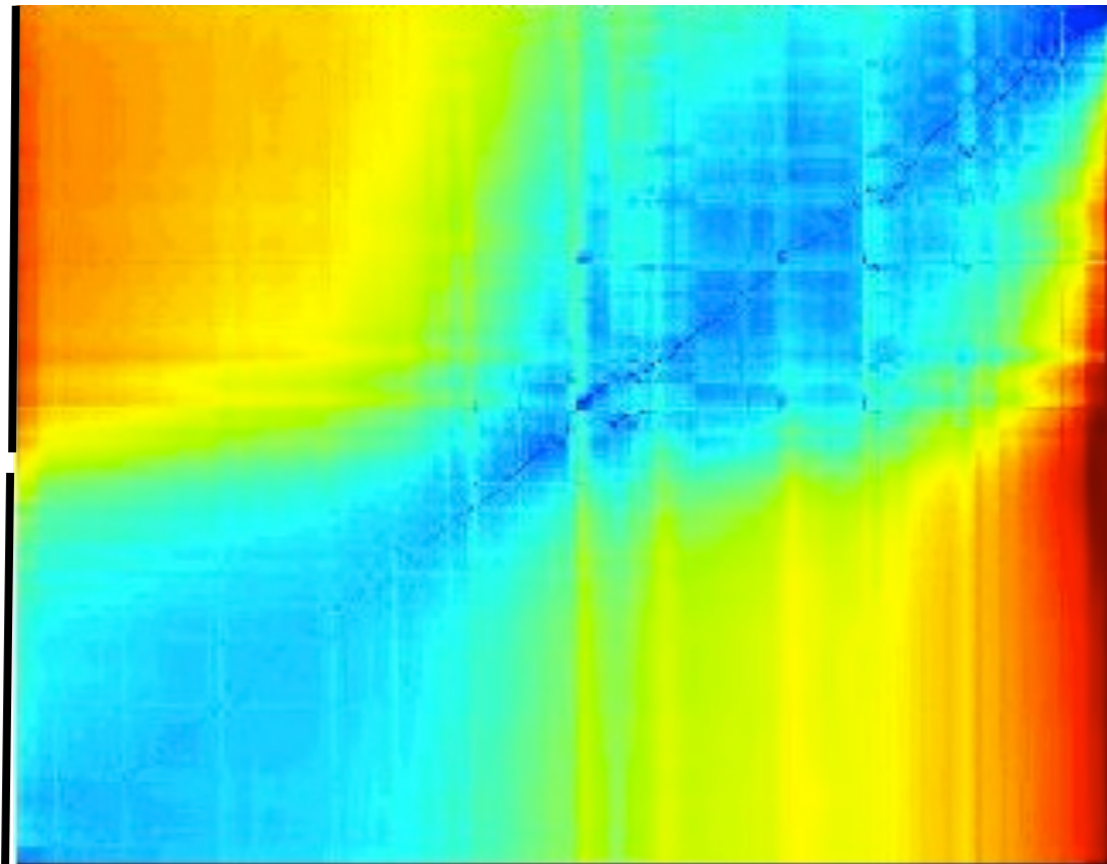




Group 1

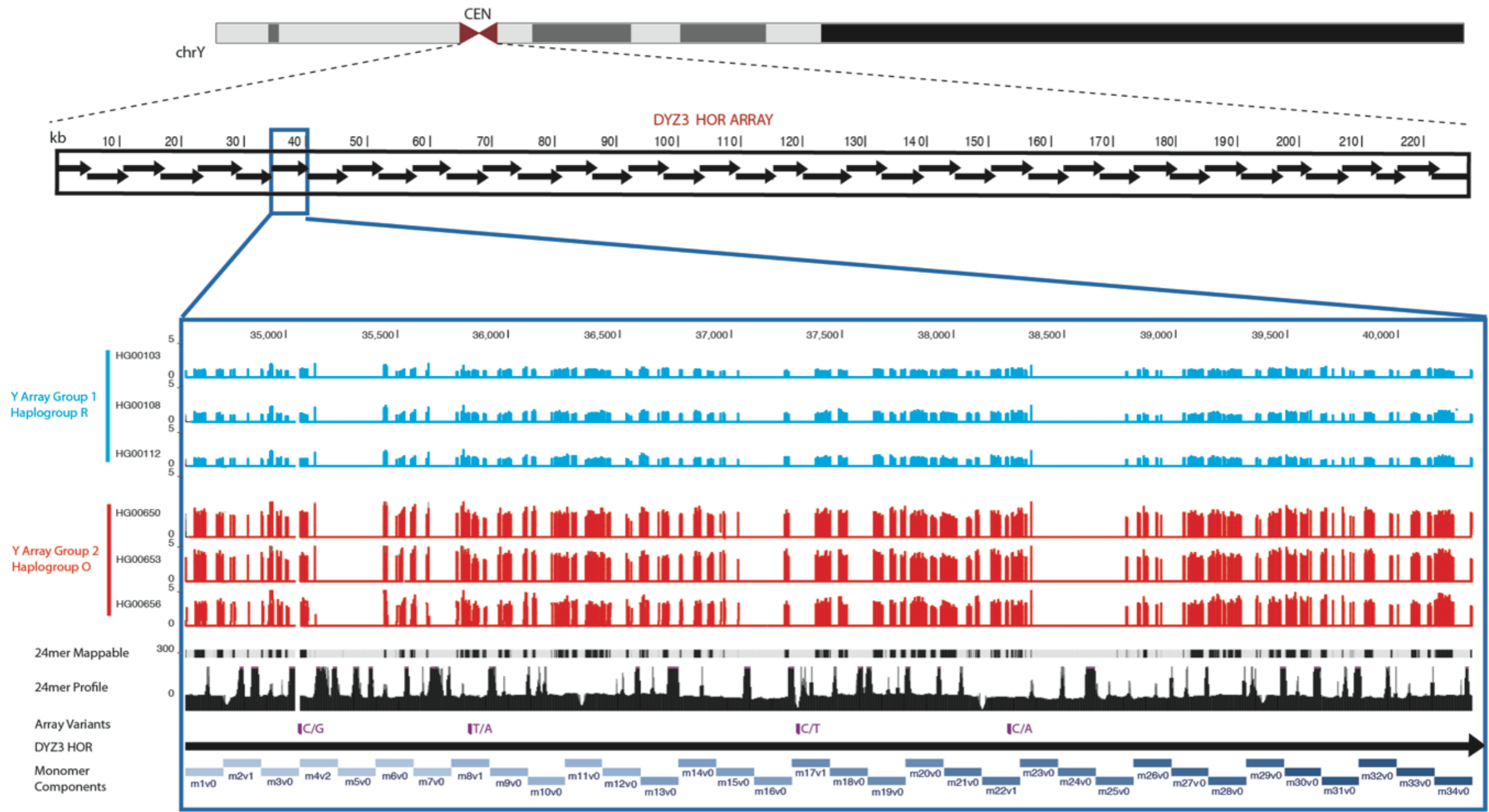


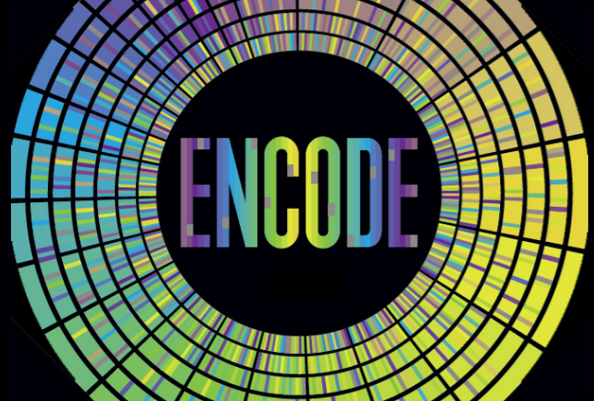
Group 2



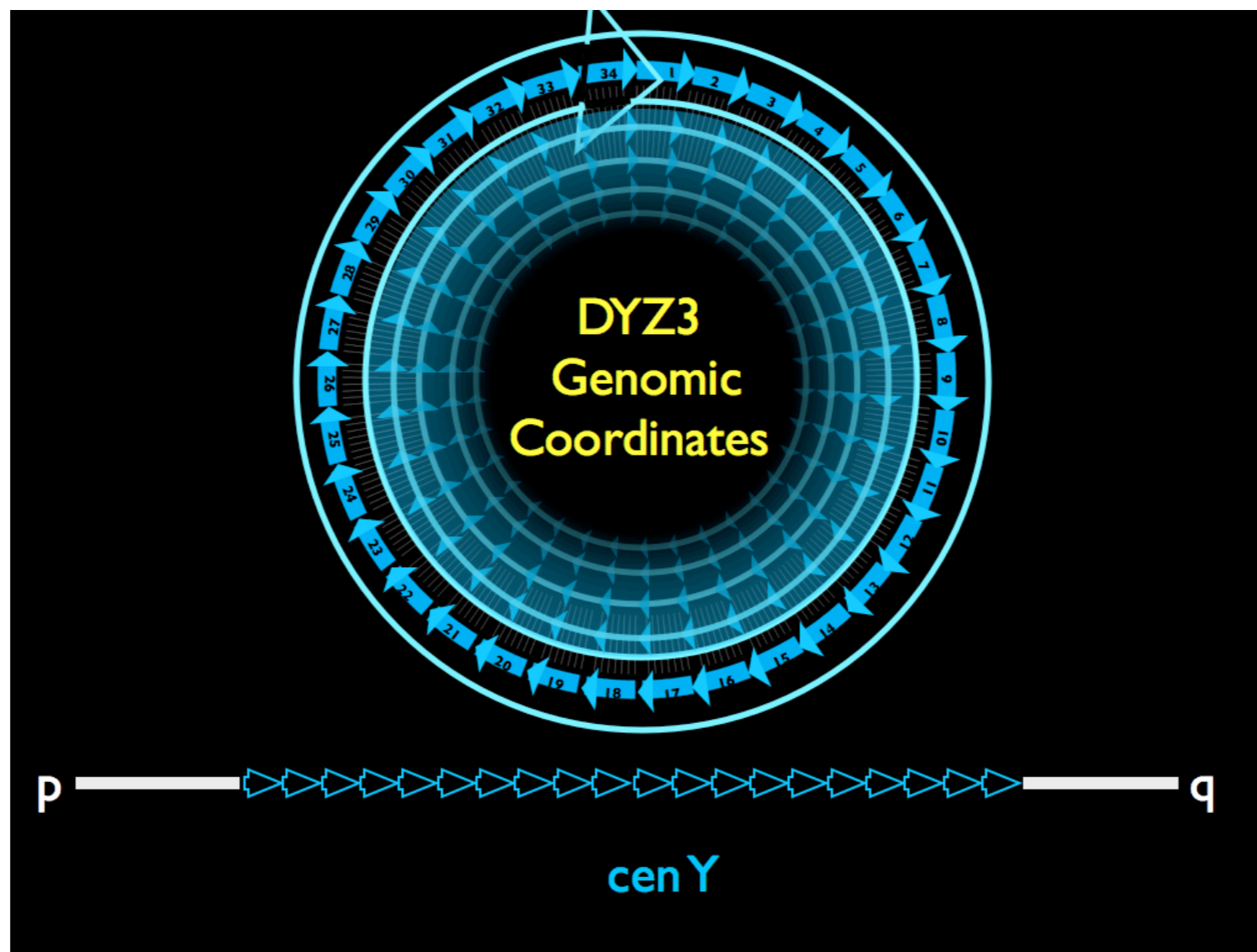
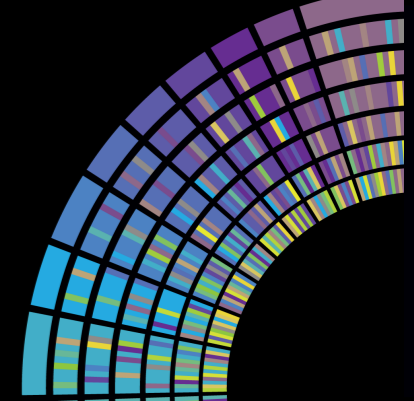
