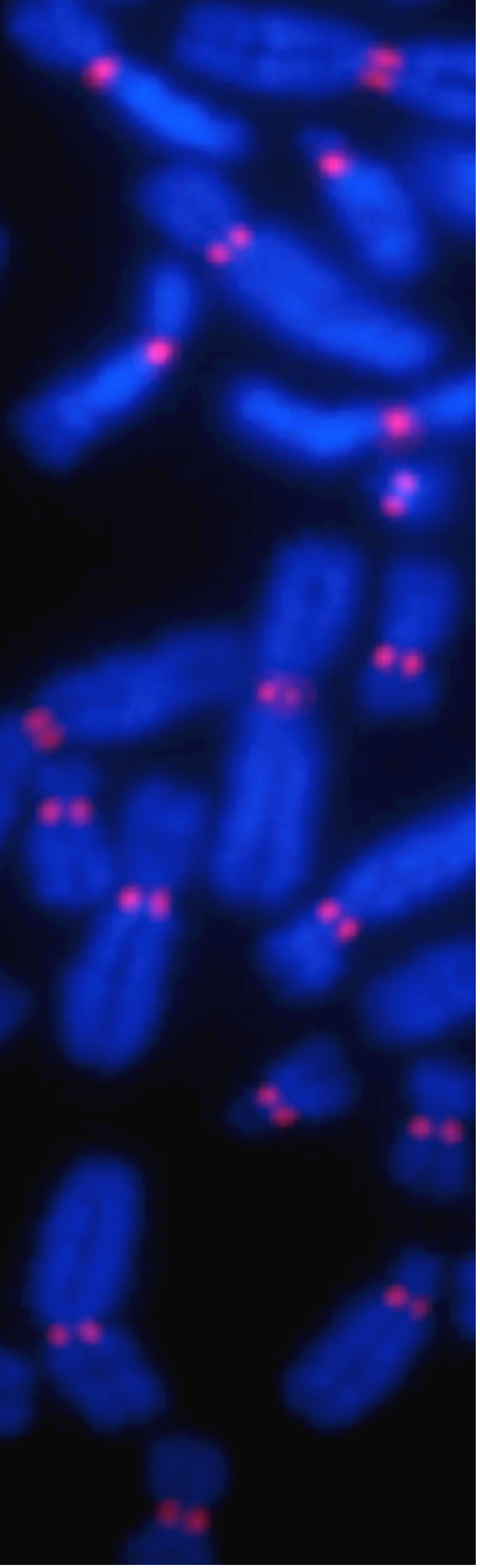


GRCh38 Centromere Reference Models

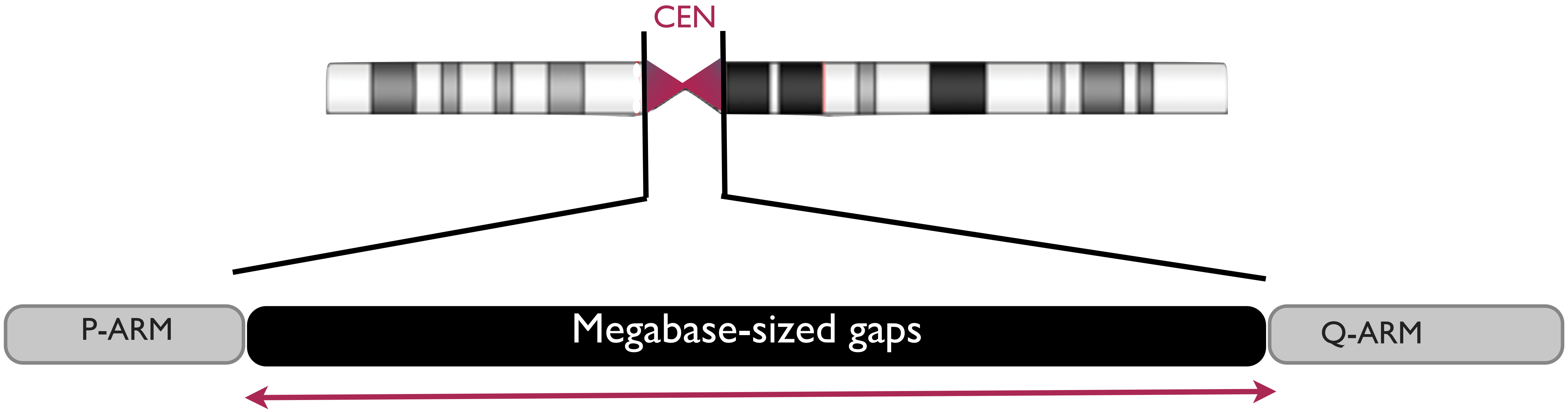
Karen H. Miga

University of California, Santa Cruz

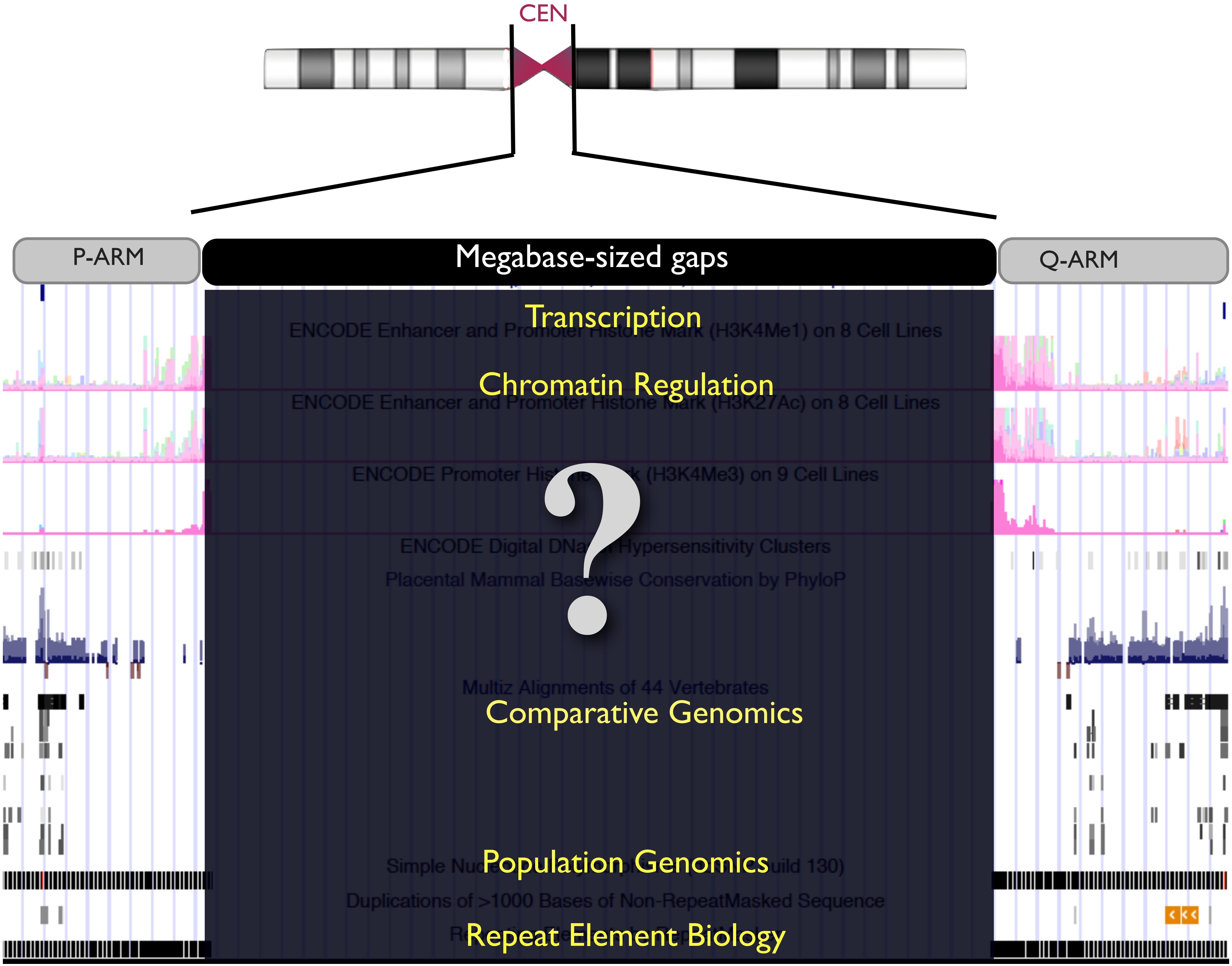
Jim Kent

Huntington F. Willard

Human centromeres are currently defined by gaps in the reference assembly



Human centromeres are currently defined by gaps in the reference assembly



CEN

P-ARM

Megabase-sized gaps

Q-ARM

Transcription

Chromatin Regulation

?

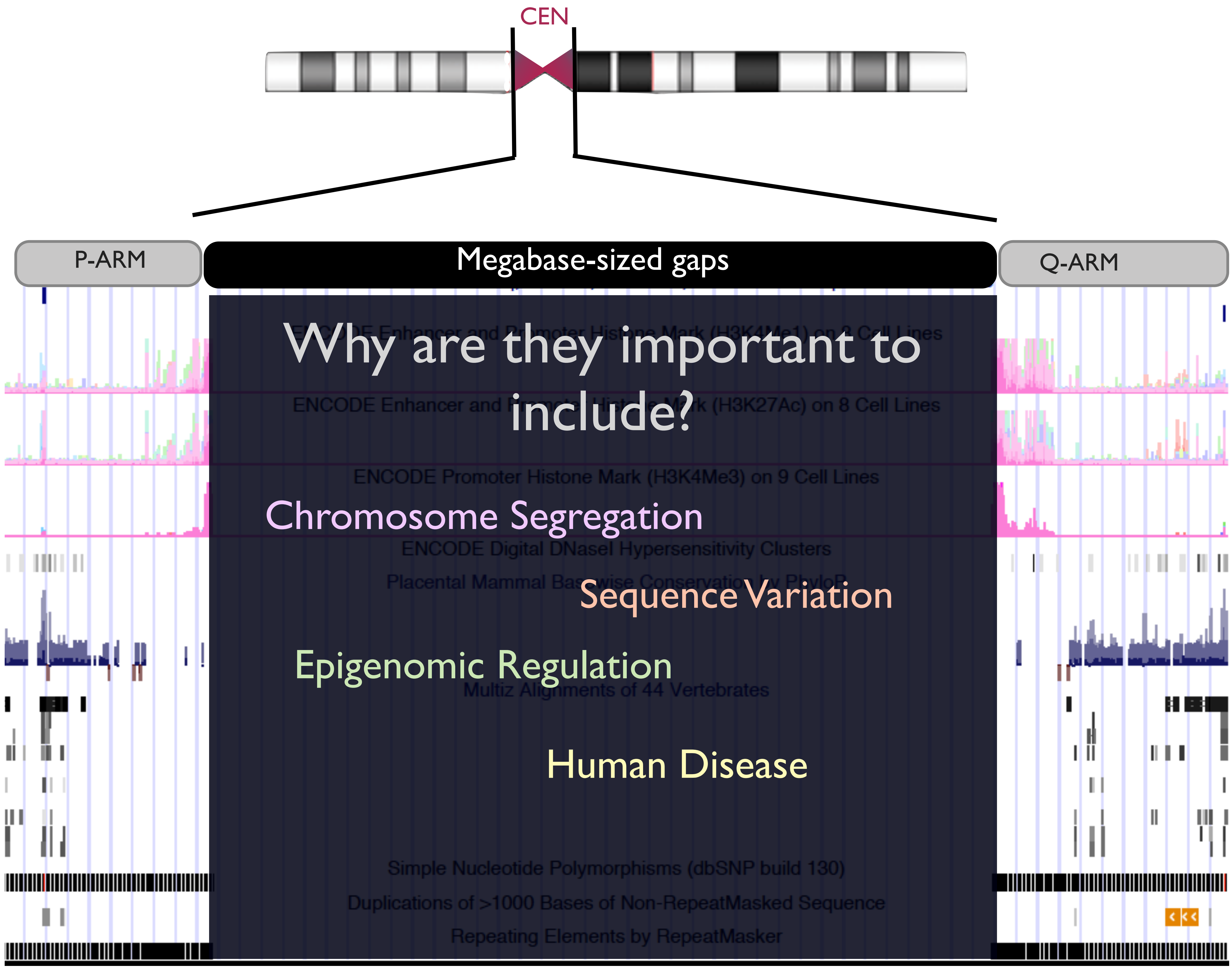
Comparative Genomics

Population Genomics

Repeat Element Biology

ENCODE Enhancer and Promoter Histone Mark (H3K4Me1) on 8 Cell Lines
ENCODE Enhancer and Promoter Histone Mark (H3K27Ac) on 8 Cell Lines
ENCODE Promoter Histone Mark (H3K4Me3) on 9 Cell Lines
ENCODE Digital DNaseI Hypersensitivity Clusters
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 44 Vertebrates
Simple Nucleotide Polymorphisms (dbSNP build 130)
Duplications of >1000 Bases of Non-RepeatMasked Sequence

Human centromeres are currently defined by gaps in the reference assembly


CEN
P-ARM
Megabase-sized gaps
Q-ARM
ENCODE Enhancer and Promoter Histone Mark (H3K4Me1) on 8 Cell Lines
ENCODE Enhancer and Promoter Histone Mark (H3K27Ac) on 8 Cell Lines
ENCODE Promoter Histone Mark (H3K4Me3) on 9 Cell Lines
ENCODE Digital DNaseI Hypersensitivity Clusters
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 44 Vertebrates
Simple Nucleotide Polymorphisms (dbSNP build 130)
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Repeating Elements by RepeatMasker

Why are they important to include?

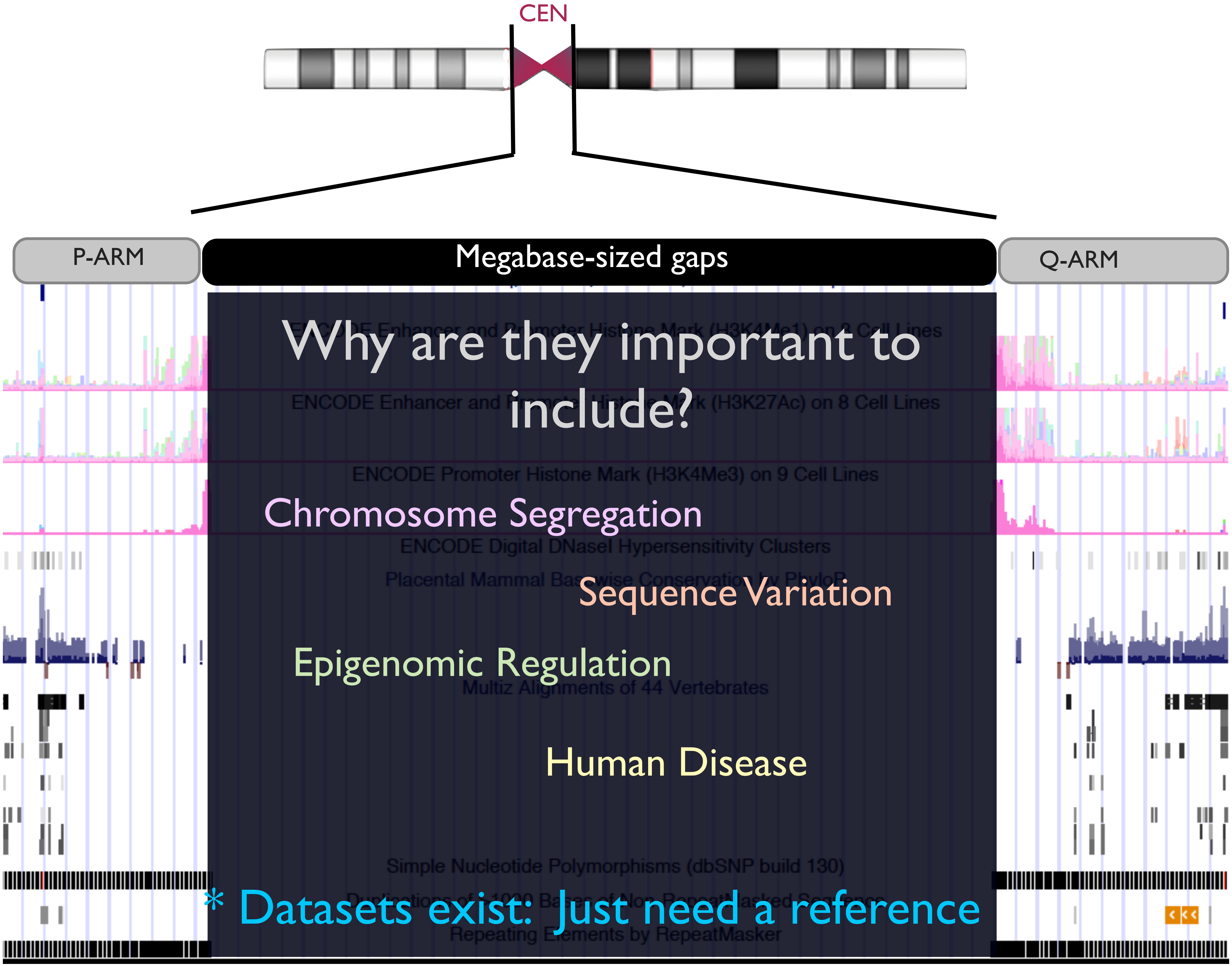
Chromosome Segregation

Sequence Variation

Epigenomic Regulation

Human Disease

Human centromeres are currently defined by gaps in the reference assembly


CEN
P-ARM
Megabase-sized gaps
Q-ARM
ENCODE Enhancer and Promoter Histone Mark (H3K4Me1) on 8 Cell Lines
ENCODE Enhancer and Promoter Histone Mark (H3K27Ac) on 8 Cell Lines
ENCODE Promoter Histone Mark (H3K4Me3) on 9 Cell Lines
ENCODE Digital DNaseI Hypersensitivity Clusters
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 44 Vertebrates
Simple Nucleotide Polymorphisms (dbSNP build 130)
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Repeating Elements by RepeatMasker

Why are they important to include?

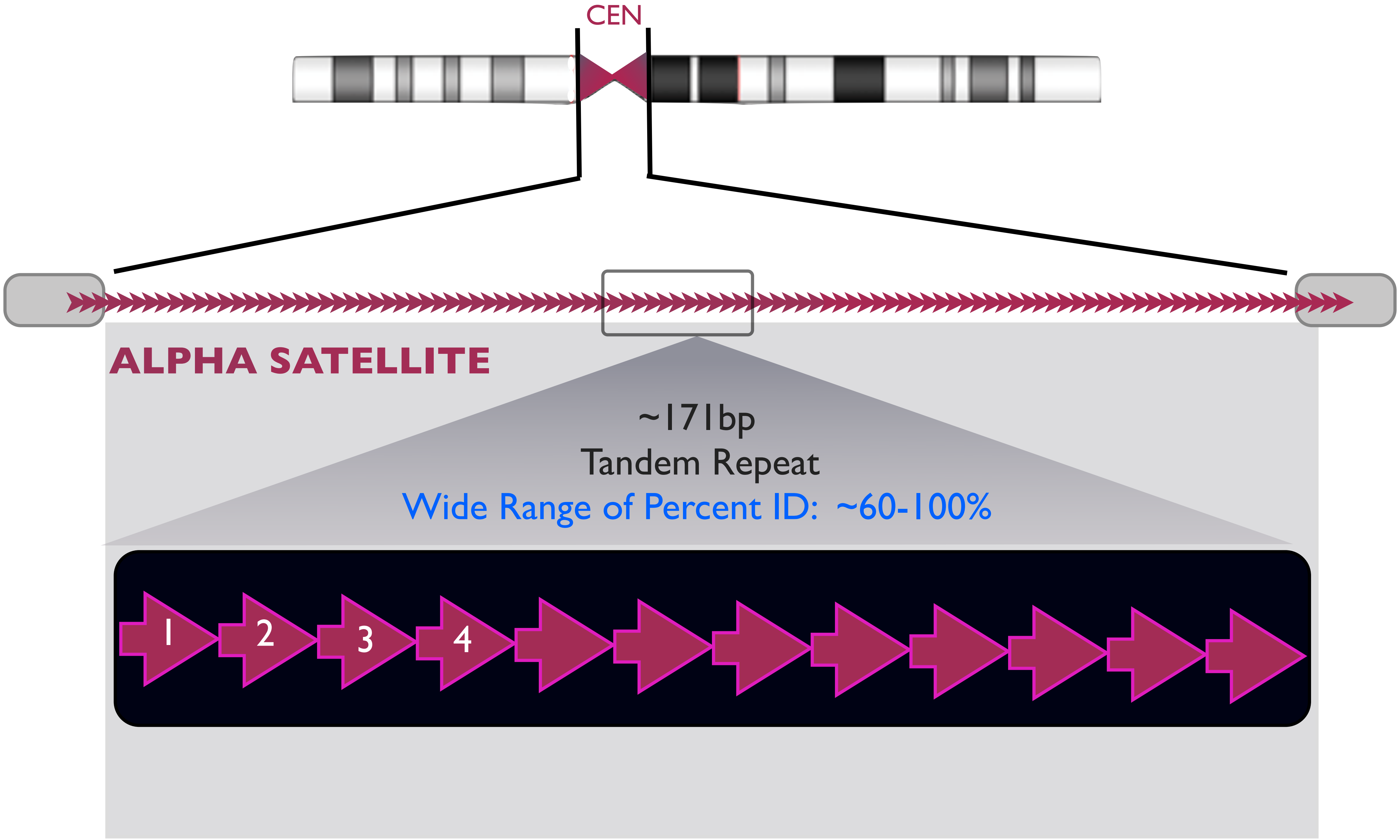
Chromosome Segregation

Sequence Variation

Epigenomic Regulation

Human Disease

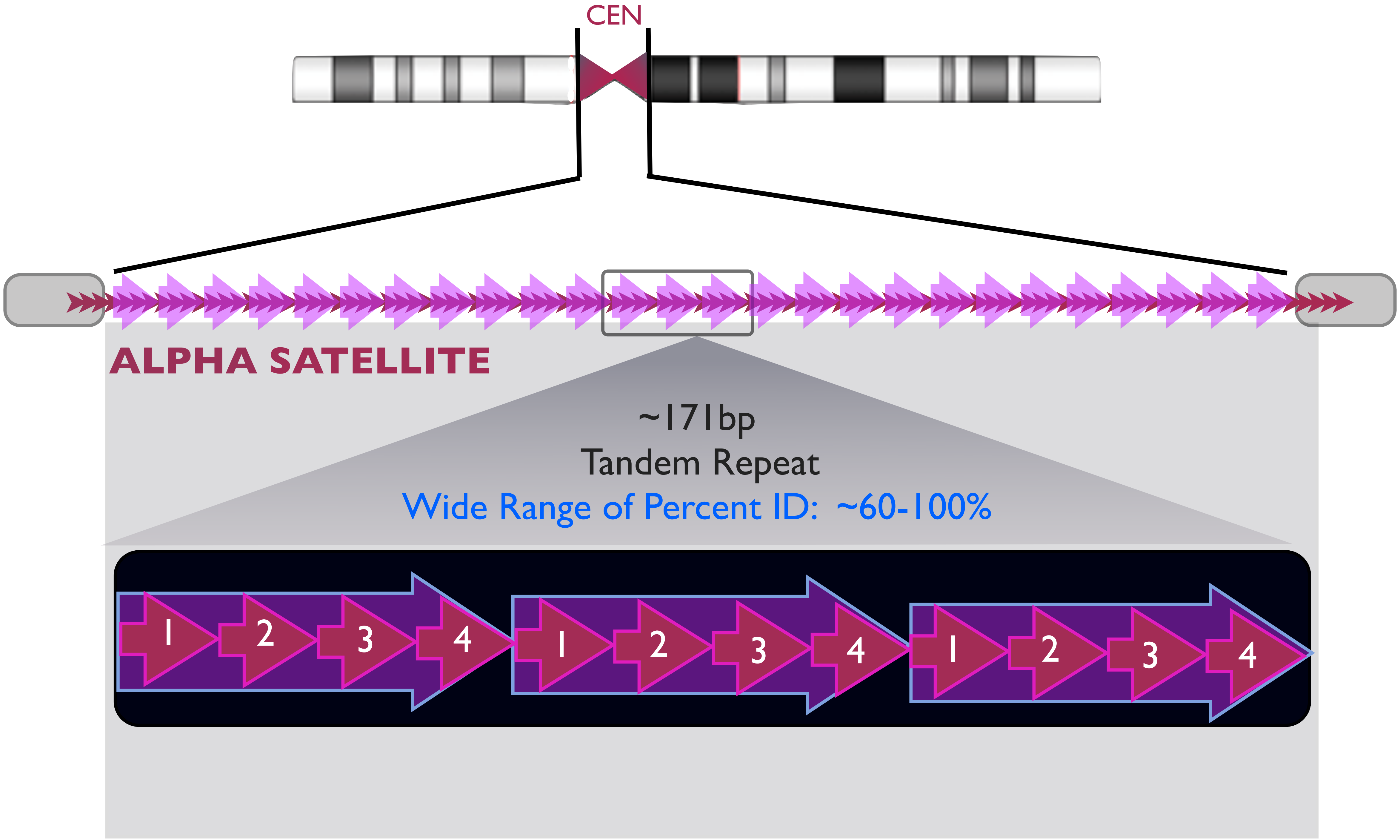
* Datasets exist: Just need a reference



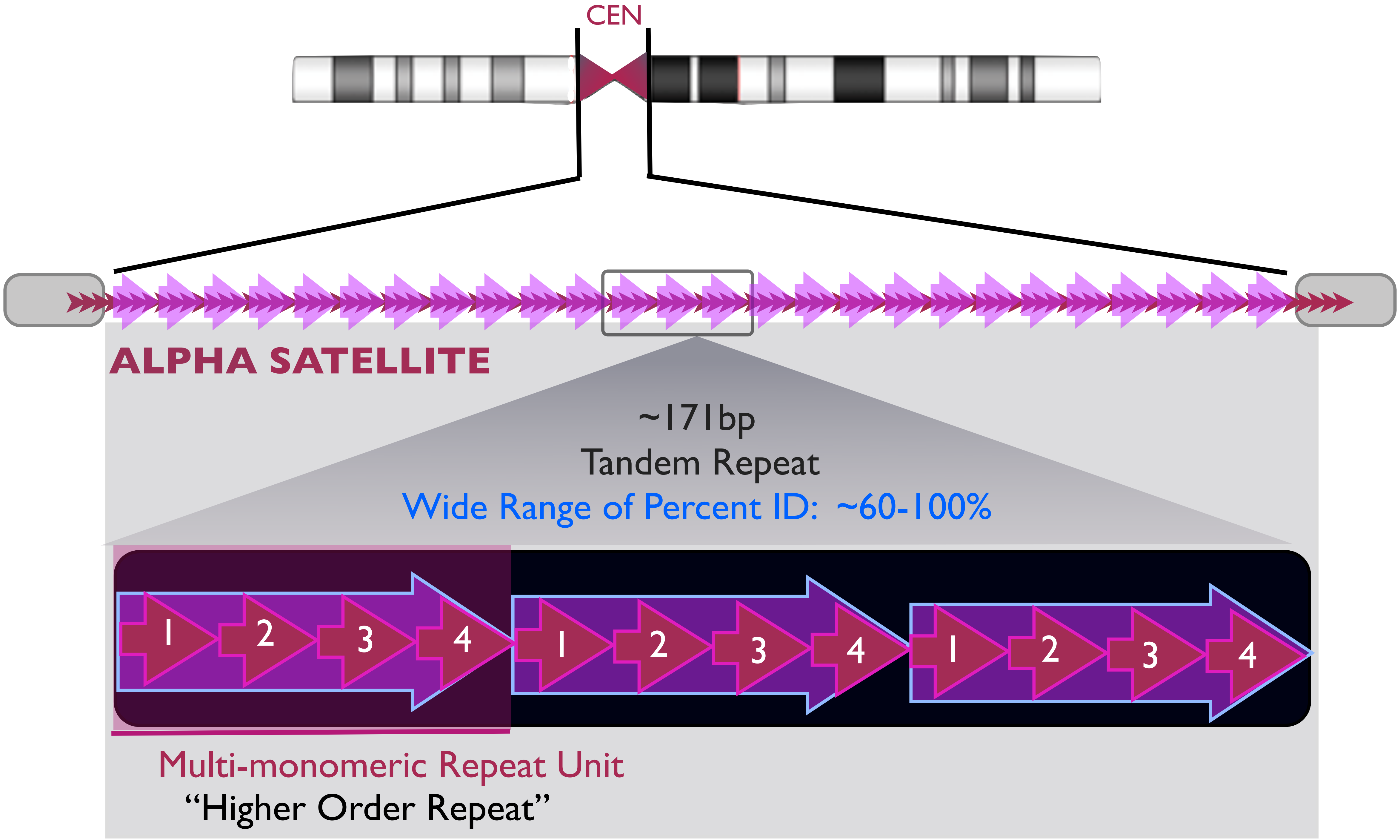
ALPHA SATELLITE

~171bp
Tandem Repeat
Wide Range of Percent ID: ~60-100%

Alpha Satellite define all normal human centromeres

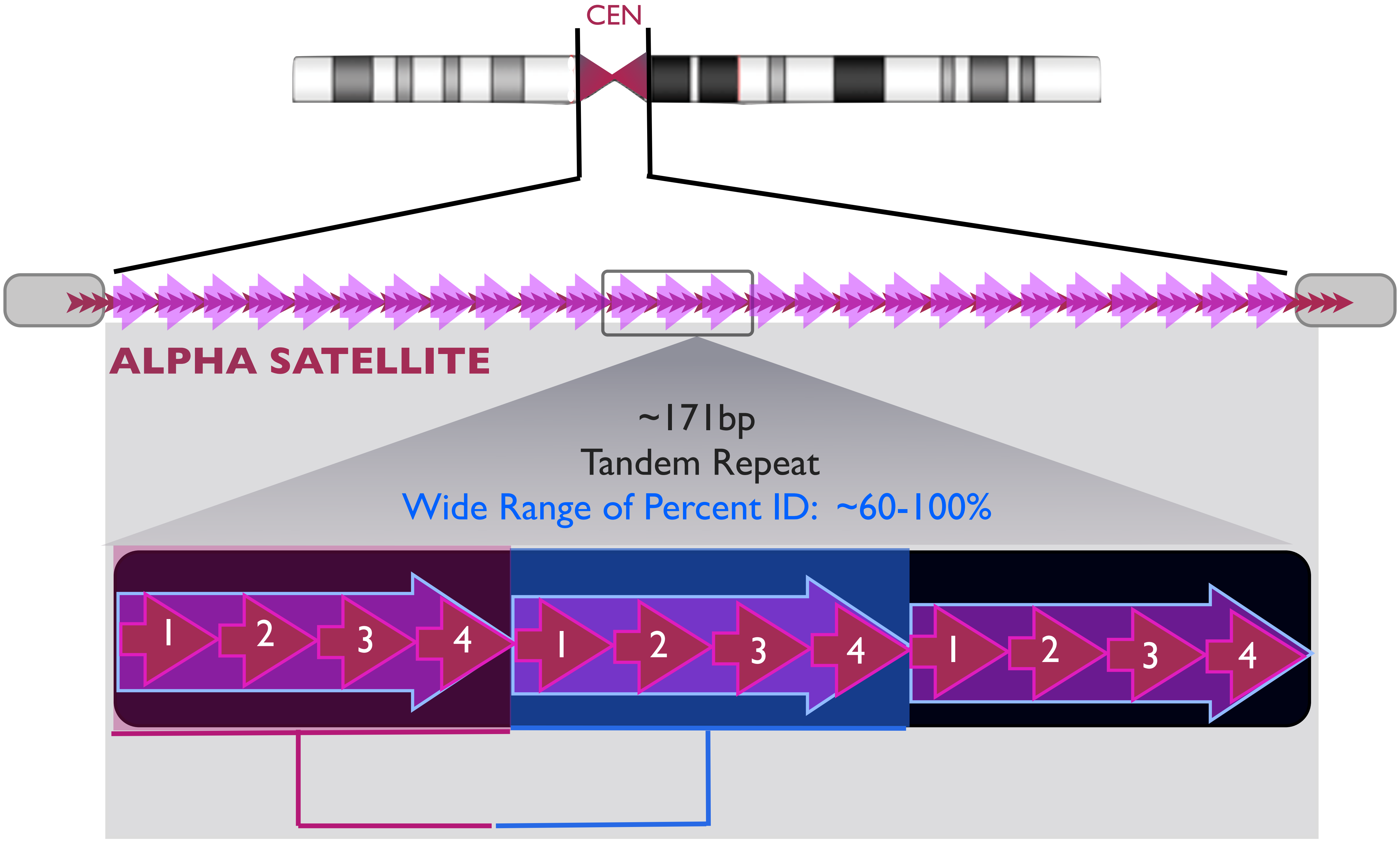


Alpha Satellite repeats (or monomers) are commonly found in long arrays of near-identical higher order repeats



