Supplemental Material for

Identification and characterization of centromeric sequences in *Xenopus laevis*

Owen K Smith[1,2], Charles Limouse[1], Kelsey A Fryer[1,3], Nicole A Teran[3], Kousik Sundararajan[1], Rebecca Heald[4], Aaron F Straight[1*]

[1]Department of Biochemistry
Stanford University School of Medicine
279 Campus Drive
Beckman Center 409
Stanford, CA 94305-5307

[2]Department of Chemical and Systems Biology
Stanford University School of Medicine
Stanford, CA 94305

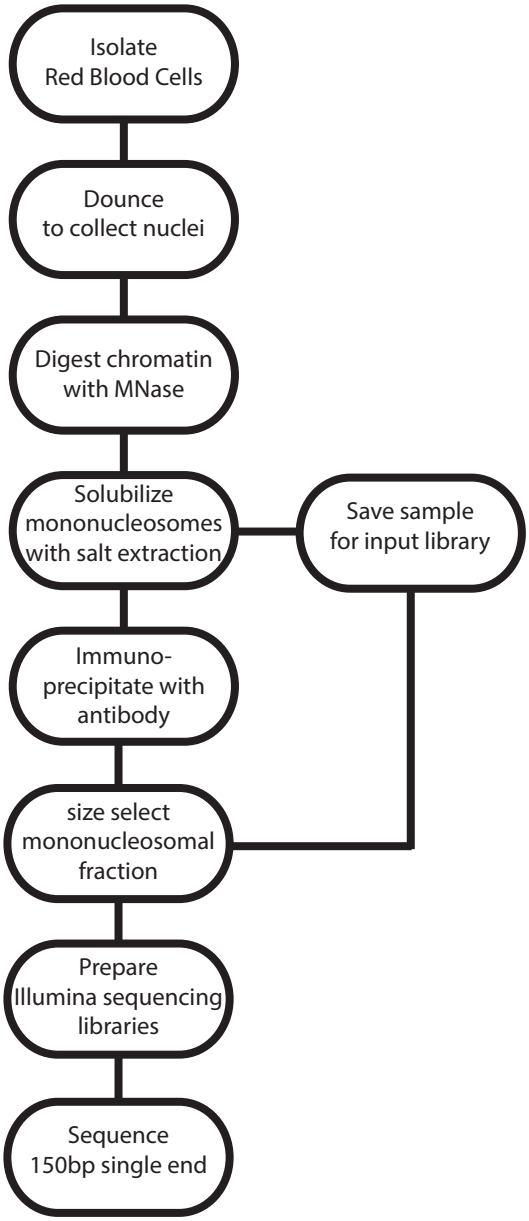
[3]Department of Genetics
Stanford University School of Medicine
Stanford, CA 94305-5120

[4]Department of Molecular and Cell Biology
University of California Berkeley
142 Life Sciences Addition #3200
Berkeley, CA 94720-3200

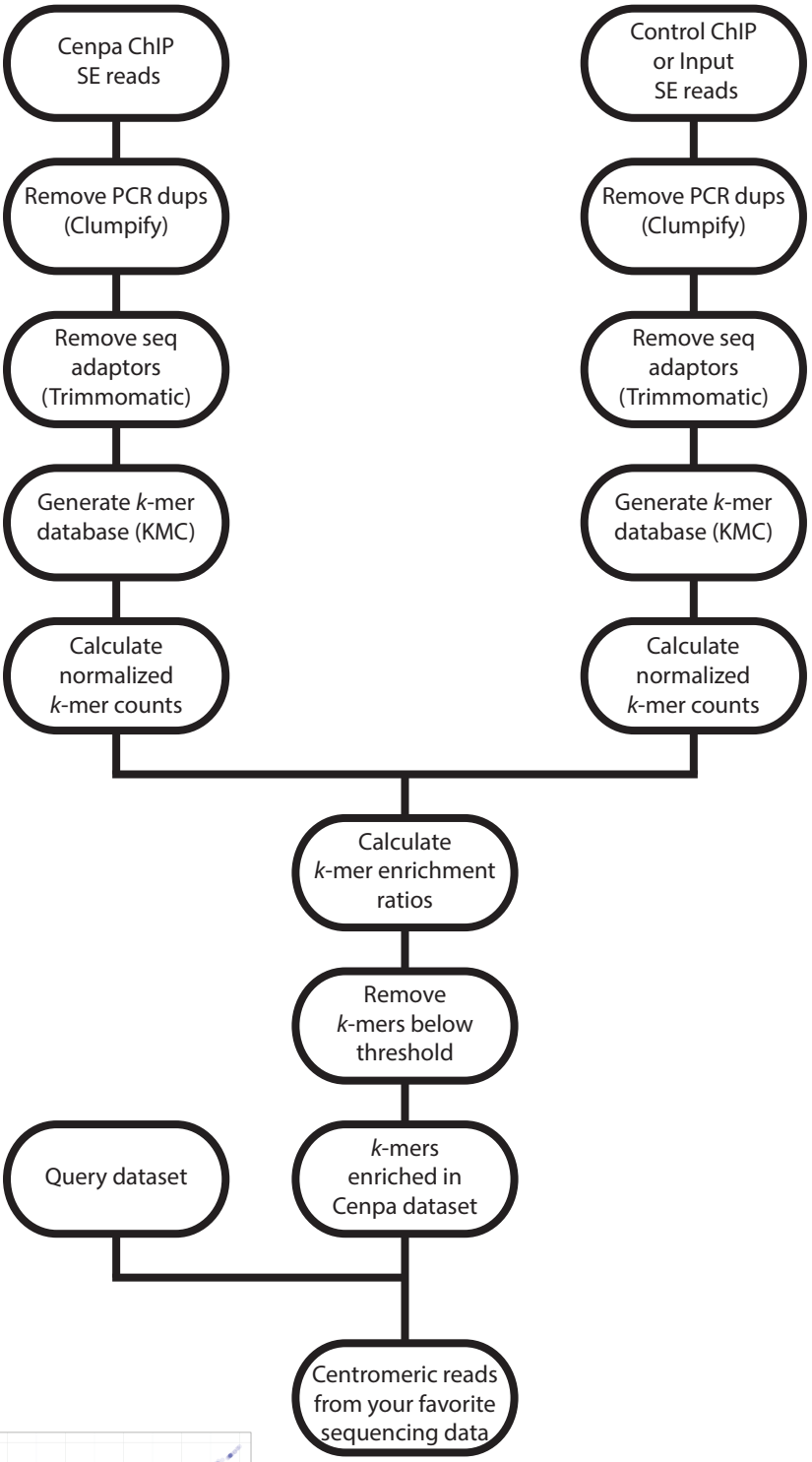
Contents:

Fig S1

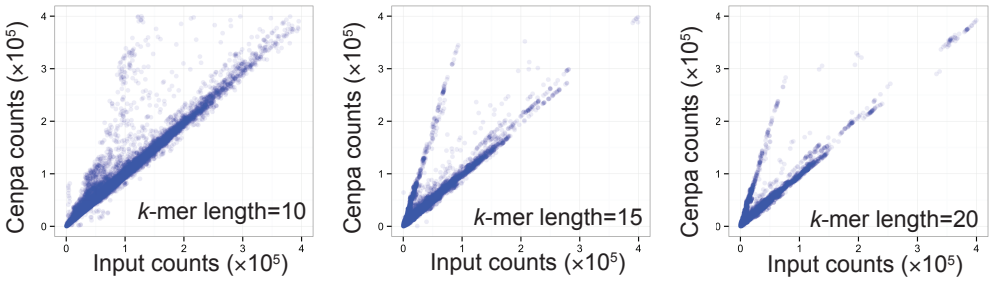
A



B



C



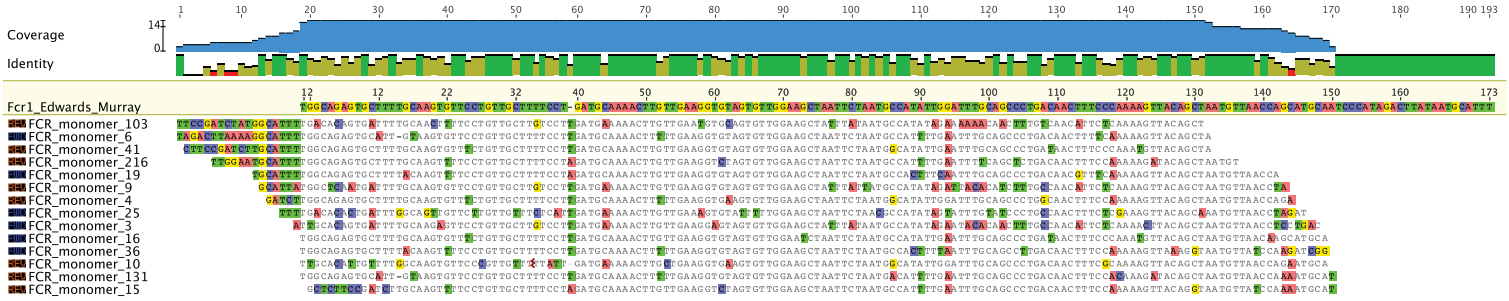
D

mad ×	% Input	% Cenpa ChIP	CA / IN
5	1.67	5.52	3.29
10	0.53	3.91	7.36
15	0.50	3.84	7.70
20	0.34	3.13	9.17

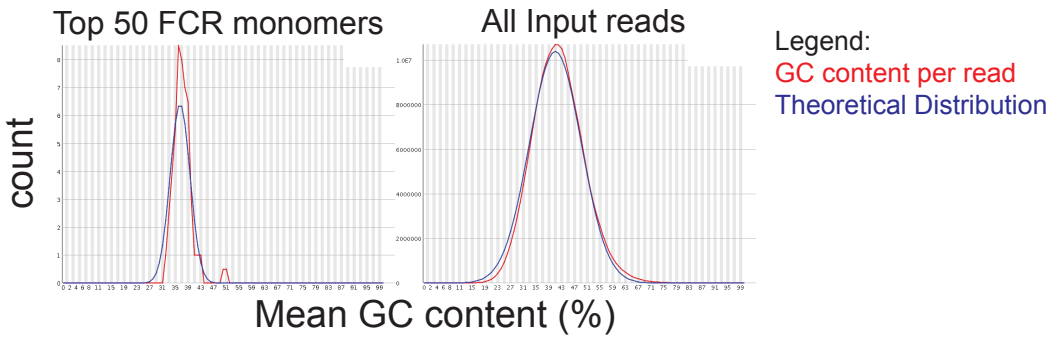
Supplemental Figure 1: A) Workflow of native MNase ChIP-seq protocol B) Overview of the pipeline to identify centromeric *k*-mers and sequences that contain centromeric *k*-mers. There are three inputs. The two at the top are reads that will be compared by *k*-mer content (i.e a ChIP dataset vs an input dataset). The third input, on the left, is a set of sequences that will be sorted by presence of centromeric *k*-mers. C) Scatter plots similar to Figure 1A. Scatter plots of normalized *k*-mer count in Cenpa ChIP (y-axis) vs input (x-axis) for *k*-mer lengths 10bp (left), 15bp (center), and 20bp (right). D) Table of the percentage of reads from Cenpa and Input sequencing libraries with increasingly stringent cutoff based on Cenpa / Input *k*-mer counting. Final column is the fold enrichment of the fraction of reads in Cenpa / Input. Multiples of median absolute deviations (MAD✕) above the median enrichment value used for defining enrichment cutoff as described for Fig 3A,B.