Catalogue a new source of human sequence variation

Survey those k-mers that are enriched in one array group

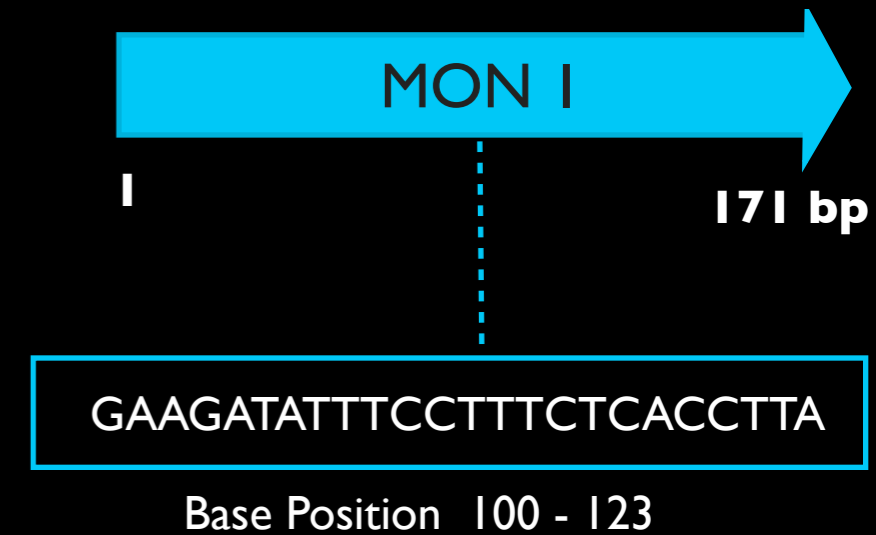
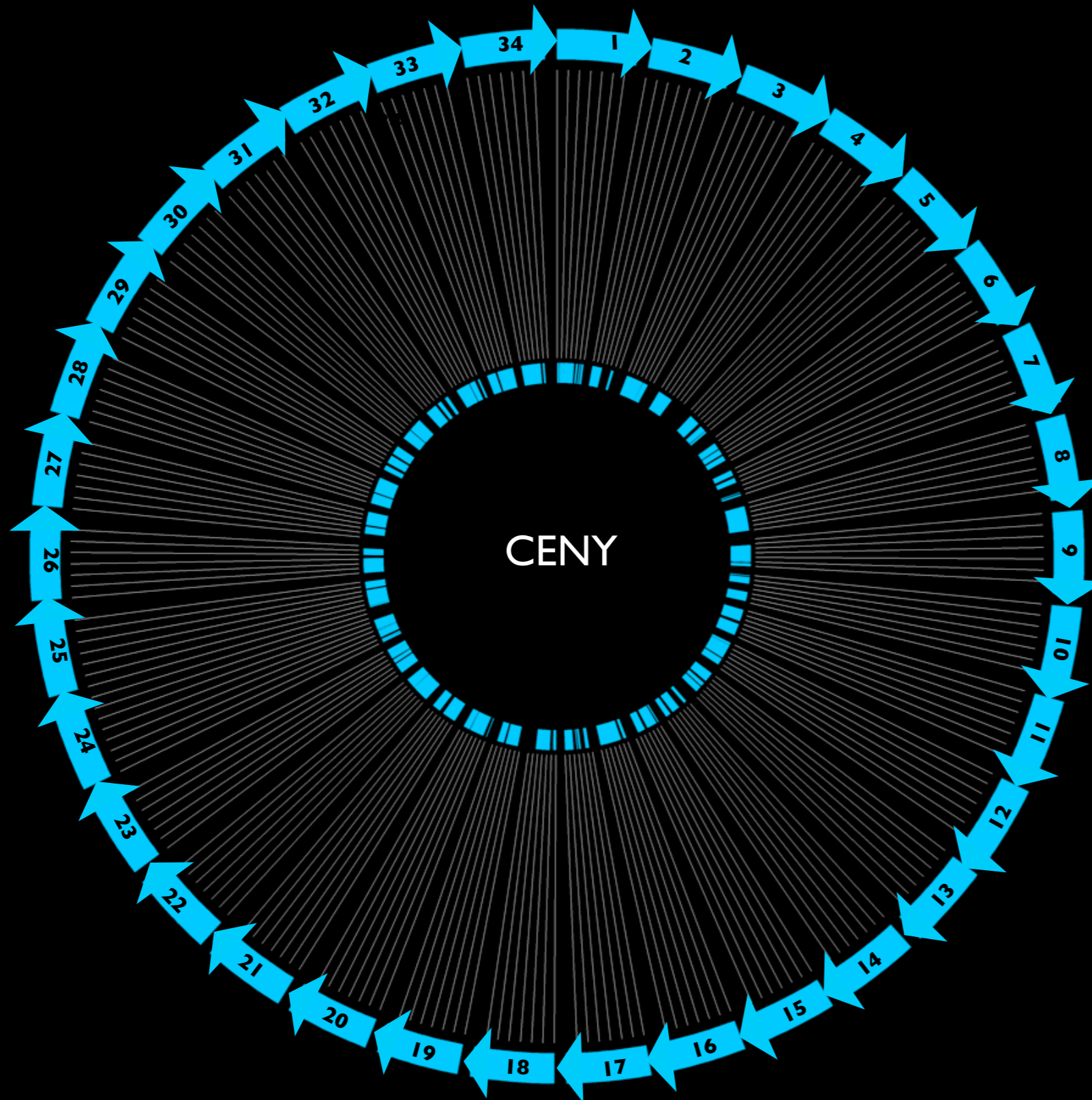