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



CEN

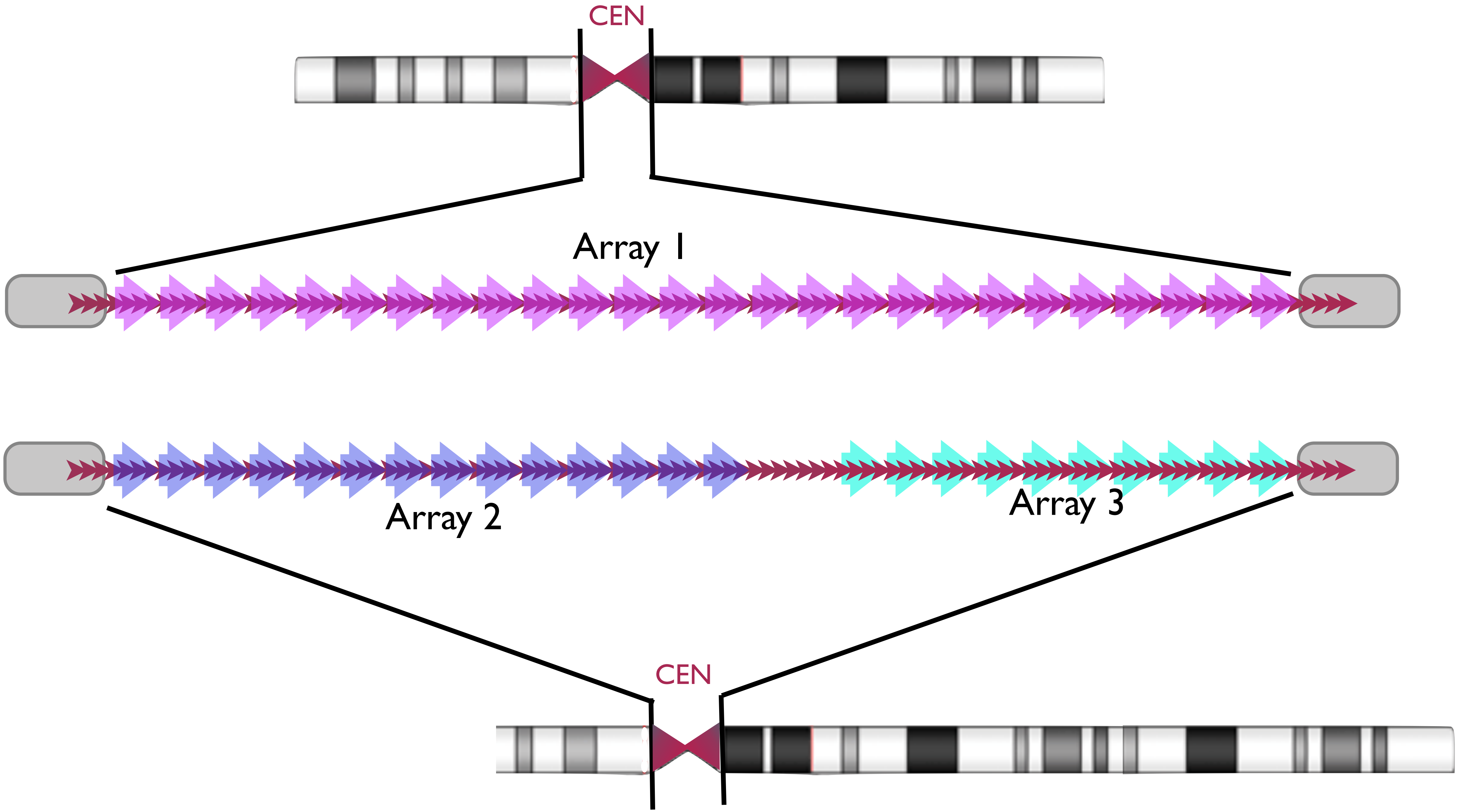
ALPHA SATELLITE

~171bp
Tandem Repeat
Wide Range of Percent ID: ~60-100%

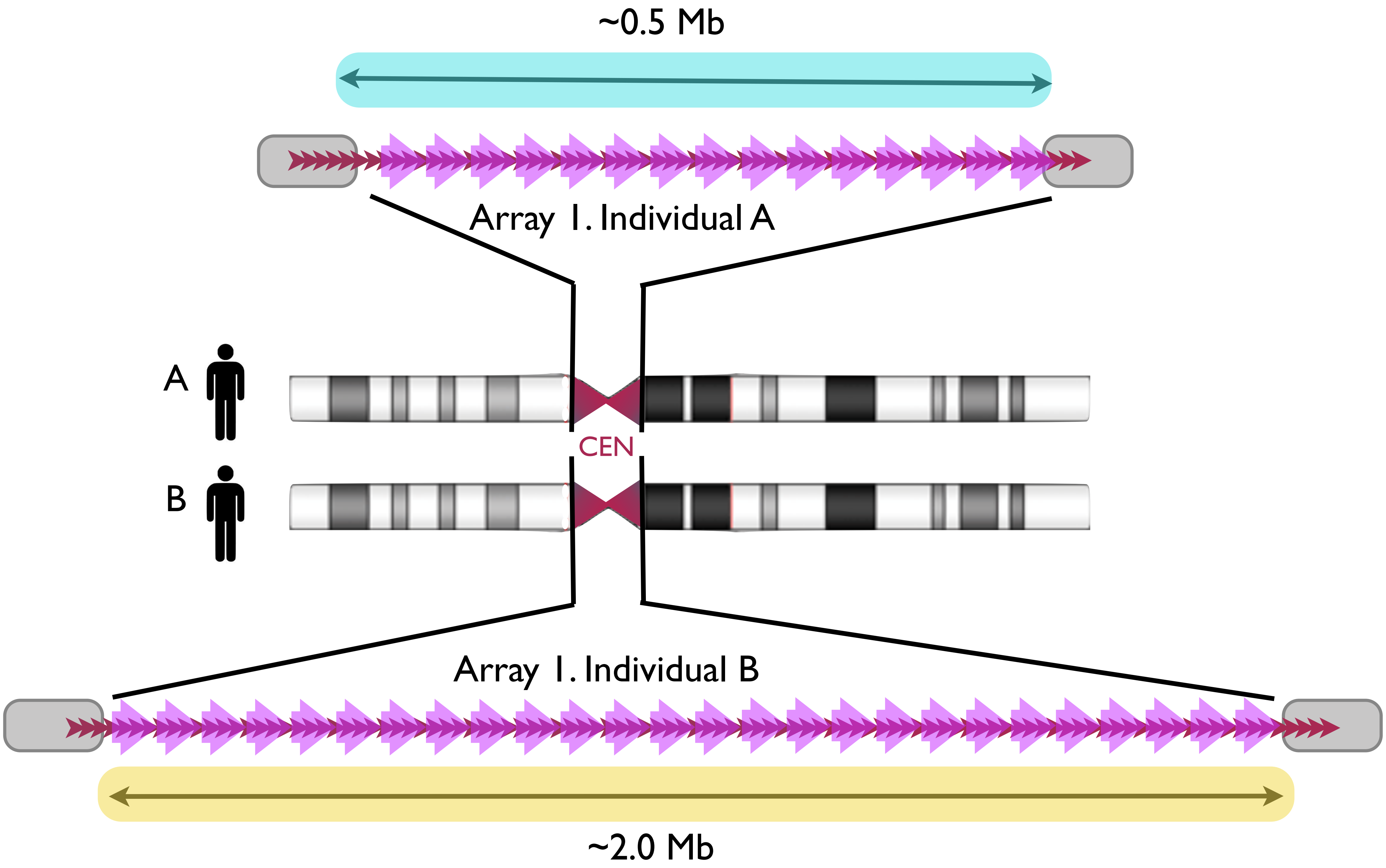
Narrow Range of Percent ID: **94% - 100%**

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

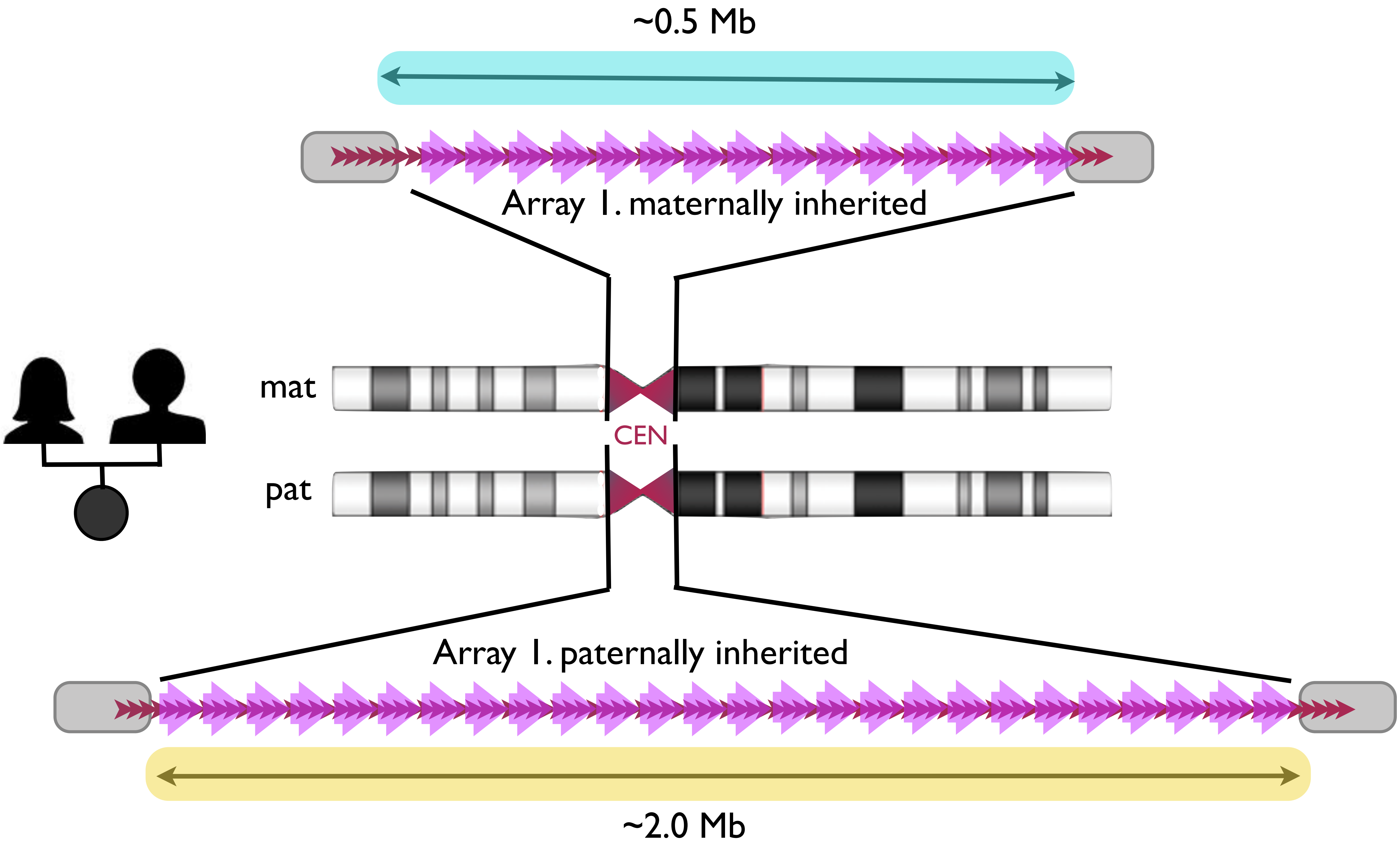
Each chromosome has a different centromeric sequences



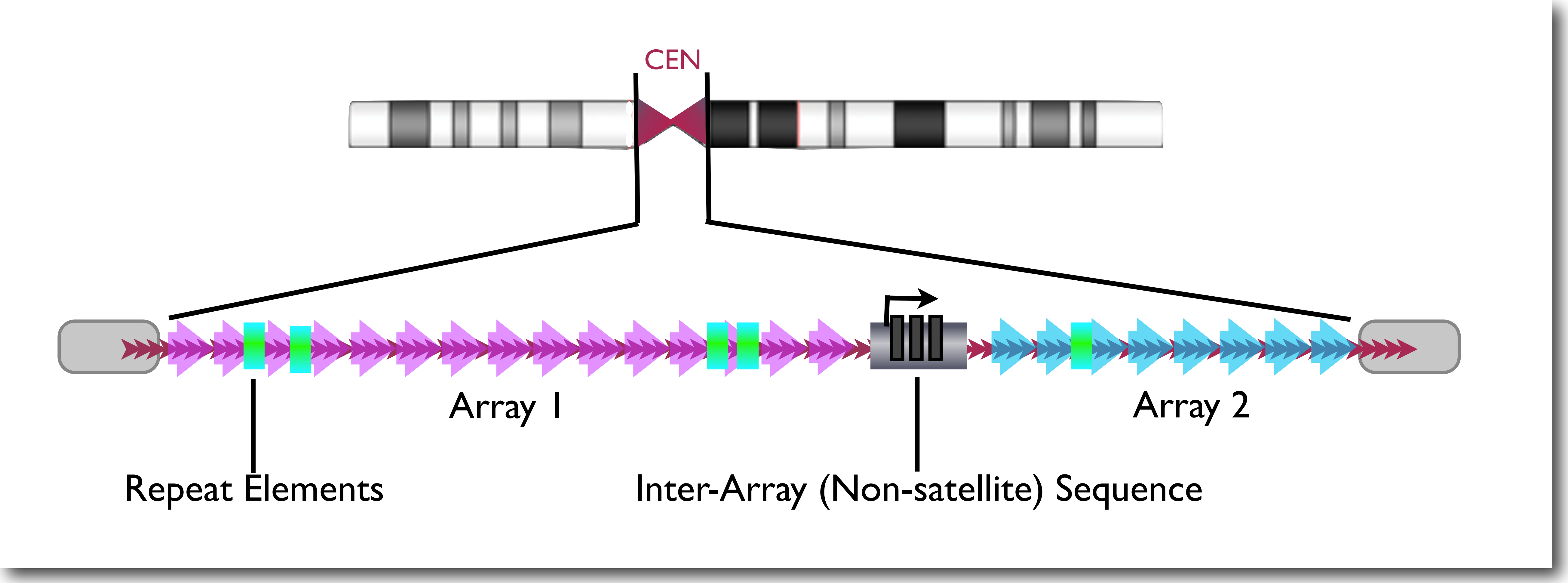
Higher-order arrays vary between individuals



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



Model of Centromere Sequence Organization



CEN

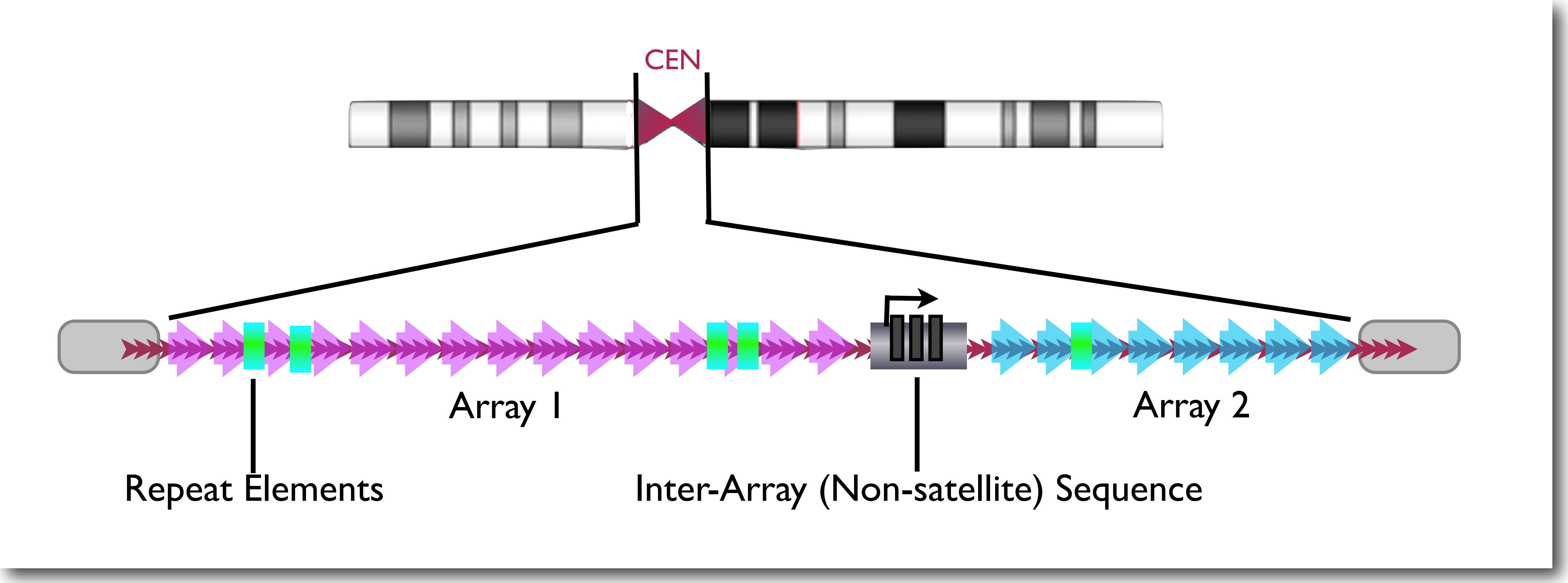
Array 1

Array 2

Repeat Elements

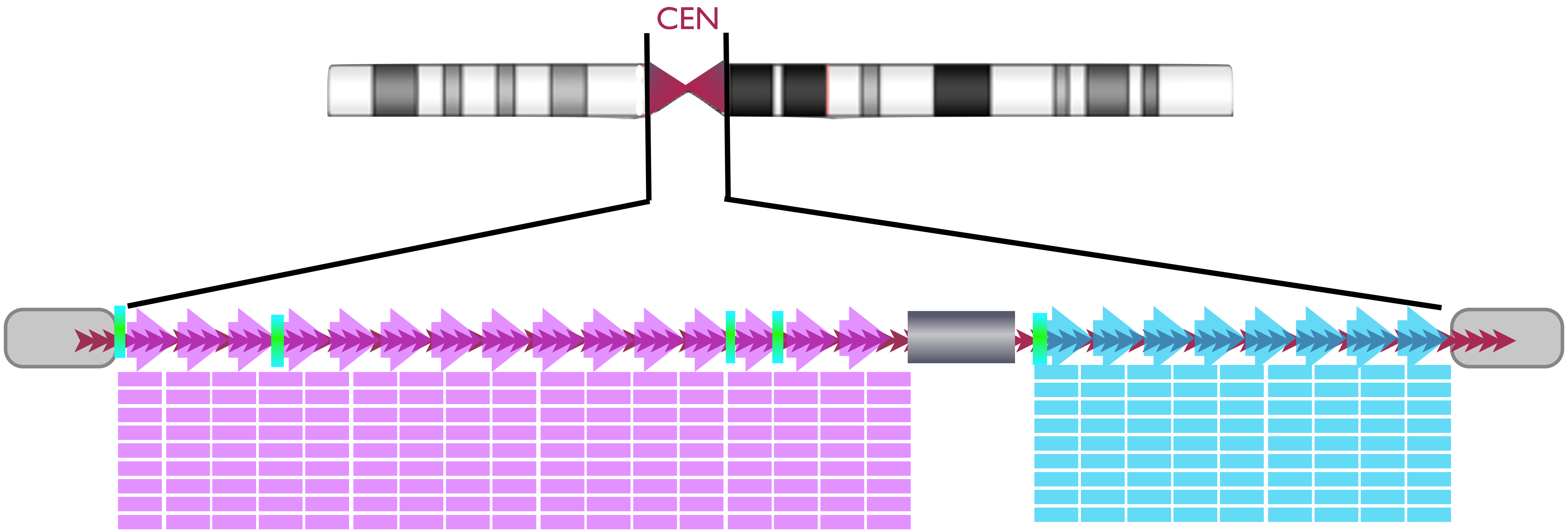
Inter-Array (Non-satellite) Sequence

Model of Centromere Sequence Organization



Goal: To generate a reference that models alpha satellite (and adjacent non-satellite) sequences within each centromeric gap

2. Reformat sequences observed in each read library into linear reference model

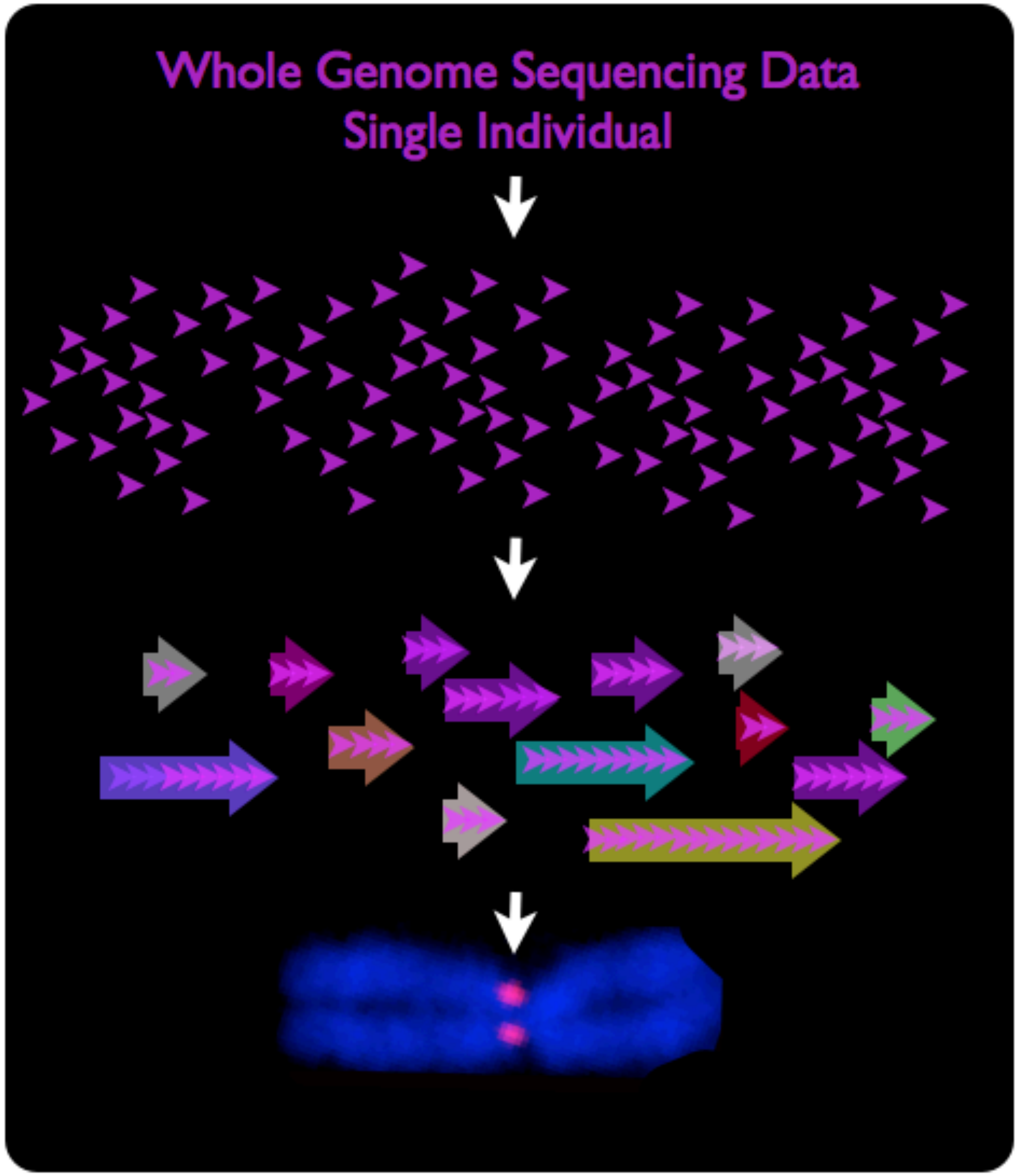


1 Constructing Read Libraries for each HOR array

2 LinearSat Software to Convert Reads to Linear Reference Models

3 Scaffold Reference Models and HuRef assembled contigs using mate pairs

1 Constructing Read Libraries for each HOR array



HuRef Genome

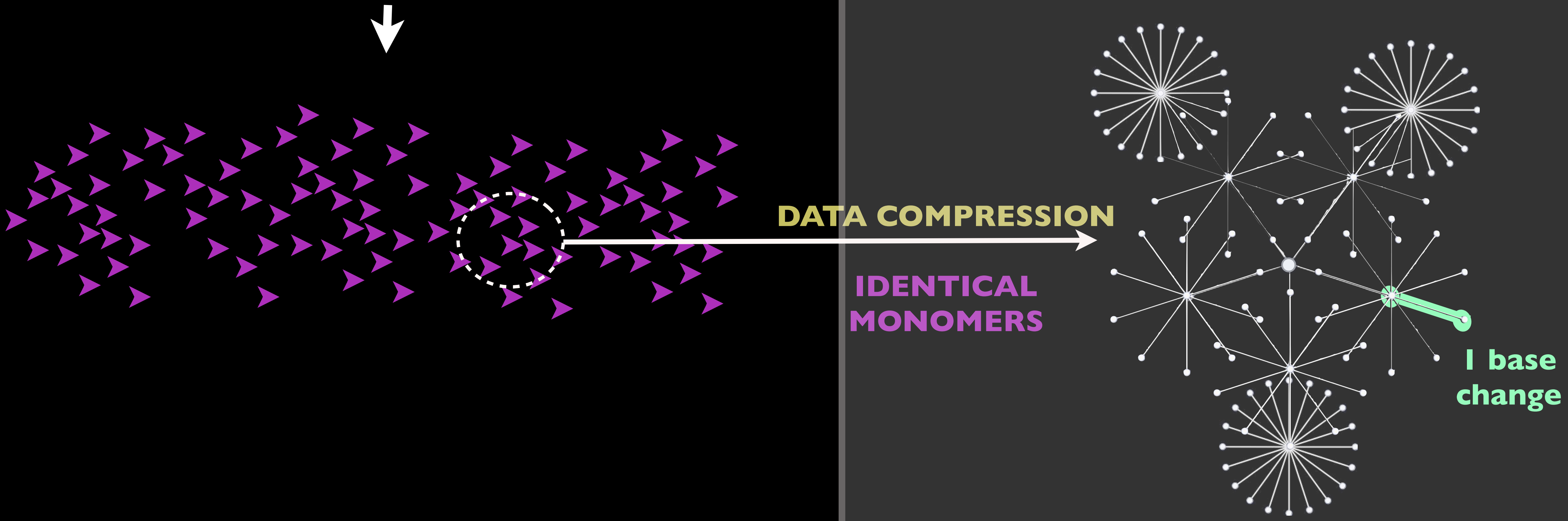
Centromeric database construction from reads containing alpha satellite repeats. (2.6% of the human genome)

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

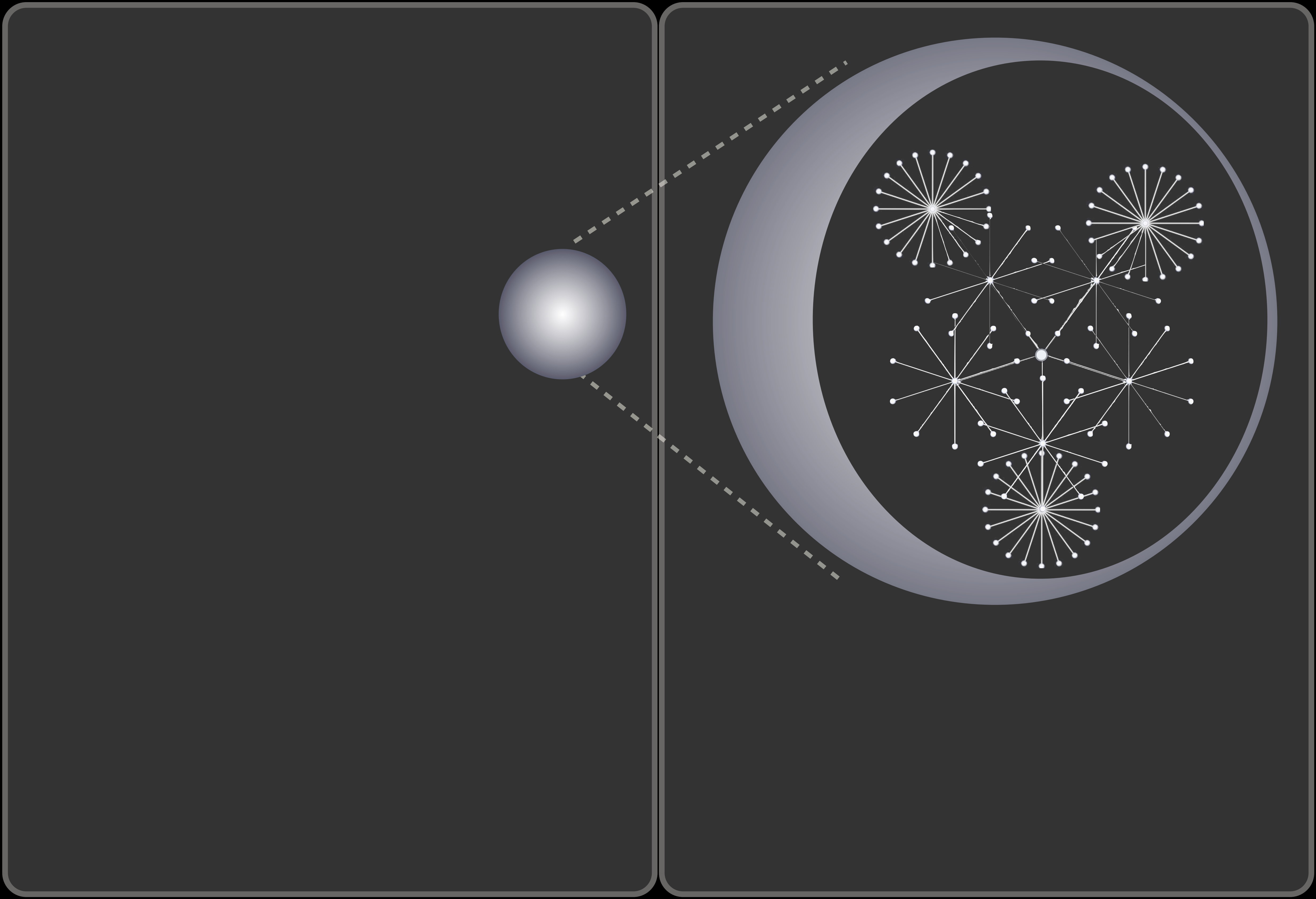
IDENTICAL MONOMERS

1 base change

Similarity Clustering:
Defining Epsilon Neighborhood

SEQUENCE RELATIONSHIP

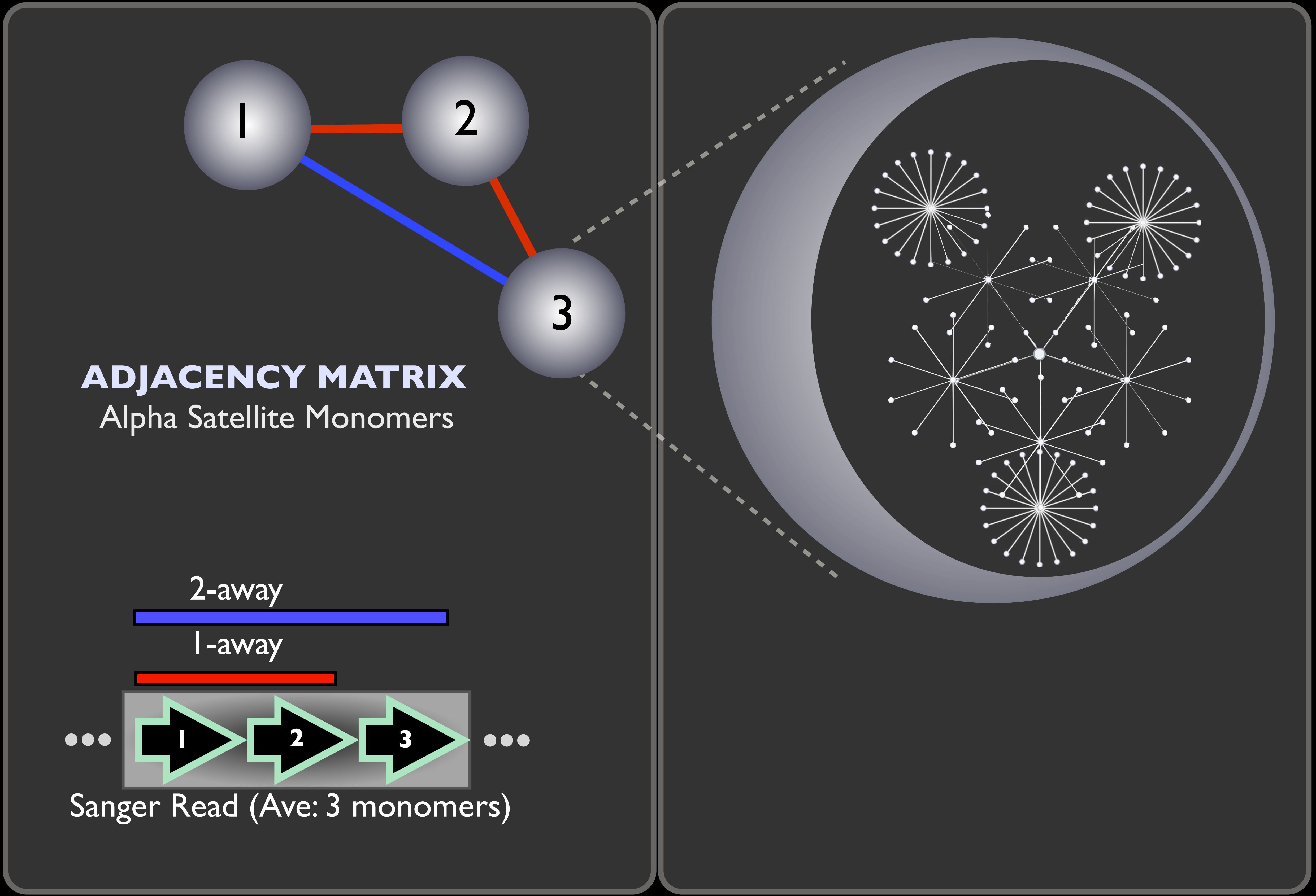
Higher Order Repeat Prediction



PHYSICAL RELATIONSHIP

SEQUENCE RELATIONSHIP

Higher Order Repeat Prediction



ADJACENCY MATRIX

Alpha Satellite Monomers

2-away

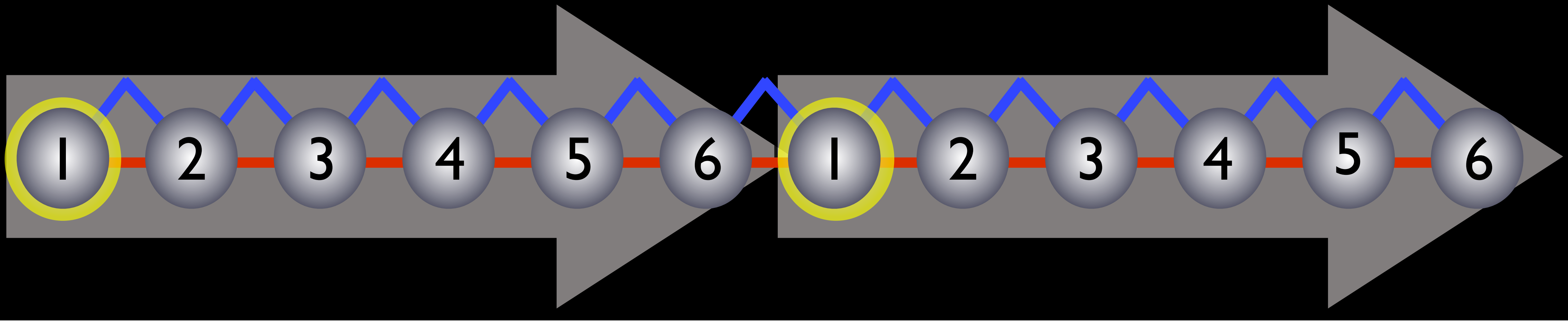
1-away

Sanger Read (Ave: 3 monomers)

