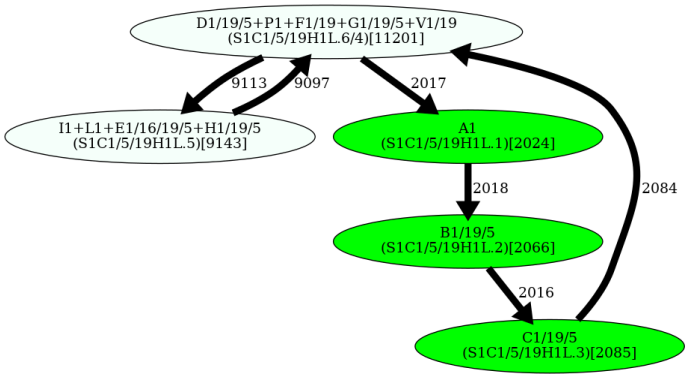
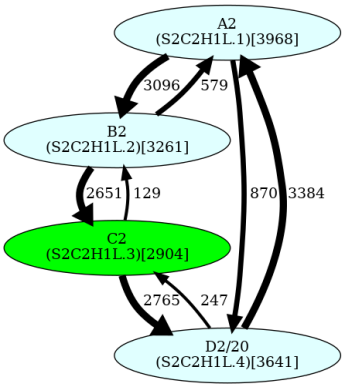
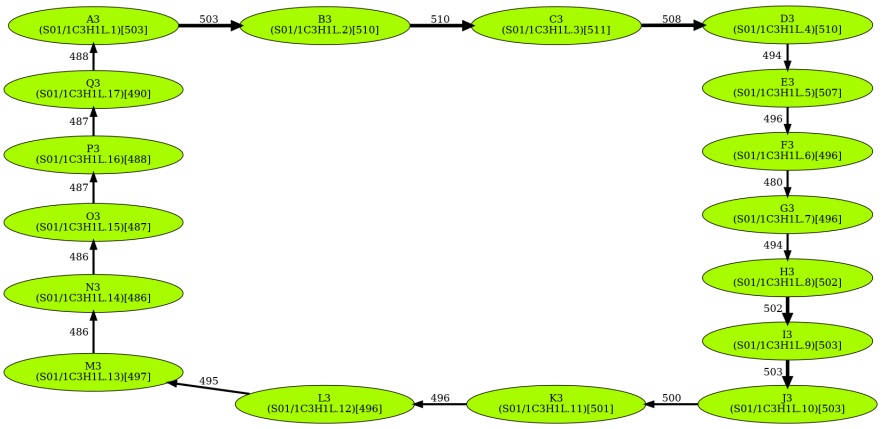
Supplementary Figures



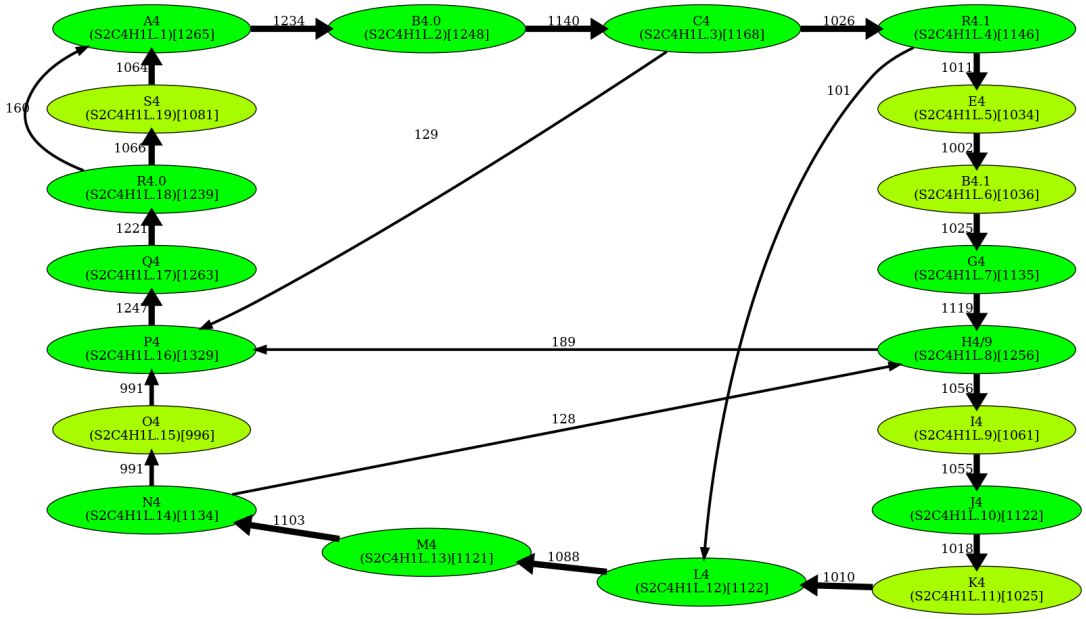
1



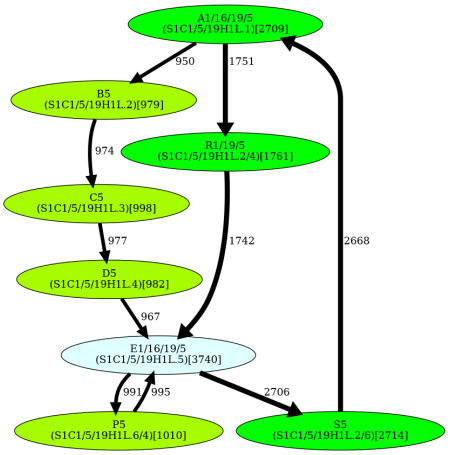
2



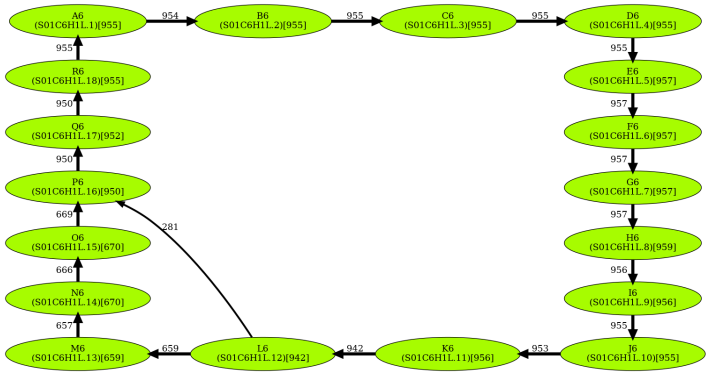
3



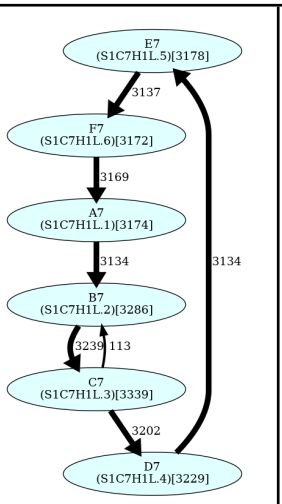
4


A1/16/19/5
(S1C1/5/19H1L.1)[2709]
950
1751
B5
(S1C1/5/19H1L.2)[979]
974
R1/19/5
(S1C1/5/19H1L.2/4)[1761]
C5
(S1C1/5/19H1L.3)[998]
977
1742
2668
D5
(S1C1/5/19H1L.4)[982]
967
E1/16/19/5
(S1C1/5/19H1L.5)[3740]
991
995
2706
P5
(S1C1/5/19H1L.6/4)[1010]
S5
(S1C1/5/19H1L.2/6)[2714]

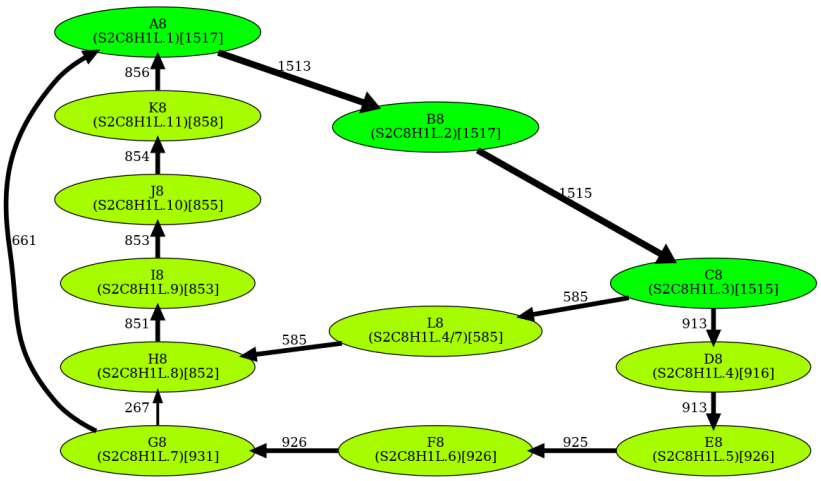
5


A6
(S01C6H1L.1)[955]
954
B6
(S01C6H1L.2)[955]
955
C6
(S01C6H1L.3)[955]
955
D6
(S01C6H1L.4)[955]
955
R6
(S01C6H1L.18)[955]
955
E6
(S01C6H1L.5)[957]
950
Q6
(S01C6H1L.17)[952]
957
F6
(S01C6H1L.6)[957]
950
P6
(S01C6H1L.16)[950]
957
G6
(S01C6H1L.7)[957]
669
281
O6
(S01C6H1L.15)[670]
957
H6
(S01C6H1L.8)[959]
666
N6
(S01C6H1L.14)[670]
956
I6
(S01C6H1L.9)[956]
657
955
M6
(S01C6H1L.13)[659]
659
L6
(S01C6H1L.12)[942]
942
K6
(S01C6H1L.11)[956]
953
J6
(S01C6H1L.10)[955]