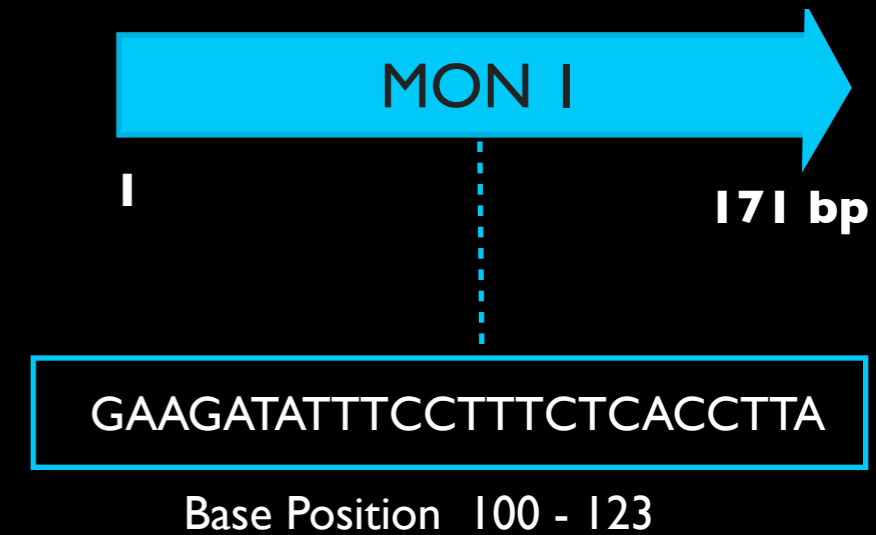
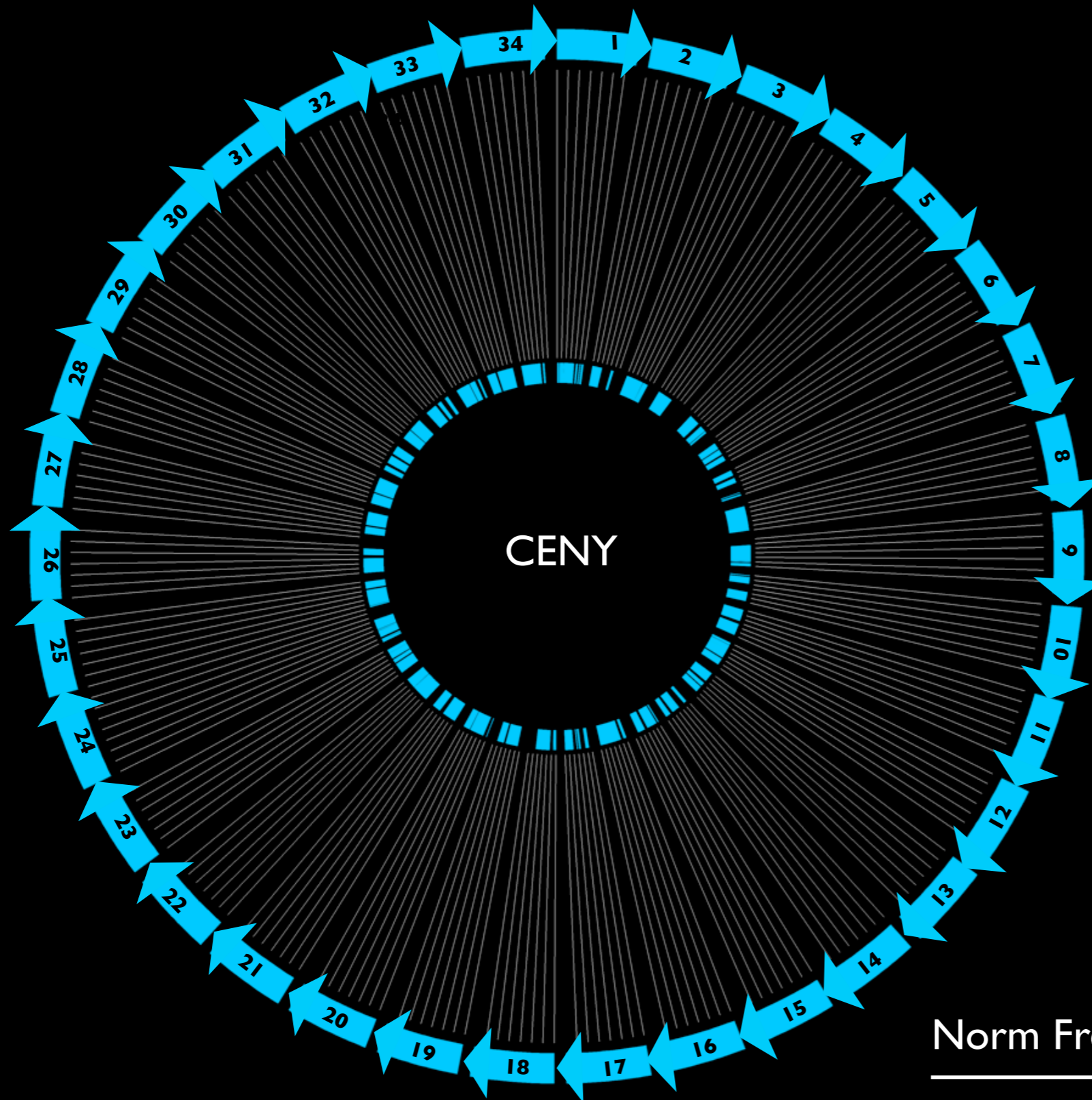
ENCODE data



ENCODE Tier I: Human Embryonic Stem Cell (HI-hESC)

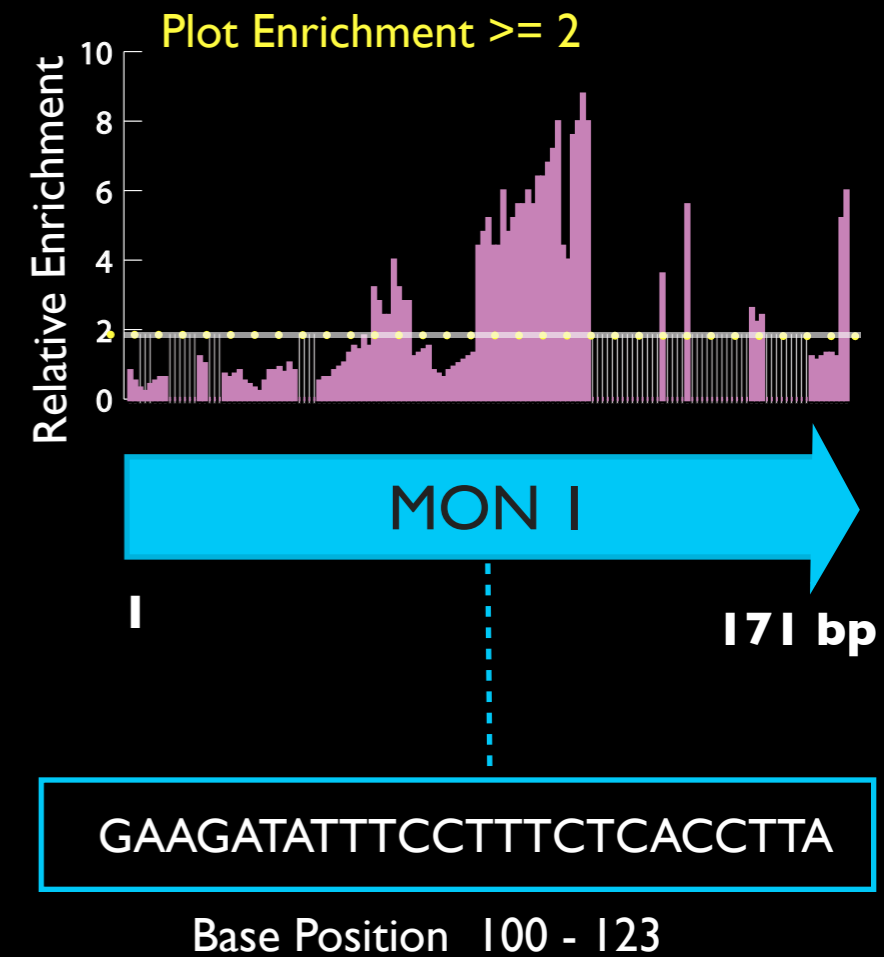
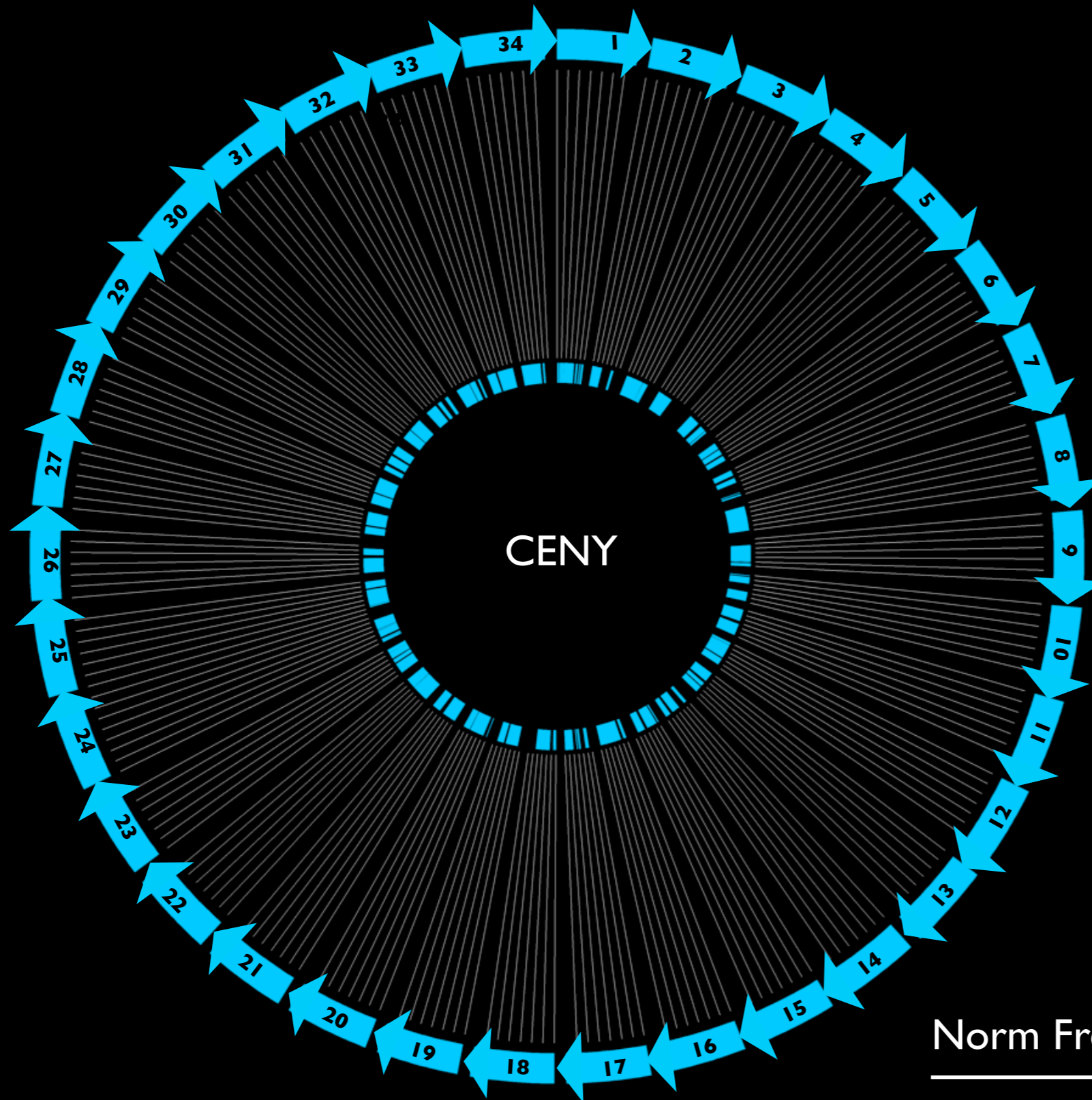