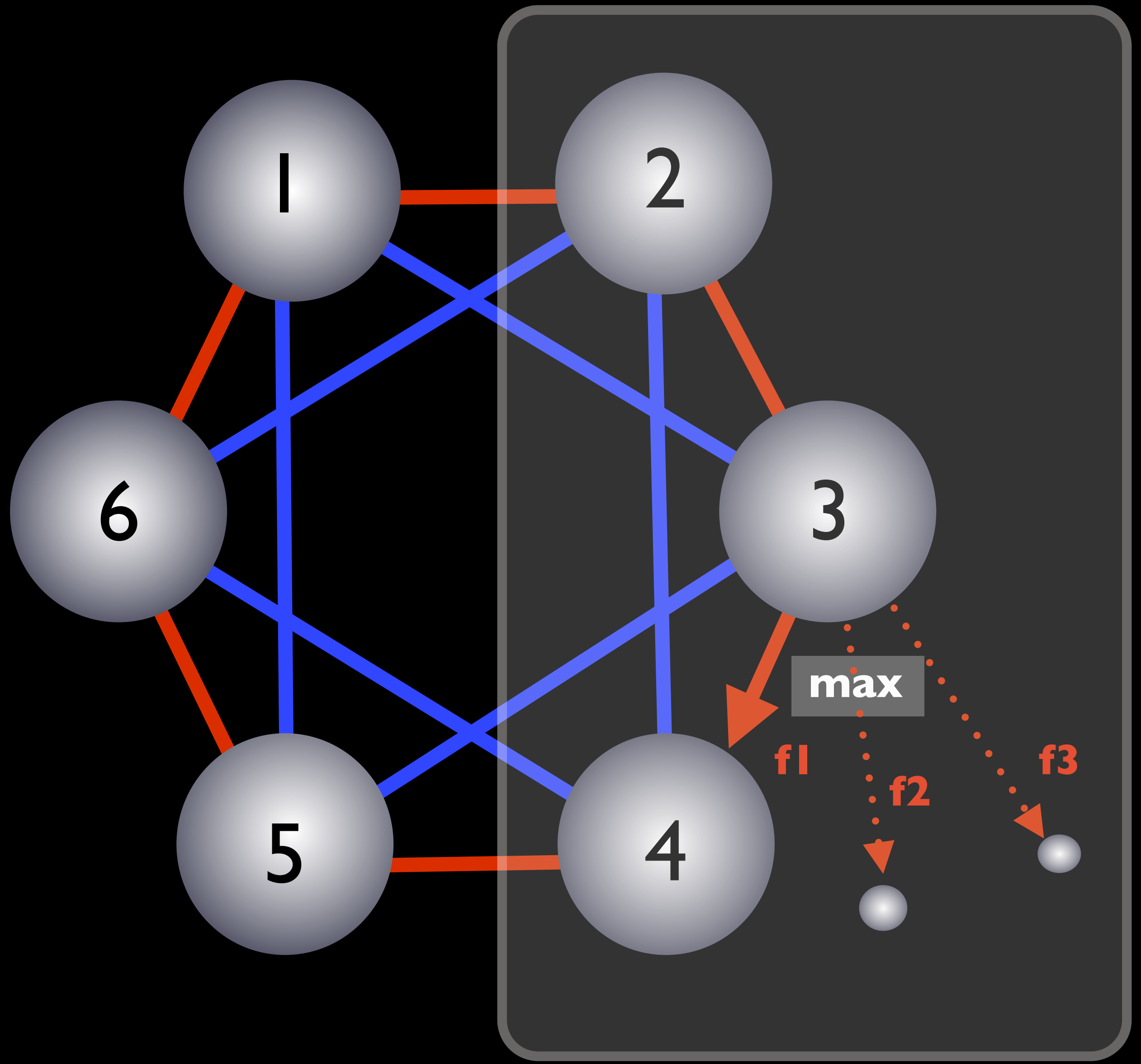
PHYSICAL RELATIONSHIP

SEQUENCE RELATIONSHIP

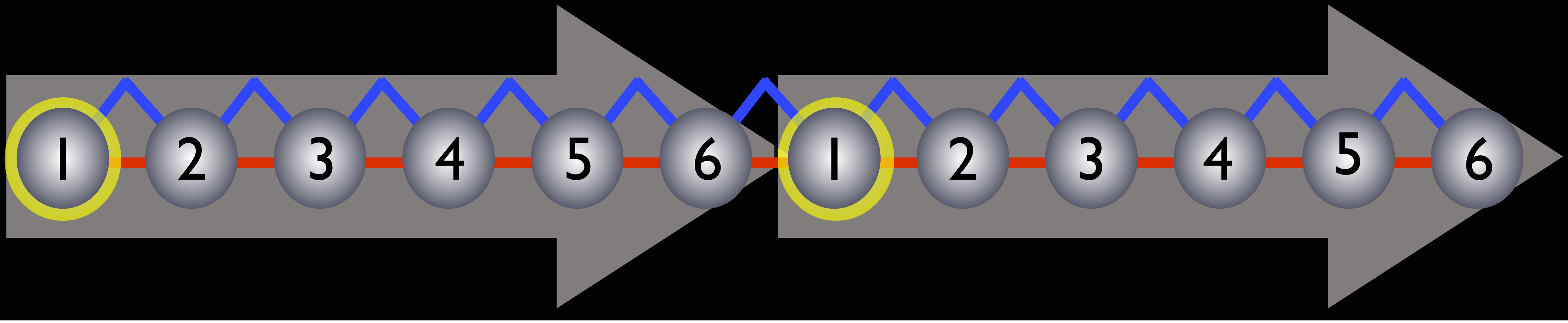
Higher Order Repeat Prediction



Higher Order Repeat Prediction

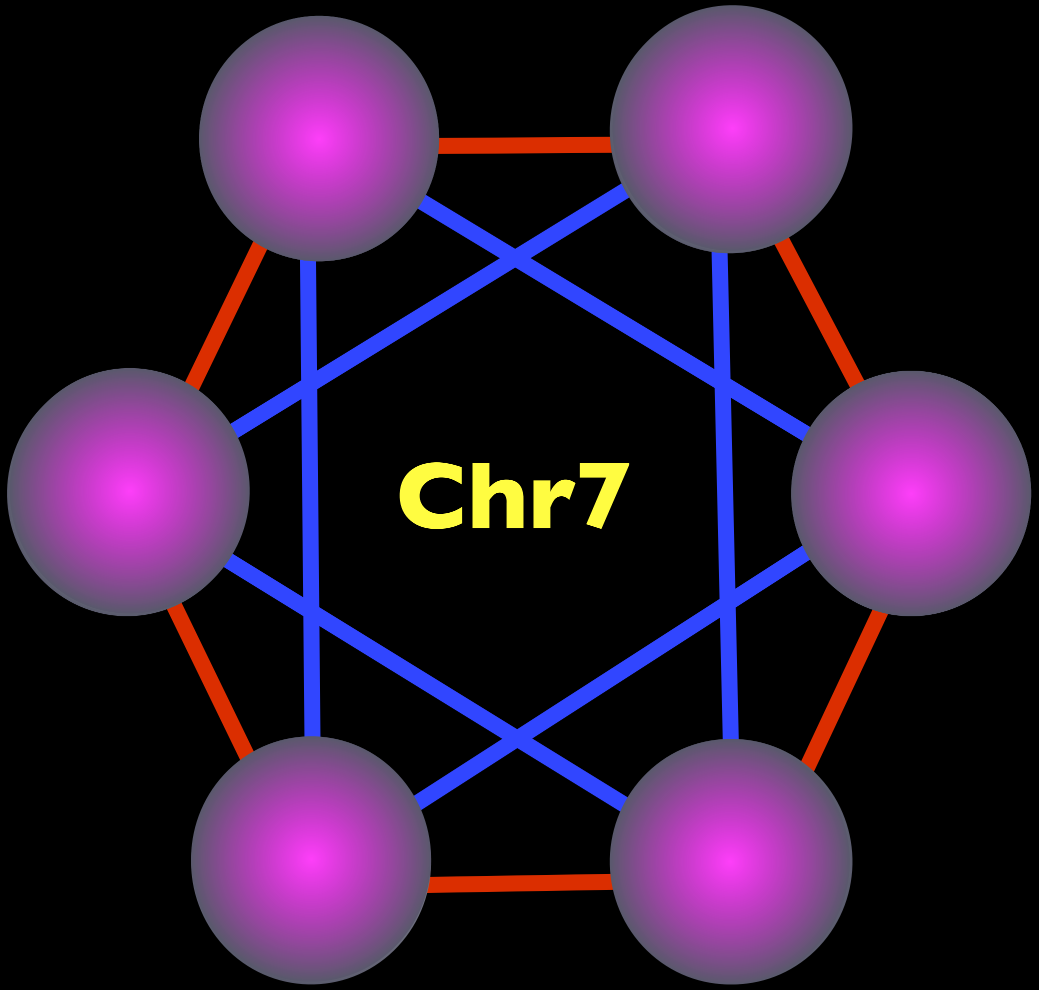


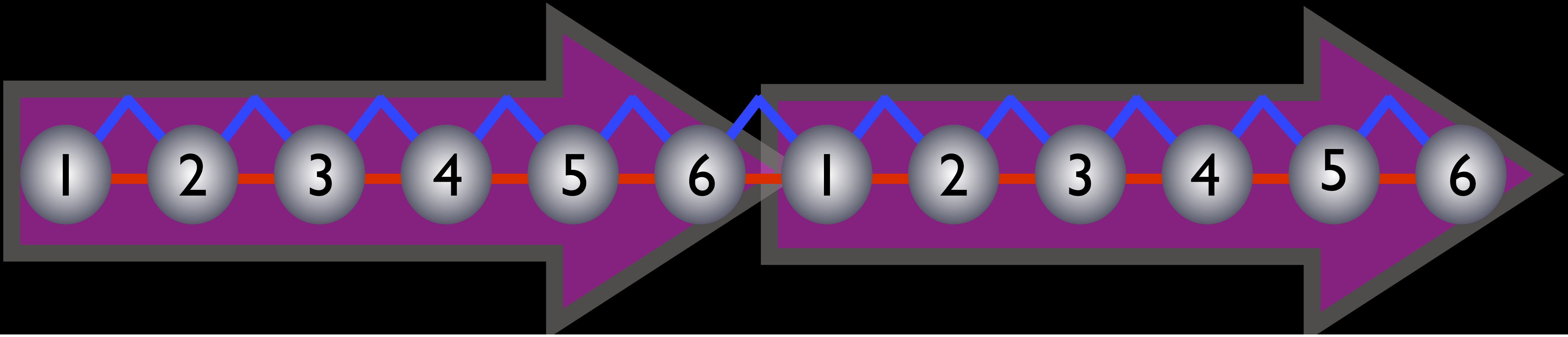
Viterbi greedy-algorithm second order markov model



Higher Order Repeat Prediction

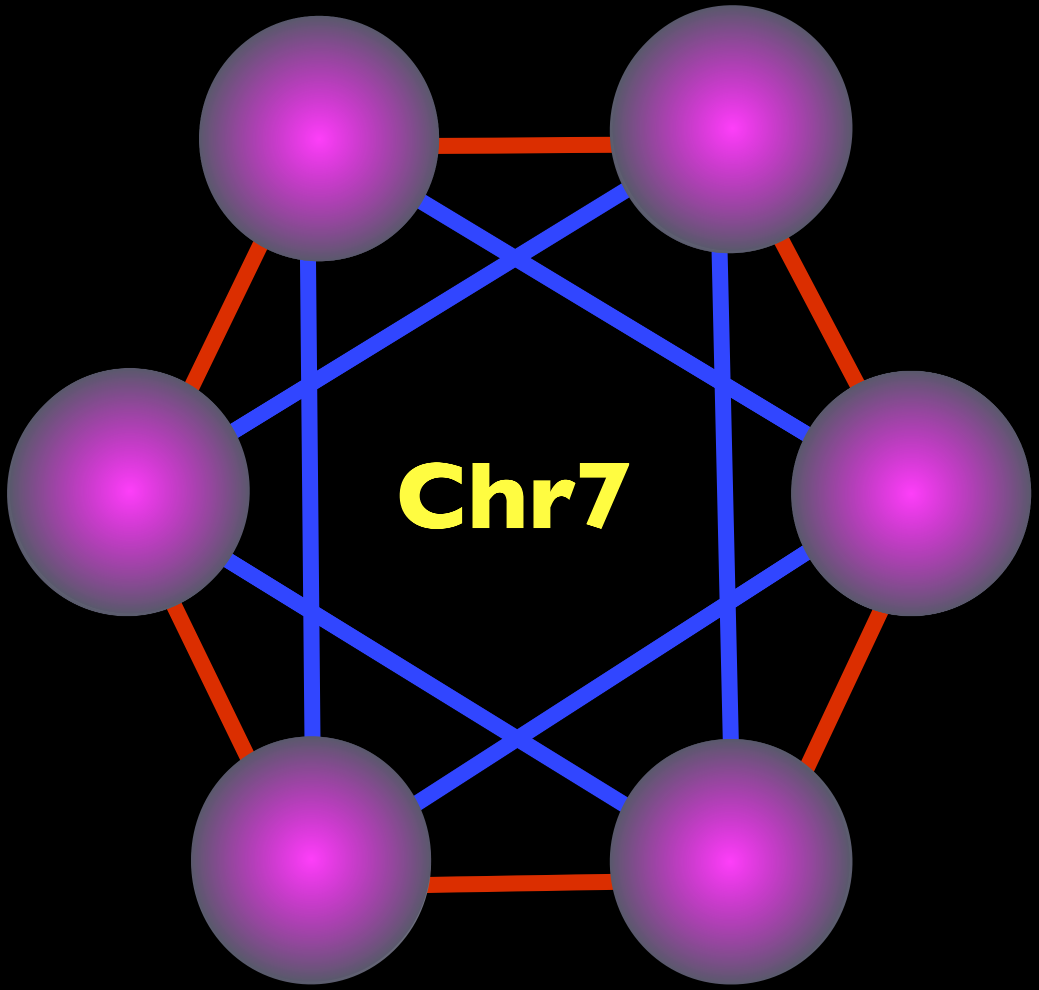
Determine Chromosome Specificity:





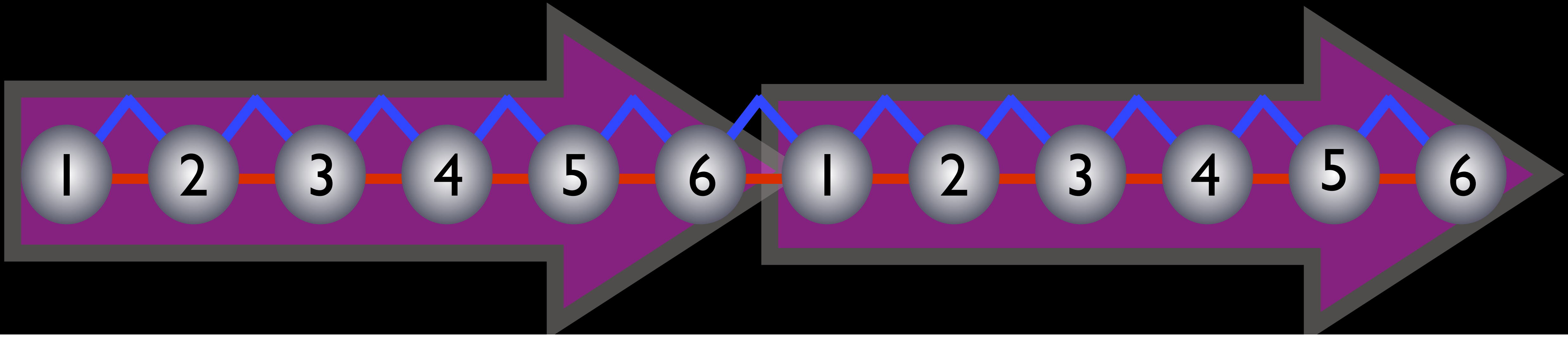
Higher Order Repeat Prediction

Determine Chromosome Specificity:



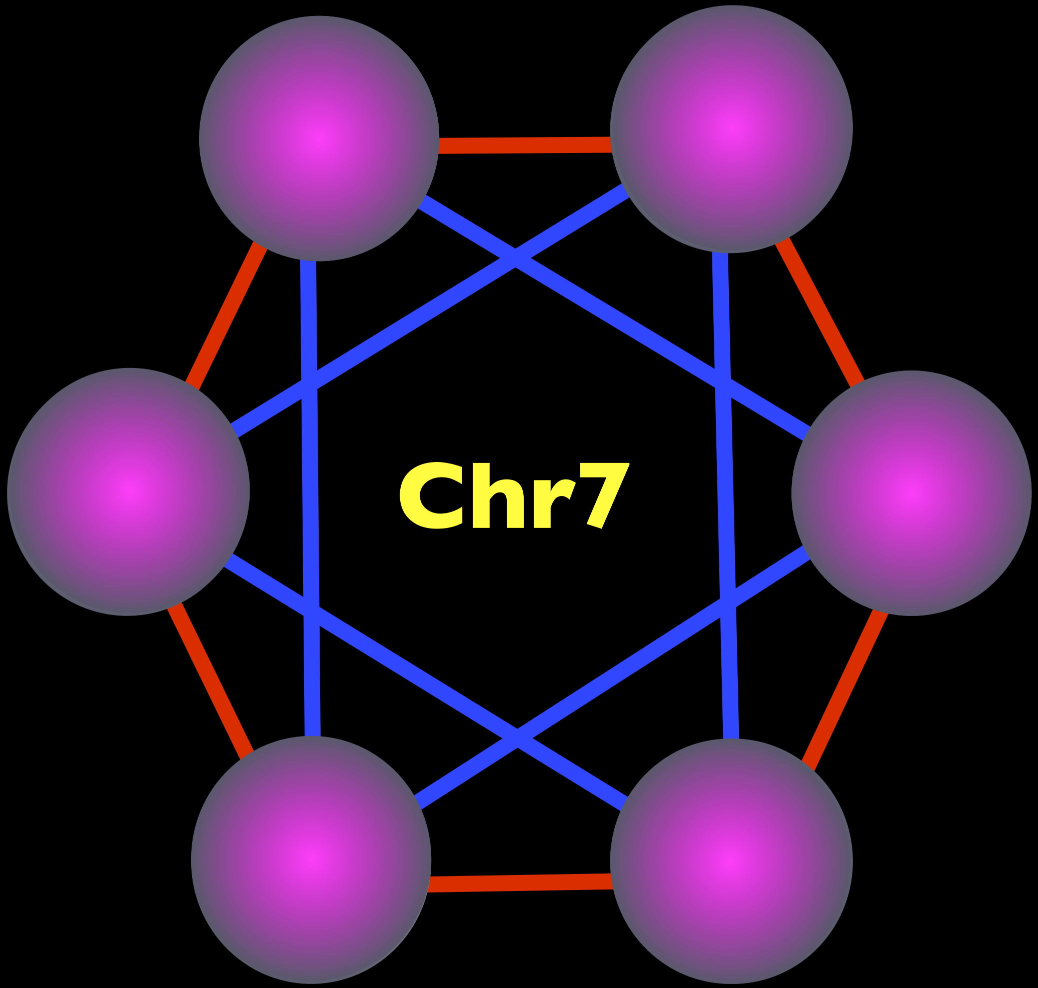
Flow Sorted Chromosome Alignment/Enrichment

344 Mb of Alpha Satellite from 15 Chromosomes



Higher Order Repeat Prediction

Determine Chromosome Specificity:

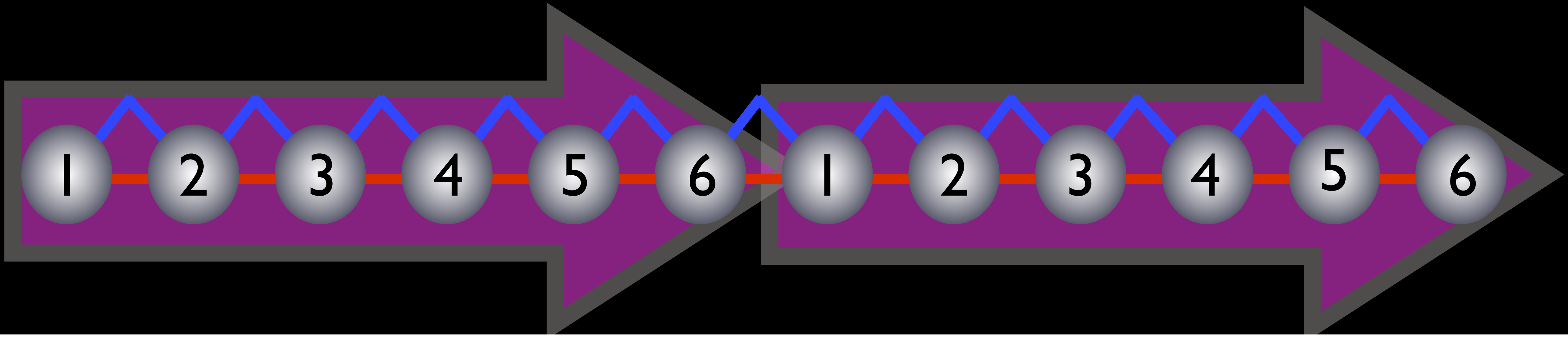


Flow Sorted Chromosome Alignment/Enrichment

344 Mb of Alpha Satellite from 15 Chromosomes

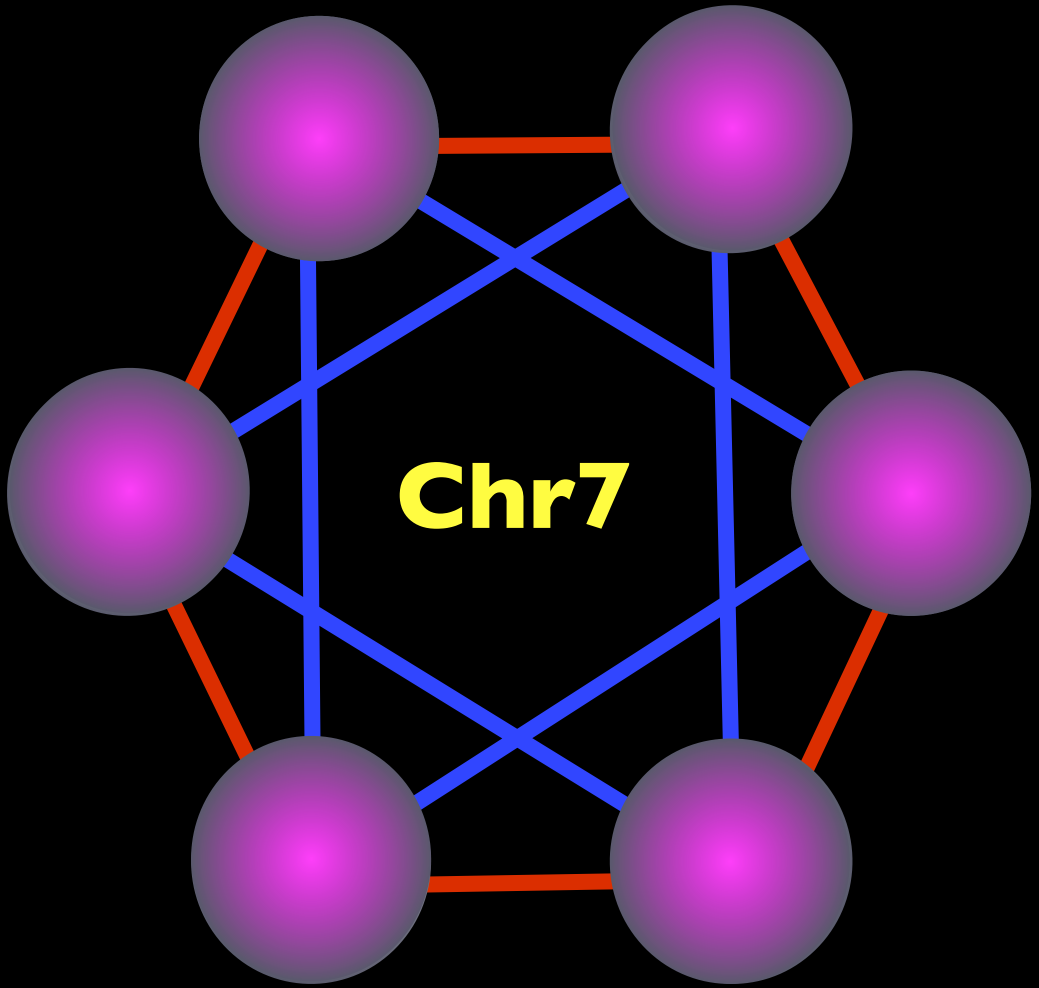
Experimental Evidence

FISH Hybridization and Screening Somatic Cell Hybrid Panel



Higher Order Repeat Prediction

Determine Chromosome Specificity:



Flow Sorted Chromosome Alignment/Enrichment

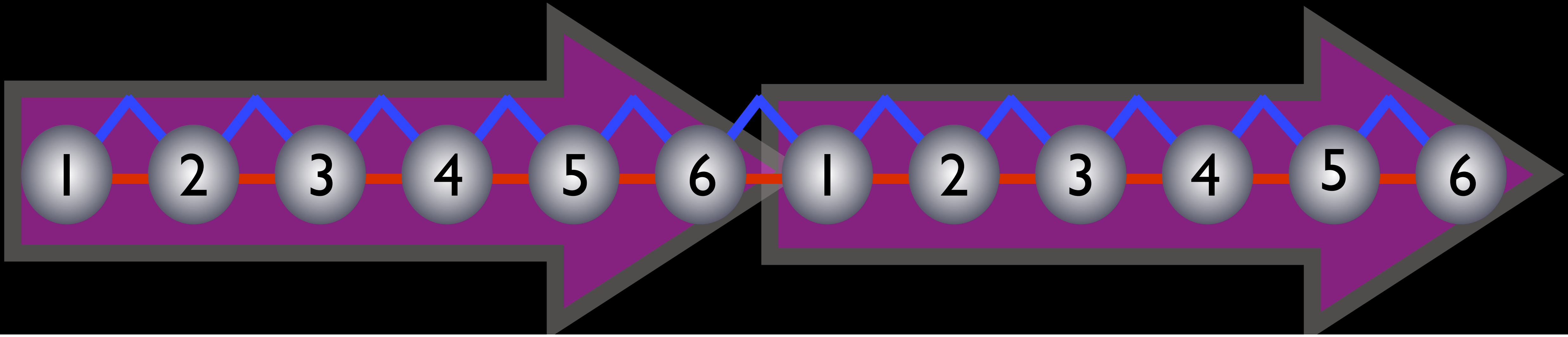
344 Mb of Alpha Satellite from 15 Chromosomes

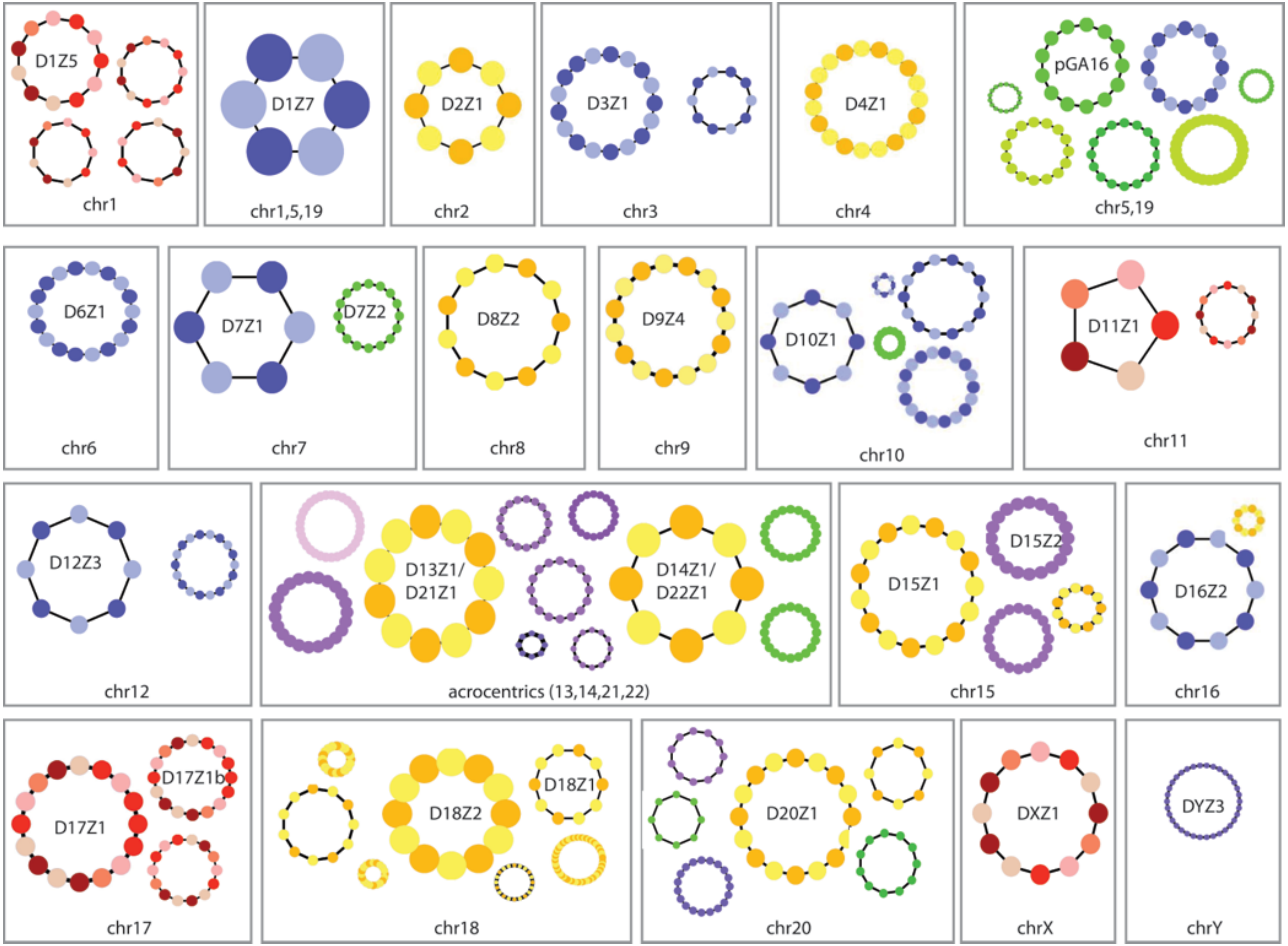
Experimental Evidence

FISH Hybridization and Screening Somatic Cell Hybrid Panel

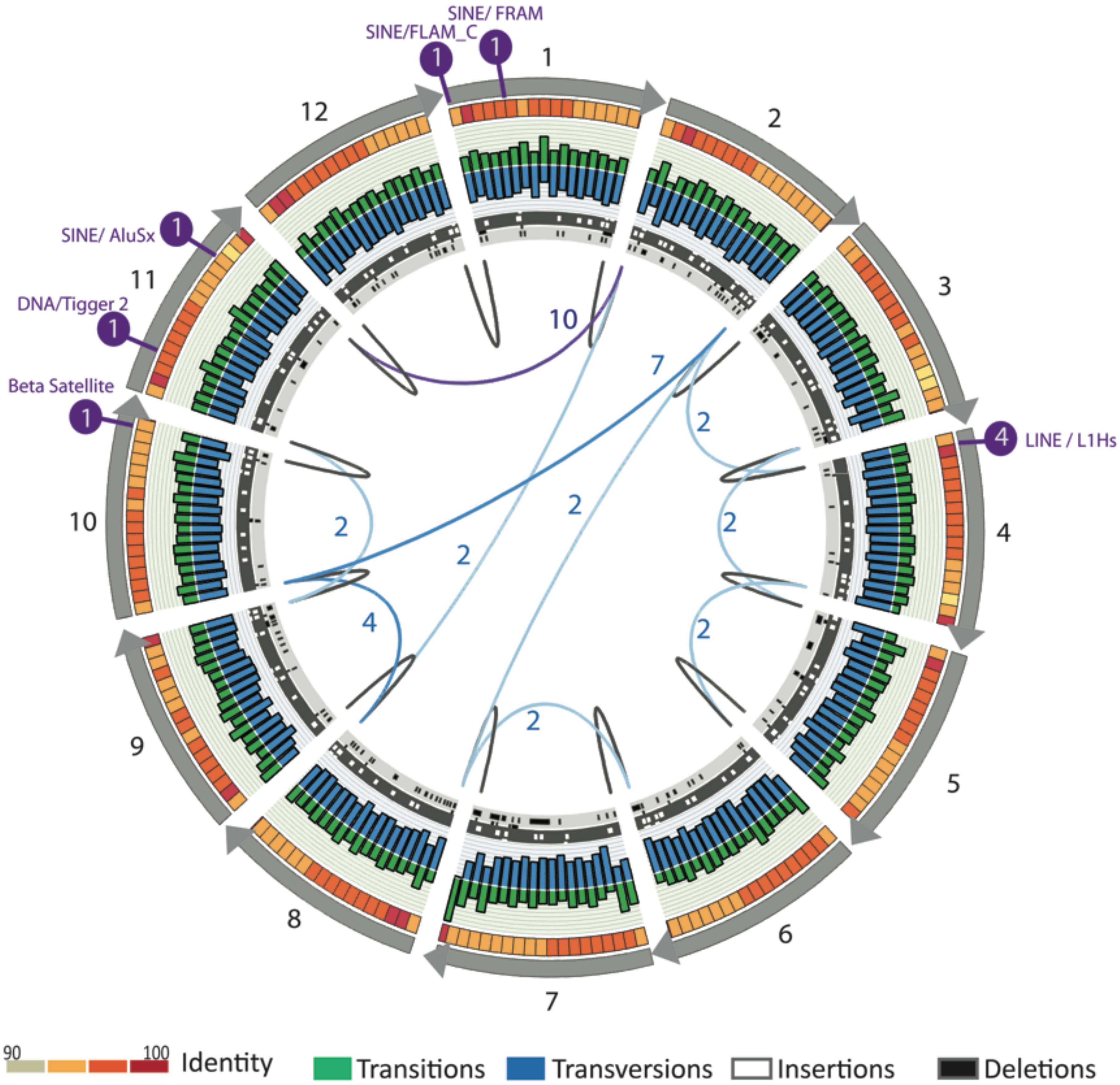
Paired Reads

“Anchor” to adjacent mapped HuRef contigs

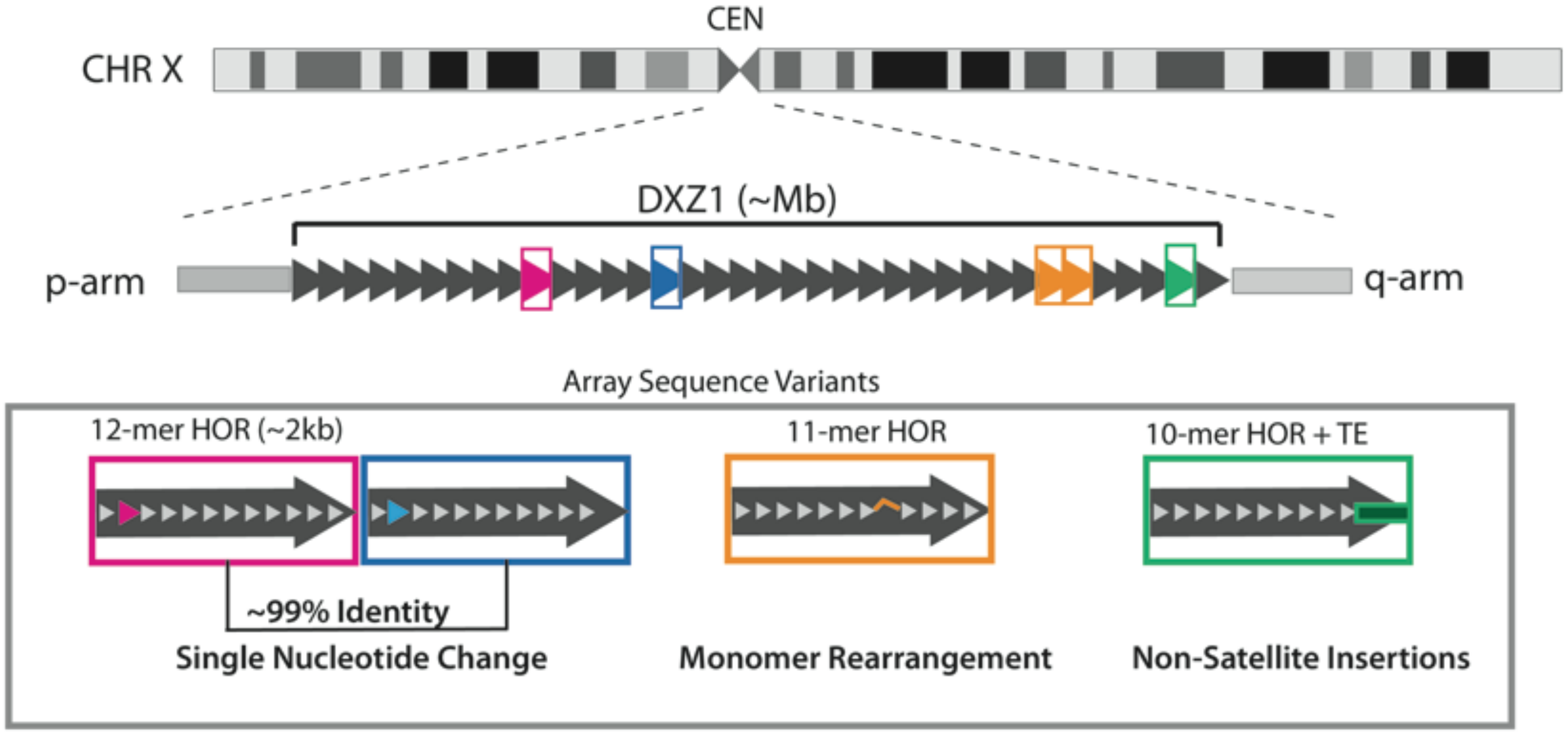



D1Z5
chr1
D1Z7
chr1,5,19
D2Z1
chr2
D3Z1
chr3
D4Z1
chr4
pGA16
chr5,19
D6Z1
chr6
D7Z1
D7Z2
chr7
D8Z2
chr8
D9Z4
chr9
D10Z1
chr10
D11Z1
chr11
D12Z3
chr12
D13Z1/
D21Z1
D14Z1/
D22Z1
acrocentrics (13,14,21,22)
D15Z1
D15Z2
chr15
D16Z2
chr16
D17Z1
D17Z1b
chr17
D18Z2
D18Z1
chr18
D20Z1
chr20
DXZ1
chrX
DYZ3
chrY