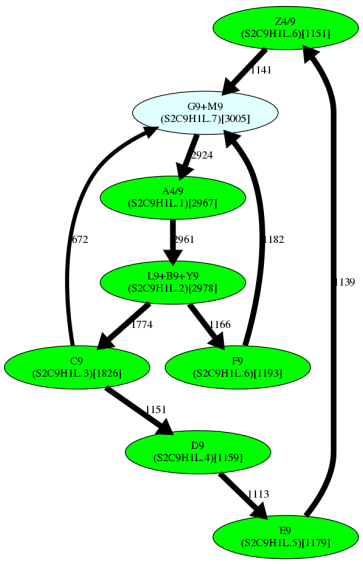
6


E7
(S1C7H1L.5)[3178]
3137
F7
(S1C7H1L.6)[3172]
3169
A7
(S1C7H1L.1)[3174]
3134
3134
B7
(S1C7H1L.2)[3286]
3239
113
C7
(S1C7H1L.3)[3339]
3202
D7
(S1C7H1L.4)[3229]

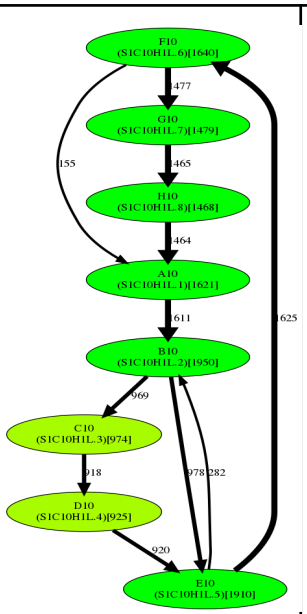
7


A8
(S2C8H1L.1)[1517]
856
1513
K8
(S2C8H1L.11)[858]
B8
(S2C8H1L.2)[1517]
854
1515
J8
(S2C8H1L.10)[855]
661
853
I8
(S2C8H1L.9)[853]
C8
(S2C8H1L.3)[1515]
585
851
L8
(S2C8H1L.4/7)[585]
913
585
H8
(S2C8H1L.8)[852]
D8
(S2C8H1L.4)[916]
267
913
G8
(S2C8H1L.7)[931]
926
F8
(S2C8H1L.6)[926]
925
E8
(S2C8H1L.5)[926]

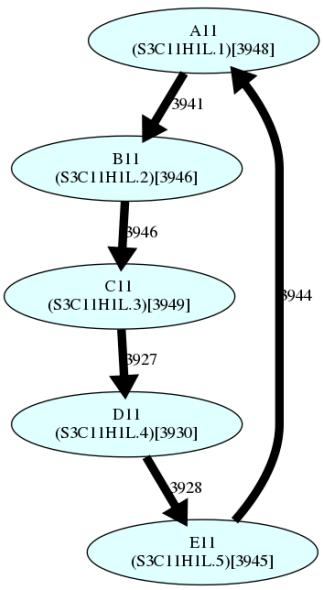
8


Z4/9
(S2C9H1L.6)[1151]
1141
G9+M9
(S2C9H1L.7)[3005]
2924
A4/9
(S2C9H1L.1)[2967]
672
2961
1182
L9+B9+Y9
(S2C9H1L.2)[2978]
1139
1774
1166
C9
(S2C9H1L.3)[1826]
F9
(S2C9H1L.6)[1193]
1151
D9
(S2C9H1L.4)[1159]
1113
E9
(S2C9H1L.5)[1179]

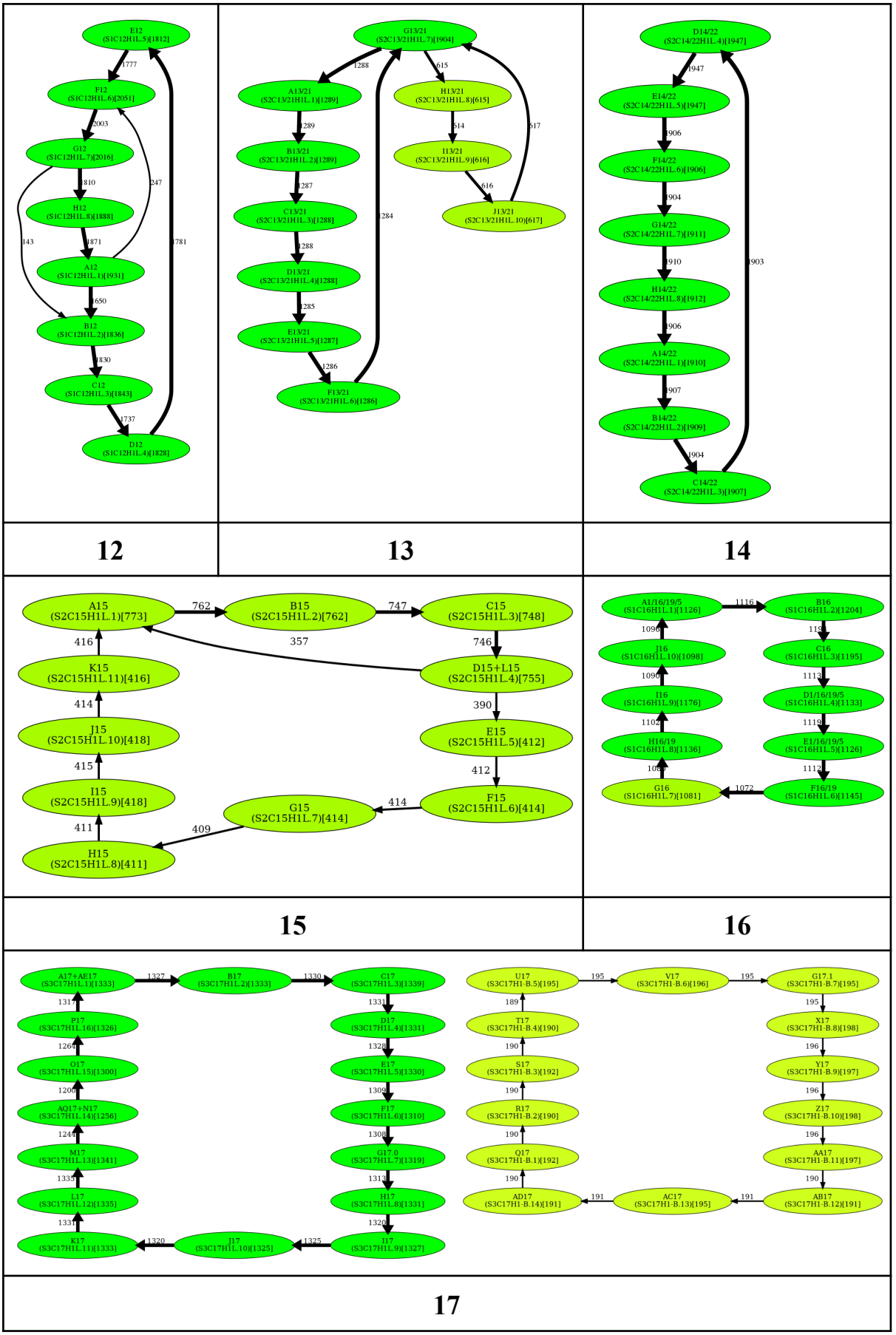
9

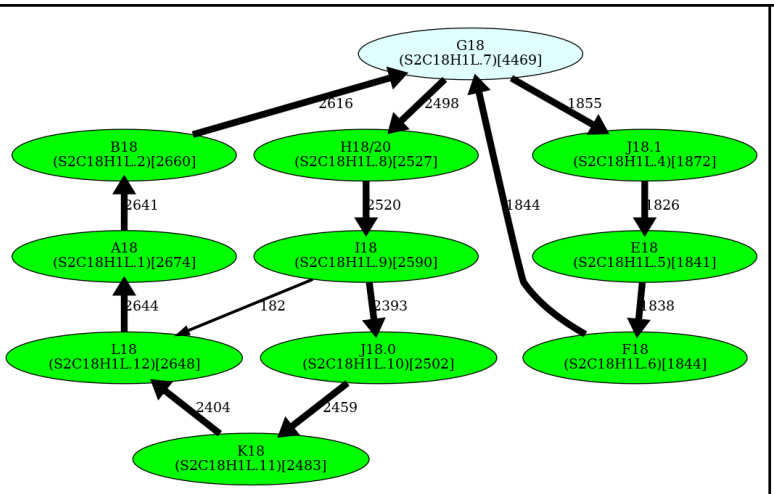

F10
(S1C10H1L.6)[1640]
477
G10
(S1C10H1L.7)[1479]
155
465
H10
(S1C10H1L.8)[1468]
464
A10
(S1C10H1L.1)[1621]
625
611
B10
(S1C10H1L.2)[1950]
969
C10
(S1C10H1L.3)[974]
978
282
918
D10
(S1C10H1L.4)[925]
920
E10
(S1C10H1L.5)[1910]