Fig S2

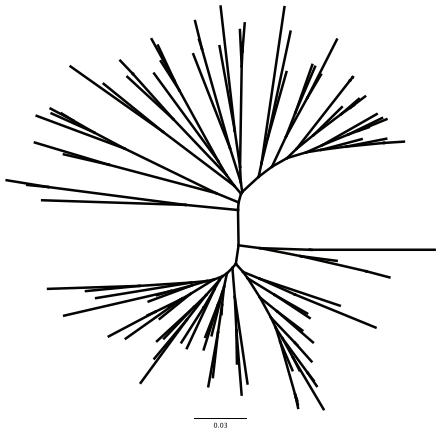
A


Coverage
Identity
Fcr1_Edwards_Murray
FCR_monomer_103
FCR_monomer_6
FCR_monomer_41
FCR_monomer_216
FCR_monomer_19
FCR_monomer_9
FCR_monomer_4
FCR_monomer_25
FCR_monomer_3
FCR_monomer_16
FCR_monomer_36
FCR_monomer_10
FCR_monomer_131
FCR_monomer_15

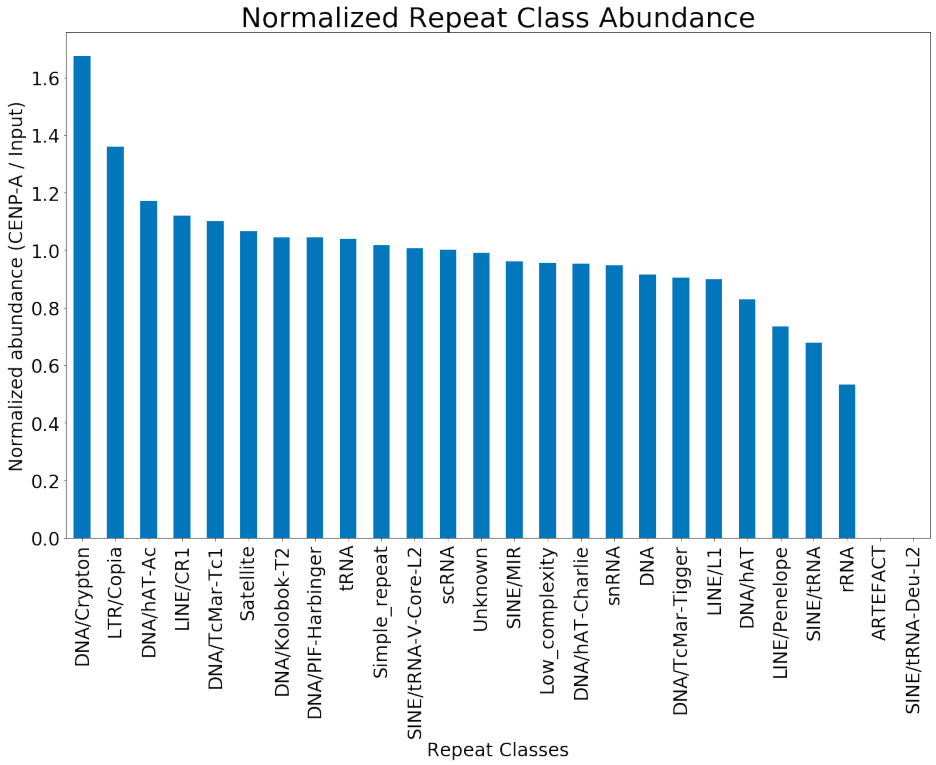
B


Top 50 FCR monomers
All Input reads
count
Mean GC content (%)
Legend:
GC content per read
Theoretical Distribution