ENCODE Tier I: Human Embryonic Stem Cell (HI-hESC)



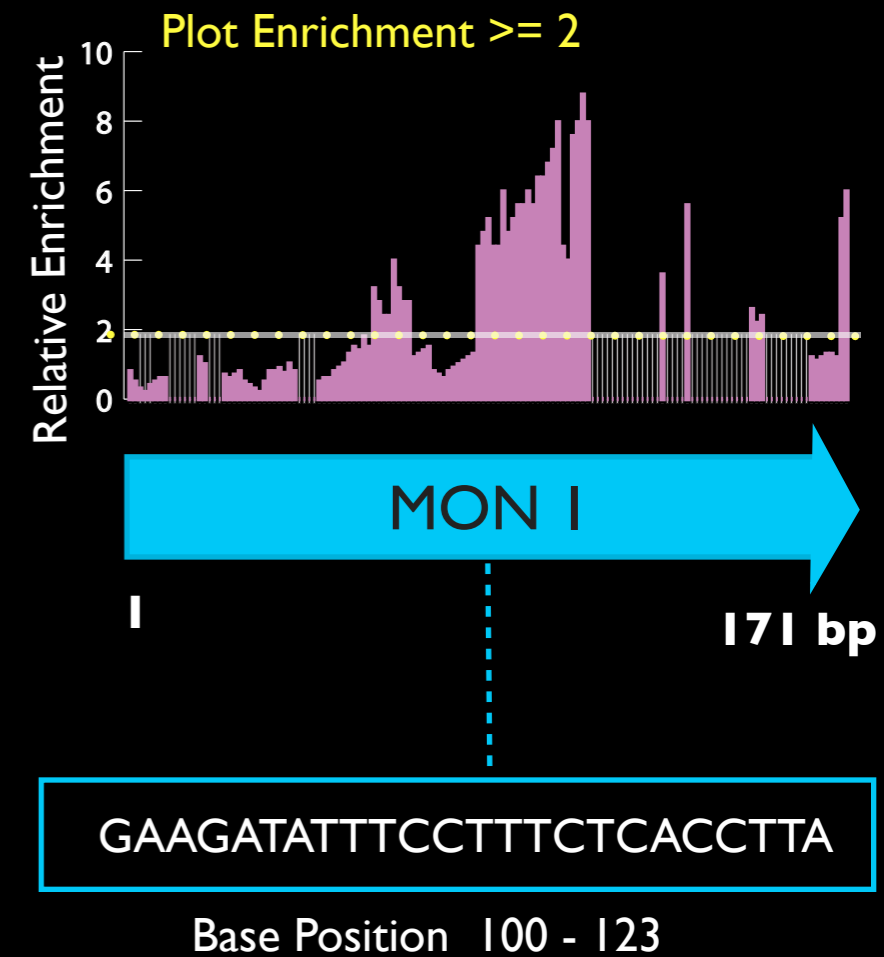
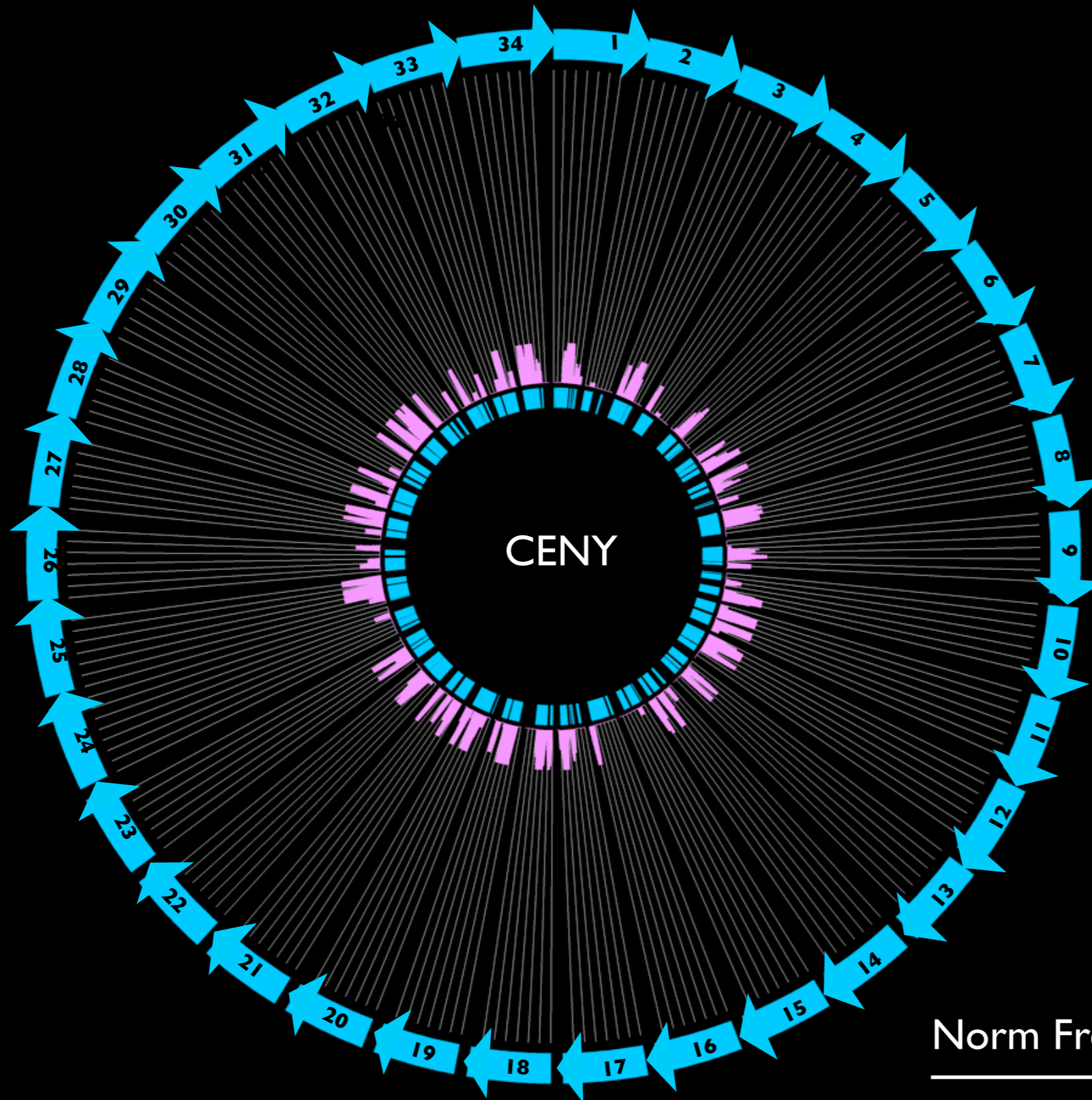
$$\frac{\text{Norm Freq H3K9me3 (IP)}}{\text{Norm Freq H3K9me3 (M)}} = \text{Relative Enrichment}$$