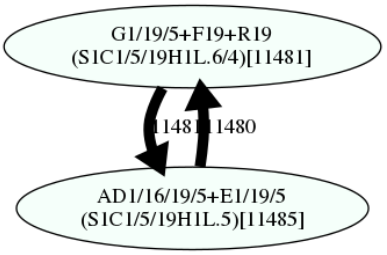
10


A11
(S3C11H1L.1)[3948]
3941
B11
(S3C11H1L.2)[3946]
3946
C11
(S3C11H1L.3)[3949]
3944
3927
D11
(S3C11H1L.4)[3930]
3928
E11
(S3C11H1L.5)[3945]

11

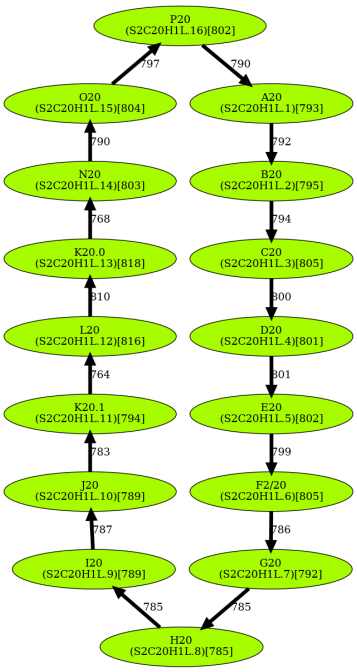

E12
(S1C12H1L.5)[1812]
1777
F12
(S1C12H1L.6)[2051]
2003
G12
(S1C12H1L.7)[2016]
1810
247
H12
(S1C12H1L.8)[1888]
143
1871
781
A12
(S1C12H1L.1)[1931]
1650
B12
(S1C12H1L.2)[1836]
1830
C12
(S1C12H1L.3)[1843]
1737
D12
(S1C12H1L.4)[1828]
12
G13/21
(S2C13/21H1L.7)[1904]
1288
615
A13/21
(S2C13/21H1L.1)[1289]
H13/21
(S2C13/21H1L.8)[615]
1289
614
617
B13/21
(S2C13/21H1L.2)[1289]
I13/21
(S2C13/21H1L.9)[616]
1287
616
C13/21
(S2C13/21H1L.3)[1288]
J13/21
(S2C13/21H1L.10)[617]
1284
1288
D13/21
(S2C13/21H1L.4)[1288]
1285
E13/21
(S2C13/21H1L.5)[1287]
1286
F13/21
(S2C13/21H1L.6)[1286]
13
D14/22
(S2C14/22H1L.4)[1947]
1947
E14/22
(S2C14/22H1L.5)[1947]
1906
F14/22
(S2C14/22H1L.6)[1906]
1904
G14/22
(S2C14/22H1L.7)[1911]
1910
1903
H14/22
(S2C14/22H1L.8)[1912]
1906
A14/22
(S2C14/22H1L.1)[1910]
1907
B14/22
(S2C14/22H1L.2)[1909]
1904
C14/22
(S2C14/22H1L.3)[1907]
14
A15
(S2C15H1L.1)[773]
762
B15
(S2C15H1L.2)[762]
747
C15
(S2C15H1L.3)[748]
416
357
746
K15
(S2C15H1L.11)[416]
D15+L15
(S2C15H1L.4)[755]
414
390
J15
(S2C15H1L.10)[418]
E15
(S2C15H1L.5)[412]
415
412
I15
(S2C15H1L.9)[418]
F15
(S2C15H1L.6)[414]
414
G15
(S2C15H1L.7)[414]
411
409
H15
(S2C15H1L.8)[411]
15
A1/16/19/5
(S1C16H1L.1)[1126]
1116
B16
(S1C16H1L.2)[1204]
1096
1115
I16
(S1C16H1L.10)[1098]
C16
(S1C16H1L.3)[1195]
1098
1113
I16
(S1C16H1L.9)[1176]
D1/16/19/5
(S1C16H1L.4)[1133]
1102
1119
H16/19
(S1C16H1L.8)[1136]
E1/16/19/5
(S1C16H1L.5)[1126]
1072
1112
G16
(S1C16H1L.7)[1081]
F16/19
(S1C16H1L.6)[1145]
16
A17+AE17
(S3C17H1L.1)[1333]
1327
B17
(S3C17H1L.2)[1333]
1330
C17
(S3C17H1L.3)[1339]
1317
1331
P17
(S3C17H1L.16)[1326]
D17
(S3C17H1L.4)[1331]
1264
1328
O17
(S3C17H1L.15)[1300]
E17
(S3C17H1L.5)[1330]
1200
1309
AQ17+N17
(S3C17H1L.14)[1256]
F17
(S3C17H1L.6)[1310]
1244
1308
M17
(S3C17H1L.13)[1341]
G17.0
(S3C17H1L.7)[1319]
1335
1312
L17
(S3C17H1L.12)[1335]
H17
(S3C17H1L.8)[1331]
1331
1320
K17
(S3C17H1L.11)[1333]
1320
J17
(S3C17H1L.10)[1325]
1325
I17
(S3C17H1L.9)[1327]
U17
(S3C17H1-B.5)[195]
195
V17
(S3C17H1-B.6)[196]
195
G17.1
(S3C17H1-B.7)[195]
189
195
T17
(S3C17H1-B.4)[190]
X17
(S3C17H1-B.8)[198]
190
196
S17
(S3C17H1-B.3)[192]
Y17
(S3C17H1-B.9)[197]
190
196
R17
(S3C17H1-B.2)[190]
Z17
(S3C17H1-B.10)[198]
190
196
Q17
(S3C17H1-B.1)[192]
AA17
(S3C17H1-B.11)[197]
190
190
AD17
(S3C17H1-B.14)[191]
191
AC17
(S3C17H1-B.13)[195]
191
AB17
(S3C17H1-B.12)[191]
17

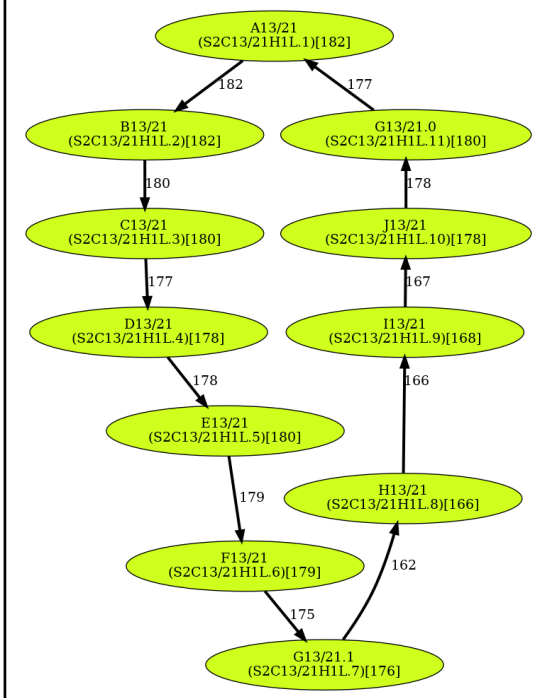

G18
(S2C18H1L.7)[4469]
2616
2498
1855
B18
(S2C18H1L.2)[2660]
H18/20
(S2C18H1L.8)[2527]
J18.1
(S2C18H1L.4)[1872]
2641
2520
1844
1826
A18
(S2C18H1L.1)[2674]
I18
(S2C18H1L.9)[2590]
E18
(S2C18H1L.5)[1841]
2644
182
2393
1838
L18
(S2C18H1L.12)[2648]
J18.0
(S2C18H1L.10)[2502]
F18
(S2C18H1L.6)[1844]
2404
2459
K18
(S2C18H1L.11)[2483]