C

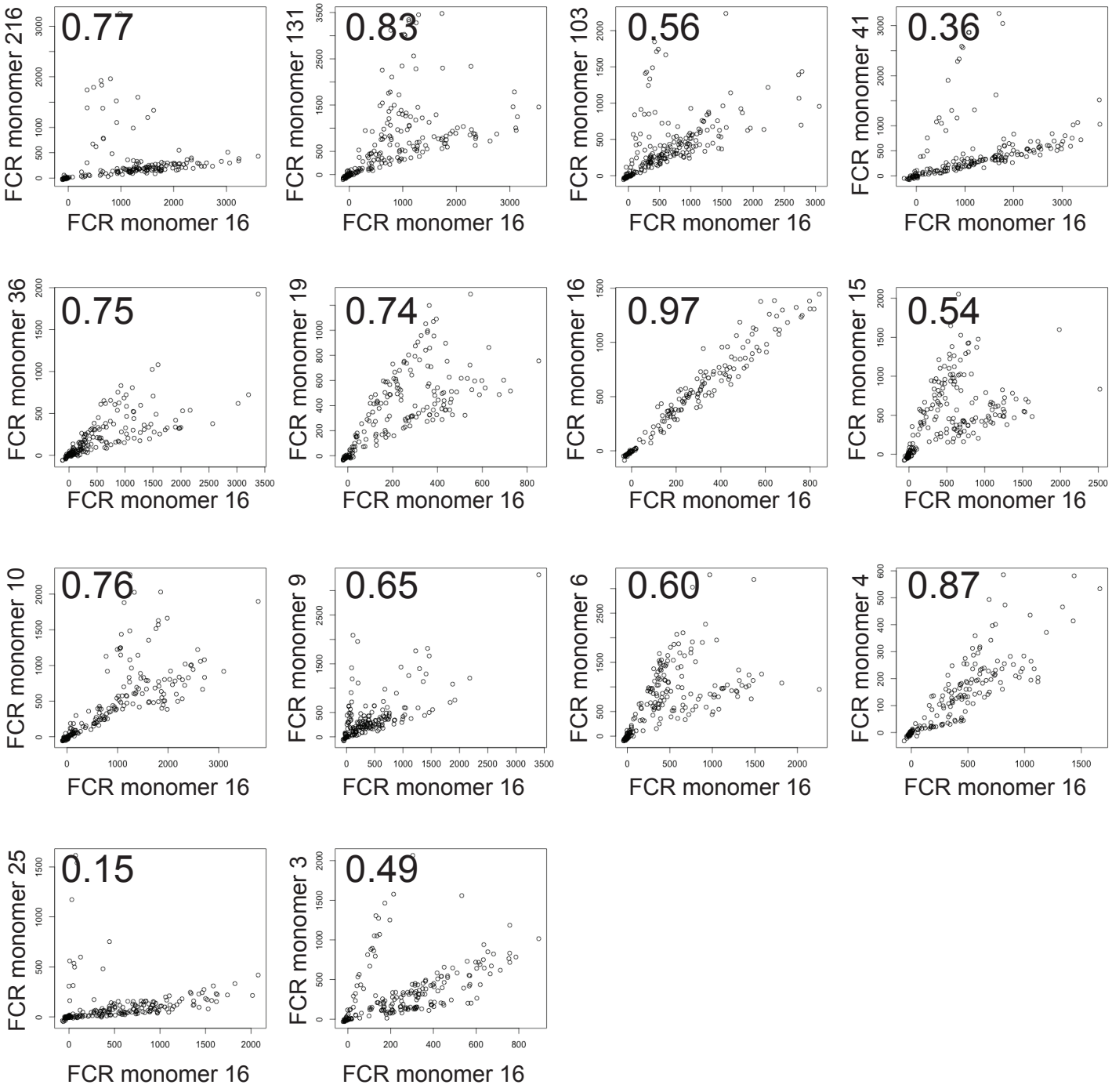


D


Normalized Repeat Class Abundance
Normalized abundance (CENP-A / Input)
Repeat Classes
DNA/Crypton
LTR/Copia
DNA/hAT-Ac
LINE/CR1
DNA/TcMar-Tc1
Satellite
DNA/Kolobok-T2
DNA/PIF-Harbinger
tRNA
Simple_repeat
SINE/tRNA-V-Core-L2
scRNA
Unknown
SINE/MIR
Low_complexity
DNA/hAT-Charlie
snRNA
DNA
DNA/TcMar-Tigger
LINE/L1
DNA/hAT
LINE/Penelope
SINE/tRNA
rRNA
ARTEFACT
SINE/tRNA-Deu-L2

Supplemental Figure 2: A) Alignment of select 150bp reads identified as FCR monomers used for FISH to the Fcr1 sequence. Colored base pairs indicate deviation from Fcr1. Coverage and percent identity of FCR monomers across the Fcr1 sequence plotted above. B) Histogram of GC-content for top 50 FCR monomers used to generate phylogram in Figure 1B (left). Histogram of GC-content for all input reads (right). GC-count per read shown in red and theoretical GC content distribution shown in blue. Plot generated using FASTQC. C) Phylogram of representative repeat monomers identified by TRF in centromeric regions from *Xenopus laevis* genome v10.2 D) Barplot of repeat classes enriched in sequencing reads from Cenpa / Input libraries. RepeatMasker was used to identify repeat classes present in 20M Cenpa and Input sequencing reads. An enrichment value for Cenpa/Input was determined for each repeat class.