ENCODE Tier I: Human Embryonic Stem Cell (HI-hESC)



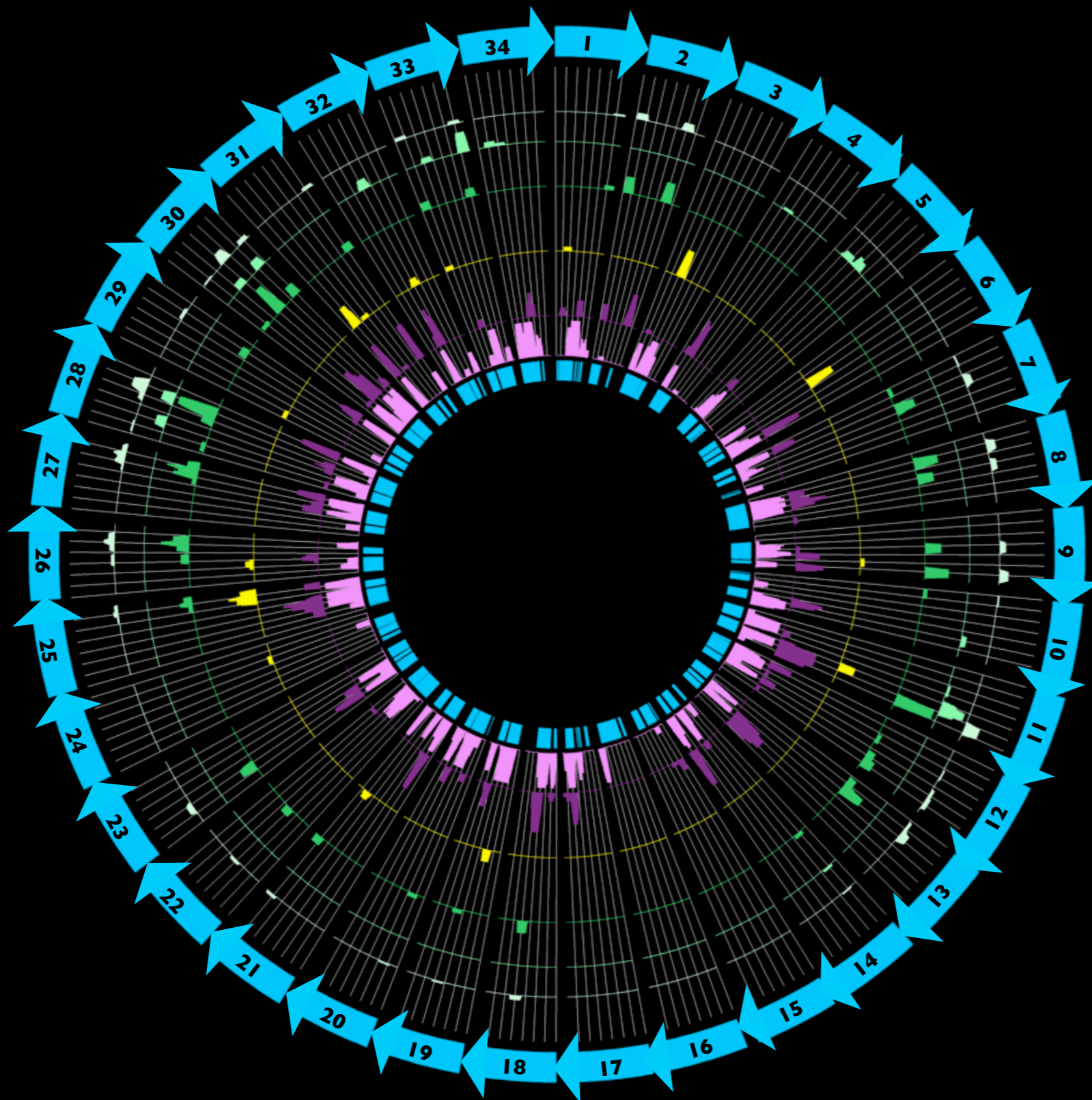
$$\frac{\text{Norm Freq H3K9me3 (IP)}}{\text{Norm Freq H3K9me3 (M)}} = \text{Relative Enrichment}$$

ENCODE Tier I: Human Embryonic Stem Cell (HI-hESC)



$$\frac{\text{Norm Freq H3K9me3 (IP)}}{\text{Norm Freq H3K9me3 (M)}} = \text{Relative Enrichment}$$