G1/19/5+F19+R19
(S1C1/5/19H1L.6/4)[11481]
11481
11480
AD1/16/19/5+E1/19/5
(S1C1/5/19H1L.5)[11485]

18

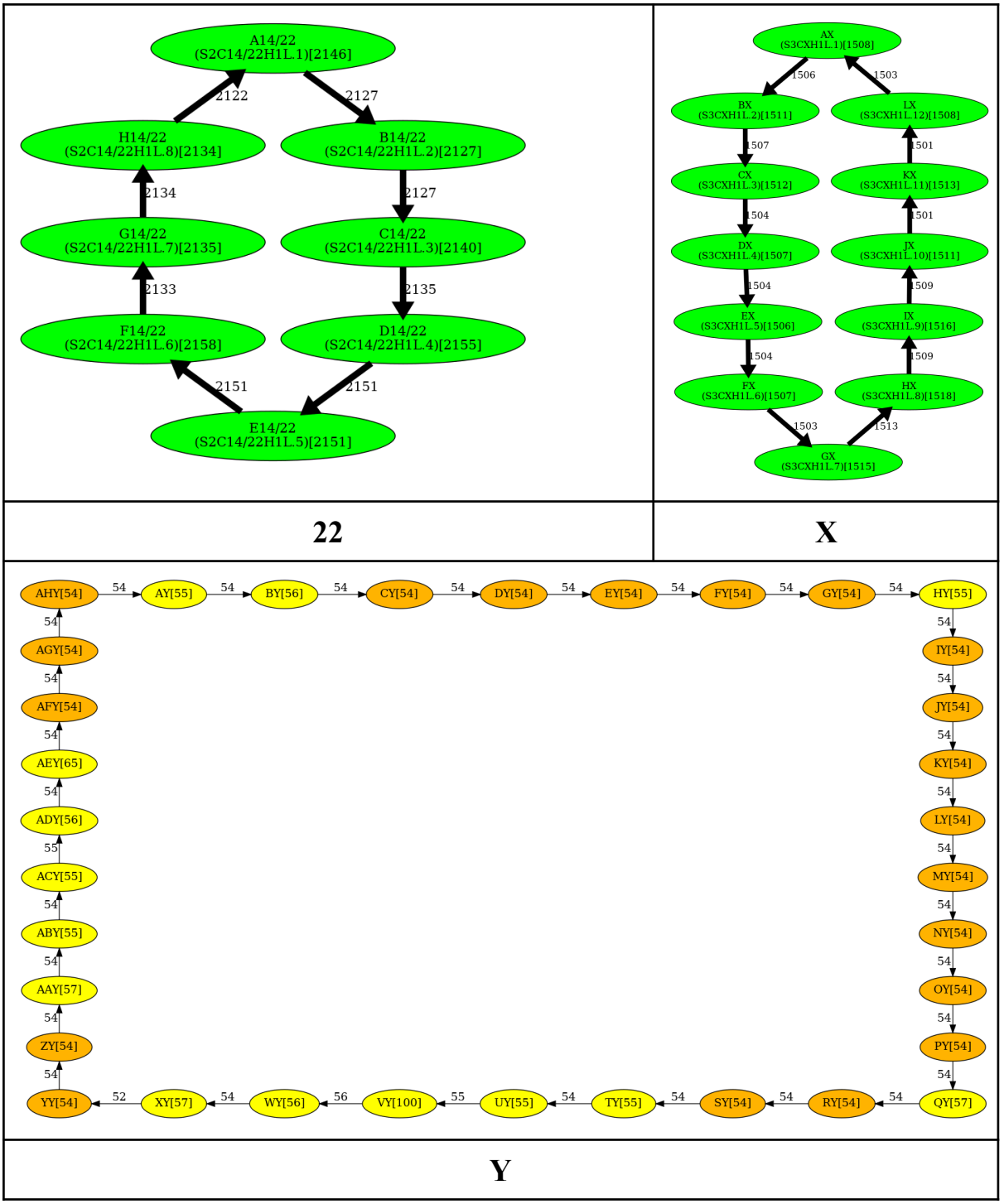
19


P20
(S2C20H1L.16)[802]
797
790
O20
(S2C20H1L.15)[804]
A20
(S2C20H1L.1)[793]
790
792
N20
(S2C20H1L.14)[803]
B20
(S2C20H1L.2)[795]
768
794
K20.0
(S2C20H1L.13)[818]
C20
(S2C20H1L.3)[805]
810
800
L20
(S2C20H1L.12)[816]
D20
(S2C20H1L.4)[801]
764
801
K20.1
(S2C20H1L.11)[794]
E20
(S2C20H1L.5)[802]
783
799
J20
(S2C20H1L.10)[789]
F2/20
(S2C20H1L.6)[805]
787
786
I20
(S2C20H1L.9)[789]
G20
(S2C20H1L.7)[792]
785
785
H20
(S2C20H1L.8)[785]


A13/21
(S2C13/21H1L.1)[182]
182
177
B13/21
(S2C13/21H1L.2)[182]
G13/21.0
(S2C13/21H1L.11)[180]
180
178
C13/21
(S2C13/21H1L.3)[180]
J13/21
(S2C13/21H1L.10)[178]
177
167
D13/21
(S2C13/21H1L.4)[178]
I13/21
(S2C13/21H1L.9)[168]
178
166
E13/21
(S2C13/21H1L.5)[180]
179
H13/21
(S2C13/21H1L.8)[166]
F13/21
(S2C13/21H1L.6)[179]
162
175
G13/21.1
(S2C13/21H1L.7)[176]

20

21



Supplemental Figure S1. The monomer-graphs constructed by HORmon for each human centromere. The label of each vertex represents the monomer ID (and the ID that is currently used by the T2T consortium that follows Uralsky et al., 2019) and its count in the monocentromere (in parentheses). The rules of monomer naming are described in the Supplementary Note 4. The label of an edge in the monomer-graph represents its multiplicity. The width of an edge (color of a vertex) reflects its multiplicity (count of a monomer).

Supplementary Tables

cen	#monomers in *MonomersNew/ MonomersFinal*	#edges in the monomer-graph	#merge/split operations	squared error distortion for *MonomersFinal*	Davies-Bouldin index for *MonomersFinal*
1	12/5	6	7/0	3.15	3.03
2	4/4	8	0/0	2.34	1.93
3	17/17	17	0/0	1.79	1.67
4	17/19	24	0/2	1.31	4.96
5	8/8	10	0/0	1.62	1.79
6	18/18	19	0/0	0.93	1.57
7	6/6	7	0/0	1.40	6.25
8	12/12	14	0/0	1.20	3.34
9	11/8	10	3/0	2.48	2.88
10	8/8	11	0/0	1.96	1.65
11	5/5	5	0/0	1.67	0.89
12	8/8	10	0/0	1.87	2.66
13	10/10	11	0/0	1.01	1.78
14	8/8	8	0/0	1.44	2.55
15	12/11	12	1/0	2.02	2.02
16	10/10	10	0/0	1.55	2.23
17	31/30	30	2/1	1.37	2.95
18	10/11	13	0/1	1.48	2.60
19	5/2	2	3/0	3.33	1.20
20	15/16	16	0/1	1.33	2.97

21	10/11	11	0/1	1.55	1.44
22	8/8	8	0/0	1.55	2.70
X	12/12	12	0/0	1.42	1.60
Y	34/34	34	0/0	2.81	2.18

Supplemental Table S1. Information about the monomer-set *MonomersFinal* for each human centromere.