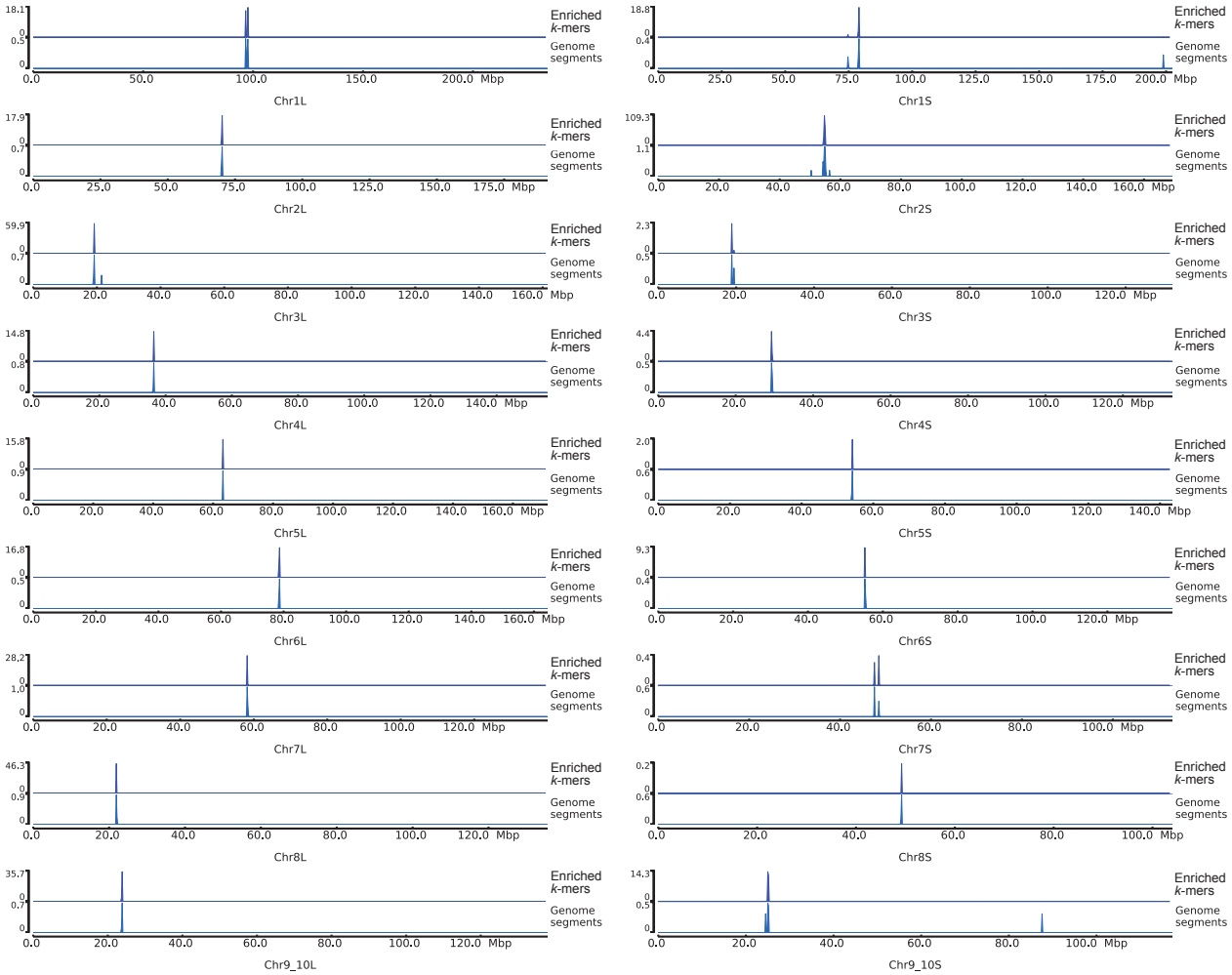
Fig S3


0.77
FCR monomer 216
FCR monomer 16
0.83
FCR monomer 131
FCR monomer 16
0.56
FCR monomer 103
FCR monomer 16
0.36
FCR monomer 41
FCR monomer 16
0.75
FCR monomer 36
FCR monomer 16
0.74
FCR monomer 19
FCR monomer 16
0.97
FCR monomer 16
FCR monomer 16
0.54
FCR monomer 15
FCR monomer 16
0.76
FCR monomer 10
FCR monomer 16
0.65
FCR monomer 9
FCR monomer 16
0.60
FCR monomer 6
FCR monomer 16
0.87
FCR monomer 4
FCR monomer 16
0.15
FCR monomer 25
FCR monomer 16
0.49
FCR monomer 3
FCR monomer 16

Supplemental Figure 3: Scatter plots similar to Figure 2A',B',C'. Scatter plots of background subtracted probe intensities for each centromere from two-color FISH experiments. Pearson coefficients are displayed in the top left corner. These scatter plots have FCR monomer 16 on the x-axis compared to other FCR monomers tested on the y-axis.

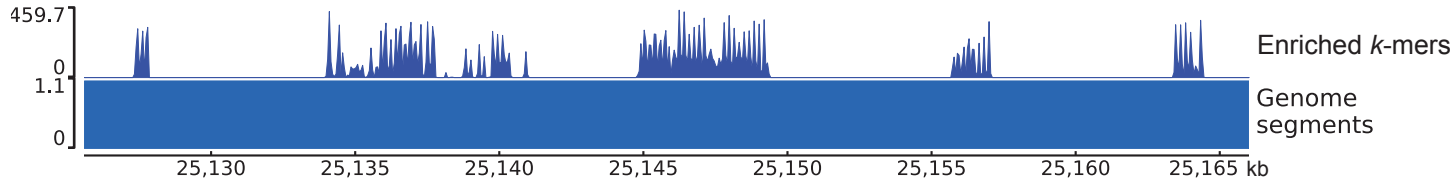
Fig S4

A



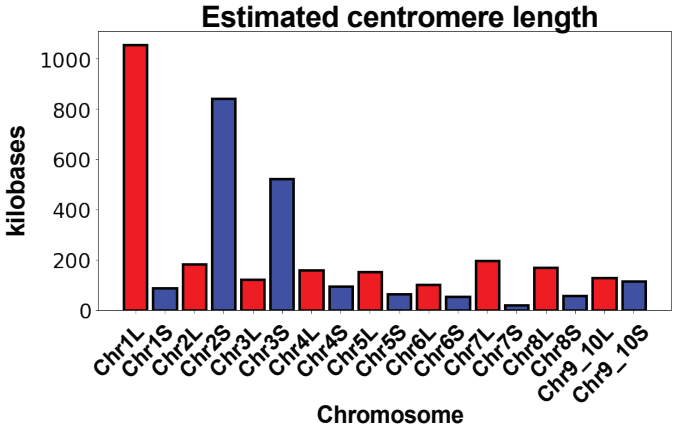
Chr1L	Chr1S	Chr2L	Chr2S	Chr3L	Chr3S	Chr4L	Chr4S	Chr5L	Chr5S	Chr6L	Chr6S	Chr7L	Chr7S	Chr8L	Chr8S	Chr9_10L	Chr9_10S
m4	m4	m103	m131	m41	m131	m4	m25	m36	m25	m16	m36	m16	m9	FCR_m19	m25	m103	m41
m41	m41	m3	m36	m16	m15	m41		m6	m19	m19	m6	m19	m41	m131	m9	m3	m19
m10	m10		m6	m4	m216	m19		m216	m41	m41	m216	m25	m16	m15		m9	m16
m16			m216		m36	m16		m15	m9	m6	m15	m41		m216			
			m15		m41			m19			m19	m9		m36			
			m19		m6									m41			
														m6			

B



Chr9_10S

C



D

Chromosome	estimated centromere length (bp)	bp with CA enriched *k*-mer	% with CA enriched *k*-mers
Chr1L	1,056,015	31,455	2.98
Chr1S	86,693	23,907	27.58
Chr2L	182,310	52,486	28.79
Chr2S	842,374	107,737	12.79
Chr3L	119,351	16,854	14.12
Chr3S	522,430	1,146	0.22
Chr4L	159,399	7,025	4.41
Chr4S	94,354	1,439	1.53
Chr5L	152,334	7,834	5.14
Chr5S	62,198	974	1.57
Chr6L	102,125	10,792	10.57
Chr6S	51,297	4,400	8.58
Chr7L	196,190	11,230	5.72
Chr7S	20,178	180	0.89
Chr8L	169,395	20,141	11.89
Chr8S	57,730	39	0.07
Chr9_10L	128,980	31,307	24.27
Chr9_10S	114,741	10,037	8.75