HiChESC Histone Profile of DYZ3 Array



Active Chromatin

- H3K4me1
- H3K4me2
- H3K4me3

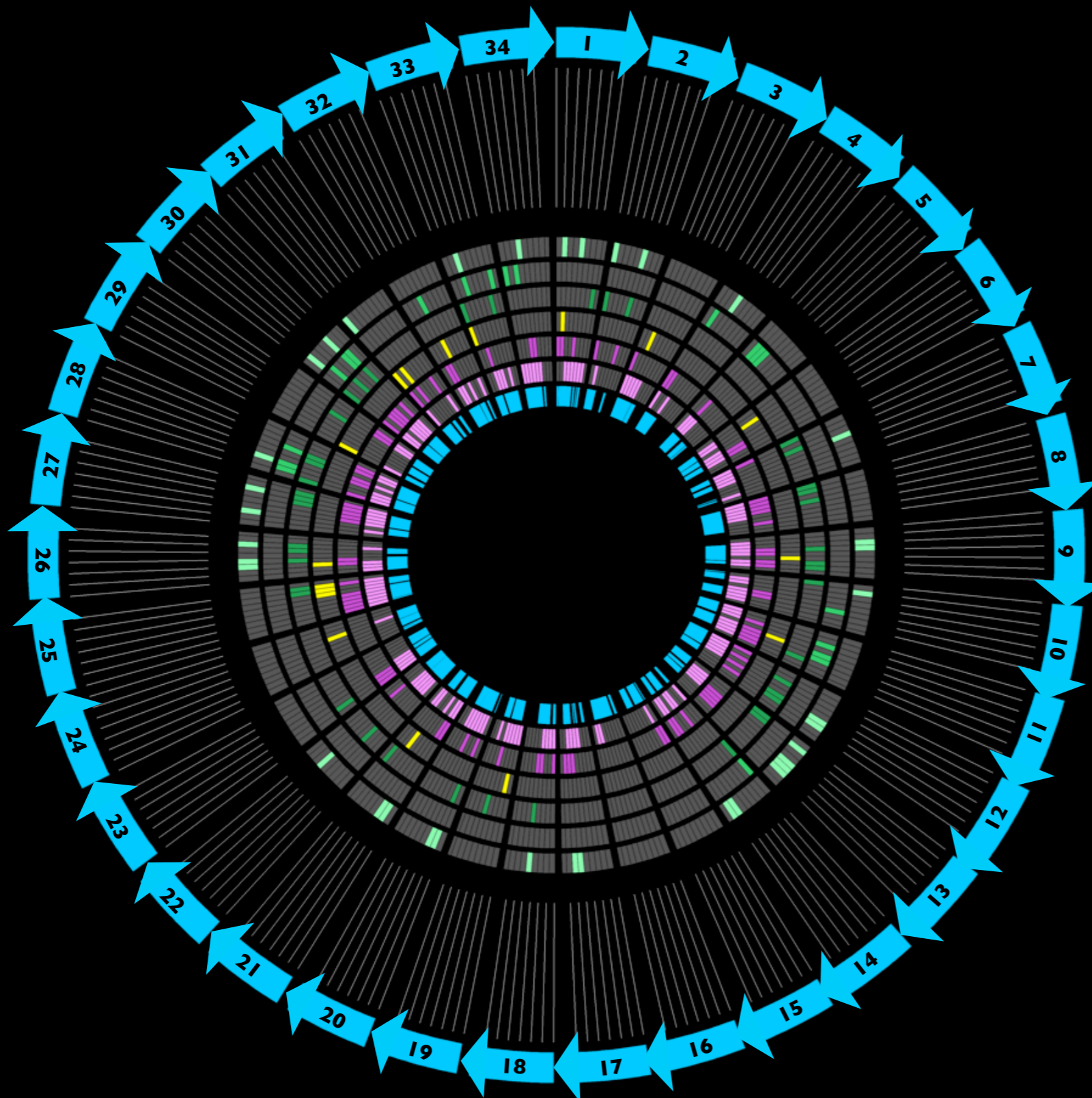
Histone Variants

- H2Az

Repressive Chromatin

- H3K27me3
- H3K9me3

HiChIP Histone Profile of DYZ3 Array



Active Chromatin

- H3K4me1
- H3K4me2
- H3K4me3

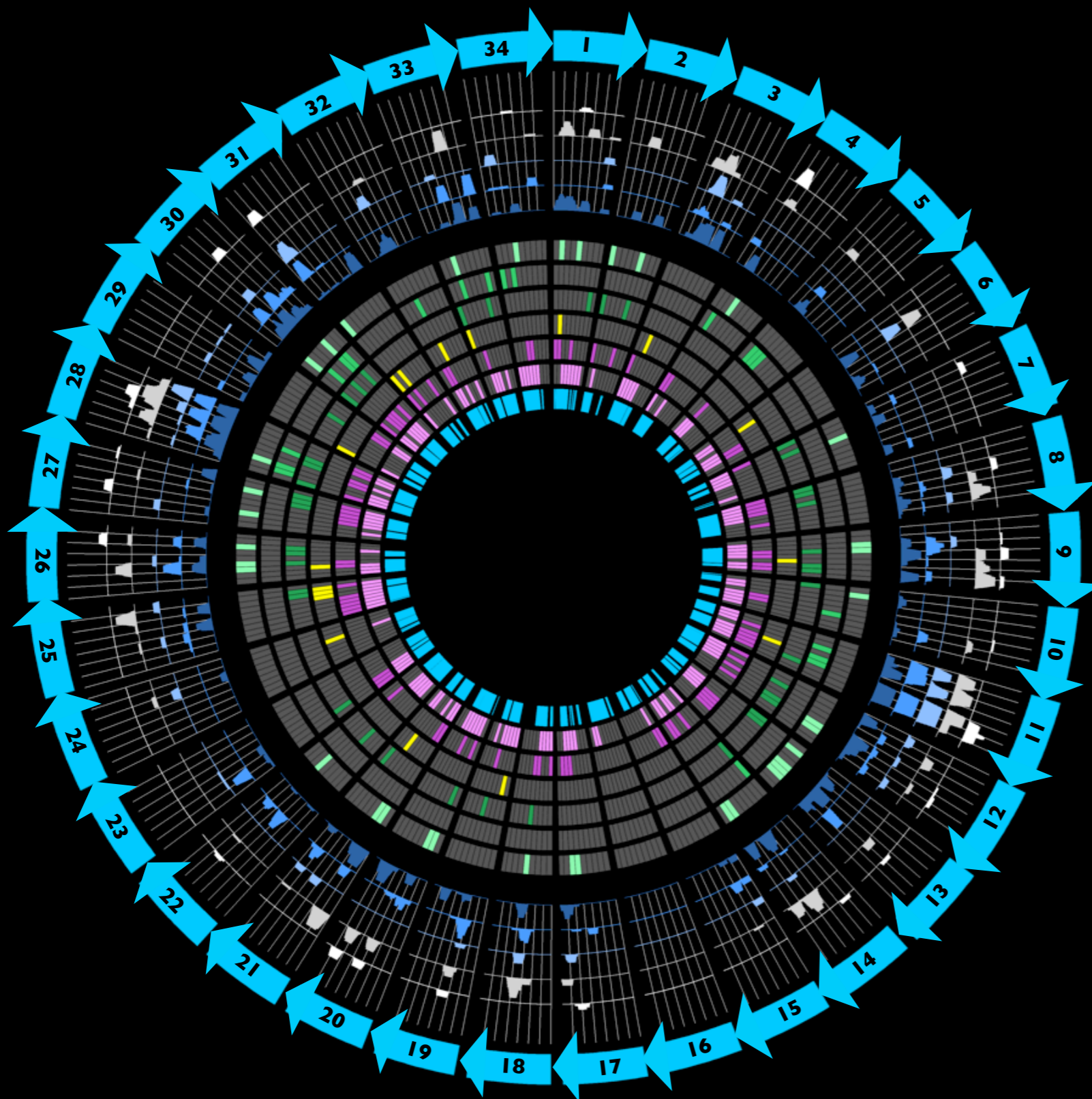
Histone Variants

- H2Az

Repressive Chromatin

- H3K27me3
- H3K9me3

HiChES Transcription Factor Enrichment Profile



Transcription Factor

- EZH2
- HDAC6
- PLU1
- JARID1A
- SUZ12

Active Chromatin

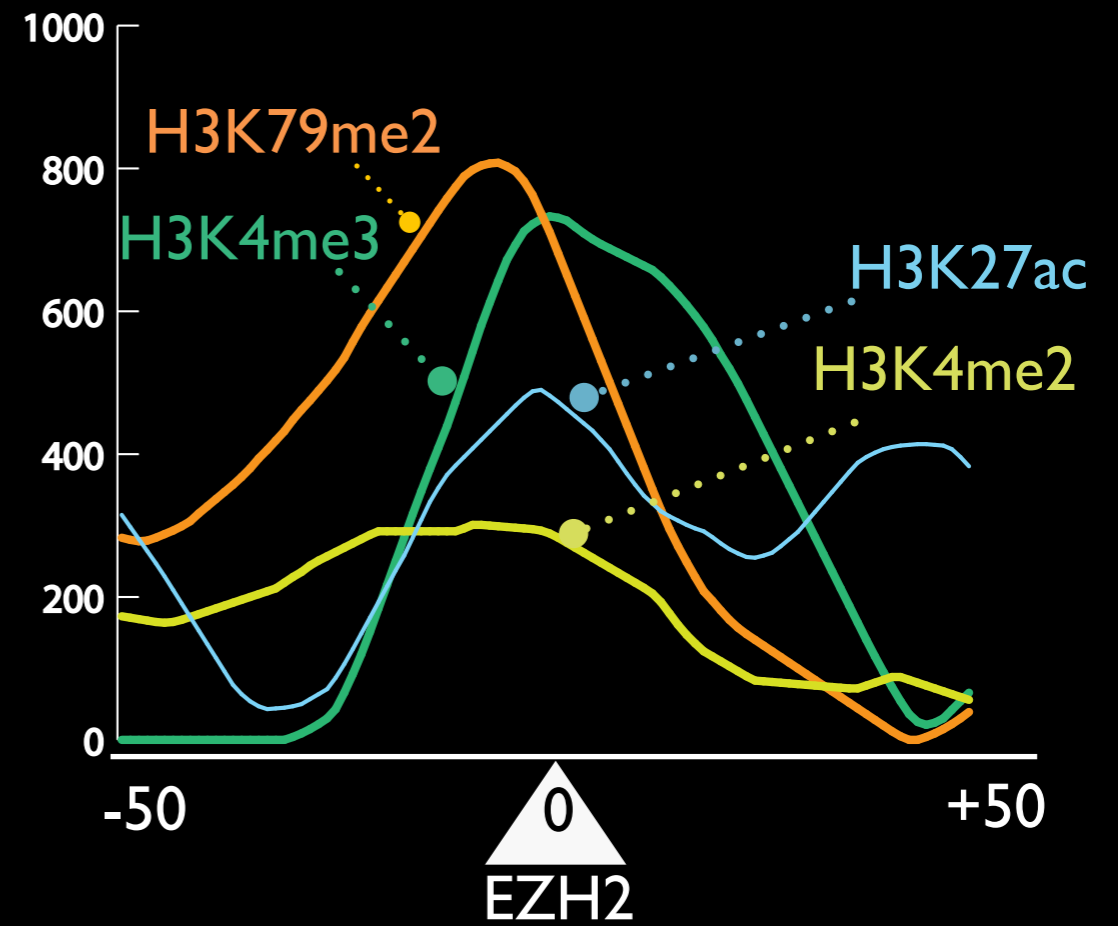
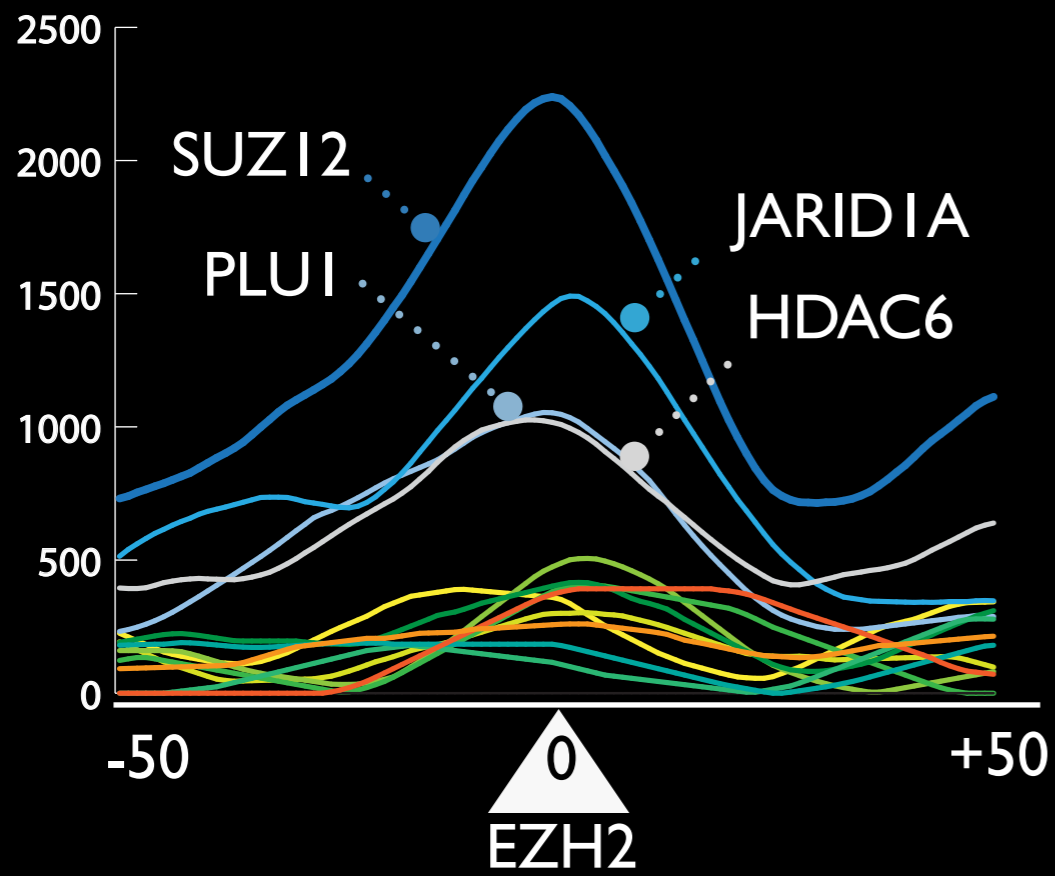
- H3K4me1
- H3K4me2
- H3K4me3