The centromere ID (first column), number of monomers in the monomer-sets *MonomersNew* and *MonomersFinal* (second column), number of edges in the monomer-graph for the monomer-set *MonomersFinal* (third column), number of merging/splitting operations performed to generate the monomer-sets *MonomersFinal* (fourth column), the squared error distortion and (fifth column), and the Davies-Bouldin index (sixth column) for the monomer-set *MonomersFinal*. Rows highlighted in blue correspond to centromeres with the monomer-set that has been affected by the split-merge transformations.

cen	# monomer/ monomer-bloks	# canonical HORs/ runs of canonical HORs	# / % monomers not covered by canonical HORs	3 most frequent non-canonical HORs/ # their occurrences	length of HOR decomposition	HORs in the *MonomersT2T* alphabet
1	12/25770	1916/1824	12434/51	p4-5/2794 p4-5'/1794 I1+L1+E1/5/16/19+H1/5/19/1271	6679	S1C1/5/19H1L.123456
2	4/13450	2665/1026	2790/20	A2/571 p4-2/323 p4-1/238	2681	S2C2H1L.1234
3	17/8252	325/27	2727/33	p8-6'/69 p6-4'/65 F3'/16	138	S01/1C3H1L.123456789(10)(11)(12)(13)(14)(15)(16)(17)
4	19/21429	766/288	6875/32	H4/9/124 p1-3/116 p17-18/95	1407	S2C4H1L.123456789(10)(11)(12)(13)(14)(15)(16)(17)(18)(19)
5	8/14843	903/889	9425/63	R1/5/19/1757 p5-1/1727 P5/999	6512	S1C1/5/19H1L.123456
6	18/16249	651/181	4531/27	p16-12/225 p17-12/46 p14-11/9	417	S1C6H1L.123456789(10)(11)(12)(13)(14)(15)(16)(17)(18)
7	6/19096	3054/209	772/4	p2-4/47 C7/33 p5-2/31	472	S1C7H1L.123456
8	12/12231	559/397	6082/49	L8/585 p1-7/348 p1-3/313	1760	S2C8H1L.123456789(10)(11)
9	8/14944	925/897	8469/56	F9/1176 p7-2/1019 p7-3/624	4192	S2C9H1L.1234567
10	8/11388	820/658	4828/42	p5-2/551 E10/360 p1-2/104	1769	S1C10H1L.12345678
11	5/19668	3904/56	148/0	p5-2/10 p5-1/8 p3-1/6	108	S3C11H1L.12345
12	8/14694	1348/529	3910/26	p6-1'/186 p2-7'/94 p6-4'/93	1437	S1C12H1L.12345678

13	11/11466	601/355	4856/42	p1-7'/654 p1-6'/9 p8-11'/9	734	S2C13/21H1L.123456789(10)(11)
14	8/15170	1835/83	490/3	G14/22'/41 p4-5'/41 p1-5'/16	249	S2C14/22H1L.12345678
15	11/5767	354/307	1873/32	p1-4'/280 p9-4'/40 p1-6'/30	728	S2C15H1L.123456789(10)(11)
16	10/11367	999/165	1377/12	I16/70 p2-6/51 p8-3/41	475	S1C16H1L.123456 789(10)
17	30/23609	188/11 1040/112	4337/18	p2-16/120 p15-13/38 P17/25	409	S3C17H1-B.123456789(10)(11)(12)(13)(14) S3C17H1L.123456789(10)(11)(12)(13)(14)(15)(16)
18	12/28041	1216/430	13448/47	G18.1/623 p7-2/613 p4-2/360	2507	S2C18H1L.123456789(10)(11)(12)
19	2/21126	10217/802	692/3	AD1/5/16/19+E1/5/19/641 G1/5/19+F19+R19/51	1430	S1C1/5/19H1L.5(6/4)
20	16/12744	728/120	1096/8	p12-13/32 p14-11/8 p7-11/6	288	S2C20H1L.123456789(10)(11)(12)(13)(14)(15)(16)
21	11/1941	159/22	192/9	p10-7'/11 p1-3'/2 p4-6'/2	52	S2C13/21H1L.123456789(10)(11)
22	8/16897	1955/164	1257/7	p1-7'/49 p7-5'/47 G14/22'/35	389	S2C14/22H1L.12345678
X	12/18030	1464/54	462/2	p11-7'/6 p7-11'/5 p12-9'/5	118	S3CXH1L.123456789(10)(11)(12)
Y	34/1914	48/2	282/15	p30-32'/4 p7-25'/1 p24-32'/1	9	-

Supplemental Table S2. Information about HOR decompositions generated by HORmon for all human centromeres. Each row presents information about the HOR decomposition of a specific monocentromere. The

second column presents information about the number of monomers/monomer-blocks. The third column presents information about the number of canonical HORs/HOR-runs of canonical HORs. The fourth column presents information about the number/percentage of monomers not covered by canonical HORs. The fifth column presents information about the three most frequent non-canonical (partial or auxiliary) HORs / the number of their occurrences (HORs with an ' character refer to reverse complemented HORs for centromeres assembled in the reverse complementary strand, i.e. 12, 13, 14, 15, 21, 22, and X, and for centromere cen1 that has a reverse complementary substring). The sixth column presents information about the length of the HOR decomposition. The seventh column specifies the HORs for the corresponding monocentromere. A single HOR corresponds to each ("live") centromere, except cen17 where the sister HOR (14-mer) resides on the live array (an epiallele) in a fraction of individuals (Supplementary Note 3).

Supplementary Notes

Supplementary Note 1: Critical analysis of the CE postulate

The CE postulate is based on the assumption that a fully-formed HOR arose in the distant past but no intermediates are present in the 'live' array. It also assumes that HORs evolve exclusively via deletions with breakpoints: i) between monomers that remove an integer number of contiguous monomers, or ii) within two different monomers that generate a hybrid monomer after the deletion. However, the human genome has several chromosomes with long stretches of a single 2-monomer HOR (Chromosomes 1, 5, and 19), that are plausibly present in the process of HOR expansion via the recruitment of new monomers through transposition. There are also other chromosomes that have regions in their centromeric arrays that are consistent with ongoing HOR expansion due to the recent recruitment of transposed monomers (e.g., Chromosomes 8, 9, 13, and 18).

Thus, some centromeres are likely undergoing HOR expansions that are precluded by the CE Postulate. This constraint forces the estimated canonical HOR to sometimes include questionable monomer splits in which different monomers differ by only one or a few base pair substitutions (see subsection "What is a HOR in cen9" and "Splitting unbreakable monomers reveals HORs in cen1, cen13, and cen18."). Such highly similar monomers likely undergo ectopic recombination between them via unequal crossover. Such recombinations may make HORs unstable, raising a concern about the CE Postulate.

Since the CE Postulate remains a subject of debate, it is important to develop a tool that has both *neutral mode* (that does not presume a specific evolutionary model of HOR evolution) and the *CE postulate mode*. Even though HORmon default mode is the CE postulate mode, the user can specify the neutral mode (parameter `--CE_disable`). In that case, HORmon will not conduct any CE postulate-specific monomer operations (merging, splitting, dehybridization, splitting unbreakable monomers) and will simply filter the set *MonomersNew* of monomers inferred by CentromereArchitect.

Supplementary Note 2: HORmon terminology

Below we summarize the main terminology used in the paper.

Monomer-block — repetitive nucleotide sequence that forms repeats of a higher order (*stacked tandem repeats*).

Monomer — sequence consensus of a cluster of similar monomer-blocks.

HOR (*higher-order repeat*) — a canonical (cyclic) order of monomers specific to each centromere. It is evolutionarily defined as the ancestral and chromosome-specific order of frequent non-hybrid monomers that has evolved into the complex organization of extant centromeres.

Hybrid monomer — a monomer obtained by concatenation of a prefix of one monomer with a suffix of another.

MonomersNew — set of frequent monomers generated by CentromereArchitect, the input set of monomers for HORmon.

MonomersNew^{+} — set of monomers generated by HORmon after the split and merge operations.

MonomersFinal — the final set of monomers generated by HORmon.

MonomersT2T — set of monomers semi-manually generated by the T2T Consortium.

Centromere — nucleotide sequence of a "live" HOR array which hosts a kinetochore.

Monocentromere — sequence of monomer-blocks that represent a decomposition of *Centromere* into the input monomer-set.

*Centromere** — monocentromere obtained by decomposition of *Centromere* into the *MonomersNew* monomer-set.

*Centromere*** — monocentromere obtained by decomposition of *Centromere* into the *MonomersFinal* monomer-set.

Monomer-graph — a directed graph of a monocentromere constructed on the vertex-set of all monomers and the edge-set formed by all pairs of consecutive monomers in this centromere. The *multiplicity* of an edge (*M,M'*) in the monomer-graph is defined as the number of times the monomer *M'* follows the monomer *M* in the centromere.

Simplified monomer-graph — a graph obtained from a monomer-graph by removing all *removable* edges.

Supplementary Note 3: Information about datasets

Information about the alpha satellite arrays from the assembly (public release v1.0) of the effectively haploid CHM13 human cell line is available at (https://github.com/nanopore-wgs-consortium/chm13#v10 (NCBI accession number GCA_009914755.2). Table S3.1 presents the coordinates of the extracted regions for all "live" human centromere arrays. The alpha satellite array of the newly assembled centromere of Chromosome X

from HG002 cell line sequenced is available at https://github.com/marbl/CHM13#hg002-chromosome-x (accession number: CP074113).