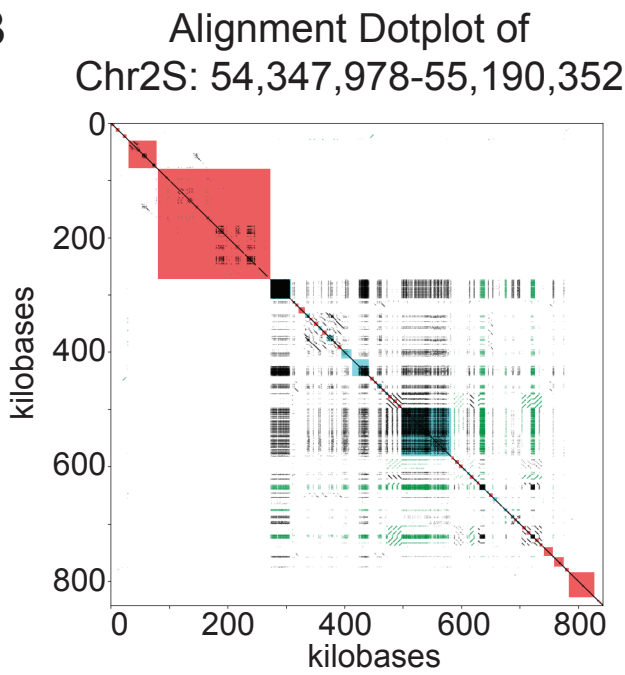
Supplemental Figure 4: A) Chromosome overviews showing the entire scaffold for each chromosome in *Xenopus laevis* genome. Top track is alignment of enriched 25bp *k*-mers, defined as 17 Median Absolute Deviations above the median of all Cenpa/Input enrichment values. Bottom track is 50kb genome segments that contain enriched *k*-mers. Most chromosomes have one location where enriched *k*-mers map. Below, table with the FCR monomers expected to be most prevalent on each chromosome based on Fig 4D. B) Expanded genome browser view of the centromeric region on Chromosome 9_10S showing that arrays of repetitive regions identified in this study are interspersed with other sequences. These intervening sequences could be repetitive or non-repetitive, but do not contain Cenpa enriched *k*-mers. C) Barplot of centromere size on each chromosome in kilobases. Exact size of the centromeric repetitive array is estimated by identifying the distance between the first and last base pairs that have a Cenpa enriched *k*-mer aligned per chromosome. These estimates include the intervening sequences between centromeric repetitive arrays shown in Fig S3B, D) Chart of the centromere length (as defined in C), base pairs with Cenpa enriched *k*-mers, and percentage of centromere that contain Cenpa enriched *k*-mers for each chromosome.

Fig S5

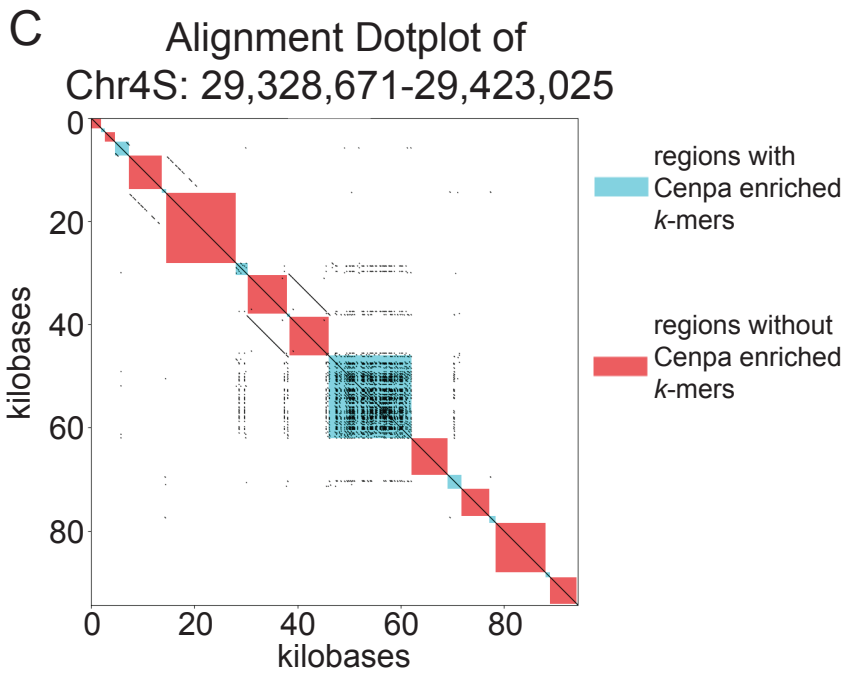
A

motif	annotation	observed	expected	fold	pvalue	qvalue
NTTCGNNNNANNCGGGN	xla_v10.2_cen	1708	2813.256	0.6073	1.00E-03	1.00E-03
NTTCGNNNNANNCGGGN	hg38_ucsc_cen	2191001	110130.436	19.8944	1.00E-03	1.00E-03