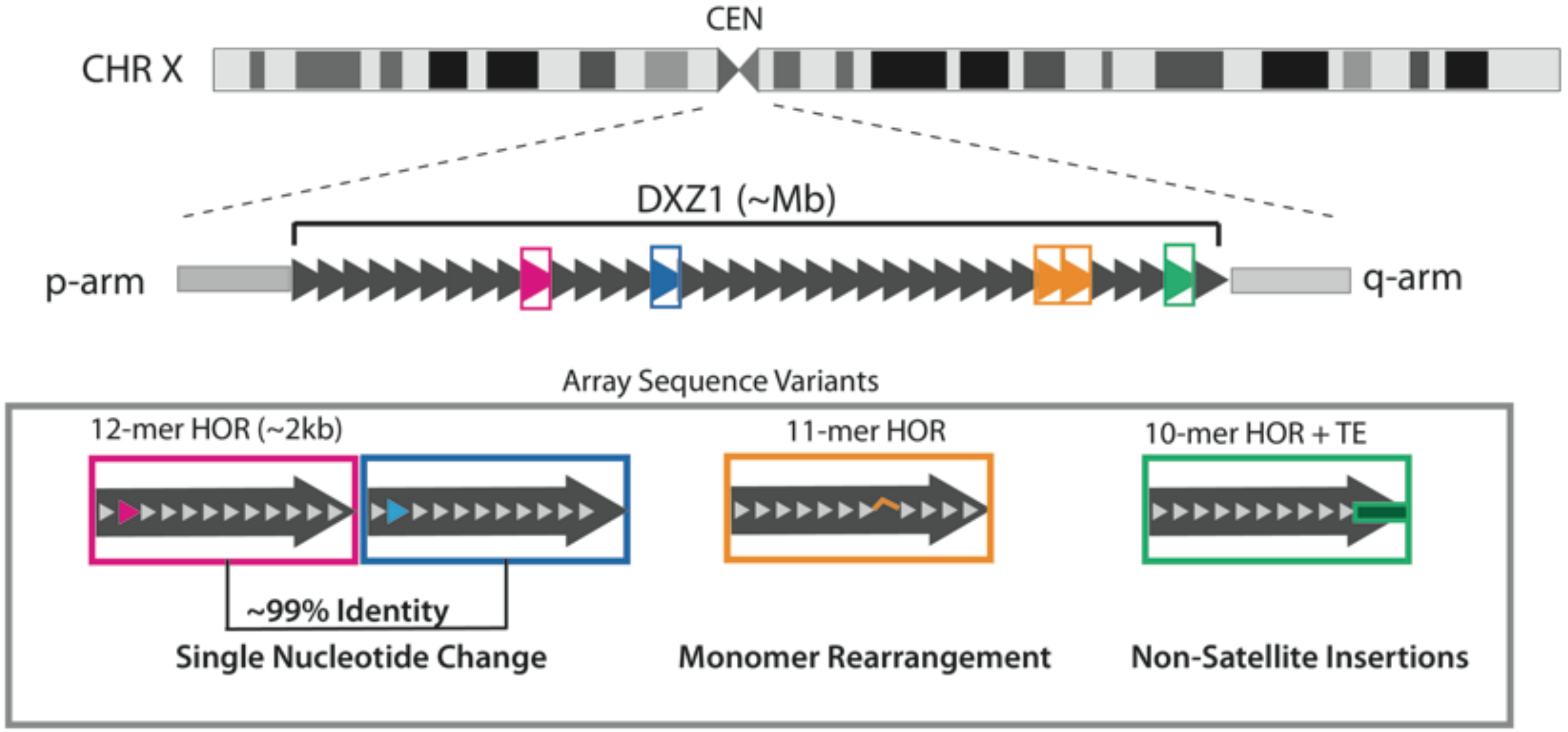
Alpha Satellite Array (DXZ1) on Chromosome X

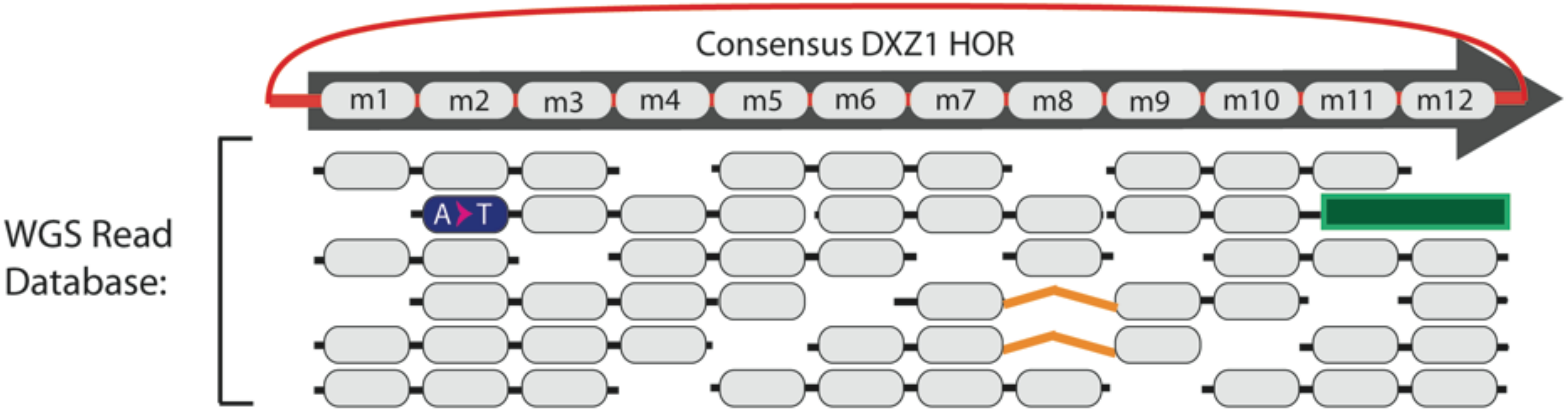


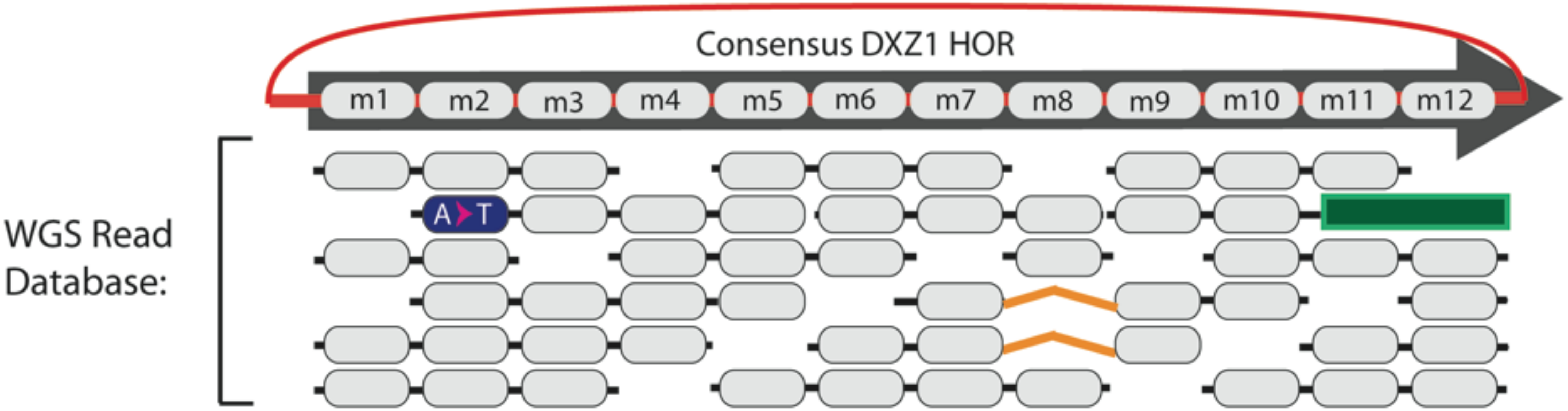
2 LinearSat Software to Convert Reads to Linear Reference Models

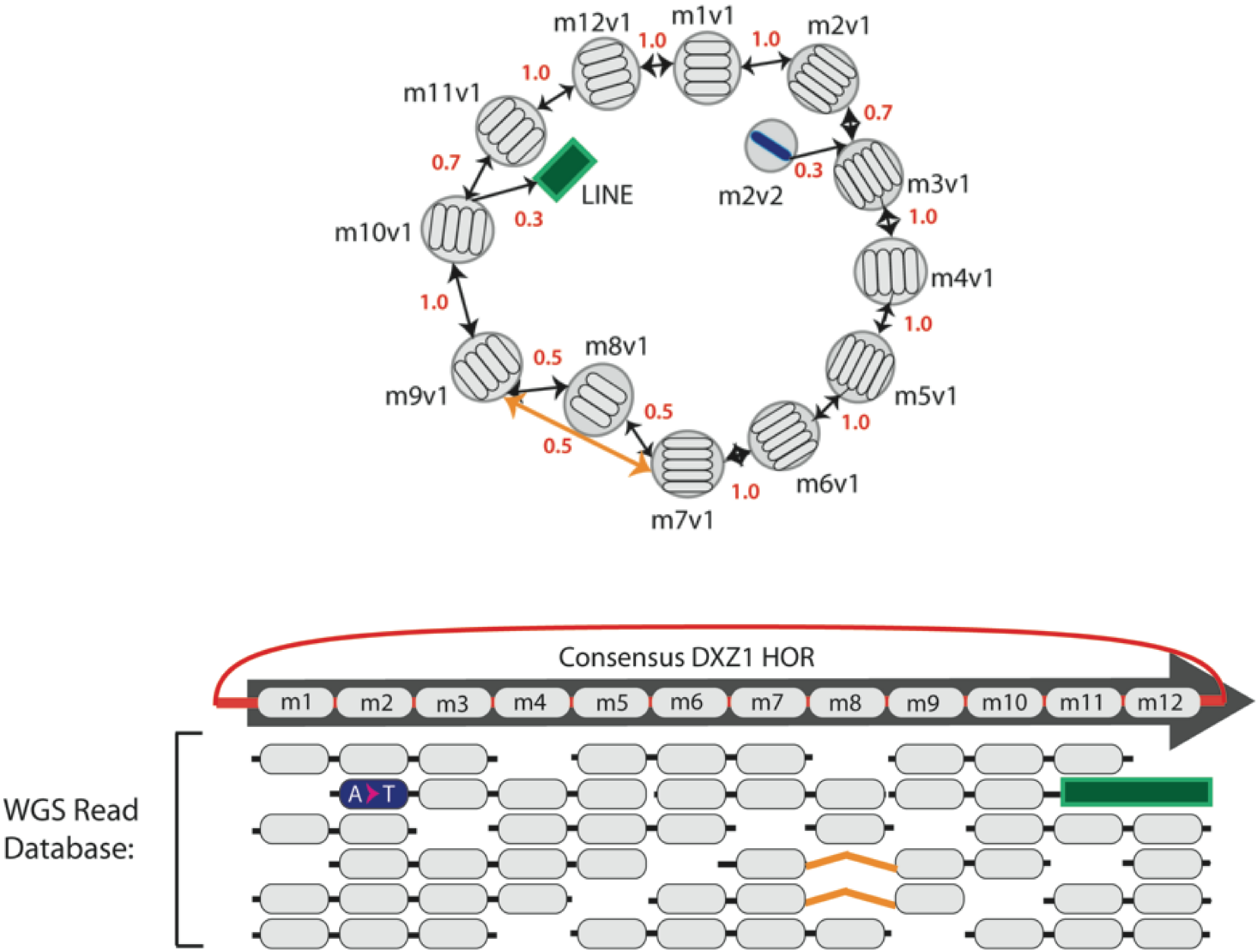


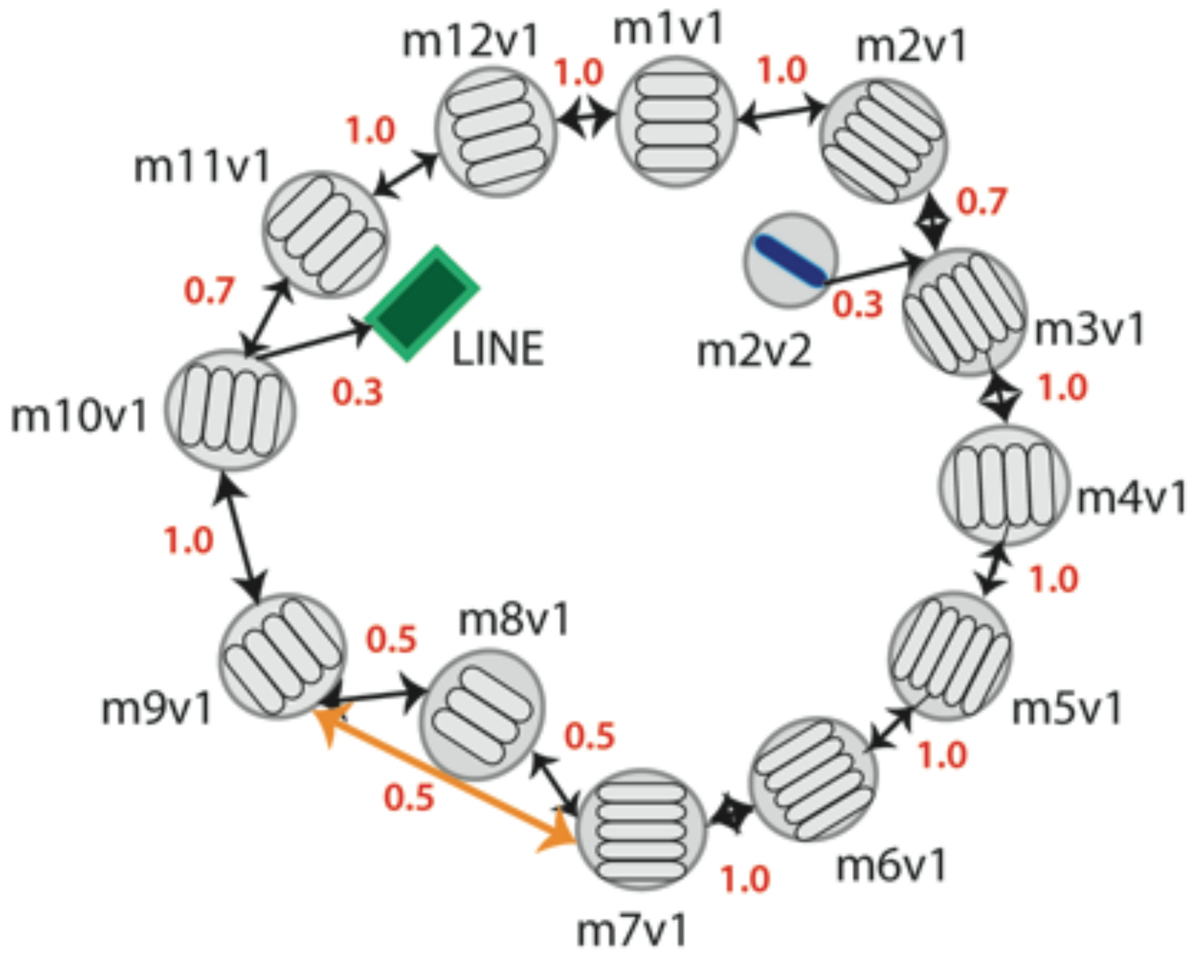
2 LinearSat Software to Convert Reads to Linear Reference Models





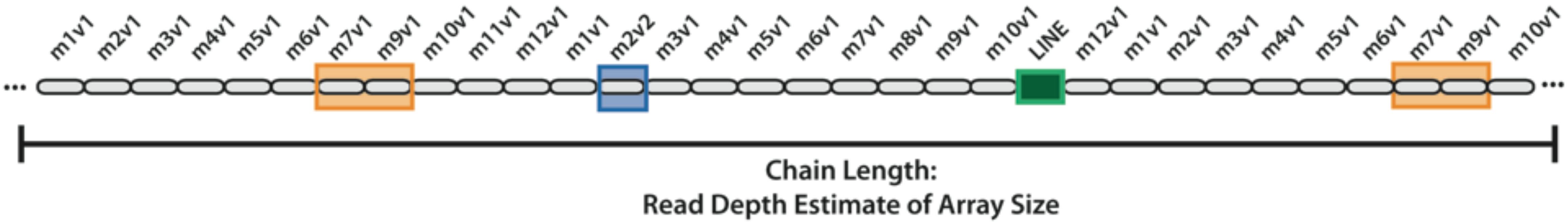

Consensus DXZ1 HOR
m1
m2
m3
m4
m5
m6
m7
m8
m9
m10
m11
m12
A
T
WGS Read Database:

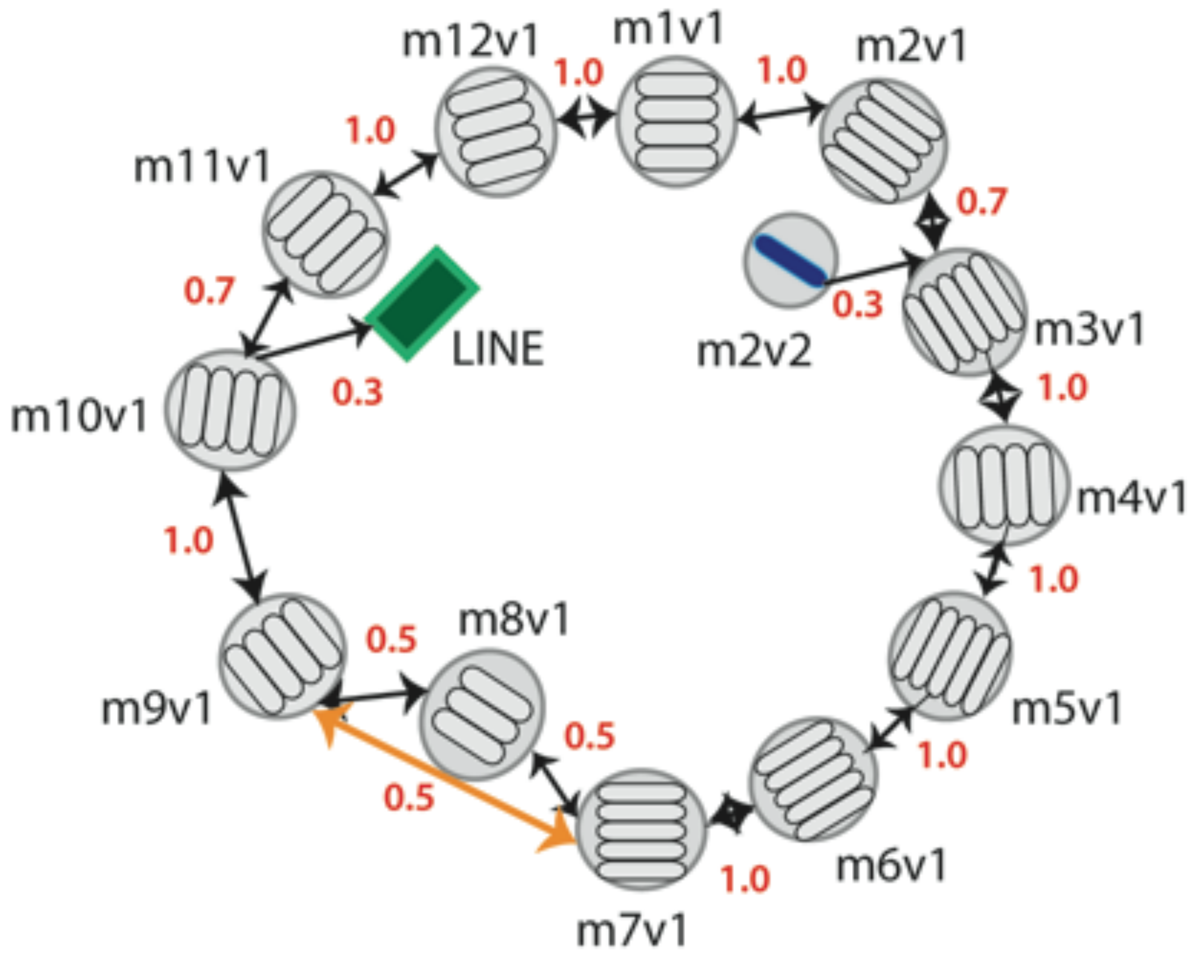

m12v1
m1v1
m2v1
1.0
1.0
1.0
m11v1
0.7
0.7
LINE
m2v2
0.3
0.3
m3v1
1.0
m10v1
m4v1
1.0
1.0
0.5
m8v1
m9v1
m5v1
0.5
1.0
0.5
m6v1
1.0
m7v1
Consensus DXZ1 HOR
m1
m2
m3
m4
m5
m6
m7
m8
m9
m10
m11
m12
A
T
WGS Read
Database:



LinearSat

- 2nd Order Markov Chain
- Length determined by Normalized Read Coverage
- Sensitive to low coverage
- Implemented with information on HOR repeat structure

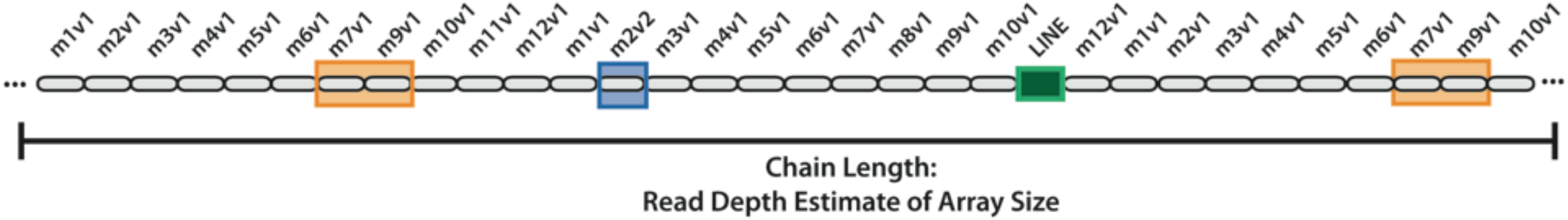


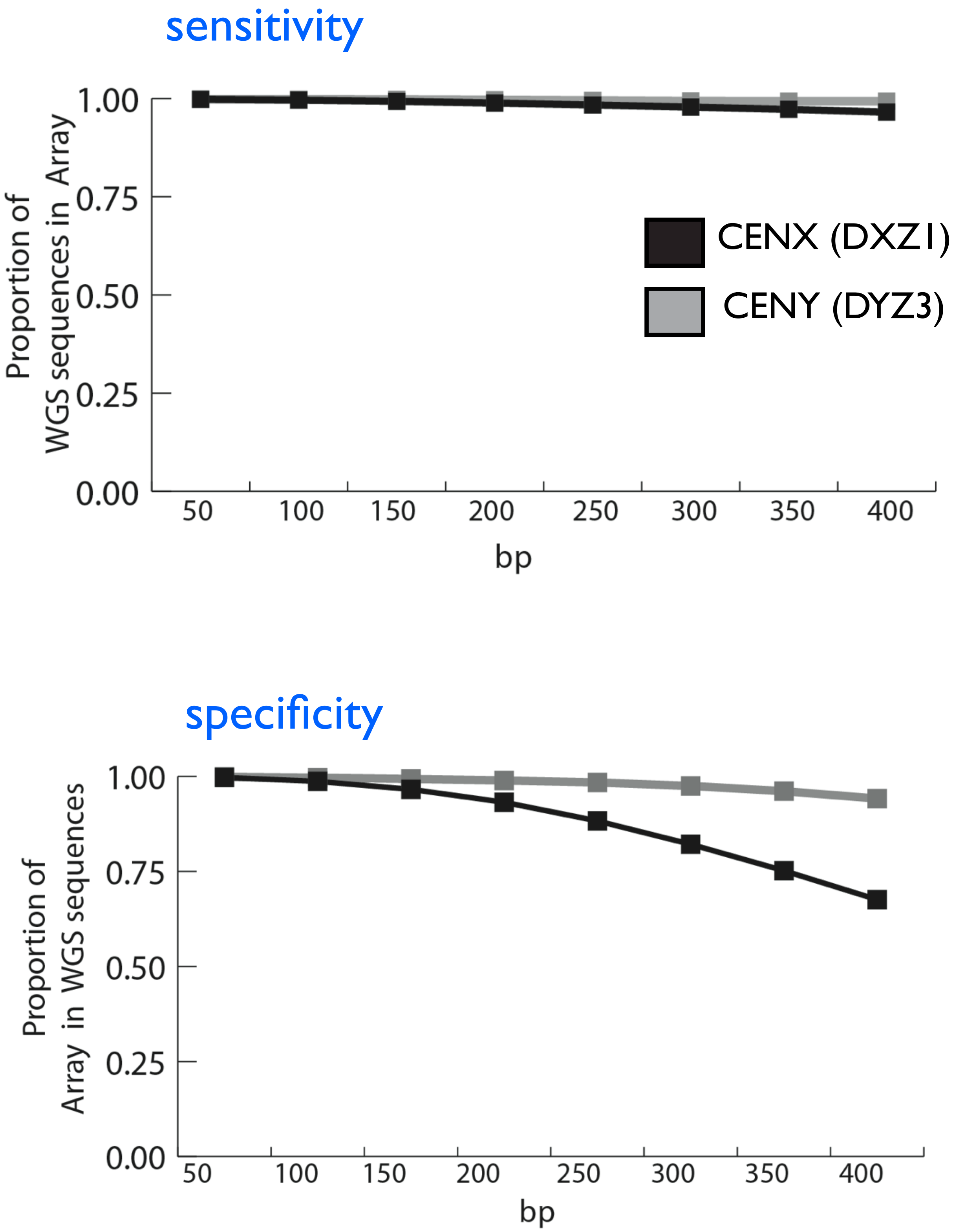


Not the "true" long-range organization, yet adequately represents the alpha satellite array sequence

LinearSat

- 2nd Order Markov Chain
- Length determined by Normalized Read Coverage
- Sensitive to low coverage
- Implemented with information on HOR repeat structure

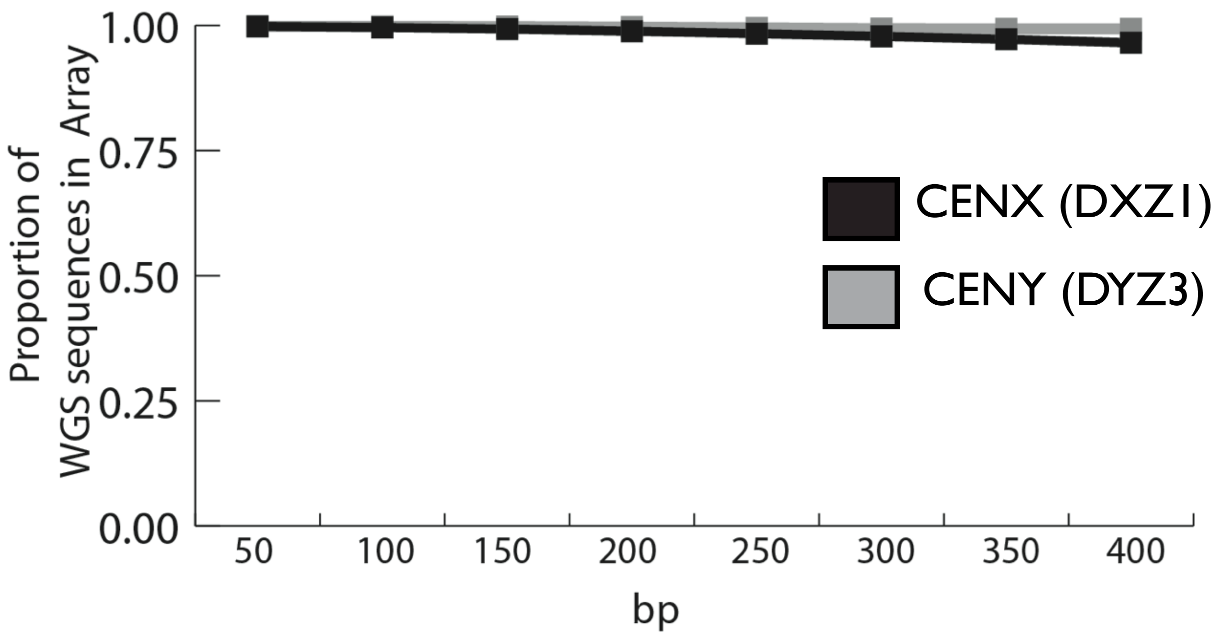




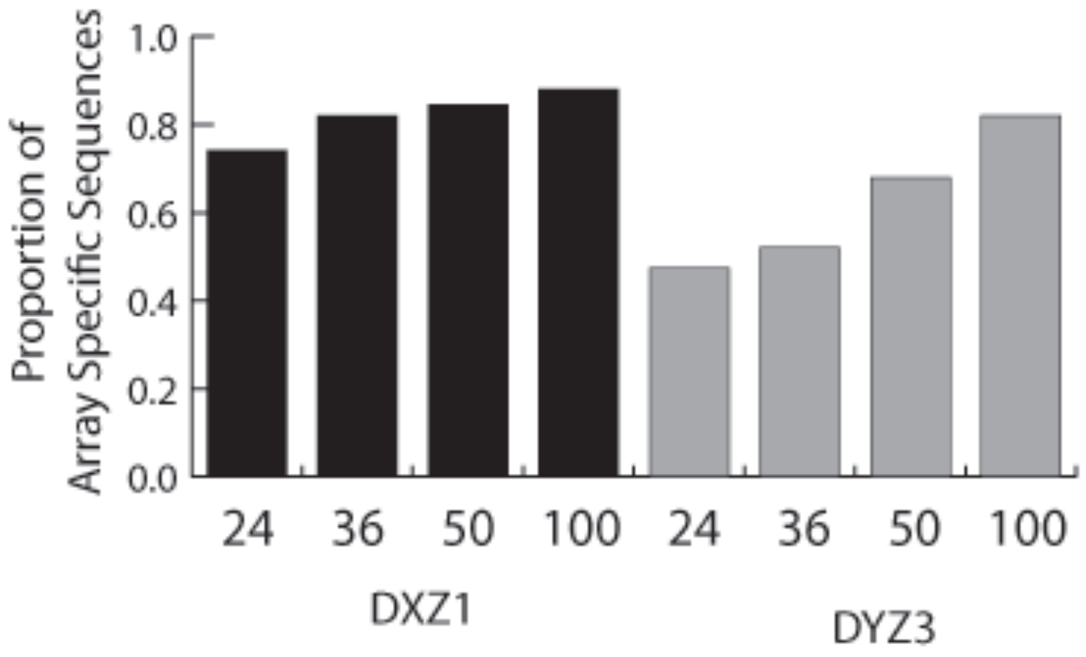
sensitivity

specificity

Test each satellite reference model to ensure that sequence variation is observed as expected within the initial read dataset sensitivity