Chromosome	start	end
1	121 796 218	126 300 656
2	92 333 539	94 673 018
3	91 738 494	92 596 313
3	92 869 954	92 903 597
3	95 863 962	96 415 434
4	49 705 249	50 433 651
4	52 115 581	54 870 604
4	54 980 385	55 199 889
5	47 039 130	47 049 658
5	47 077 198	49 596 620
6	58 286 939	61 058 622
7	60 414 370	63 714 496
8	44 243 543	46 325 076
9	44 952 789	47 582 587
10	39 633 785	41 664 580
11	51 061 950	54 413 485
12	34 620 831	37 202 143

13	16 220 361	18 171 058
14	10 149 798	12 766 096
15	17 263 917	18 275 855
16	35 854 534	37 793 358
17 S3C17H1-B+S3C17H1L	23 433 664	27 487 230
18	15 971 634	20 740 248
19	25 846 349	29 749 516
20	26 925 846	29 099 648
21	11 699 867	12 031 015
22	12 816 949	15 739 833
X	57 820 108	60 927 196
Y	10 562 678	10 888 333

Table S3.1 Coordinates of the alpha satellite arrays in human chromosomes. Chromosomes 3 (4; 5) contains three (three; two) alpha satellite arrays that are separated by non-monomeric regions of lengths 274 kb and 2960 kb (1682 kb and 110 kb; 27 kb). The coordinates are modified from Dvorkina et al., 2021 to include only the live HOR arrays (without sister HORs). A single exception is the S3C117H1L (D17Z1) array where the sister HOR S3C17H1-B (D17Z1-B) was included. This sister HOR has been shown to represent the live array (an epiallele) in a fraction of individuals (McNulty and Sullivan, 2018).

Supplementary Note 4: HORmon monomer naming

Below we describe the HORmon rules for monomer naming. Table S4.1 provides information about the monomer naming generated by HORmon, the classical naming, and naming in Altemose

et al., 2022.

- Each monomer from the monomer-set *MonomersNew,* that is not affected by heuristics inspired by the CE Postulate (split/merge transformations or a dehybridization), is named as $\mathbf{SC_1/../C_n}$, where the identifier **S** is typically a letter from A to Z (if there are more than 26 monomers, HORmon names them using 2-letter strings from AA to ZZ), and $\mathbf{C_1/../C_n}$ is the list of chromosomes, where this monomer was found by CentromereArchitect. For example, **A1** is a monomer A found in cen1, **CX** is a monomer C from cenX, and **R1/5/19** is a monomer R found in three centromeres 1, 5, and 19. No two monomers within a single centromere can share the same identifier.

- Hybrid monomers have notation $\mathbf{SC_1/../C_n(M_1/M_2)}$. The notation $\mathbf{SC_1/../C_n}$ is described above, and M_1 and M_2 are monomers that constitute the hybrid monomer **S**. For example, **NX(K/J)** is a hybrid monomer N from cenX, constructed as a concatenate of a prefix of the monomer KX with a suffix of the monomer JX. In order to describe a hybrid monomer $\mathbf{SC_1/../C_n(M_1/M_2)}$ in more detail, we occasionally represent it as $\mathbf{M_1C_1/../C_n}(r)\mathbf{/M_2C_1/../C_k}(l)$ to show that **S** was constructed as a concatenate of monomers $\mathbf{M_1C_1/../C_n}$ and $\mathbf{M_2C_1/../C_k}$ using the prefix of $\mathbf{M_1C_1/../C_n}$ of length *r* and the suffix of $\mathbf{M_2C_1/../C_k}$ of length *l*. Note that Dvorkina et al., 2021 used "+" sign instead of "/" to denote hybrid. We decided to reassign the hybrid operation to "/" and reserve "+" for the merge operation.

- Monomers obtained by a merge operation are represented as a sum of merged monomers, i.e. $\mathbf{S_1C_1/../C_n + .. + S_mC_1/../C_n}$. For example, monomer **I1+L1+E1/5/16/19+H1/5/19** is a result of the merge operation of four initial monomers **I1**, **L1**, **E1/5/16/19**, and **H1/5/19**.

Monomers obtained by a split operation of a single monomer $\mathbf{SC_1/../C_n}$ are represented as $\mathbf{SC_1/../C_n.0}$ and $\mathbf{SC_1/../C_n.1}$. For example, the initial monomer H18 was split into monomers **H18.0** and **H18.1**.

cen	classical HOR name	Altemose et al., 2022 HOR name	HOR length (bp)	Altemose et al., 2022 monomer name	HORmon monomer name
1	D1Z7	S1C1/5/19H1L	1019	1	A1
				2	B1/5/19
				3	C1/5/19
				4	D1/19/5+P1+F1/19+G1/5/19+V1/19.1
				5	I1+L1+E1/5/16/19+H1/5/19
				6	D1/5/19+P1+F1/19+G1/5/19+V1/19.0
2	D2Z1	S2C2H1L	680	1	A2
				2	B2
				3	C2
				4	D2/20
3	D3Z1	S01/1C3H1L	2891	1	A3
				2	B3
				3	C3
				4	D3
				5	E3
				6	F3
				7	G3
				8	H3
				9	I3
				10	J3
				11	K3
				12	L3
				13	M3
				14	N3
				15	O3
				16	P3
				17	R3
4	D4Z1	S2C4H1L	3232	1	A4
				2	B4.0
				3	C4
				4	R4.1
				5	E4

				6	B4.1
				7	G4
				8	H4/9
				9	I4
				10	J4
				11	K4
				12	L4
				13	M4
				14	N4
				15	O4
				16	P4
				17	Q4
				18	R4.0
				19	S4
5	D5Z2	S1C1/5/19H1L	1019/1020	1	A1/5/16/19
				2	B5
				3	C5
				4	D5
				5	E1/5/16/19
				6	F5
6	D6Z1	S01C6H1L	3057	1	A6
				2	B6
				3	C6
				4	D6
				5	E6
				6	F6
				7	G6
				8	H6
				9	I6
				10	J6
				11	K6
				12	L6
				13	M6
				14	N6
				15	O6
				16	P6

				17	R6
7	D7Z1	S1C7H1L	1022	1	A7
				2	B7
				3	C7
				4	D7
				5	E7
				6	F7
8	D8Z2	S2C8H1L	1868	1	A8
				2	B8
				3	C8
				4	D8
				5	E8
				6	F8
				7	G8
				8	H8
				9	I8
				10	J8
				11	K8
9	D9Z4	S2C9H1L	1194/1192	1	A4/9
				2	L9+B9+Y9
				3	C9
				4	D9
				5	E9
				6	Z4/9
				7	G9+M9
10	D10Z1	S1C10H1L	1357	1	A10
				2	B10
				3	C10
				4	D10
				5	E10
				6	F10
				7	G10
				8	H10
11	D11Z1	S3C11H1L	850	1	A11
				2	B11

				3	C11
				4	D11
				5	E11
12	D12Z3	S1C12H1L	1359	1	A12
				2	B12
				3	C12
				4	D12
				5	E12
				6	F12
				7	G12
				8	H12
13	D13Z1	S2C13/21H1L	1870	1	A13/21
				2	B13/21
				3	C13/21
				4	D13/21
				5	E13/21
				6	F13/21
				7	G13/21.0
				8	H13/21
				9	I13/21
				10	J13/21
				11	G13/21.1
14	D14Z9	S2C14/22H1L	1364	1	A14/22
				2	B14/22
				3	C14/22
				4	D14/22
				5	E14/22
				6	F14/22
				7	G14/22
				8	H14/22
15	D15Z3	S2C15H1L	1877	1	A15
				2	B15
				3	C15
				4	D15+L15
				5	E15

				6	F15
				7	G15
				8	H15
				9	I15
				10	J15
				11	K15
16	D16Z2	S1C16H1L	1699	1	A1/5/16/19
				2	B16
				3	C16
				4	D1/5/16/19
				5	E1/5/16/19
				6	F16/19
				7	G16
				8	H16/19
				9	I16
				10	J16
17	D17Z1	S3C17H1L	2715	1	A17+AE17
				2	B17
				3	C17
				4	D17
				5	E17
				6	F17
				7	G17.0
				8	H17
				9	I17
				10	J17
				11	K17
				12	L17
				13	M17
				14	AQ17+N17
				15	O17
				16	P17
17	D17Z1B	S3C17H1-B	2379	1	Q17
				2	R17
				3	S17
				4	T17

				5	U17
				6	V17
				7	G17.1
				8	X17
				9	Y17
				10	Z17
				11	AA17
				12	AB17
				13	AC17
				14	AD17
18	D18Z1	S2C18H1L	2035	1	A18
				2	B18
				3	G18.0
				4	J18.1
				5	E18
				6	F18
				7	G18.1
				8	H18/20
				9	I18
				10	J18.0
				11	K18
				12	L18
19	D19Z3	S1C1/5/19H1L	340	5	AD1/5/16/19+E1/5/19
				6/4	G1/5/19+F19+R19
20	D20Z2	S2C20H1L	2719	1	A20
				2	B20
				3	C20
				4	D20
				5	E20
				6	F2/20
				7	G20
				8	H20
				9	I20
				10	J20
				11	K20.1
				12	L20