Histone Variants

- H2Az

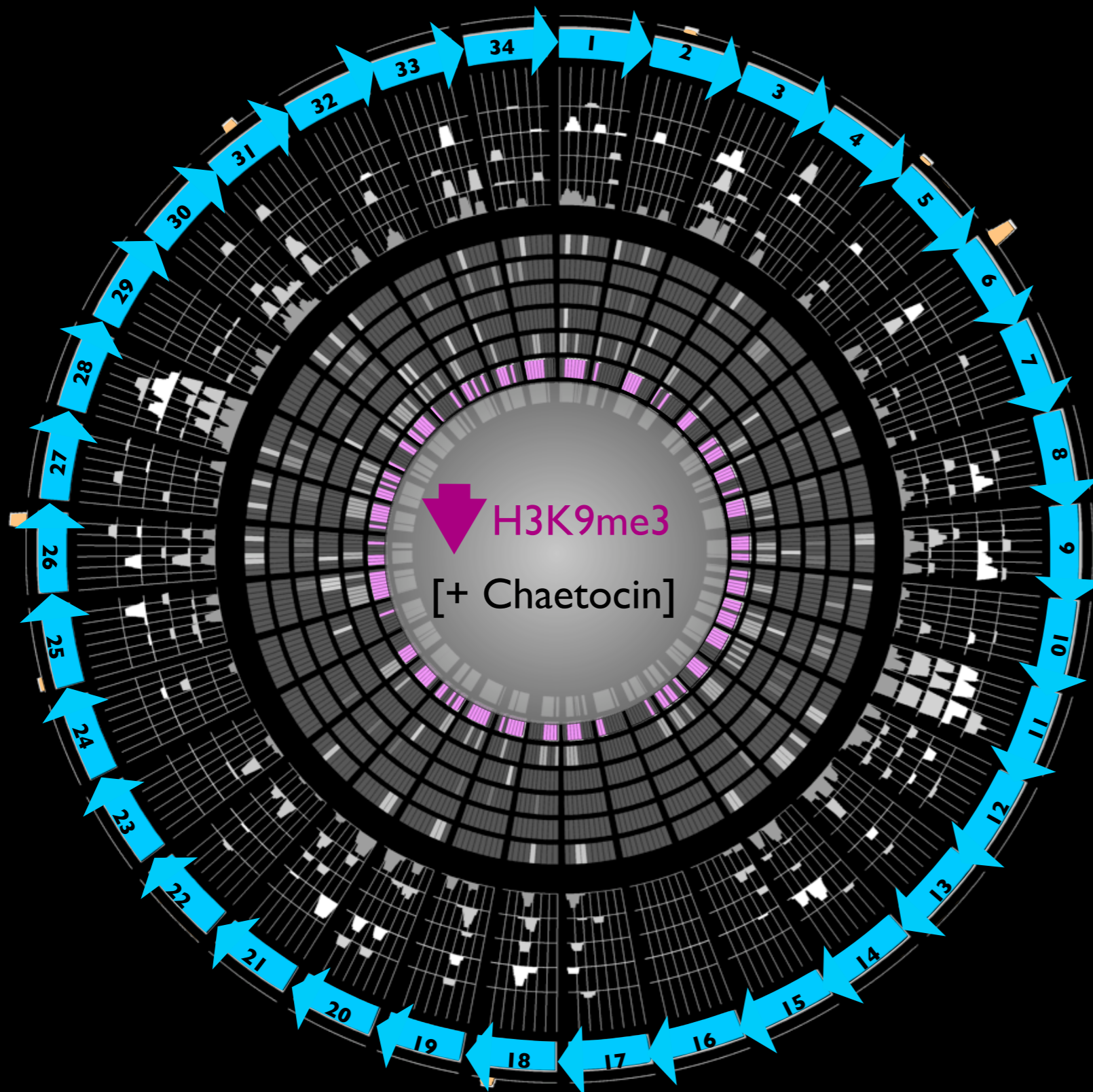
Repressive Chromatin

- H3K27me3
- H3K9me3



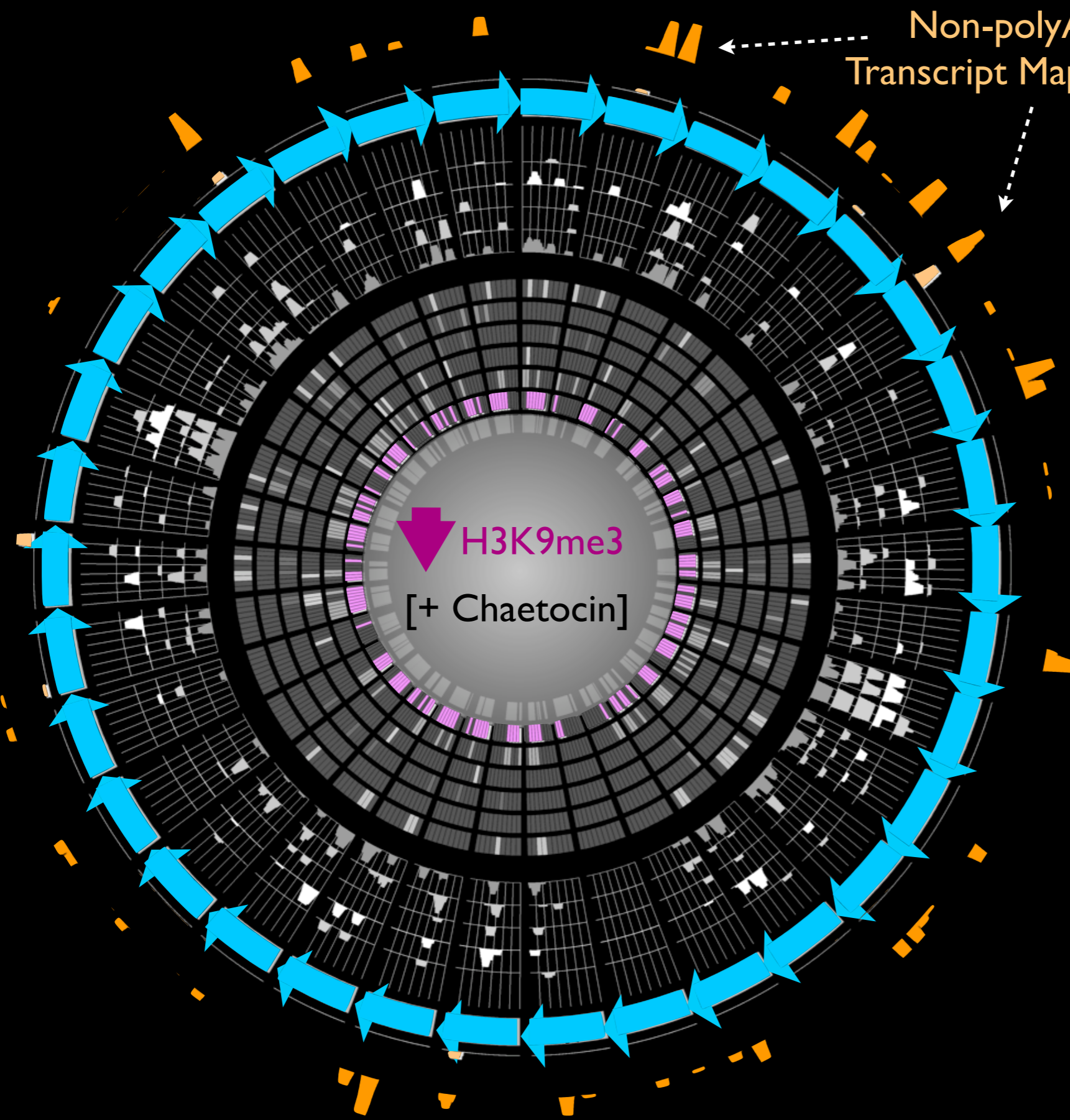
- PHF8
- JMJD2
- P300
- CHD7
- SIRT6
- RBBP5
- HDAC2
- CTCF
- CHD1