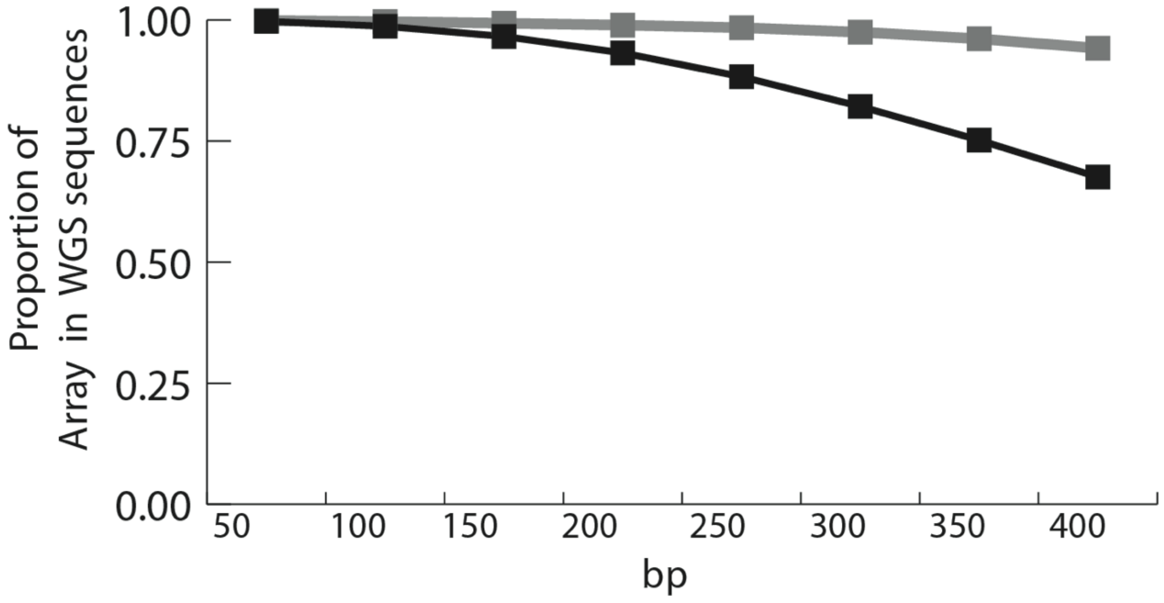


Mappability



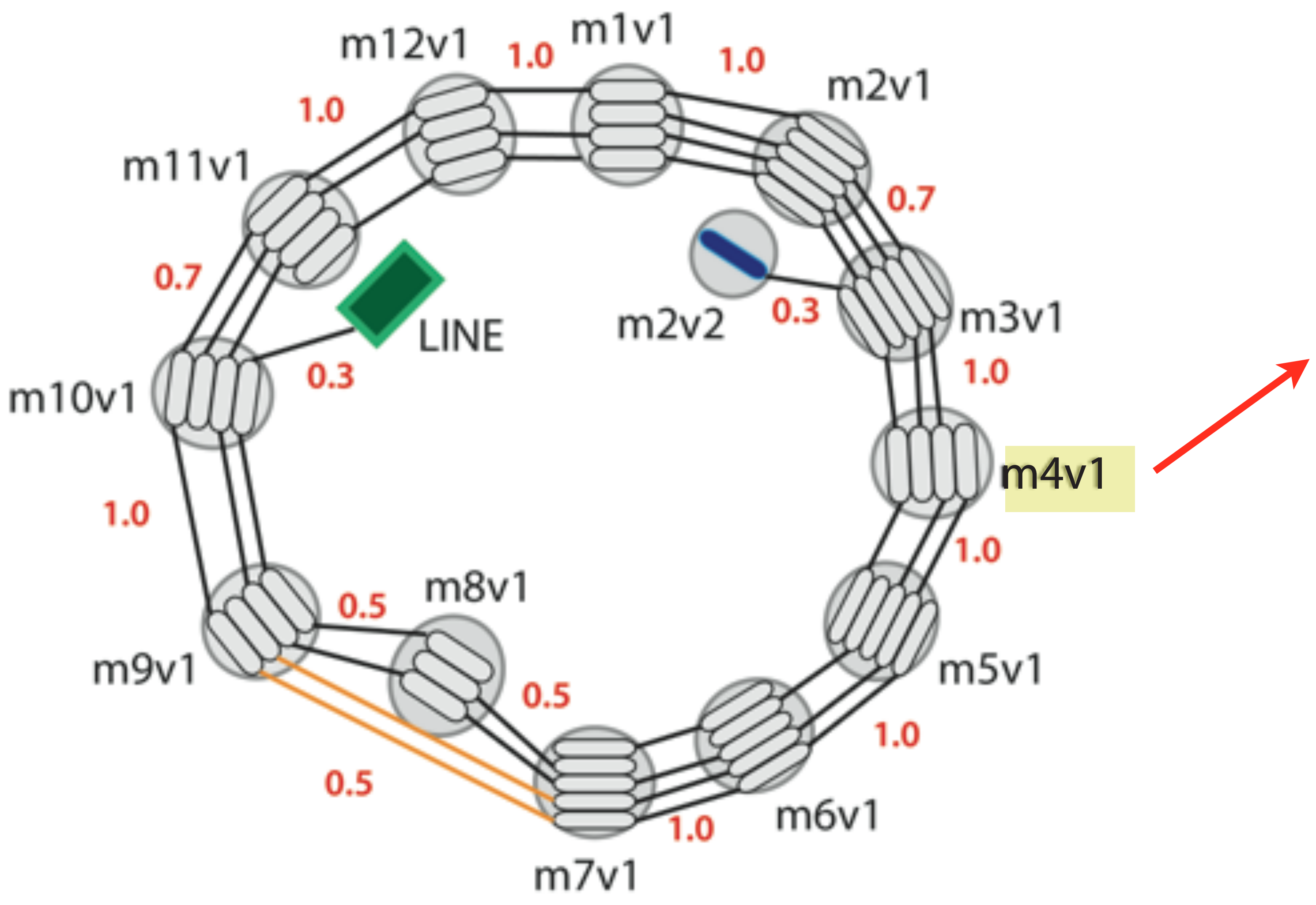
Evaluate as a Potential Read Mapping Target

specificity



Test each satellite reference model to ensure that sequence variation is observed as expected within the initial read dataset