B



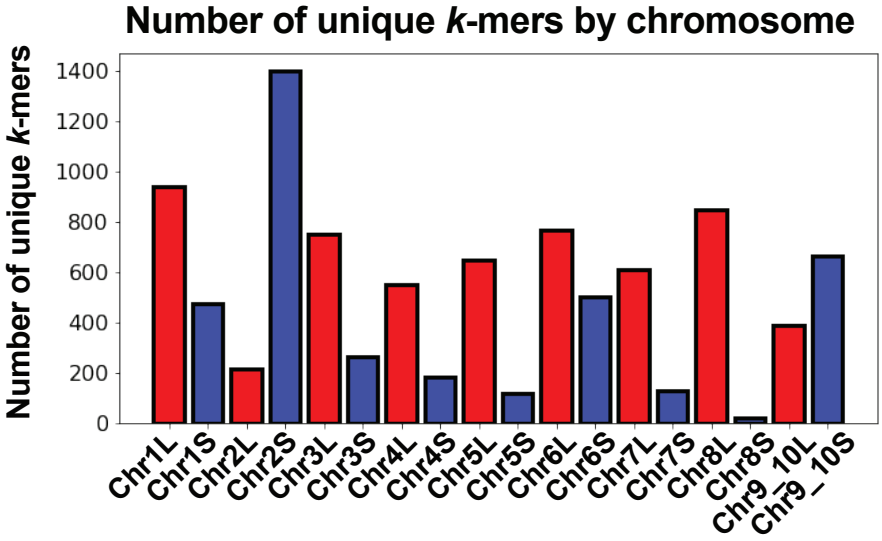
C



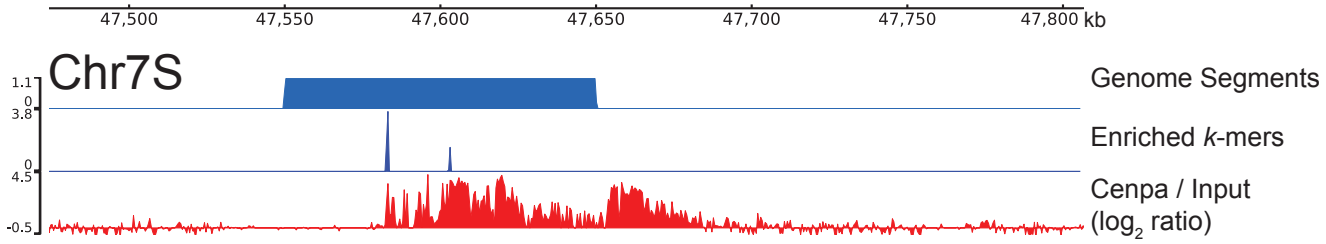
Supplemental Figure 5: A) Table showing the observed, expected, and ratio of observed/expected ("fold") for the number of occurrences of the CENPB motif within the regions defined as the centromere, based on the presence of Cenpa enriched *k*-mers (Fig S4D), in *Xenopus laevis* v10.2 genome and in human hg38 genome based on UCSC centromere annotations. B,C) Self dotplots of centromeric regions identified by Cenpa enriched *k*-mers on Chr2S (B) and Chr4S (C). Window size is 150bp. Matches are depicted black and reverse matches are depicted in green. Off-diagonal signal indicates non-unique sequence. Cyan boxes overlay regions with Cenpa enriched *k*-mers and red boxes overlay regions without Cenpa enriched *k*-mers. X and y axis coordinates are bp relative to the start of each centromeric region.

Fig S6

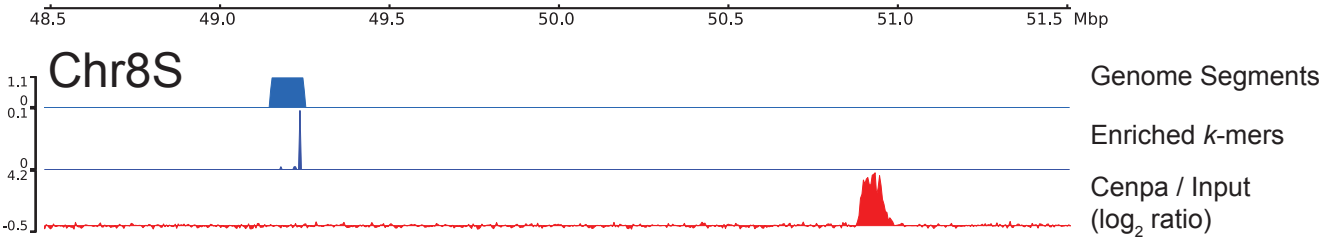
A


Number of unique k-mers by chromosome
Number of unique k-mers
1400
1200
1000
800
600
400
200
0
Chr1L
Chr1S
Chr2L
Chr2S
Chr3L
Chr3S
Chr4L
Chr4S
Chr5L
Chr5S
Chr6L
Chr6S
Chr7L
Chr7S
Chr8L
Chr8S
Chr9_10L
Chr9_10S

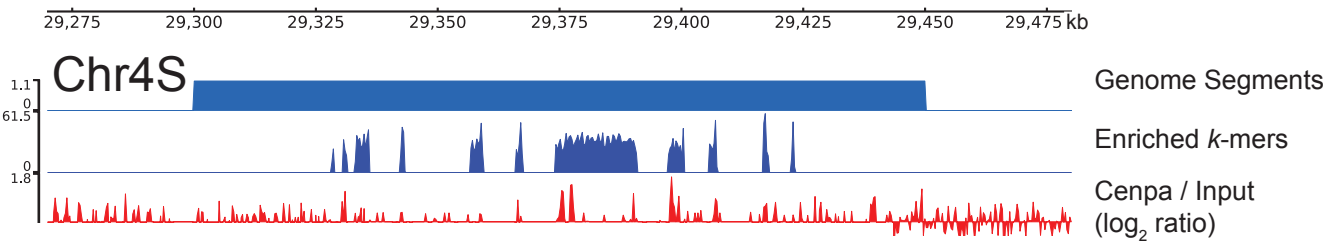
B


47,500
47,550
47,600
47,650
47,700
47,750
47,800 kb
Chr7S
Genome Segments
Enriched k-mers
Cenpa / Input
(log2 ratio)