Adding Custom Datasets or “Tracks”



- Transcription Factor
 - EZH2
 - HDAC6
 - PLU1
 - JARID1A
 - SUZ12
- Active Chromatin
 - H3K4me1
 - H3K4me2
 - H3K4me3
- Histone Variants
 - H2Az
- Repressive Chromatin
 - H3K27me3
 - H3K9me3

Non-polyA
Transcript Mapping



Transcription Factor

- EZH2
- HDAC6
- PLU1
- JARID1A
- SUZ12

Active Chromatin

- H3K4me1
- H3K4me2
- H3K4me3

Histone Variants

- H2Az

Repressive Chromatin

- H3K27me3
- H3K9me3



UCSC: Centromere Annotation and Tool Development

Home Genomes Genome Browser Tools Mirrors Downloads My Data About Us Help

CentromereY (Human Centromere Reference Models) TEST Genome Browser Gateway

The UCSC Genome Browser was created by the [Genome Bioinformatics Group of UC Santa Cruz](#).
Software Copyright (c) The Regents of the University of California. All rights reserved.

group	genome	assembly	position	search term
Other	CentromereY	2013 GrCh38	DYZ3:1-5,785	enter position or search terms

[Click here to reset](#) the browser user interface settings to their defaults.

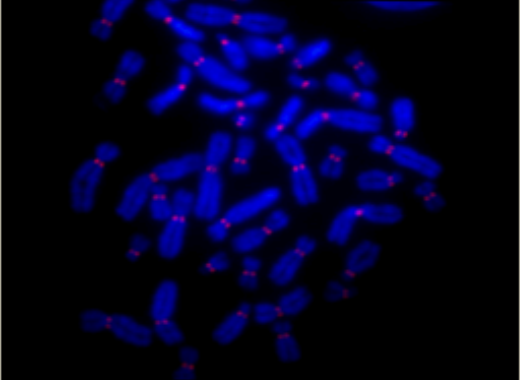
WARNING: This is our development and test site. It usually works, but it is filled with tracks in various stages of construction, and others of little interest to people outside of our local group. It is usually slow because we are building databases on it. The documentation is poor. More data than usual is flat out wrong. Maybe you want to go to genome.ucsc.edu instead.

CentromereY Genome Browser – centromers1 assembly ([sequences](#))

Karen Miga's reconstructed centromer reference sequence, with ENCODE annotations mapped to them. This is part of the the 2013 GrCH38 reference genome sequence. In this browser, it is represented as one long sequence composed of monomers.

Search the assembly:

- **By position or search term:** Use the "position or search term" box to find areas of the genome associated with many different attributes, such as a specific chromosomal coordinate range; mRNA, EST, or STS marker names; or keywords from the GenBank description of an mRNA. [More information](#), including sample queries.
- **By track type:** Click the "track search" button to find Genome Browser tracks that match specific selection criteria. [More information](#).



Reconstructed Centromeres
([Karen Miga](#))



UCSC: Centromere Annotation and Tool Development

Genomes Genome Browser Tools Mirrors Downloads My Data About Us View Help

UCSC TEST Genome Browser on CentromereY 2013 GrCh38 Assembly (centromeres1)

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

DYZ3:1-5,785 5,785 bp. enter position or search terms go

move start < 2.0 > Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. move end < 15.0 >

default tracks default order hide all add custom tracks track hubs configure reverse resize refresh

collapse all Use drop-down controls below and press refresh to alter tracks displayed. expand all
Tracks with lots of items will automatically be displayed in more compact modes.

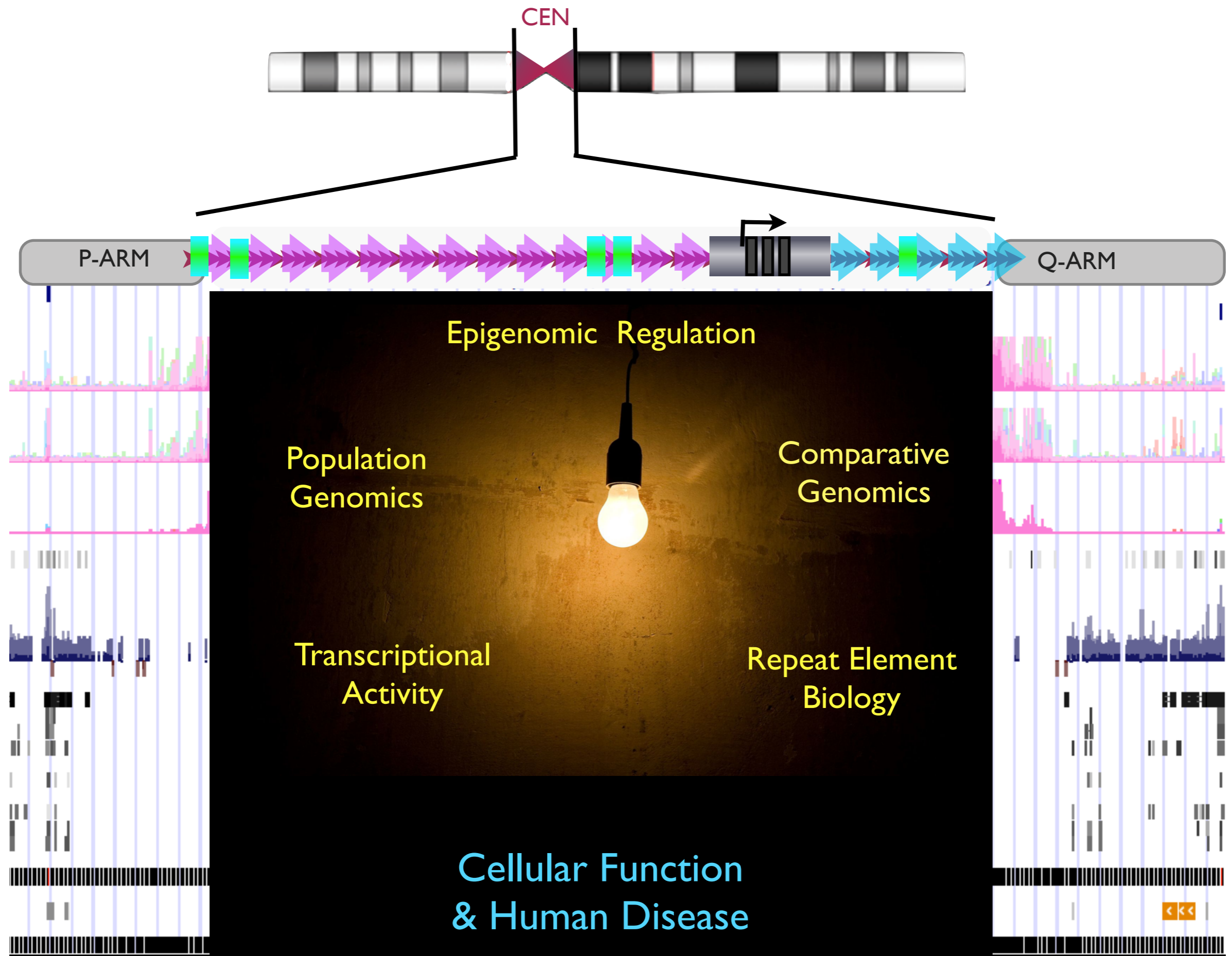
Mapping and Sequencing Tracks refresh

Base Position Assembly Gap GC Percent Short Match Restr Enzymes

hide full dense hide hide hide

Genes and Gene Prediction Tracks refresh

Human centromeric regions are currently defined by gaps in the reference assembly



A vertical strip on the left side of the slide shows a fluorescence microscopy image of chromosomes. The DNA is stained blue, and specific regions are highlighted in red, likely representing centromeres or specific bands of interest. The chromosomes are arranged in a somewhat organized pattern, typical of a karyotype or a specific stage of cell division.

Acknowledgements

Jim Kent

Hunt Willard

Max Haeussler

Dave Greenberg

Kevin Brown

Brittany Wellence