GRCh38 Data Structure
Level 1: Repeat Components



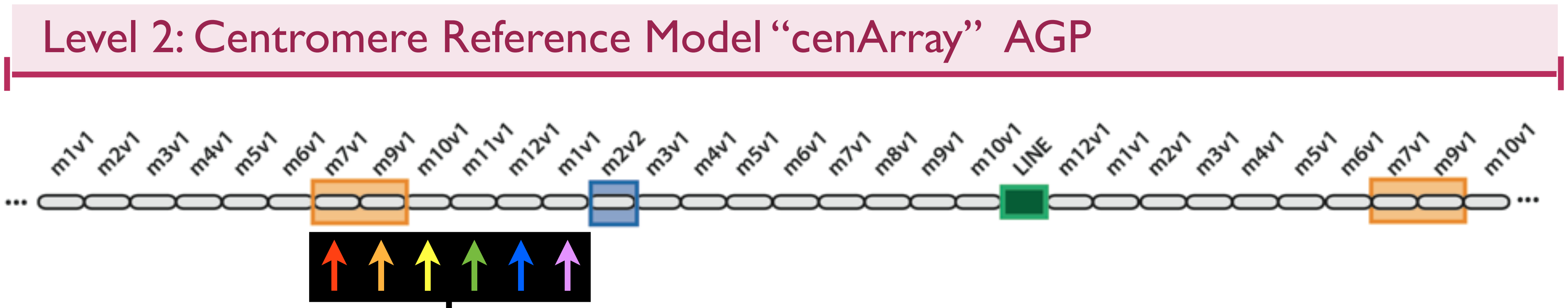
Database all unique sequence in each array graph

```
>m4v1 4 identical monomers

CACTTGCAGATTCTACAAAAAGAGTGCTTCAAAAC
TGCTCTGTCAAAAGGAAGGTTCAACTCTGTTACTT
GAGTACACACATCACAAGGAAGTTTCTGAGAATGC
TTCTGTCTGGTTTTTAGGAGAAGATATTTCCTTTT
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

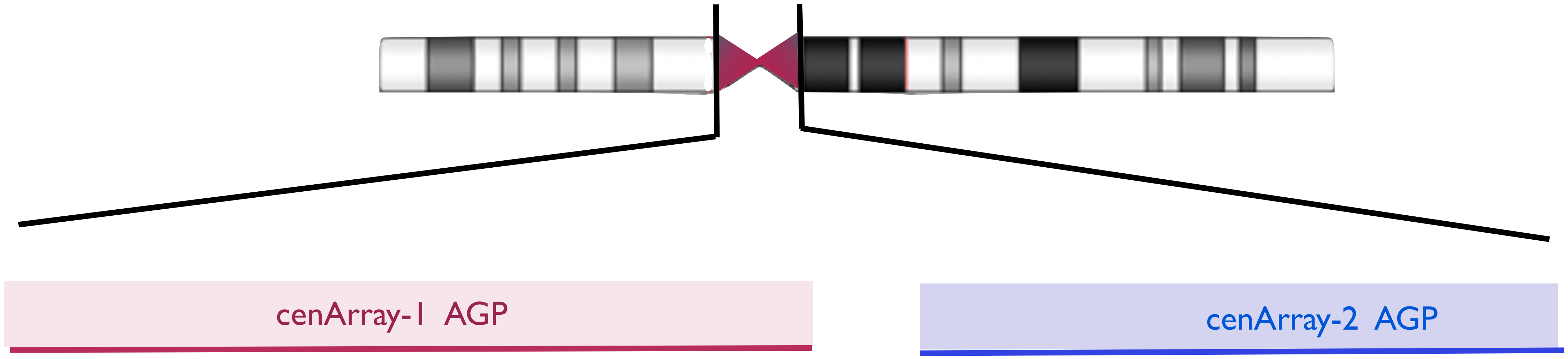


	Array Name	Array Start	Array End	UID	UID	Level I Entry	LI Start	LI End	Level I Ori
→	cenArray	129970	130138	759	O	m7v1	1	169	+
→	cenArray	130139	130309	760	O	m9v1	1	171	+
→	cenArray	130310	130608	761	O	m10v1	1	170	+
→	cenArray	130609	130708	762	N	m11v1	1	171	+
→	cenArray	130709	130878	763	O	m12v1	1	170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

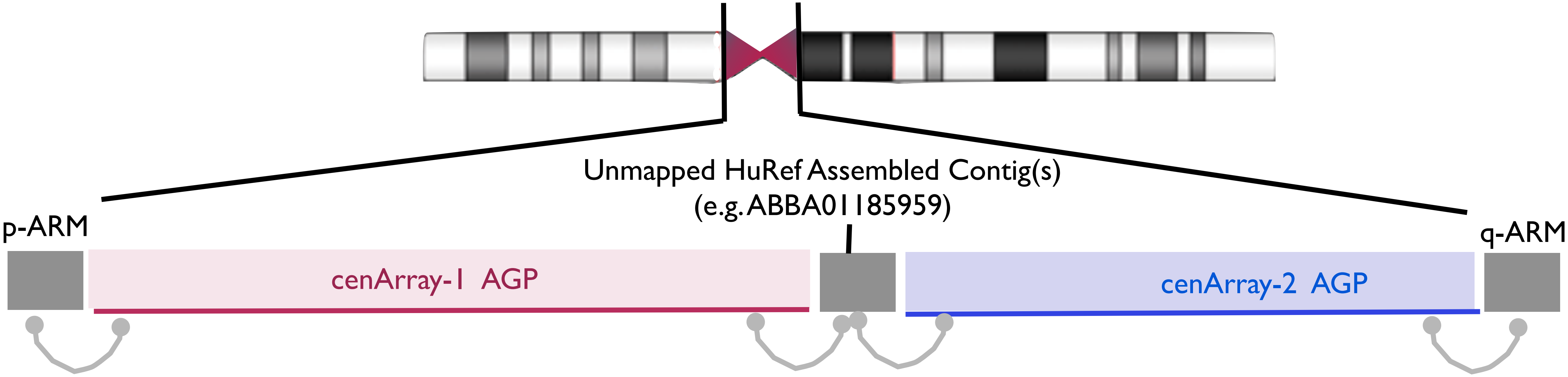


cenArray-1 AGP

cenArray-2 AGP

Single centromeric gap can contain more than one array

❸ Scaffold Reference Models and HuRef assembled contigs using mate pairs

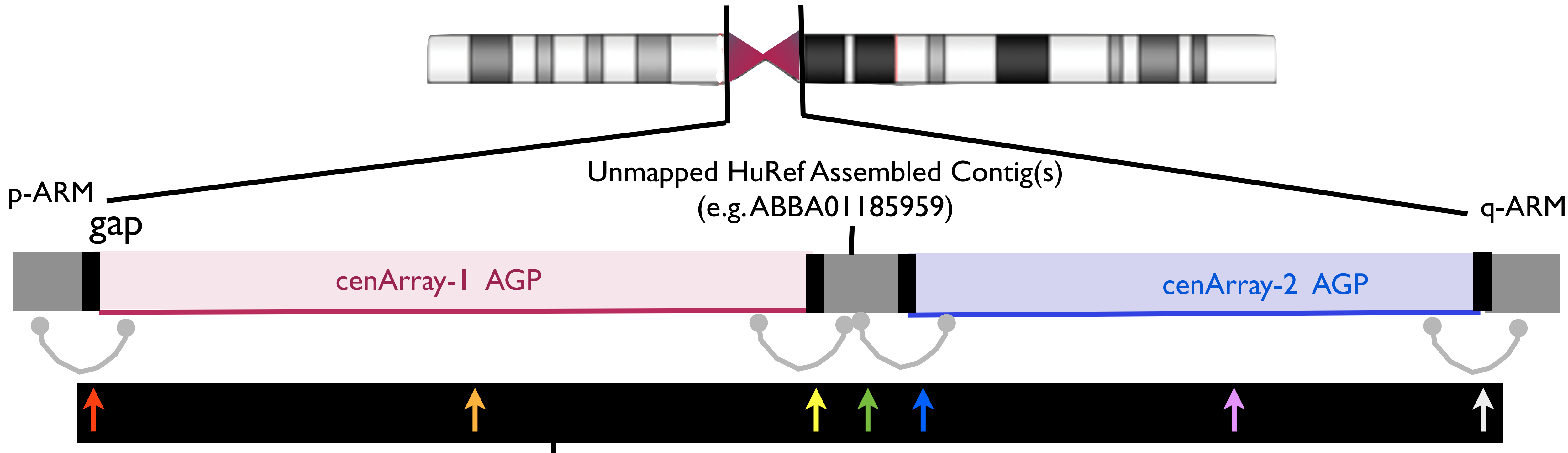


Single centromeric gap can contain more than one array

Scaffolding Order: Weighted by Mate Pairs

-- Bundled paired read information informs array component order

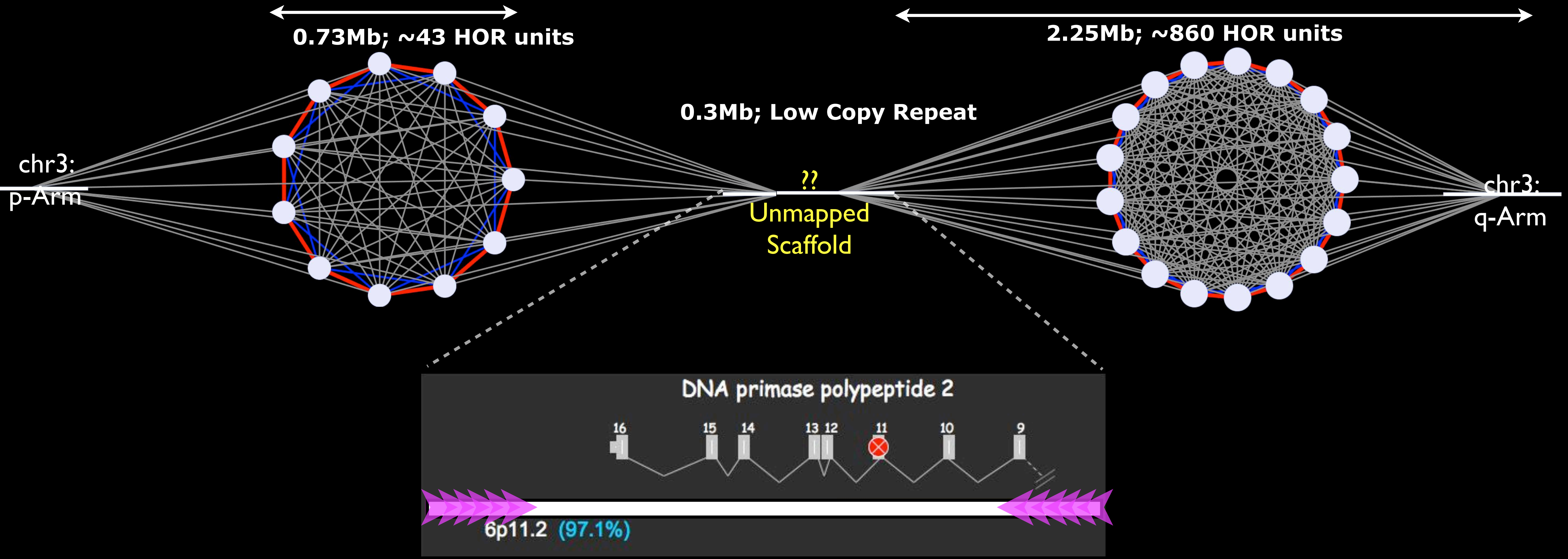
GRCh38 Data Structure
Level 3: AGP describing the order of Array components



	Array Name	Array Start	Array End	UID	UID	Level 1 Entry	L1 Start	L1 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



0.73Mb; ~43 HOR units

2.25Mb; ~860 HOR units

0.3Mb; Low Copy Repeat

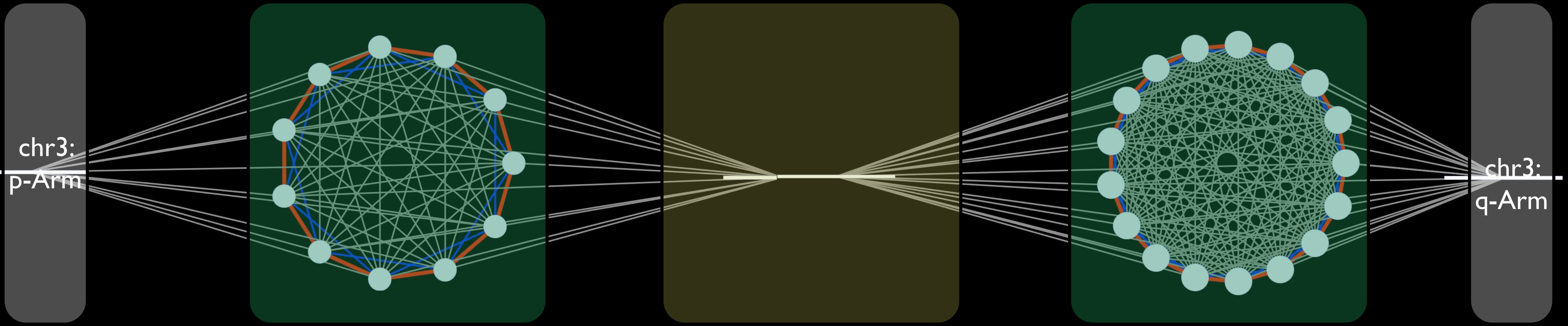
chr3:
p-Arm

??
Unmapped
Scaffold

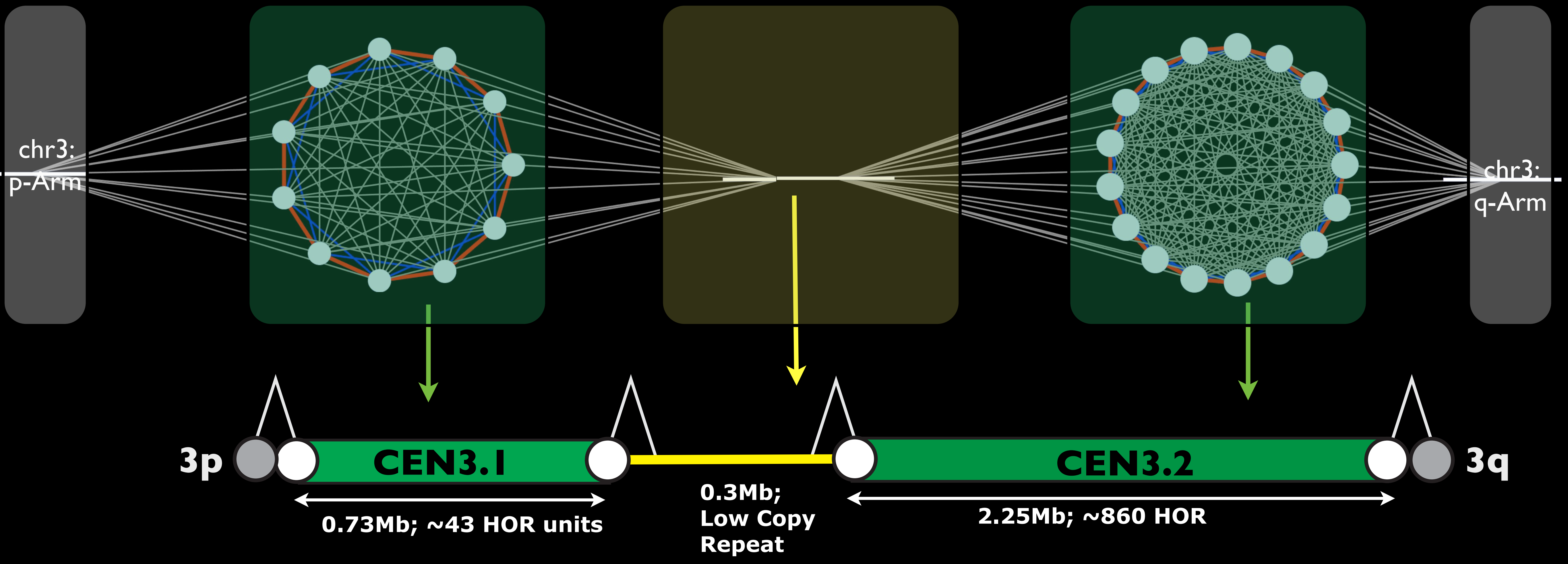
chr3:
q-Arm

DNA primase polypeptide 2

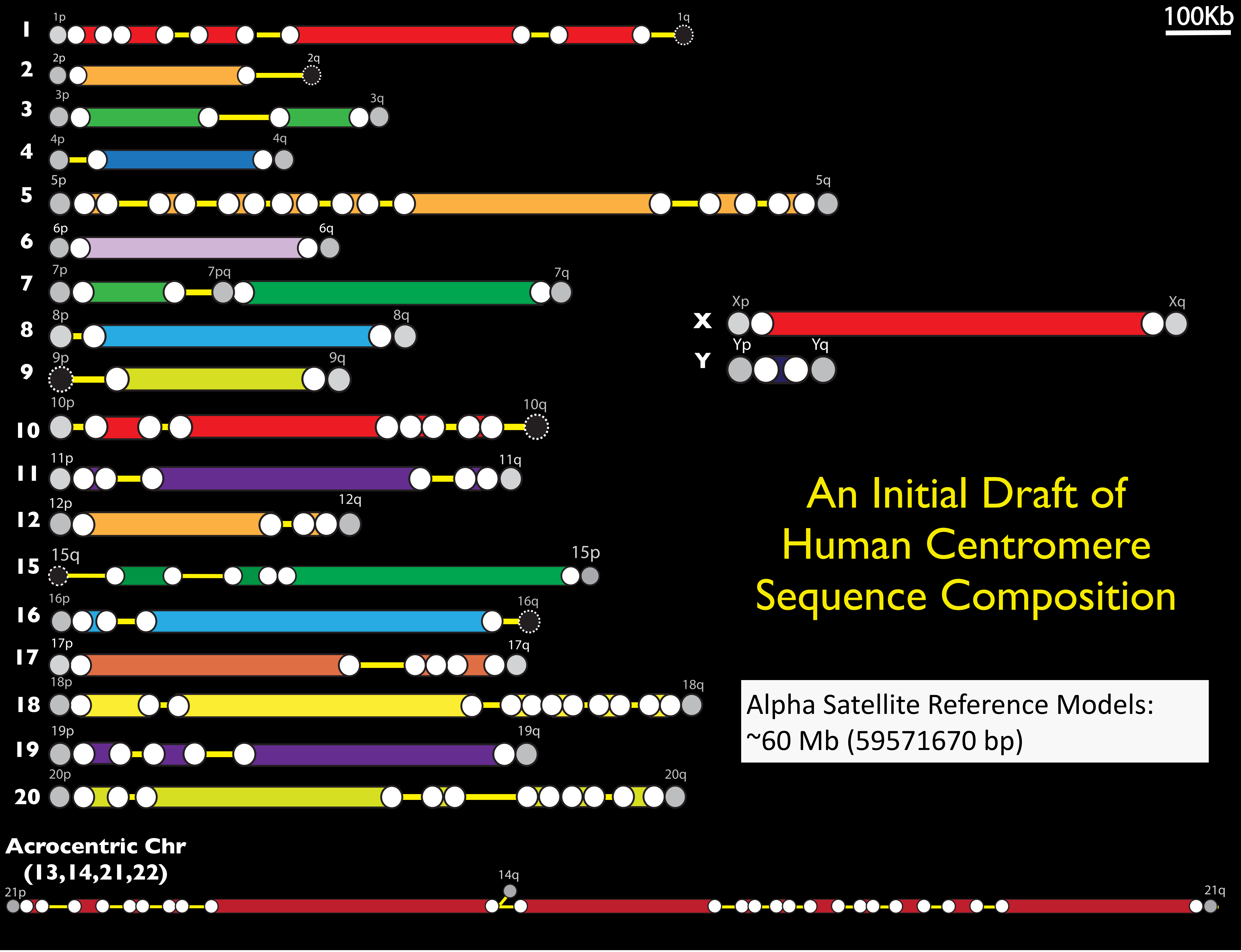
Scaffolding Problem: Order Elements by Paired Reads



Scaffolding Problem:
Order Elements by Paired Reads

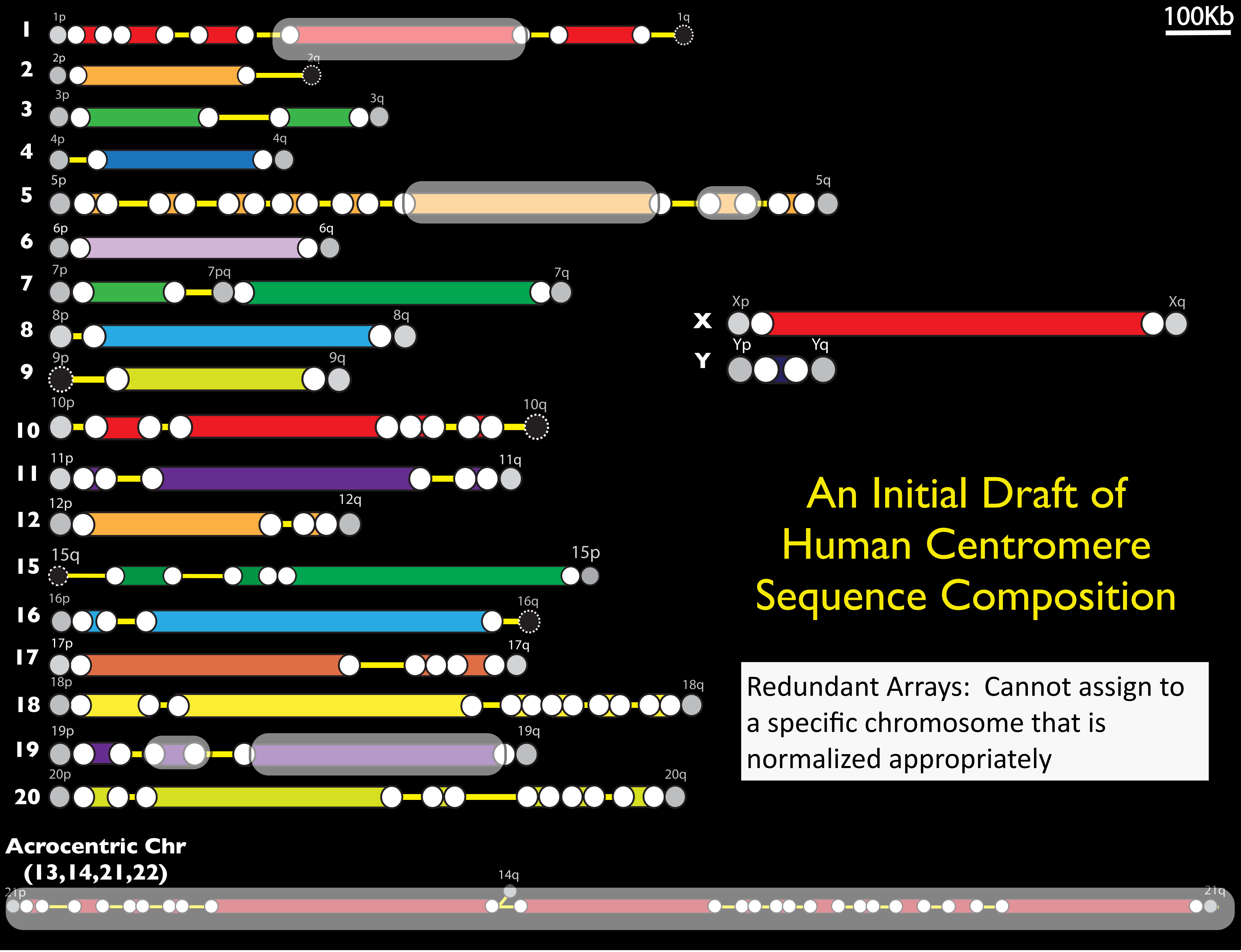


An Initial Draft of Human Centromere Sequence Composition



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

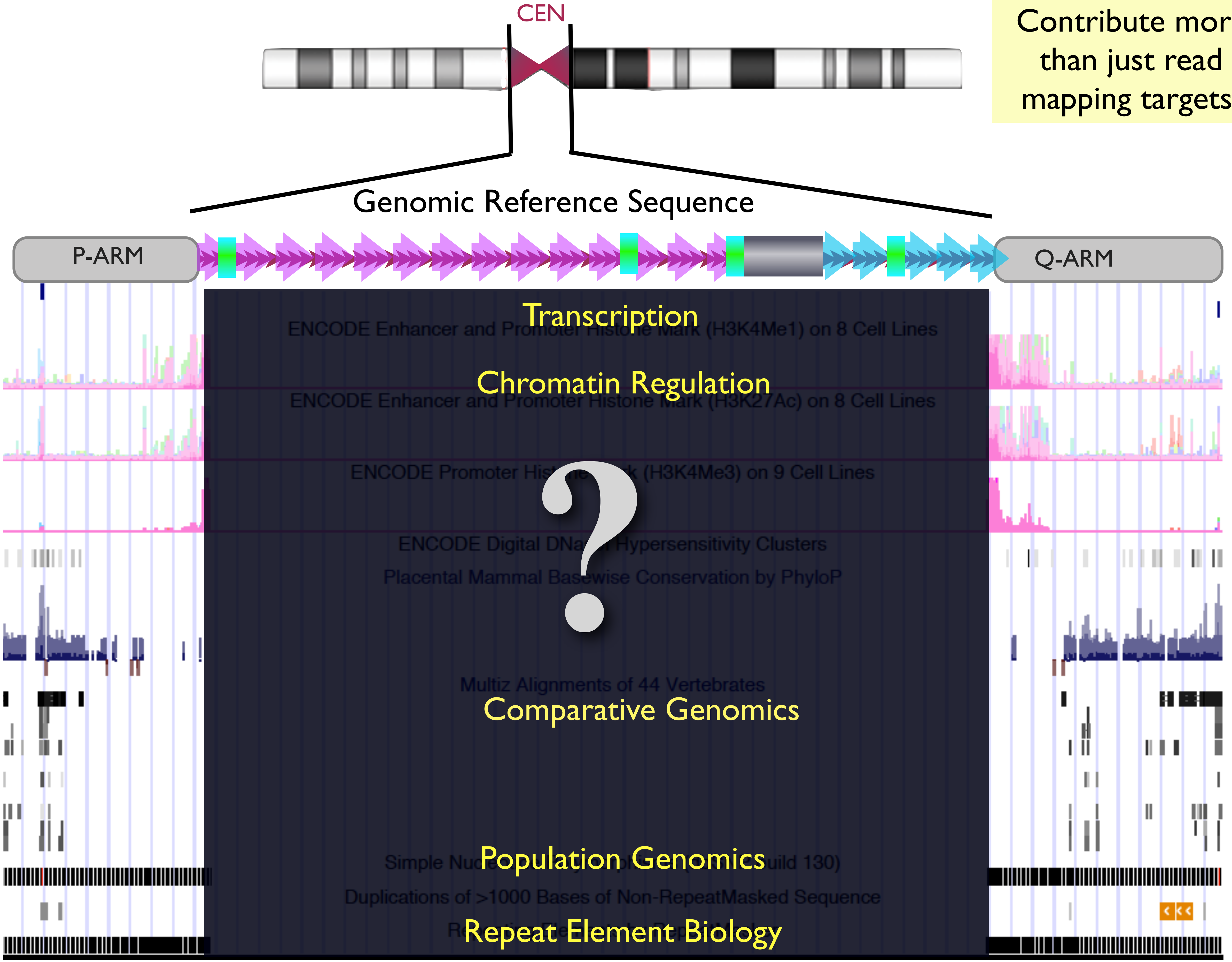
An Initial Draft of Human Centromere Sequence Composition



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

Adding Genome Annotation in Centromere Regions

Contribute more than just read mapping targets!


CEN
Genomic Reference Sequence
P-ARM
Q-ARM
ENCODE Enhancer and Promoter Histone Mark (H3K4Me1) on 8 Cell Lines
ENCODE Enhancer and Promoter Histone Mark (H3K27Ac) on 8 Cell Lines
ENCODE Promoter Histone Mark (H3K4Me3) on 9 Cell Lines
ENCODE Digital DNaseI Hypersensitivity Clusters
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 44 Vertebrates
Simple Nucleotide Polymorphisms (dbSNP build 130)
Duplications of >1000 Bases of Non-RepeatMasked Sequence
?

Transcription

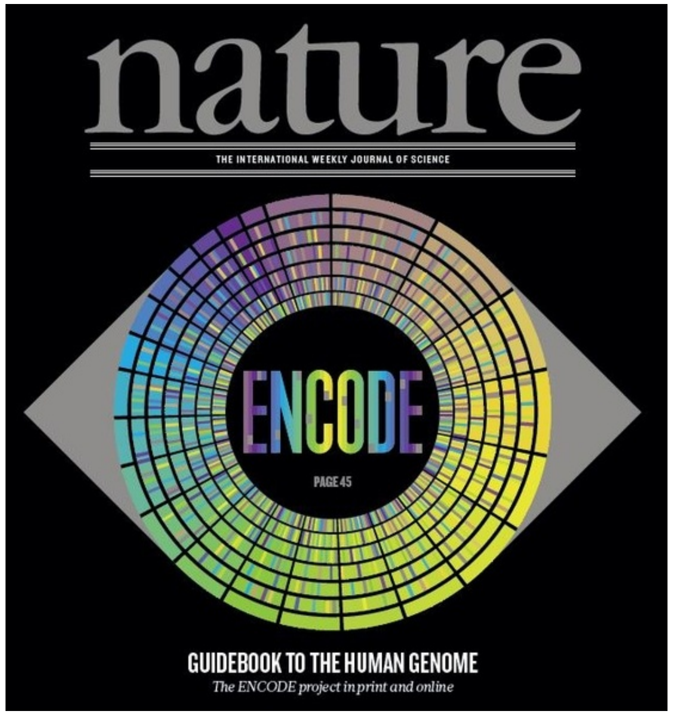
Chromatin Regulation

Comparative Genomics

Population Genomics

Repeat Element Biology

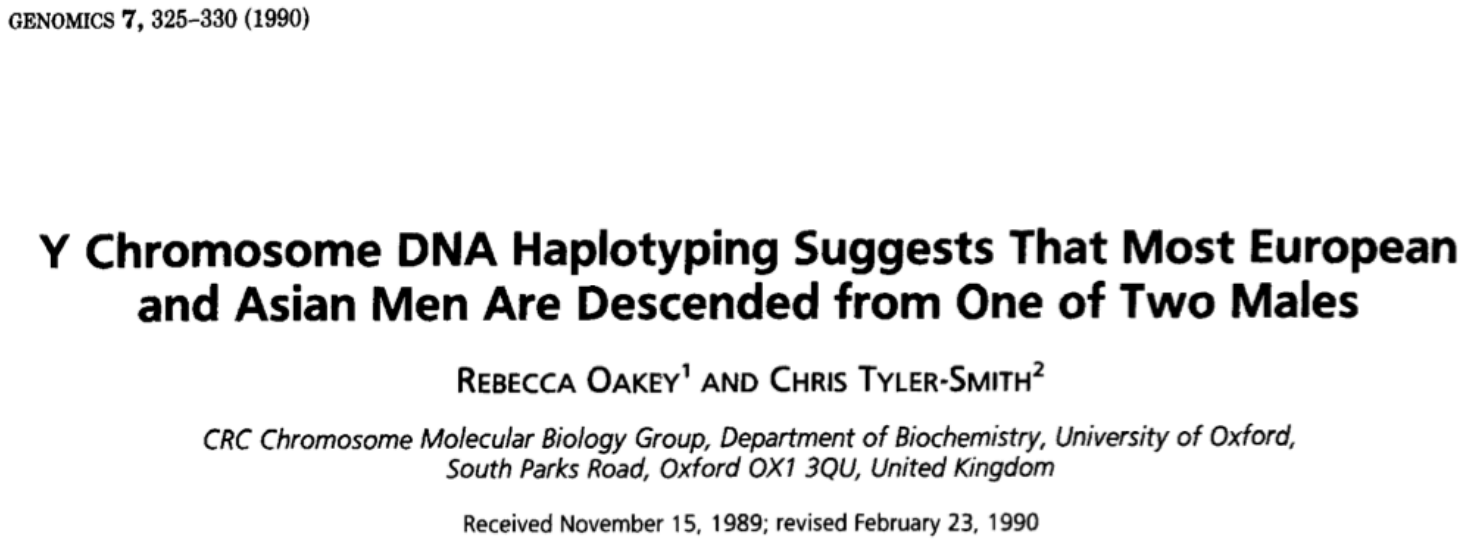
Query existing datasets that contain centromeric sequence


nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
ENCODE
PAGE 45
GUIDEBOOK TO THE HUMAN GENOME
The ENCODE project in print and online

K-mer frequency comparison and confident assignment of annotation back to the reference coordinates

1000 Genomes

A Deep Catalog of Human Genetic Variation



GENOMICS 7, 325–330 (1990)

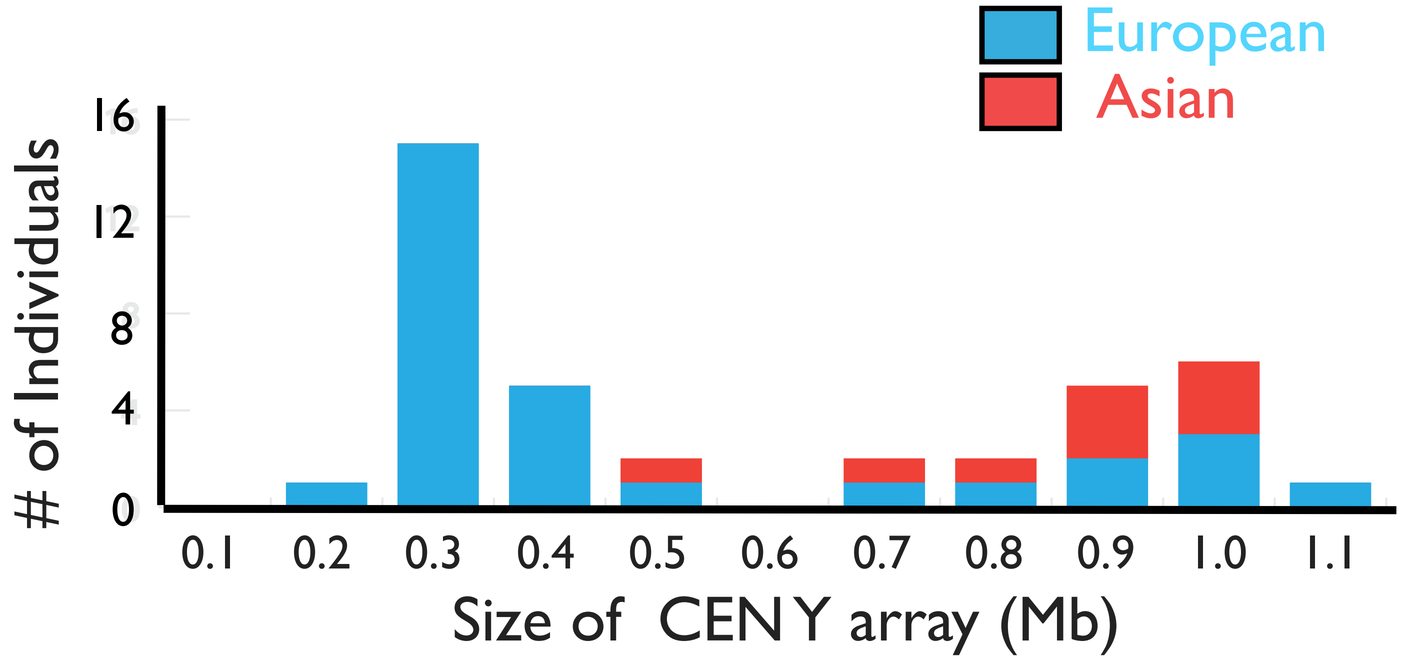
Y Chromosome DNA Haplotyping Suggests That Most European and Asian Men Are Descended from One of Two Males

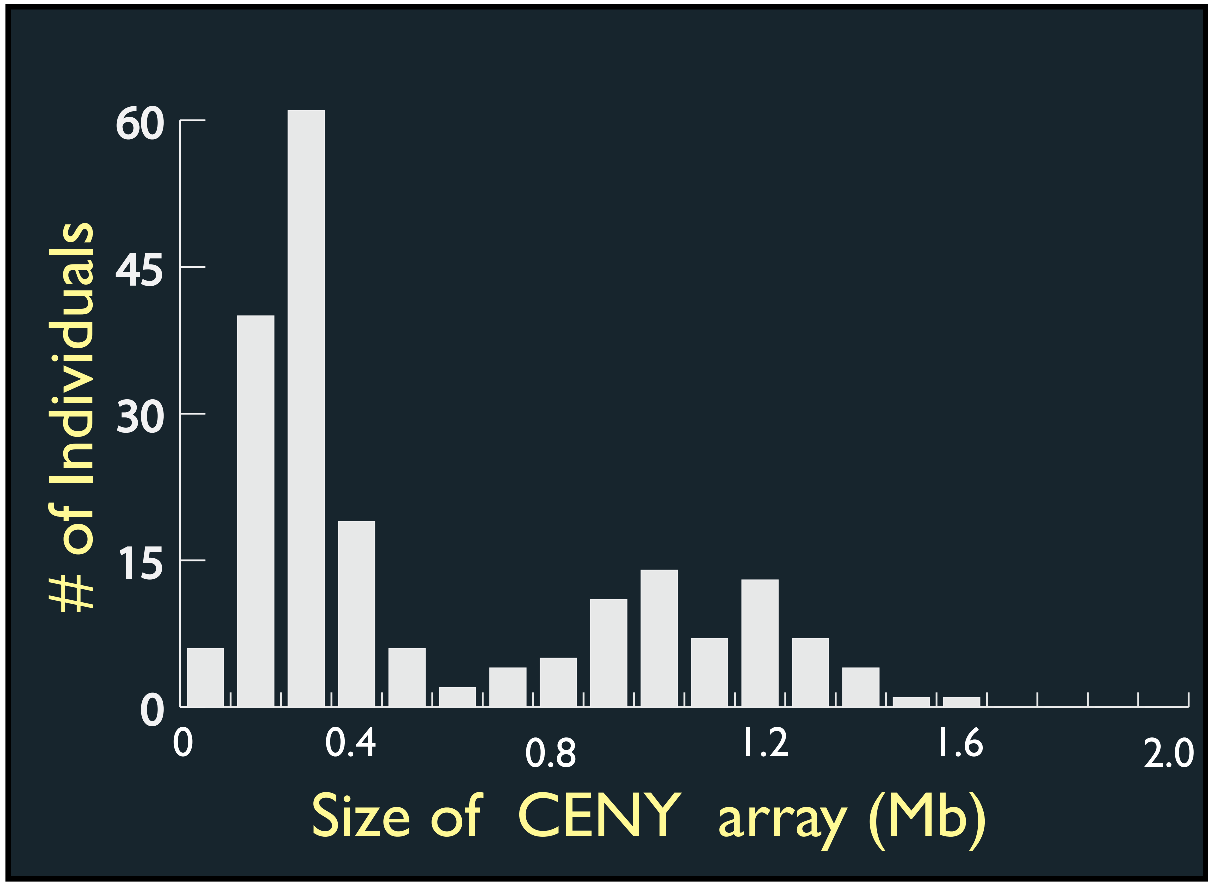
REBECCA OAKEY[1] AND CHRIS TYLER-SMITH[2]

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

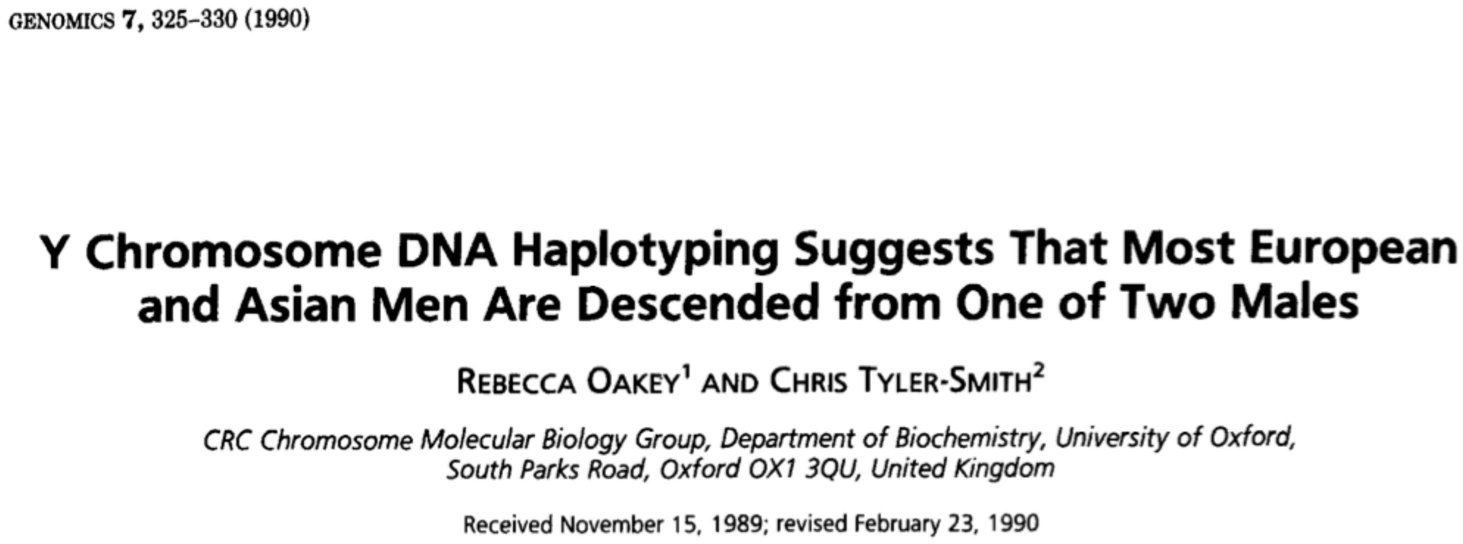
HuRef k-mers (24mers) useful in predicting array length across ~400 male individuals





1000 Genomes

A Deep Catalog of Human Genetic Variation



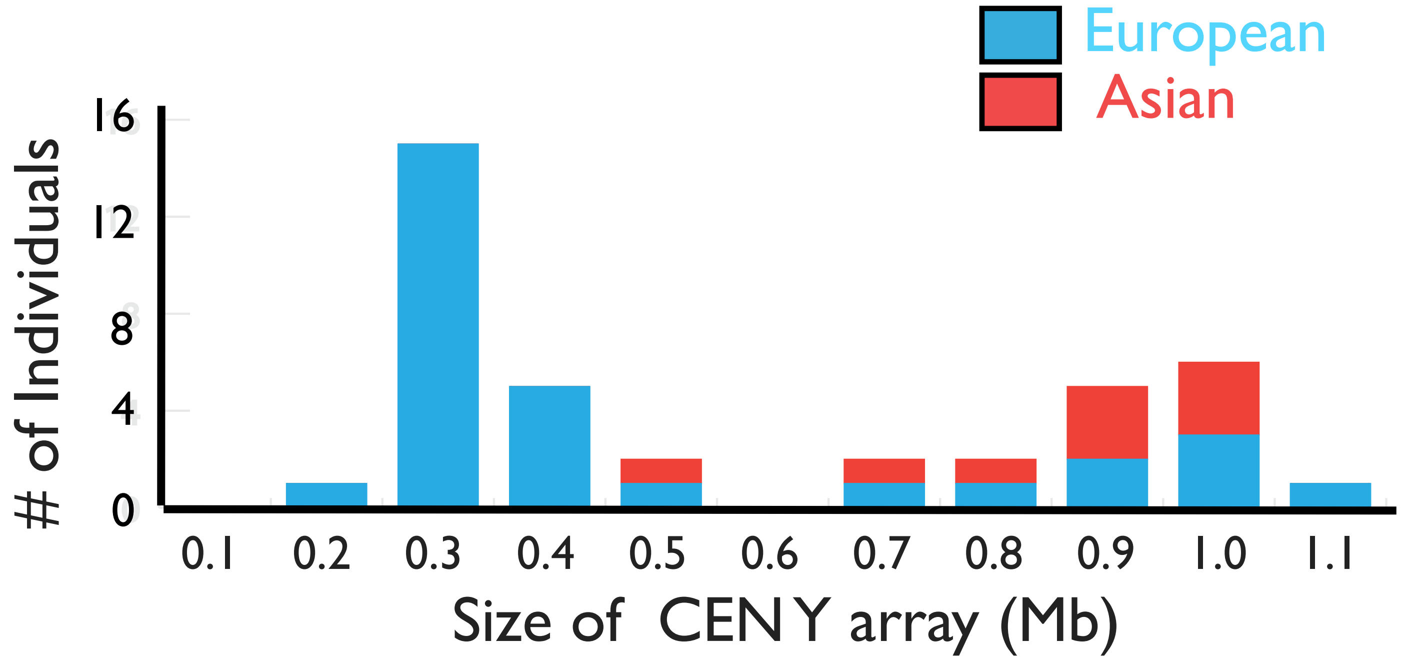
GENOMICS 7, 325–330 (1990)

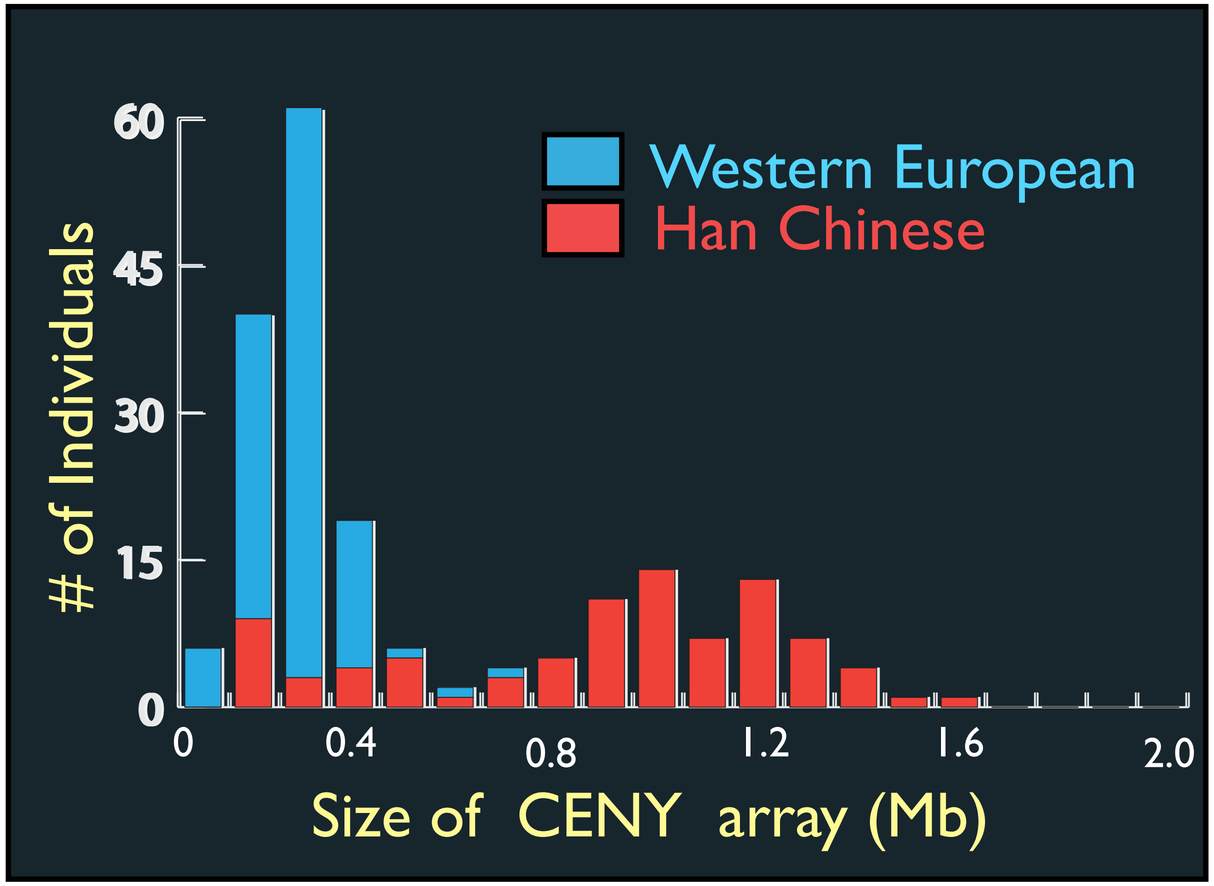
Y Chromosome DNA Haplotyping Suggests That Most European and Asian Men Are Descended from One of Two Males

REBECCA OAKEY[1] AND CHRIS TYLER-SMITH[2]

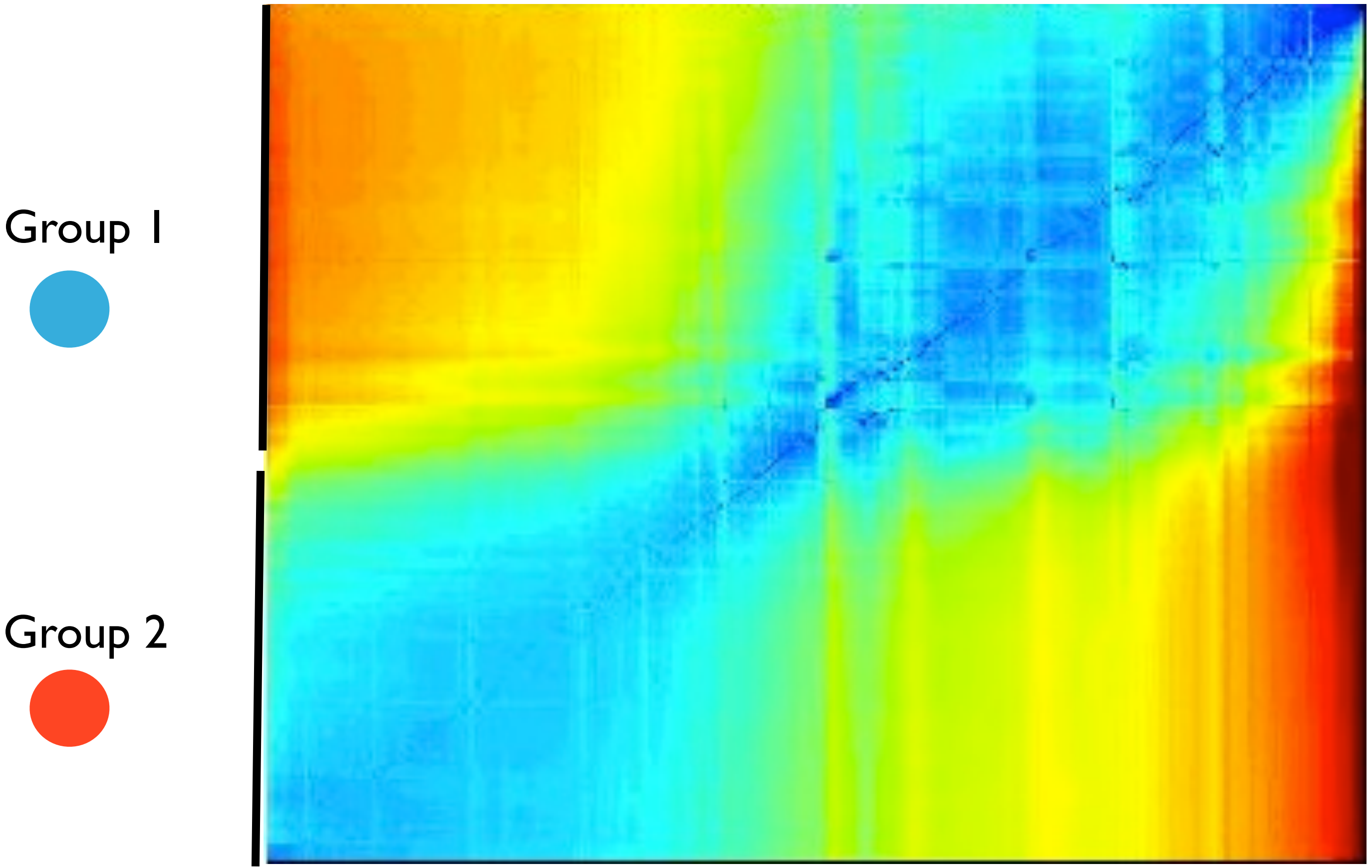
CRC Chromosome Molecular Biology Group, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, United Kingdom

Received November 15, 1989; revised February 23, 1990

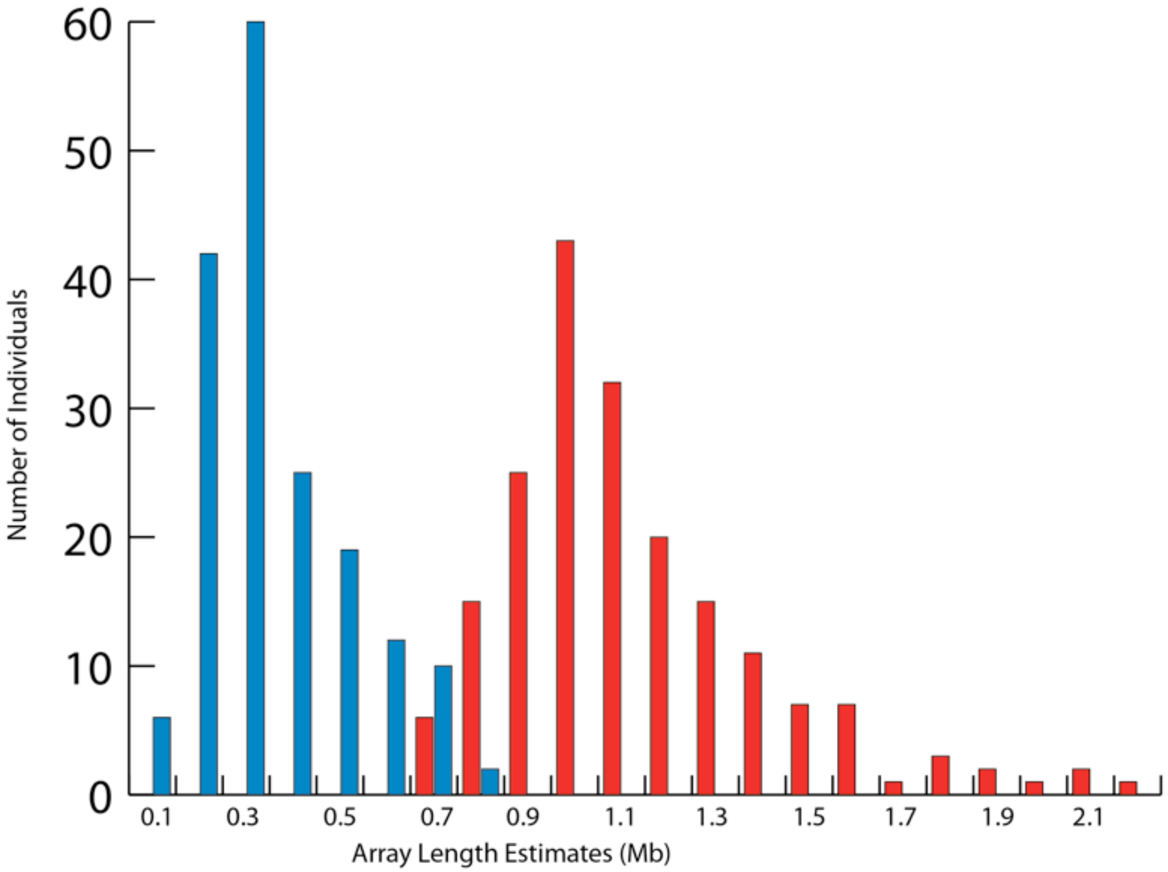


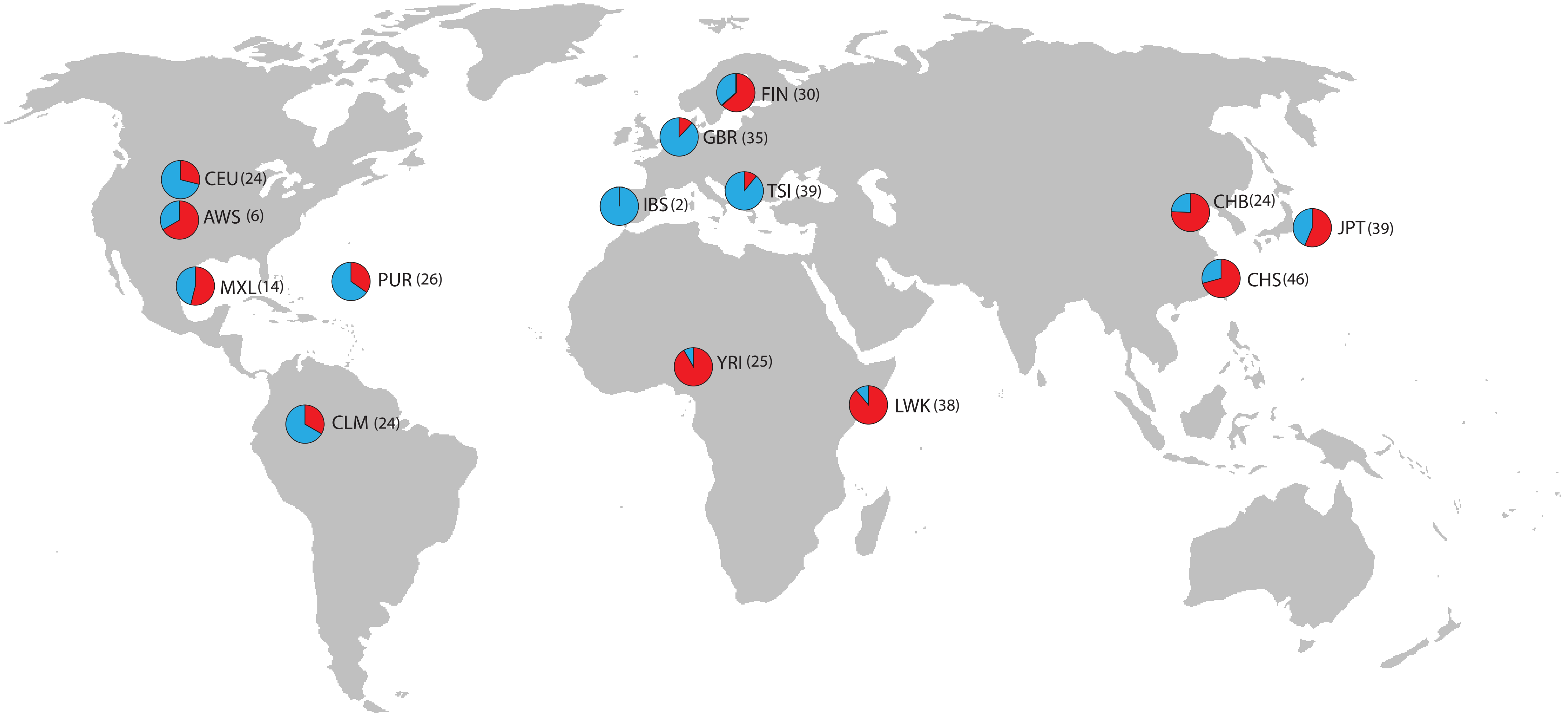


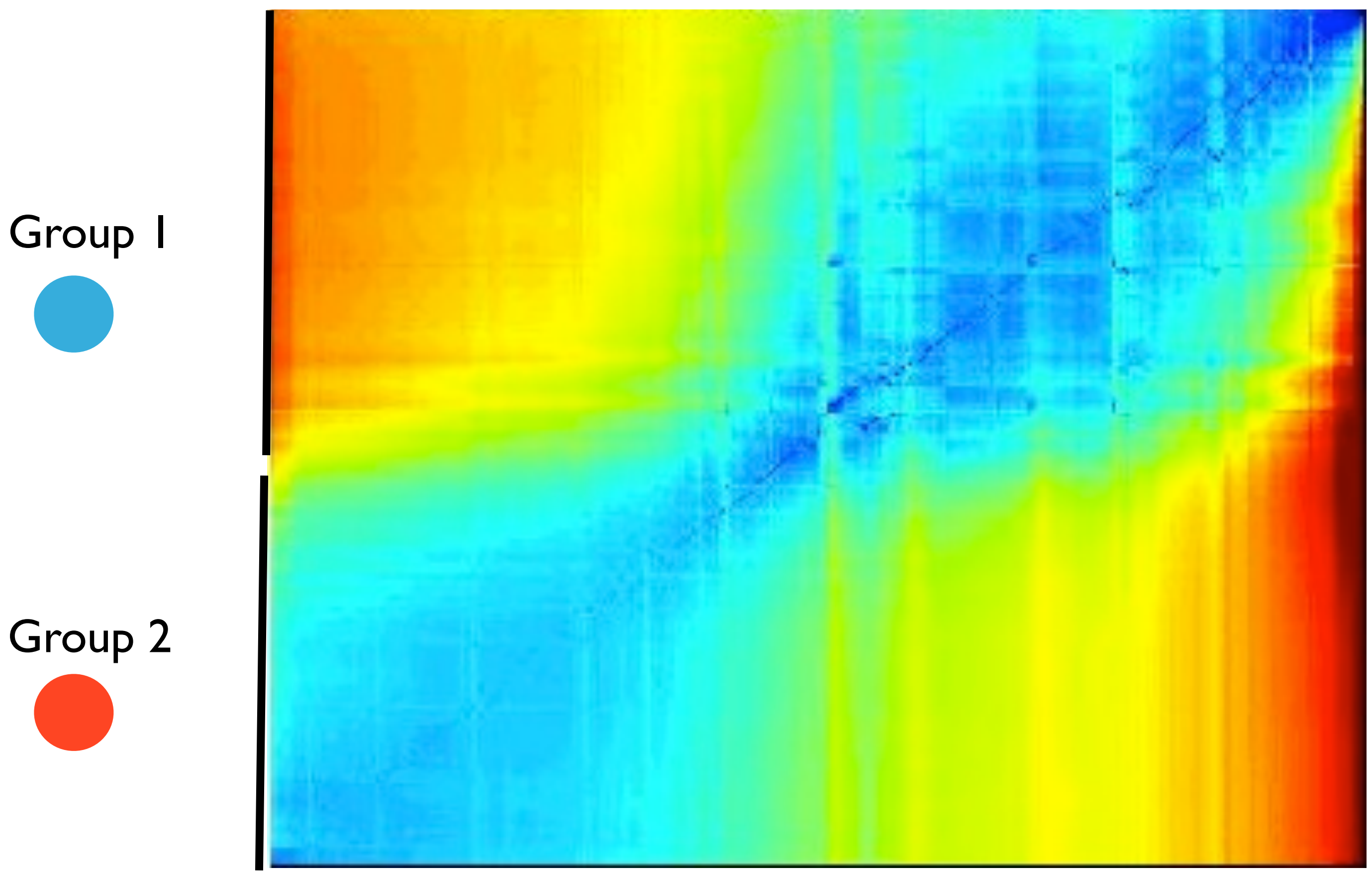
HuRef k-mer profiles are useful in predicting array classification across ~400 male individuals into two distinct groups



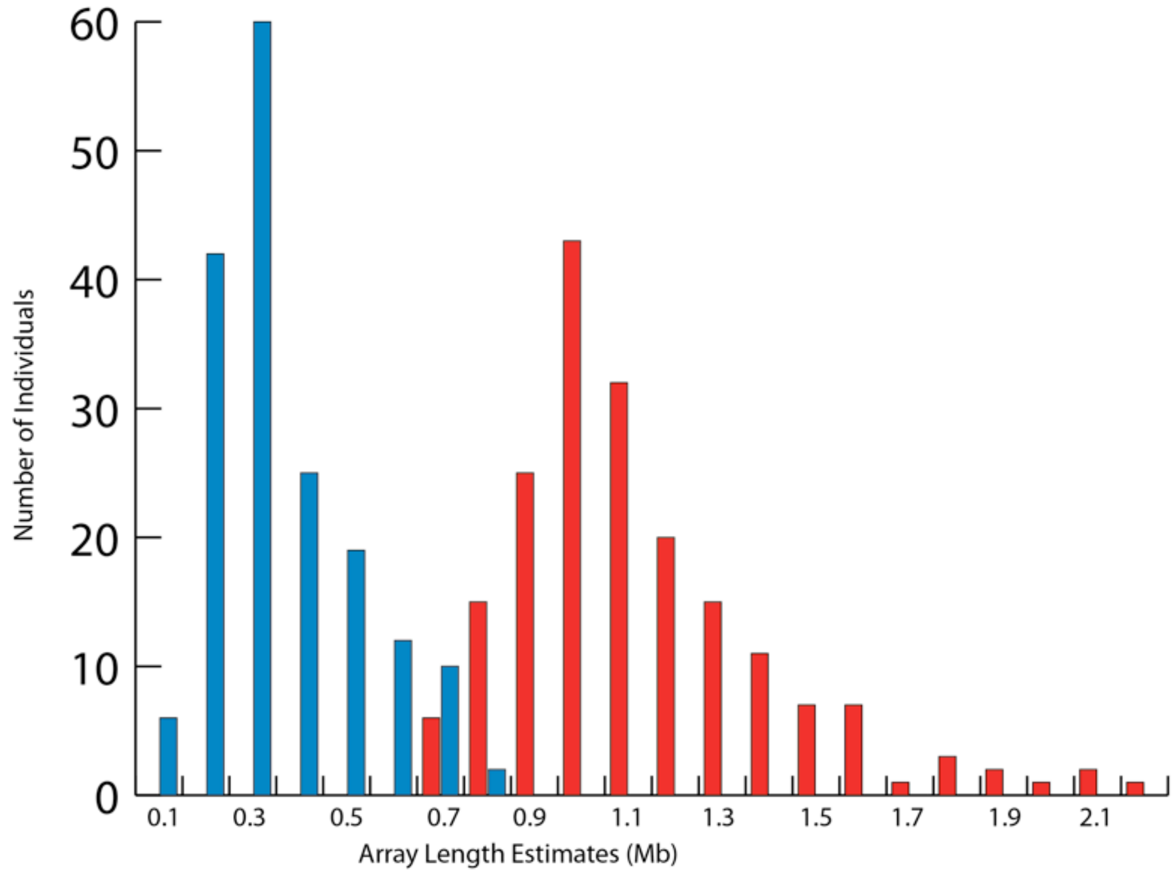
Clustergram: K-mer Identity Matrix between Male Individuals

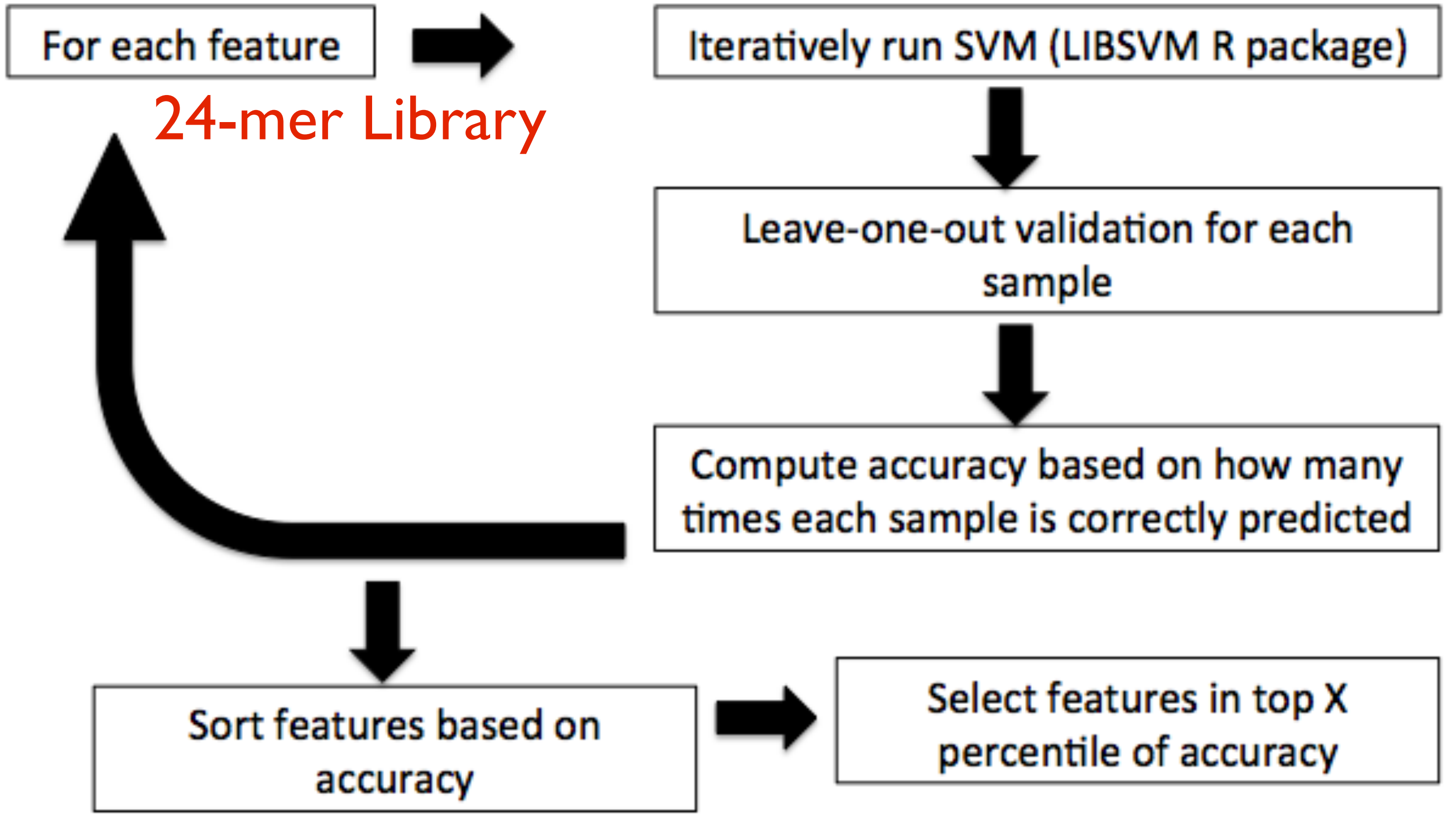






Clustergram: K-mer Identity Matrix between Male Individuals





For each feature

24-mer Library

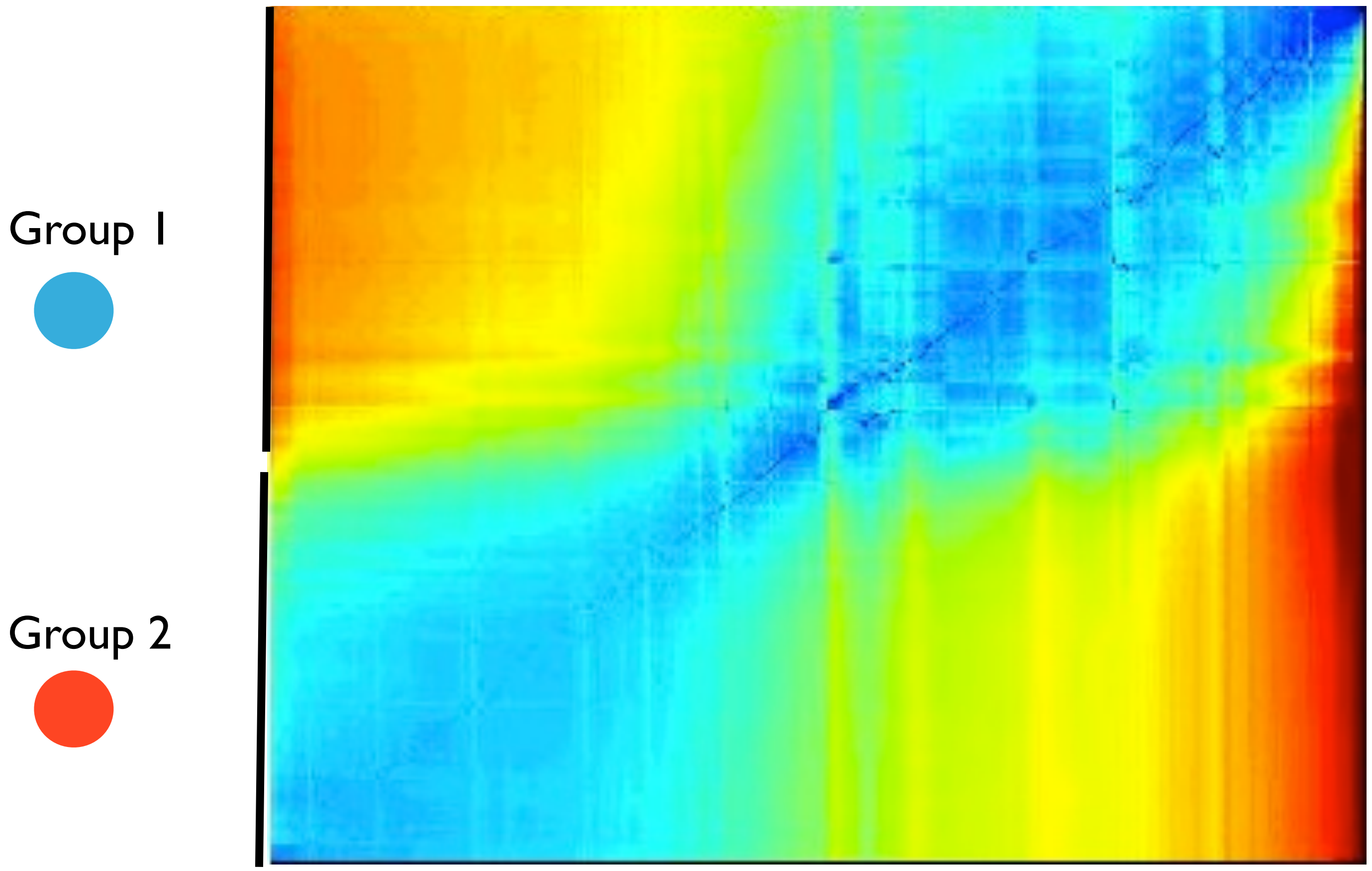
Iteratively run SVM (LIBSVM R package)

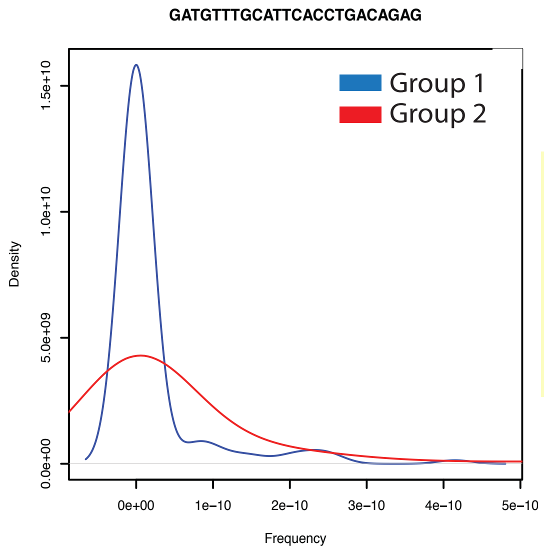
Leave-one-out validation for each sample

Compute accuracy based on how many times each sample is correctly predicted

Sort features based on accuracy

Select features in top X percentile of accuracy

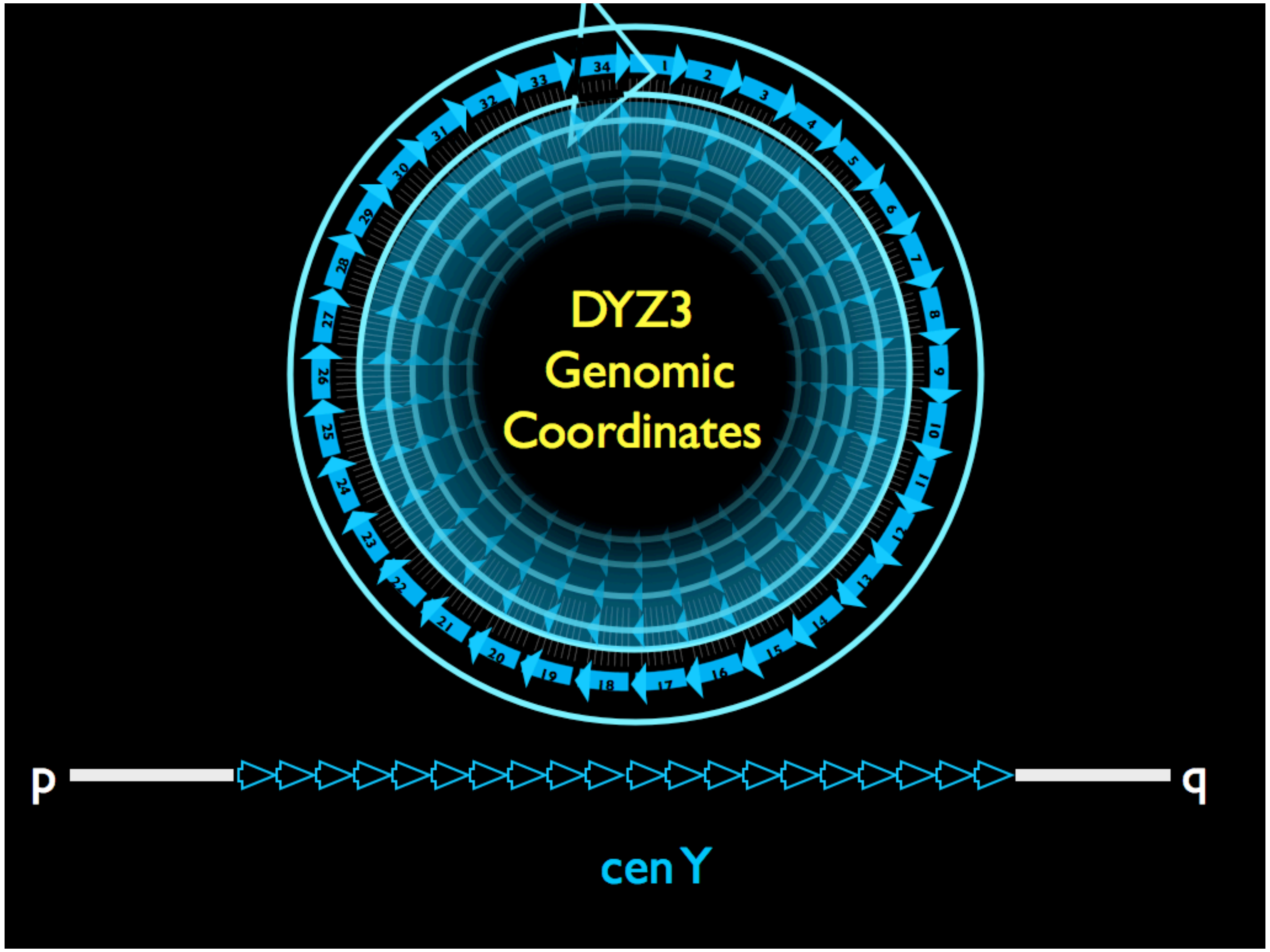




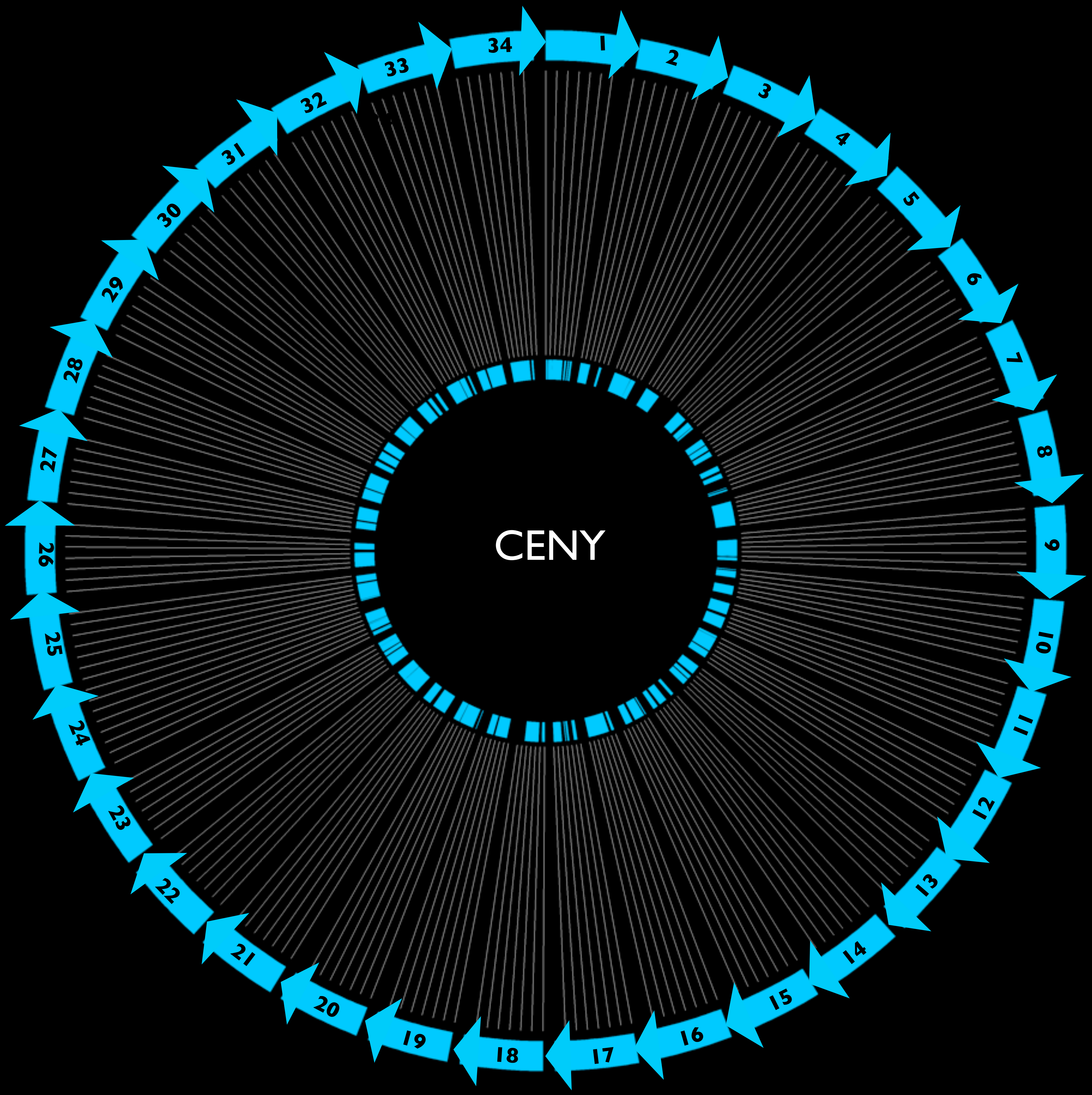
Survey those k-mers that are enriched in one array group

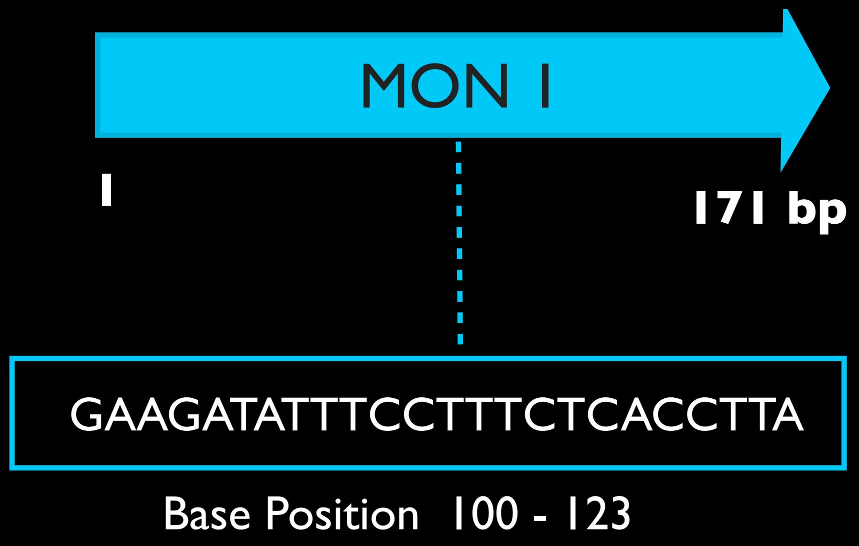
Catalogue a new source of human sequence variation

ENCODE data


DYZ3
Genomic
Coordinates
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
p
q
cen Y

ENCODE Tier 1: Human Embryonic Stem Cell (H1-hESC)

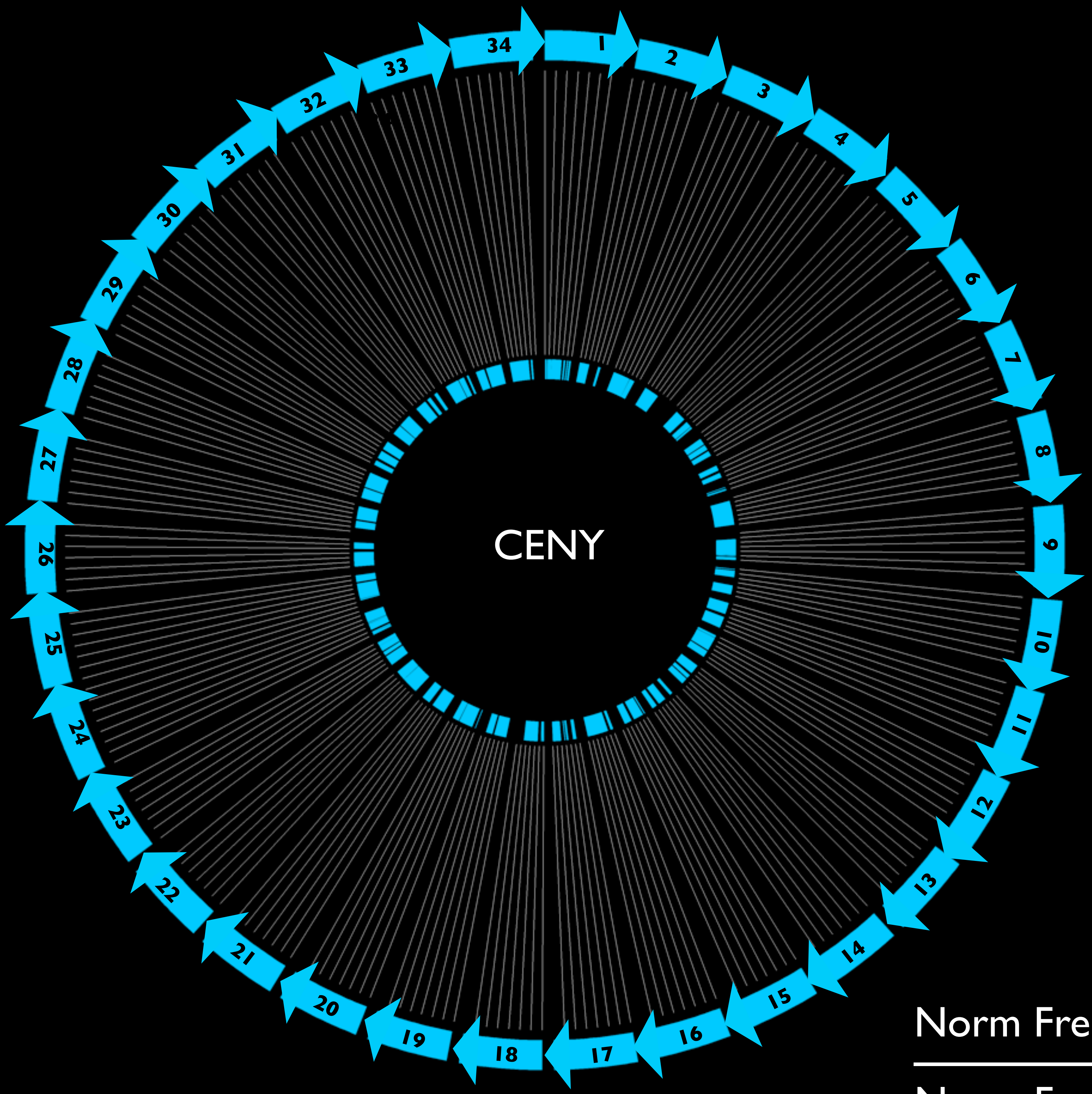


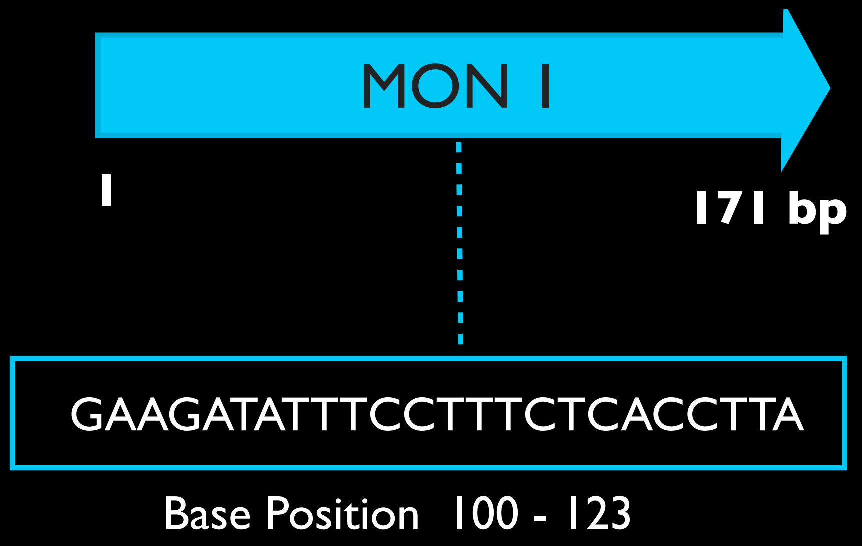


GAAGATATTTCCTTTCTCACCTTA

Base Position 100 - 123

ENCODE Tier 1: Human Embryonic Stem Cell (H1-hESC)



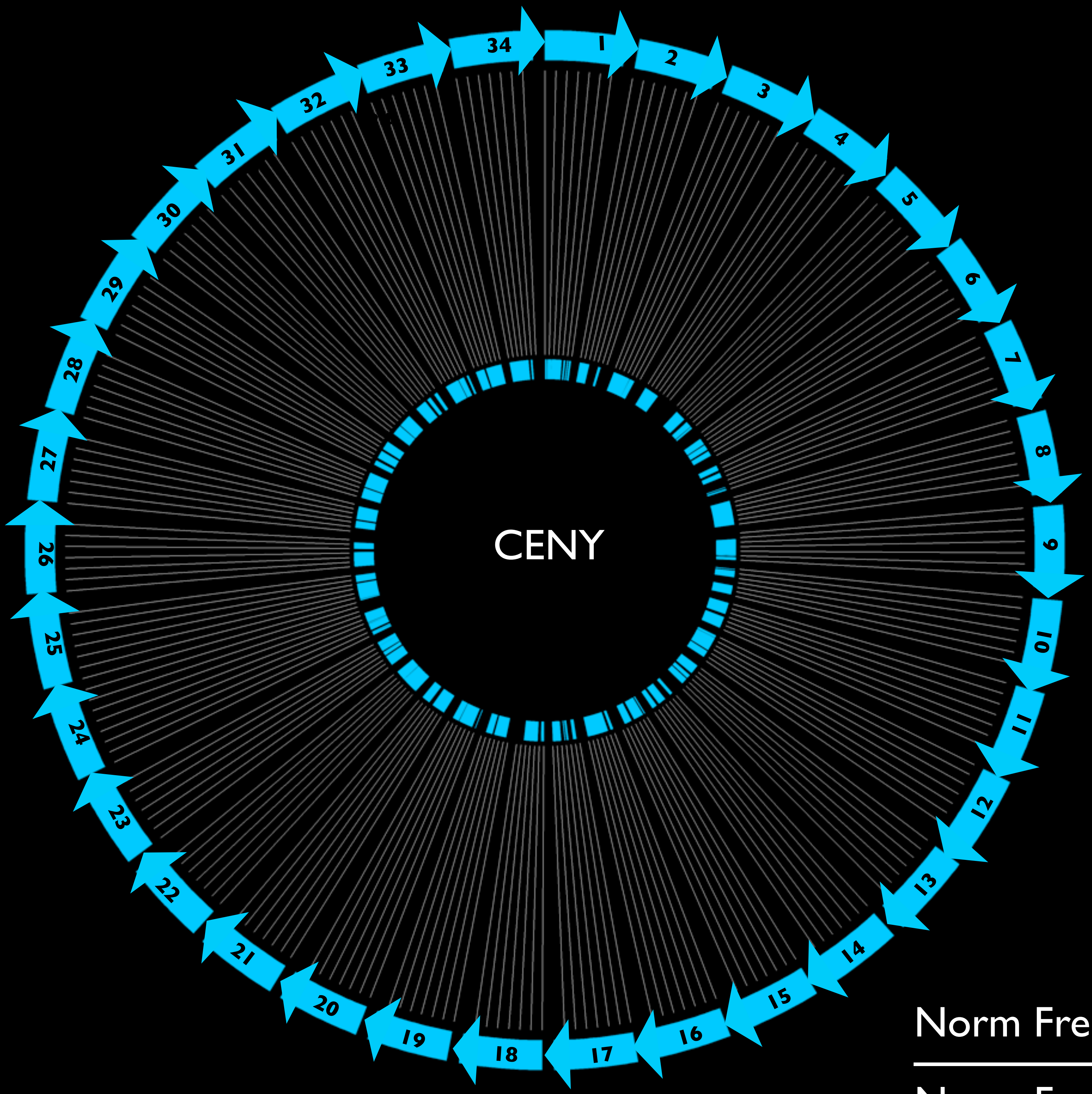


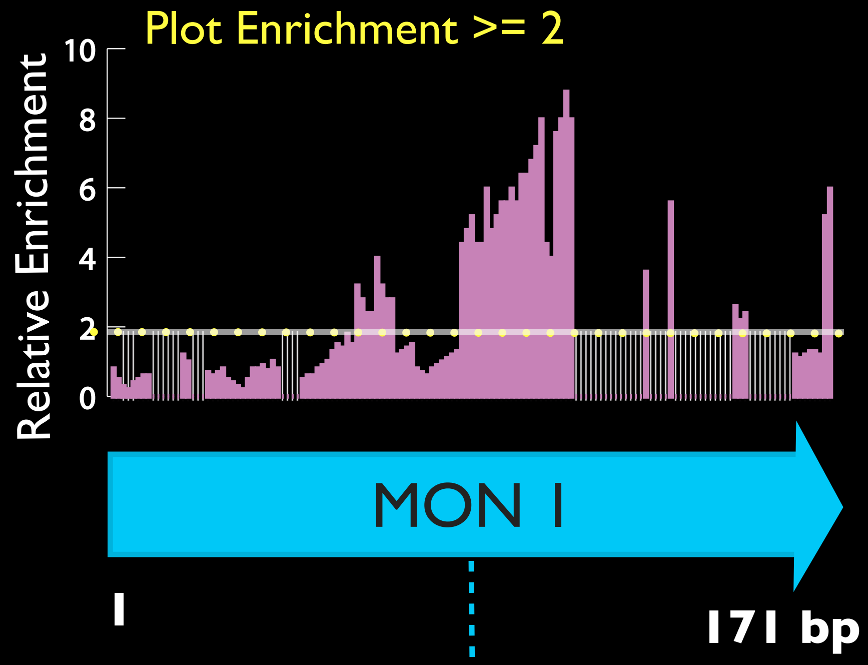
GAAGATATTTCCTTTCTCACCTTA

Base Position 100 - 123

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

ENCODE Tier 1: Human Embryonic Stem Cell (H1-hESC)