				13	K20.0
				14	N20
				15	O20
				16	P20
21	D21Z1	S2C13/21H1L	1870	1	A13/21
				2	B13/21
				3	C13/21
				4	D13/21
				5	E13/21
				6	F13/21
				7	G13/21.1
				8	H13/21
				9	I13/21
				10	J13/21
				11	G13/21.0
22	D22Z1	S2C14/22H1L	1364/1365	1	A14/22
				2	B14/22
				3	C14/22
				4	D14/22
				5	E14/22
				6	F14/22
				7	G14/22
				8	H14/22
X	DXZ1	S3CXH1L	2057	1	AX
				2	BX
				3	CX
				4	DX
				5	EX
				6	FX
				7	GX
				8	HX
				9	IX
				10	JX
				11	KX
				12	LX

Table S4.1. Information about the monomer naming generated by HORmon, the classical naming, and

naming in Altemose et al., 2022. Each row corresponds to a frequent non-hybrid monomer for a specific centromere. The first column corresponds to the centromere. The second column corresponds to the classical HOR naming based on Alexandrov et al., 2001, Shepelev et al., 2015, and McNulty and Sullivan 2018. The third column corresponds to the HOR naming in Altemose et al., 2022. The fourth column shows the length of the nucleotide consensus of HOR. The fifth column corresponds to the monomer numbering with respect to the rules specified in Uralsky et al., 2019 (see also Altemose et al., 2022). The sixth column provides the monomer names generated by HORmon. In centromeres 1, 2, 5, and 15, we report a different number of monomers than McNulty and Sullivan 2018. We hypothesize that these minor differences are due to the absence of a complete genome assembly in prior studies.

Supplementary Note 5: Evaluating the monomer-sets

Since HORmon utilizes the monomer-set constructed by CentromereArchitect, it is important to compare this monomer-set with the manually-derived monomer-sets. Below we show that the monomer-set automatically constructed by CentromereArchitect marginally improves on the currently known (manually constructed) monomer-set.

Given a monomer-set *Monomers,* we partition a centromere *Centromere* into monomer-blocks *Blocks*. We define the *average radius* of a monomer M (denoted as $r(M)$) as the average distance between M and all M-blocks in *Blocks*. Given strings *S'* and *S"*, we denote the edit distances between them as *distance*(*S',S"*).

From the clustering perspective, the monomer-set represents the *centers* of the *data points* formed by the monomer-blocks. We use the *squared error distortion* and the *Davies-Bouldin* index (Davies and Bouldin, 1979) to evaluate the clustering quality. The squared error distortion is defined as follows:

$$distortion(Monomers, Blocks) =$$

$$1/|Blocks| * \sum_{\text{each monomer } M \text{ in } Monomers} \sum_{\text{each } M\text{-block } Block \text{ in } Blocks} distance(M, Block)^2.$$

The *Davies-Bouldin* index is defined as follows:

$$DBI(Monomers,Blocks) =$$

$$1/|Monomers| * \sum_{\text{each monomer } M} \max_{\text{each monomer } M' \neq M} (r(M)+r(M')) / distance(M, M').$$

The *count* of a monomer *M* (referred to as *count*(*M*)=*count*(*M,Centromere**)) is the number of its occurrences in the monocentromere *Centromere**. HORmon orders all monomers by their decreasing counts and refers to the *i*-th most frequent monomer in *Centromere** as M_i. Given the set $Monomers_i = \{M_1, \ldots M_i\}$ of *i* most frequent monomers in *Centromere*,* we define $count(Monomers_i)$ as the total count of all monomers in this monomer-set. We identify the *imin* as the minimum value of *i* such that $count(Monomers_i)$ exceeds the threshold $MinFraction \cdot |Centromere^*|$, where $|Centromere^*|$ is the length of the monocentromere (the default value *MinFraction* = 0.9). HORmon constructs the set of *frequent* monomers as $Monomers_{imin}$ complemented by monomers M_{imin+1}, M_{imin+2}, ... with counts exceeding $MinExtension \cdot count(M_{imin})$ (the default value *MinExtension* = 0.7). The resulting human monomer-set is referred to as *MonomersNew*.

The sets *MonomersNew* turned out to contain more monomers than the set *MonomersT2T* for all centromeres except for centromeres 3, 4, 10, 13, 17, 18, 20, and 21. Since two clustering solutions of the same set of data points are usually compared for the case when these solutions have the same number of centers, we attempted to select a subset of monomers from the set *MonomersNew* to make it comparable with the set *MonomersT2T* (for each centromere). For each monomer in *MonomersT2T*, we thus identified the closest monomer in *MonomersNew* and constructed the monomer-set *MonomersNew** of the same size as *MonomersT2T* (in this case, we define *MonomersT2T** = *MonomersT2T*). Similarly, if the set *MonomersT2T* contains more monomers than the set *MonomersNew*, for each monomer in *MonomersNew,* we identified the closest monomer in *MonomersT2T* and constructed the monomer-set *MonomersT2T** of the same size as *MonomersNew* (in this case, we define *MonomersNew* = MonomersNew)*.

Table S5.1 compares the monomer-sets *MonomersT2T** and *MonomersNew**. To ensure that this comparison is adequate (i.e., compares two equally-sized sets of centers for the same data points), we define the set of monomer-blocks *SharedBlocks* that are shared between both monomer-sets. For each monomer-block *B'* (in the centromere decomposition into monomers from *MonomersNew**), and an overlapping monomer-block *B"* (in the centromere decomposition into monomers from

*MonomersT2T**), a new block *B* is formed by taking the overlap between *B'* and *B''*. We add *B* to the set *SharedBlocks* if its length exceeds the threshold *MinSharedLength* (the default value *MinSharedLength* = 150). Since the block *B* in *SharedBlocks* is typically shorter than blocks *B'* or *B''*, we modify the definition of the distance between *B* and any monomer *M* as the minimum distance between *B* and all substrings of the *M*-consensus, where the *M-consensus* is defined as the consensus of the multiple alignment of all *M*-blocks.

Table S5.1 illustrates that the set *MonomersNew** results in a marginally better clustering of monomer-blocks (with respect to the squared error distortion) than the set *MonomersT2T** (*MonomersT2T** resulted in a lower squared error distortion for centromeres 15, 16, 20, and 21). Both sets demonstrate similar performance with respect to the Davies-Bouldin index.

In order to compare the set *MonomersNew* generated by CentromereArchitect to the final monomer-set *MonomersFinal* generated by HORmon after merging/splitting and hybrid monomer decomposition, we similarly define the monomer-sets *MonomersNew** and *MonomersFinal* of the same size. Table S5.2 illustrates that these sets result in a comparable clustering of monomer-blocks with respect to the analyzed clustering metrics. Specifically, the Davies-Bouldin index (squared error distortion) for *MonomersFinal** does not exceed the same metric for *MonomersNew** for all centromeres except centromeres 4, 18, 20, 21 (1, 18, 19). These results are not surprising since the CE Postulate-guided monomer transformations that HORmon conducts over the set *MonomersNew* are not coordinated with the objective clustering metrics, but are rather dictated by the biological model of a canonical HOR as an ancestral unit.

cen	**# monomers in *MonomersT2T/ MonomersNew****	**# blocks for *MonomersT2T* / MonomersNew* / SharedBlocks***	**squared error distortion for *MonomersT2T* / MonomersNew****	**Davies-Bouldin index for *MonomersT2T* / MonomersNew****	**# monomers reported by Centromere Architect**	**# monomers in *MonomersT2T***	**# monomers in *MonomersNew***
1	12	26504/26504/2648	2.37/**1.57**	**4.22**/7.07	23	15	12

		6					
2	4	13744/13744/13744	**2.33/2.33**	1.96/**1.95**	10	6	4
3	17	8485/8485/8485	1.82/**1.77**	**1.42**/1.44	24	19	17
4	17	21715/21715/21711	1.72/**1.65**	**2.56**/2.70	25	19	17
5	8	14893/14893/14893	1.69/**1.62**	**1.73**/1.80	19	14	8
6	18	16315/16315/16313	**0.87/0.87**	1.49/**1.47**	19	18	18
7	6	19375/19375/19373	**1.34/1.34**	**4.23/4.23**	17	6	6
8	12	12247/12247/12243	**1.03/1.03**	**1.19/1.19**	12	16	12
9	10	15456/15456/15455	3.51/**2.15**	**2.72**/3.36	26	10	11
10	8	11967/11967/11967	**1.95/1.95**	**1.67**/1.69	36	19	8
11	5	19718/19718/19718	1.70/**1.67**	**0.89/0.89**	14	6	5
12	8	15204/15204/15204	**1.86/1.86**	2.72/**2.71**	21	8	8
13	10	11478/11478/11476	1.01/**0.92**	**1.01**/1.05	14	11	10
14	8	15349/15349/15349	1.44/**1.43**	**2.55**/2.57	14	8	8
15	11	5941/5941/5941	**1.76**/1.87	**2.11**/2.19	16	11	12
16	10	11393/11393/11391	**0.99**/1.47	**1.45**/1.65	16	10	10
17	16	23849/23849/23846	**2.68/2.68**	**1.96/1.96**	43	16	31