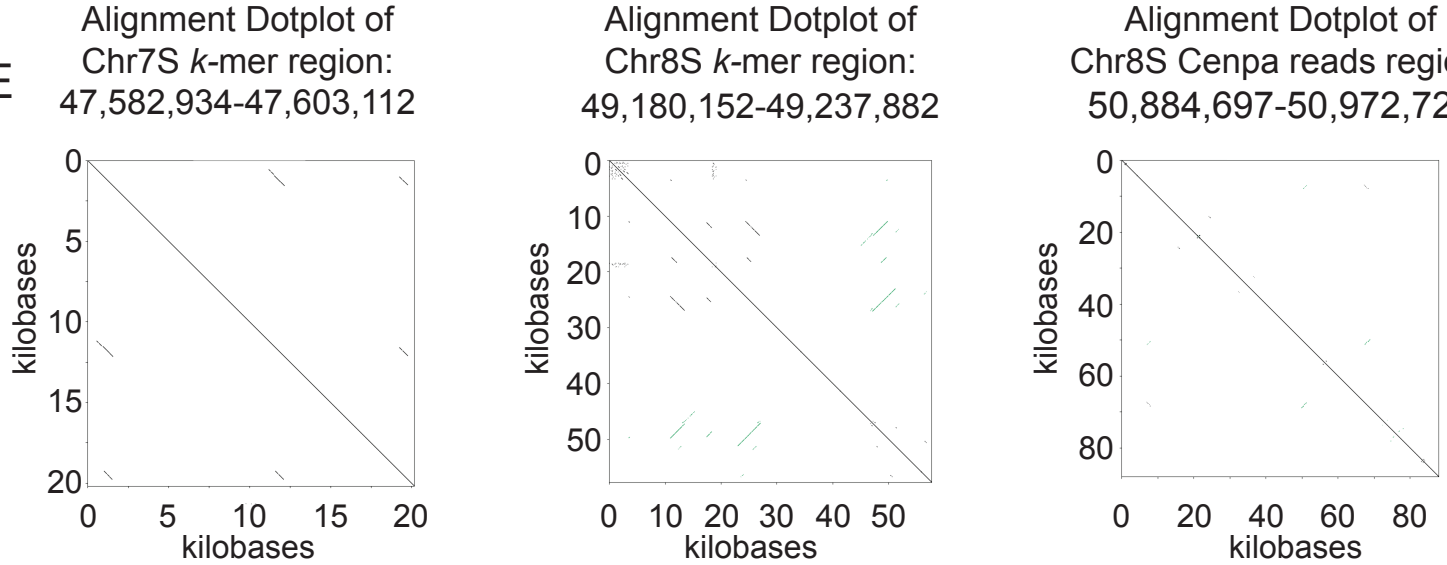
C


48.5
49.0
49.5
50.0
50.5
51.0
51.5 Mbp
Chr8S
Genome Segments
Enriched k-mers
Cenpa / Input
(log2 ratio)

D


29,275
29,300
29,325
29,350
29,375
29,400
29,425
29,450
29,475 kb
Chr4S
Genome Segments
Enriched k-mers
Cenpa / Input
(log2 ratio)

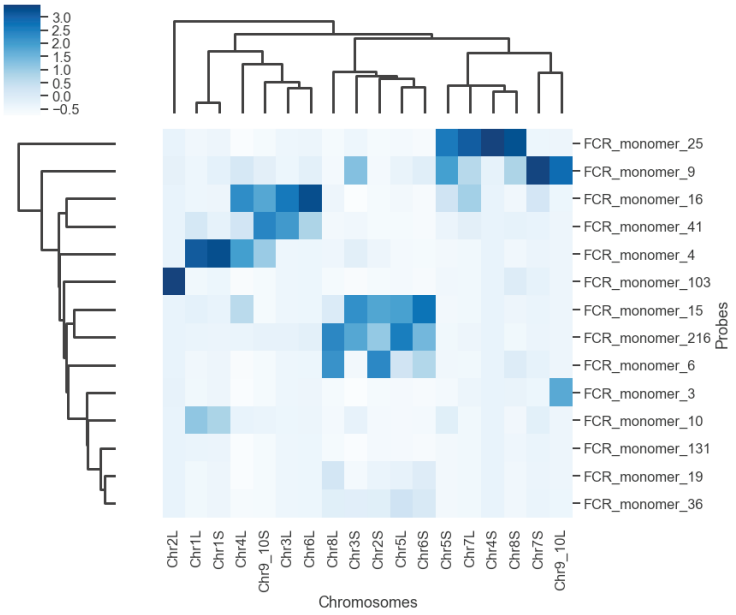
E


Alignment Dotplot of
Chr7S k-mer region:
47,582,934-47,603,112
Alignment Dotplot of
Chr8S k-mer region:
49,180,152-49,237,882
Alignment Dotplot of
Chr8S Cenpa reads region:
50,884,697-50,972,722
kilobases
kilobases
kilobases
kilobases
kilobases
kilobases

Supplemental Figure 6: A) Bar plot of the total number of unique *k*-mers found on each chromosome based on centromeric genome segments. Chromosomes from the L subgenome are shown in red and from the S subgenome are shown in blue. B, C, D) Genome browser views depicting location of genome segments containing Cenpa enriched *k*-mers (top track), alignment of Cenpa enriched *k*-mers (middle track), and a bigwig track depicting the $\log_2$ ratio of Cenpa / Input reads that were only allowed a single alignment (bottom track) for Chr7S (B), Chr8S (C), and Chr4S (D). E) Self dot plot as for Fig S5B,C with 50bp window size. Region with Cenpa enriched *k*-mers on Chr7S (left), region with Cenpa enriched *k*-mers on Chr8S (middle), and region with Cenpa enriched reads on Chr8S (right).

Fig S7

Clustered heatmap of chromosomes and FCR monomers by alignment of specific *k*-mers



Supplemental Figure 7: Heatmap similar to Figure 4C. Clustered heatmap of counts reported from Bowtie of the number of times any unique *k*-mer from each FCR monomer aligns to each chromosomal contig. Alignment counts were normalized by the number of *k*-mers unique to each FCR monomer.