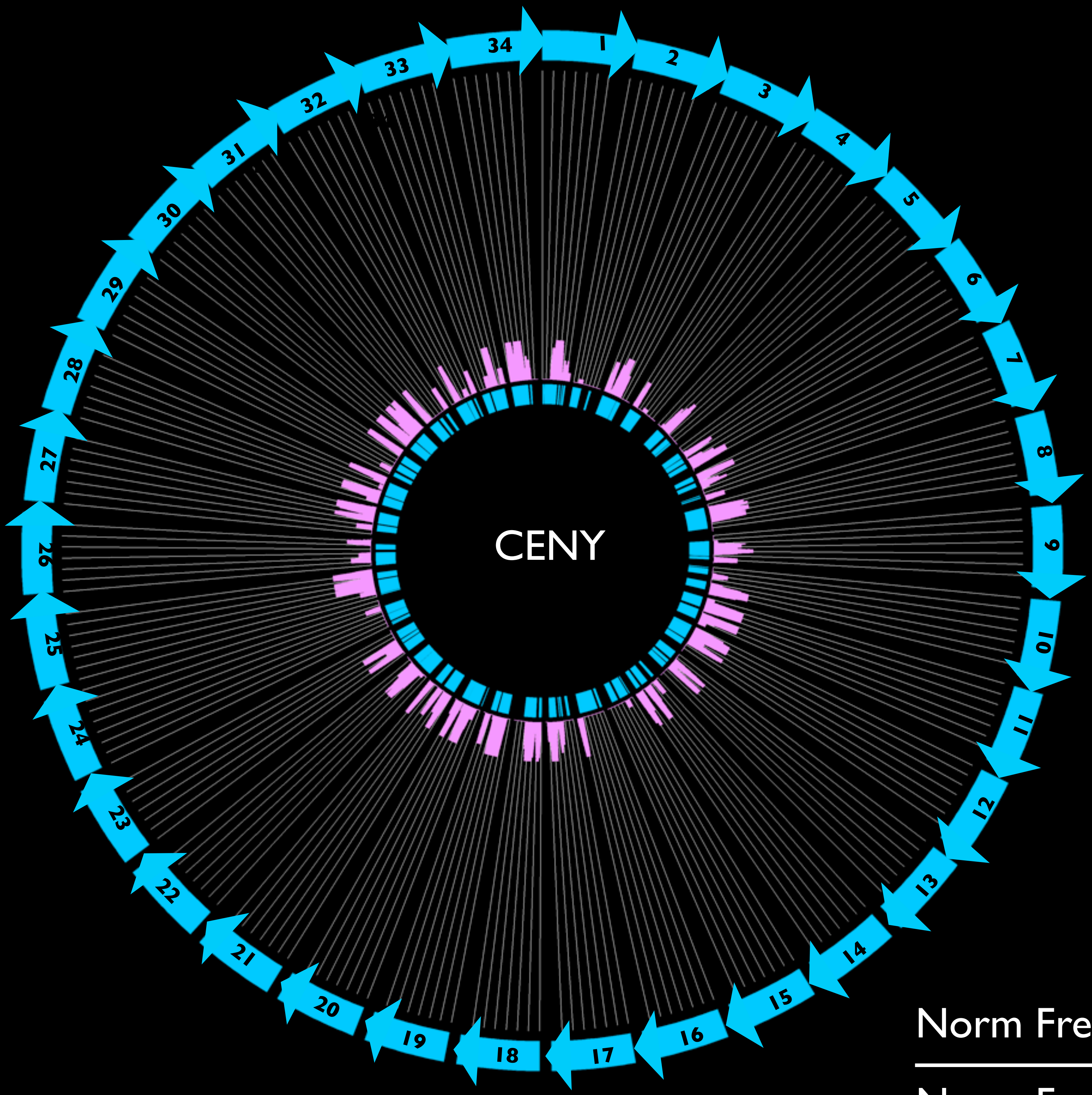


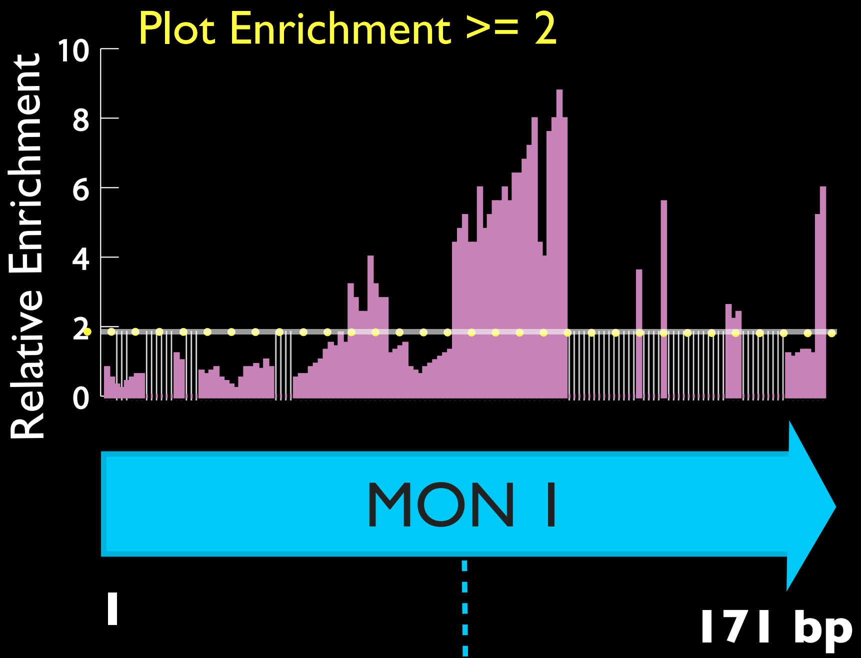
GAAGATATTTCCTTTCTCACCTTA

Base Position 100 - 123

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

ENCODE Tier 1: Human Embryonic Stem Cell (H1-hESC)



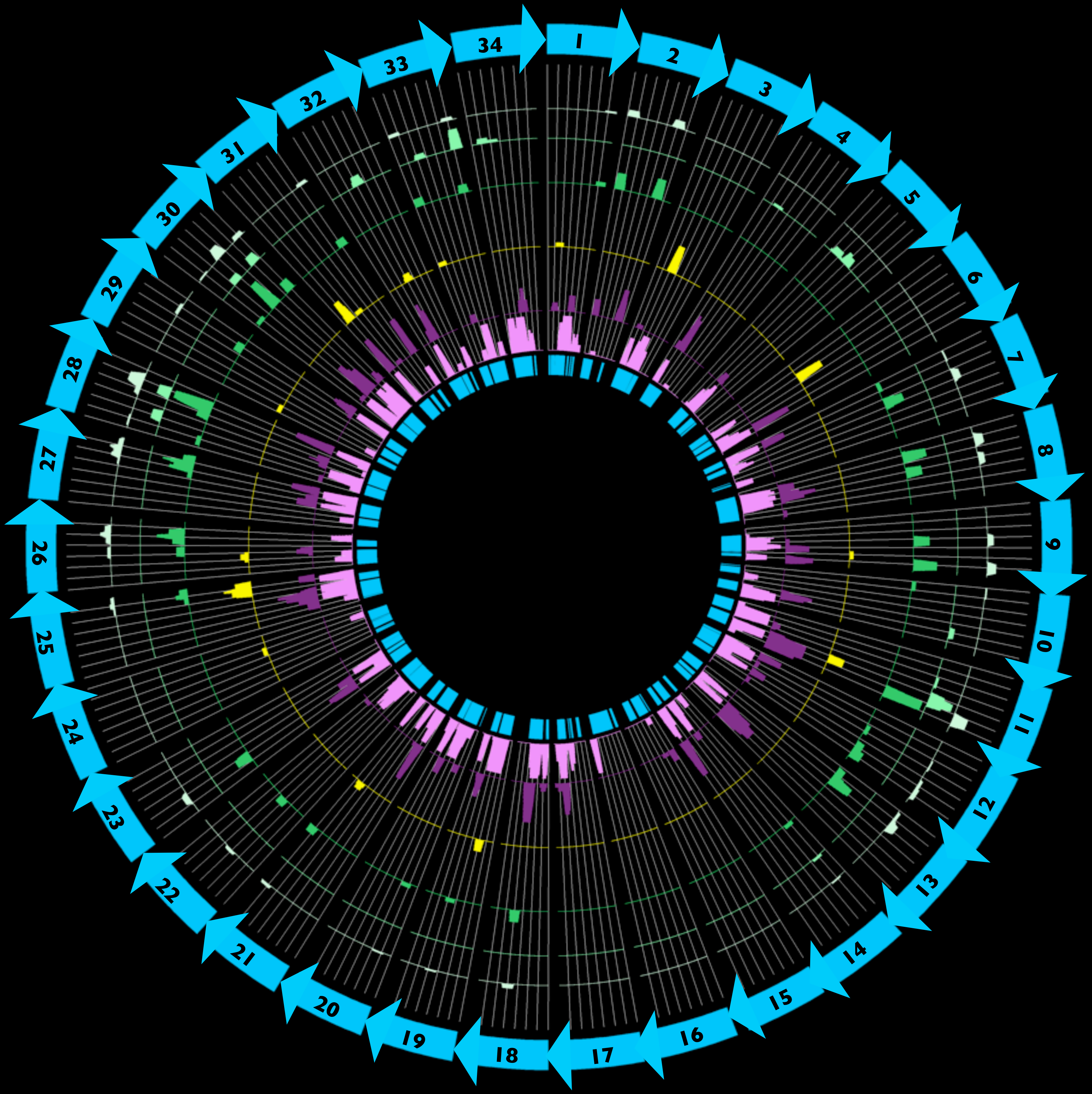


GAAGATATTTCCTTTCTCACCTTA

Base Position 100 - 123

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

H1hESC Histone Profile of DYZ3 Array



Active Chromatin

● H3K4me1

● H3K4me2

● H3K4me3

Histone Variants

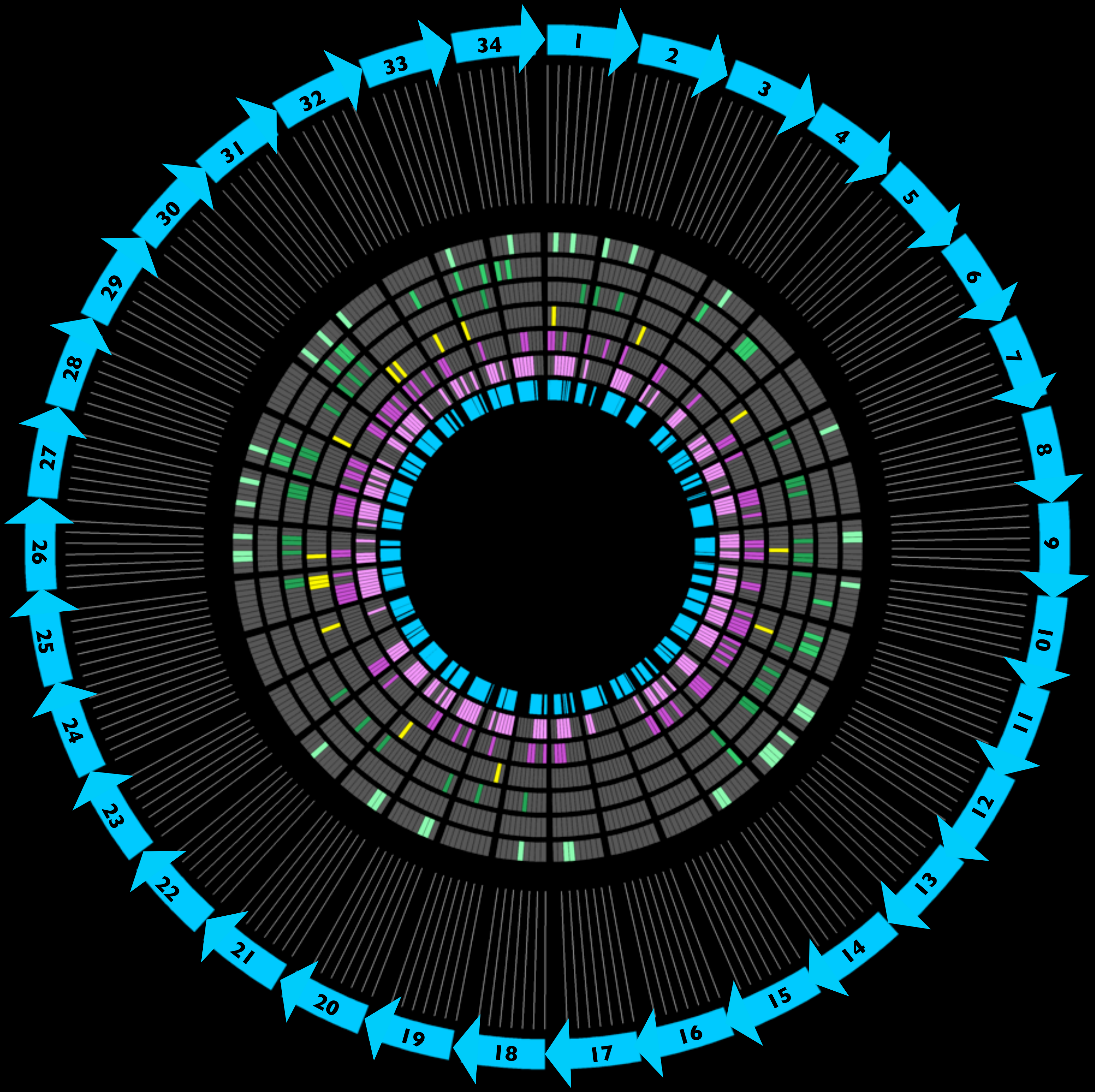
● H2Az

Repressive Chromatin

● H3K27me3

● H3K9me3

H1hESC Histone Profile of DYZ3 Array



Active Chromatin

● H3K4me1

● H3K4me2

● H3K4me3

Histone Variants

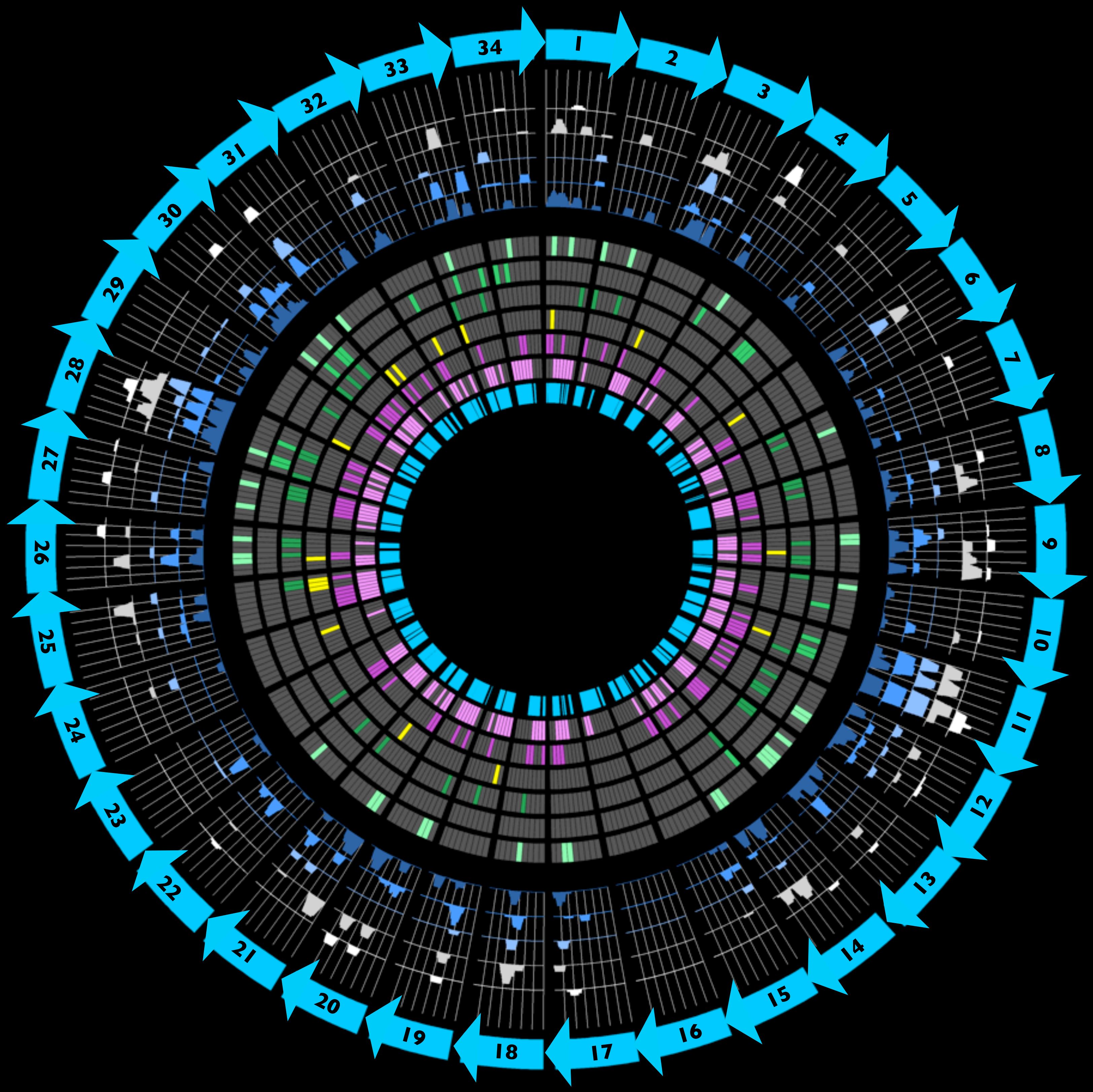
● H2Az

Repressive Chromatin

● H3K27me3

● H3K9me3

H1hESC Transcription Factor Enrichment Profile



Transcription Factor

- ● EZH2
- ● HDAC6
- ● PLU1
- ● JARID1A
- ● SUZ12

Active Chromatin

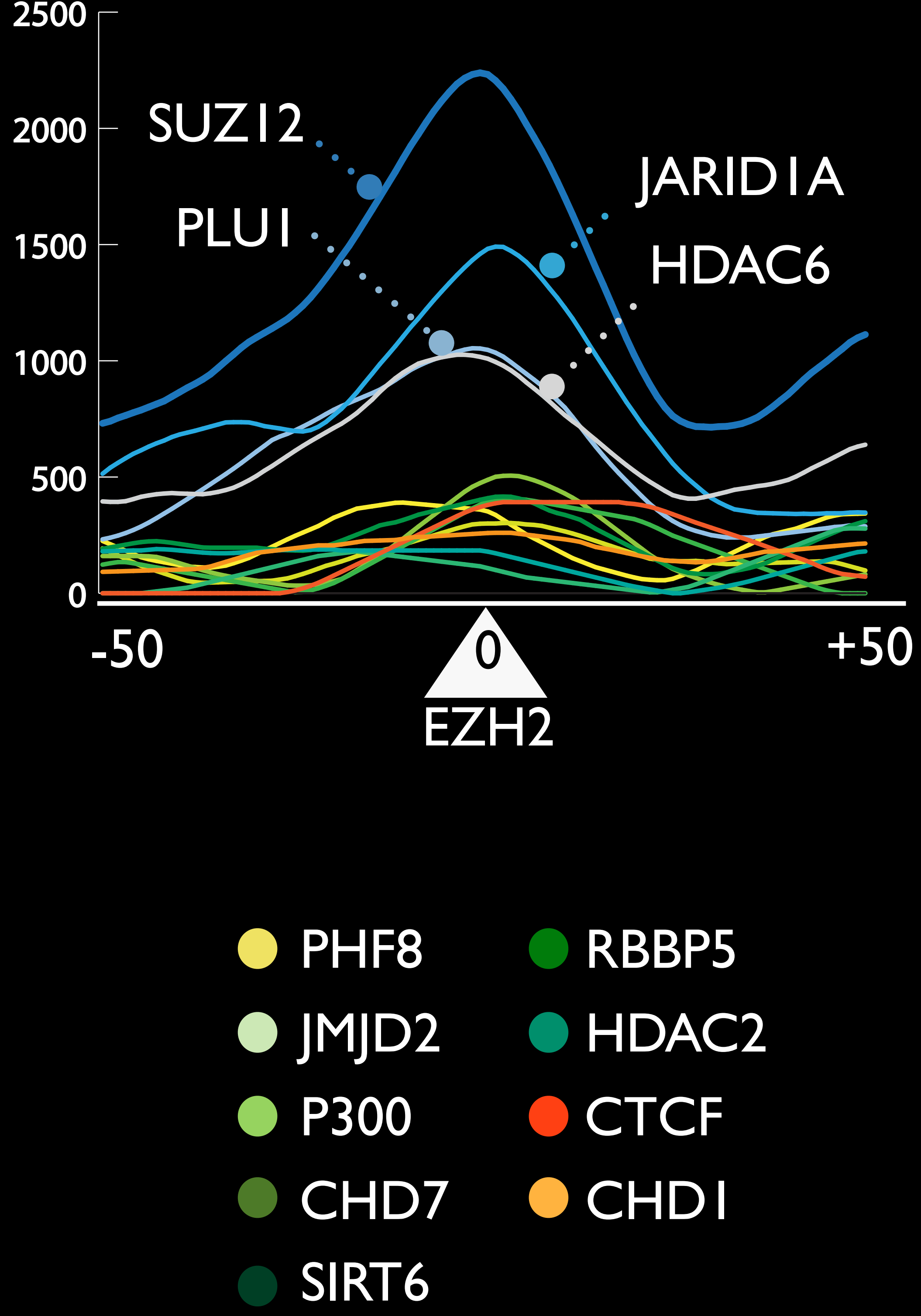
- ● H3K4me1
- ● H3K4me2
- ● H3K4me3

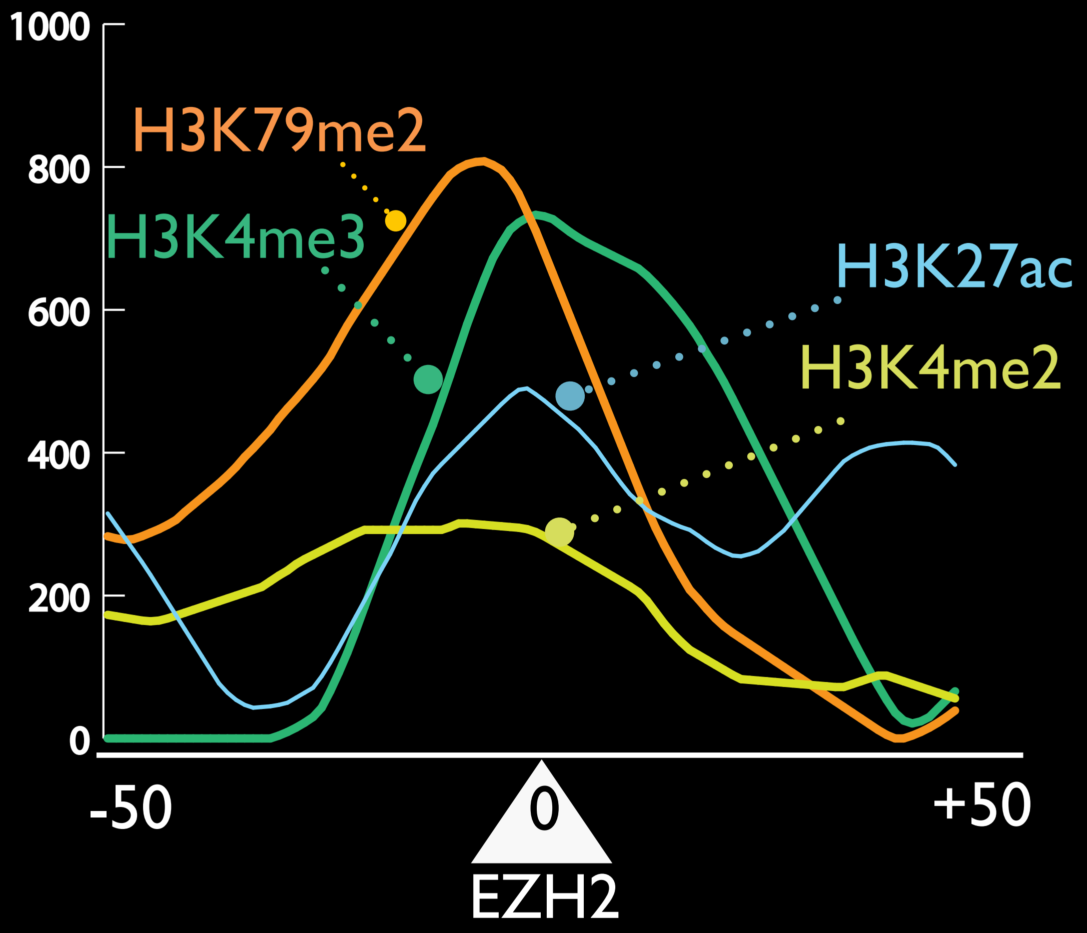
Histone Variants

- ● H2Az

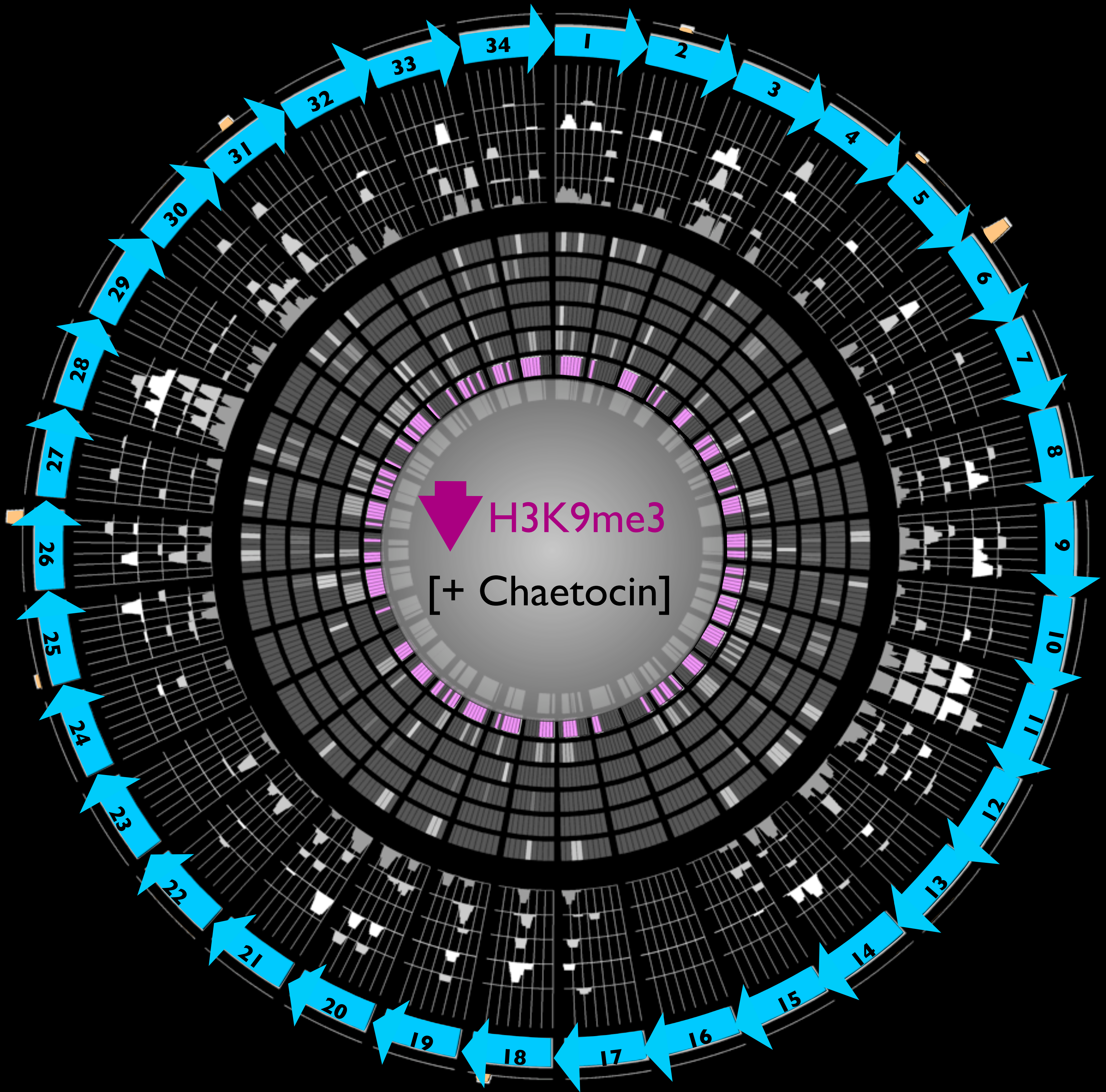
Repressive Chromatin

- ● H3K27me3
- ● H3K9me3


2500
2000
1500
1000
500
0
SUZ12
PLU1
JARID1A
HDAC6
-50
0
+50
EZH2
● PHF8
● JMJD2
● P300
● CHD7
● SIRT6
● RBBP5
● HDAC2
● CTCF
● CHD1


1000
800
600
400
200
0
H3K79me2
H3K4me3
H3K27ac
H3K4me2
-50
0
+50
EZH2

Adding Custom Datasets or “Tracks”



Transcription Factor

● EZH2

● HDAC6

● PLU1

● JARID1A

● SUZ12

Active Chromatin

● H3K4me1

● H3K4me2

● H3K4me3

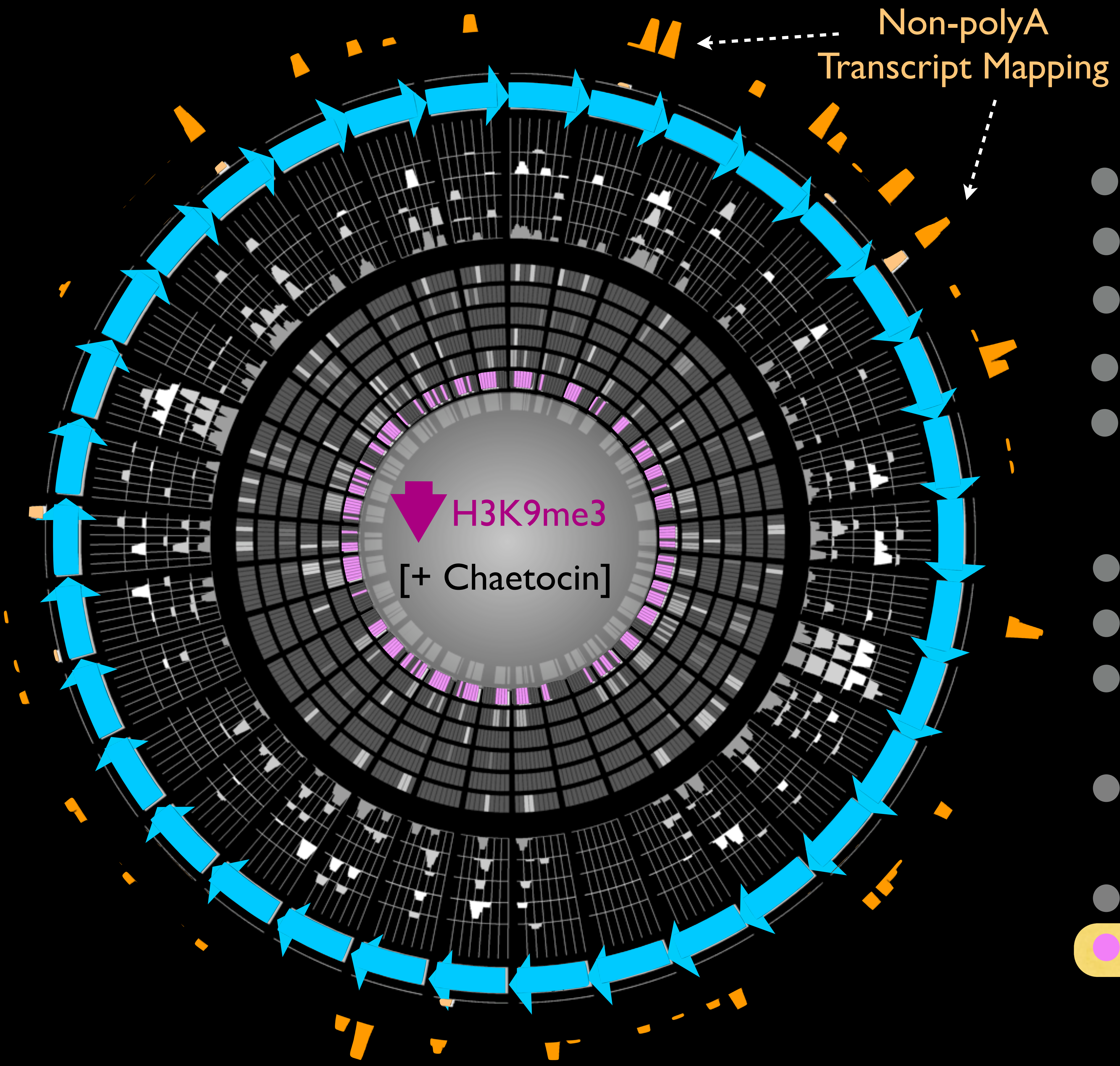
Histone Variants

● H2Az

Repressive Chromatin

● H3K27me3

● H3K9me3



Transcription Factor

- ● EZH2
- ● HDAC6
- ● PLU1
- ● JARID1A
- ● SUZ12

Active Chromatin

- ● H3K4me1
- ● H3K4me2
- ● H3K4me3

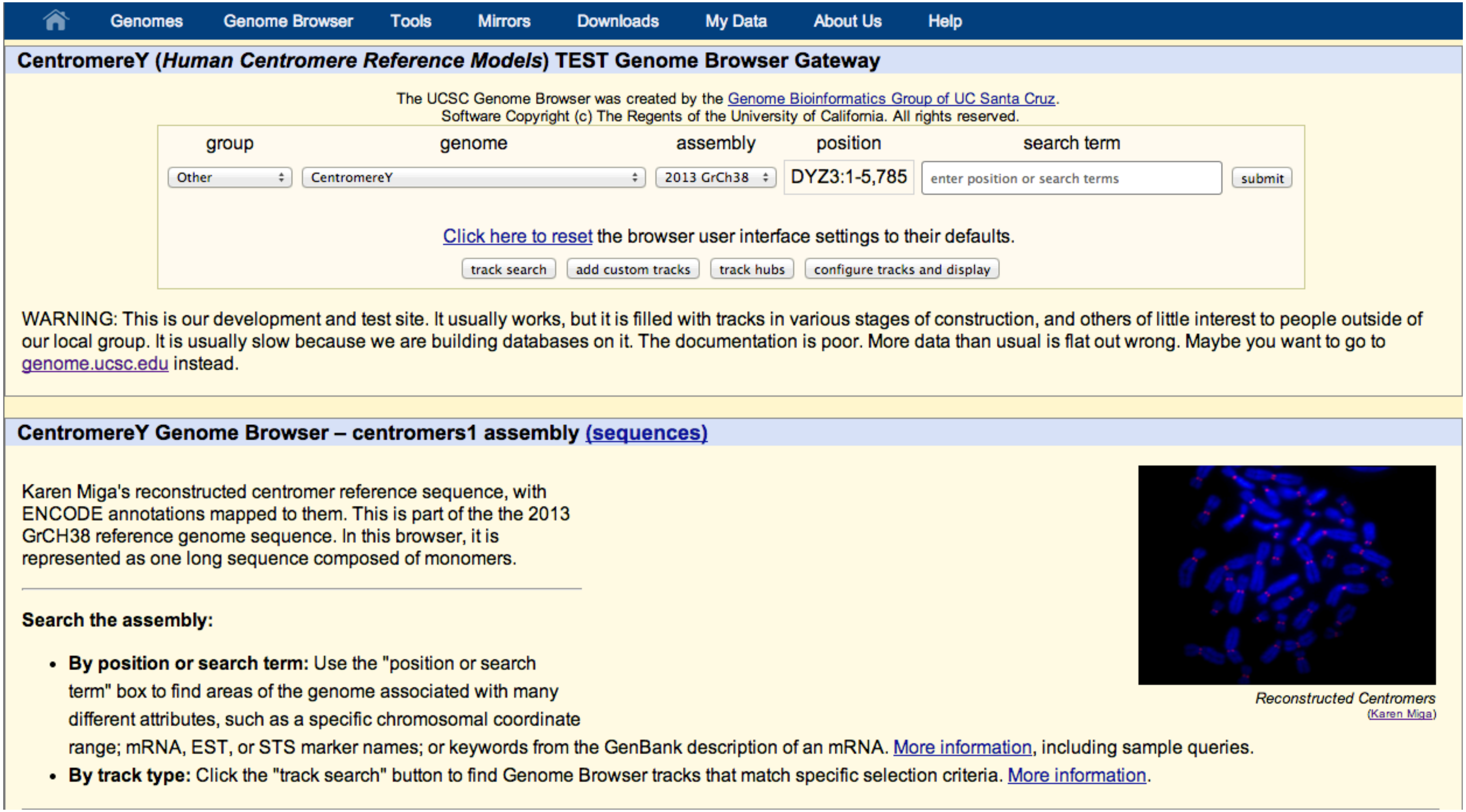
Histone Variants

- ● H2Az

Repressive Chromatin

- ● H3K27me3
- ● H3K9me3

UCSC: Centromere Annotation and Tool Development



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	submit

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)

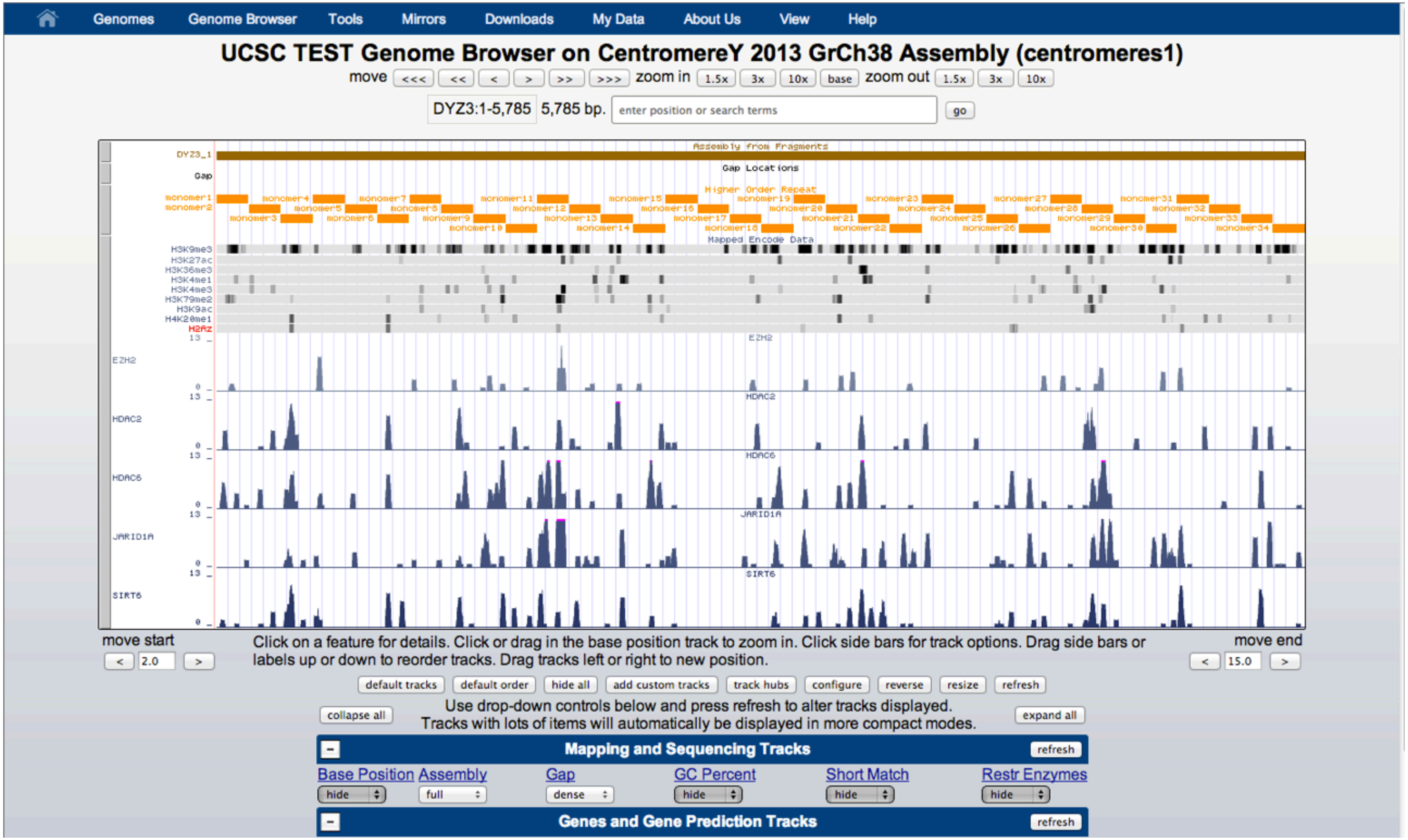
Karen Miga's reconstructed centromer reference sequence, with ENCODE annotations mapped to them. This is part of the the 2013 GrCH38 reference genome sequence. In this browser, it is represented as one long sequence composed of monomers.

Reconstructed Centromers
(Karen Miga)

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.

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.
Tracks with lots of items will automatically be displayed in more compact modes.

expand all

− 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

Acknowledgements

Baskin Engineering
UC SANTA CRUZ

Jim Kent
David Haussler
Max Hauesslar
Miten Jain
Yulia Newton
Dave Greenberg

DUKE INSTITUTE FOR
GENOME
SCIENCES & POLICY

Hunt Willard
Nick Altemose

GRC Genome Reference Consortium

Deanna Church
Valerie Schneider
Karen Clarke