18	10	28110/28110/28110	1.65/**1.59**	2.28/**2.11**	20	13	10
19	5	22964/22964/22963	2.87/**2.82**	11.36/**10.81**	31	12	5
20	15	12793/12793/12786	**1.48**/1.53	**2.46**/2.53	17	16	15
21	10	1948/1948/1948	**1.26**/1.67	**1.03**/1.11	14	11	10
22	8	17146/17146/17146	**1.54/1.54**	2.85/**2.76**	13	8	8
X	12	18095/18095/18095	**1.41/1.41**	**1.58/1.58**	14	17	12

Table S5.1. Comparison of the monomer-set identified by CentromereArchitect (Dvorkina et al., 2021) with the previously inferred monomer-set (Shepelev et al., 2015, Uralsky et al., 2019). The centromere ID (first column), the number of monomers in the monomer-sets *MonomersT2T** and *MonomersNew** (second column), # blocks for *MonomersT2T**/*MonomersNew**/*SharedBlocks* (third column), the squared error distortion for the monomer-sets *MonomersT2T** and *MonomersNew** (fourth column), the Davies-Bouldin index for the monomer-sets *MonomersT2T** and *MonomersNew** (fifth columns), the number of monomers in the monomer-set identified by CentromereArchitect (sixth column), the number of monomers in the monomer-set *MonomersT2T* that were inferred by Uralsky et al., 2019 (seventh column), the number of monomers in the monomer-set *MonomersNew* that represent frequent monomers in the CentromereArchitect output (eight column).

cen	**# monomers in *MonomersNew*/ MonomersFinal****	**# blocks for *MonomersNew*/ MonomersFinal*/ SharedBlocks**	**squared error distortion for *MonomersNew/ MonomersFinal****	**Davies-Bouldin index for *MonomersNew* / MonomersFinal****	**# monomers in *MonomersFinal***
1	6	26504/26504/26486	**2.55**/2.60	3.44/**3.15**	6
4	19	21715/21715/21711	1.55/**1.25**	**2.63**/3.05	19
9	8	15456/15456/15456	2.70/**2.49**	2.91/**2.89**	8
15	11	5941/5941/5941	1.87/**1.77**	2.19/**2.07**	11

17	30	23849/23849/23846	1.37/**1.34**	2.34/**2.07**	30
18	12	28110/28110/28110	**1.39**/1.46	**2.32**/4.58	12
19	2	22964/22964/22963	**3.28**/3.32	**1.22/1.22**	2
20	16	12793/12793/12786	1.52/**1.29**	**2.56**/2.83	16
21	11	1948/1948/1948	1.67/**1.55**	**1.08**/1.30	11

Table S5.2. Comparison of the monomer-set identified by CentromereArchitect (Dvorkina et al., 2021) with the monomer-set generated by HORmon. The centromere ID (first column), the number of monomers in the monomer-sets *MonomersNew** and *MonomersFinal** (second column), # blocks for *MonomersNew**/*MonomersFinal**/*SharedBlocks* (third column), the squared error distortion for the monomer-sets *MonomersNew** and *MonomersFinal** (fourth column), the Davies-Bouldin index for the monomer-sets *MonomersNew** and *MonomersFinal** (fifth columns), the number of monomers in the monomer-set identified by HORmon (sixth column). Since the monomer-sets for centromeres 2, 3, 5, 6, 7, 8, 10, 11, 12, 13, 14, 16, 22, and X were not affected by HOR-guided split-merge transformation, we omit the corresponding rows.

Supplementary Note 6: HORmon parameters

Selecting HORmon parameters is a complex challenge since only a single human genome has been assembled up to date. Moreover, the choice of "ground truth" to benchmark HORmon is limited since (1) the concept of a HOR is computationally poorly defined, (2) the CE Postulate has not been statistically assessed yet, and (3) the nucleotide sequences of the manually extracted HORs have been generated decades ago at the dawn of the sequencing era. Below we describe the rationale behind selecting HORmon parameters and limitations for their selection.

- *Generating the monomer-set.* HORmon transforms the monomer-set generated by CentromereArchitect into the set *MonomersNew.* This transformation relies on parameters *MinFraction* (default value = 0.9) and *MinExtension* (default value = 0.7) that were chosen by analyzing the counts of monomers in the set monomer-set generated by CentromereArchitect. Filtering out infrequent monomers was straightforward: even using a single parameter

MinFraction (formally, setting *MinExtension* = 1 under the assumption that no two monomers share the same count) was sufficient to generate a reasonable monomer-set *MonomersNew* for most centromeres. However, in rare cases, we observed that HORmon filters out monomers that are included in the manually constructed monomer-sets. To ensure comparability with previous studies on all centromeres, we introduced an additional parameter *MinExtension* that allowed us to include these monomers, otherwise, filtered monomers in the monomer-set *MonomersNew*.

- *Split and merge transformations.* Splitting a monomer relies on a single parameter *splitValue* (default value 1/8). Merging monomers relies on parameters *minPI* (default value 94%) and *minPosSim* (default value 0.4). Since the length of all monomer-blocks is close to 171 bp, *minPI* corresponds to the maximum edit distance 10 between two monomers. In selecting all these parameters, we tried to ensure that the resulting monomer graphs are topologically close to "a cycle with a few chords". Typically, a lower (higher) value of *splitValue* ensures more relaxed (strict) conditions for splitting monomers. Setting an extreme value *splitValue* = 0 results in splitting each monomer with in-degree n and out-degree m into $n \cdot m$ monomers. On the other hand, setting an extreme value *splitValue* = 1 does not alter the initial monomer-set. A similar rationale is applicable to parameters *minPI* and *minPosSim*. Extensive benchmarking on additional assemblies, once they are available, is required to rule out potential overfit.

- *Monomer graphs*. Construction of the monomer-graphs relies on parameters the *MinEdgeMultiplicity* (default value 100) and *minCountFraction* (default value 0.9). It is possible to eliminate these parameters (formally, select *MinEdgeMultiplicity* = *minCountFraction* = 0) and include edges of all multiplicities into the monomer-graph. However, since the resulting monomer graphs often contain many low multiplicity edges, the included low-multiplicity edges will complicate the analysis of the main "trends" in the centromere architecture (Figure 3). Introduction of both parameters rather than a single parameter *MinEdgeMultiplicity* is necessary for shorter centromeres. For example, cen21,

which is only 331,148 bp long, contains only ~2000 monomer-blocks. Selection of these parameters does not substantially affect HORmon's ability to extract HORs from monomer-graphs, and is mostly needed to ensure reasonable filtering of low multiplicity edges.

- *Comparison of monomer-sets.* The selection of a single parameter *MinSharedLength* (default value 150 bp) does not substantially affect the comparison of the monomer-sets. Typically, a higher (lower) value of *MinSharedLength* results in a stricter (looser) selection of blocks for the set *SharedBlocks*.

Supplementary Note 7: Comparison of centromere decomposition generated by HORmon and traditional approaches

Previous approaches to centromere annotation and inference typically defined a HOR as a linear string of monomers. In order to do that, these approaches implicitly selected a certain monomer in the HOR as its *first* monomer. Then any "structural variant" of a HOR can be reported as a certain "modification" of the string corresponding to such HOR. Note that we do not rigorously define a "structural variant" of a HOR as it is challenging and limit ourselves to the following example. Let a HOR consist of monomers ABCDE and monomer A is chosen as the first monomer. Then this HOR can be rewritten as 12345. If a toy *Monocentromere* is

ABCDEABDEABDEABCDEABC

(or 12345124512451245123 45123)

the traditional way to decompose *Monocentromere* into HORs is

… 12345 — 1245 — 1245 — 12345 — 123 …

where 1245 is often referred to as a "structural variant" of the HOR.

Arguably a more natural model of a HOR is not a linear but a cyclic string of monomers that does not have a notion of a "first" monomer. In HORmon notation, a HOR (as a cyclic string) corresponds to a cycle in the monomer-graph. Thus, HORmon does not rely on the notion of a first monomer, and decomposes the above monocentromere as

… ABCDE — AB — DEAB — DEABC — DEABC …

HORmon presents this decomposition concisely as $c_A\, p_{AB}\, p_{DB}\, c_D^2$ that represents a traversal through a monomer-graph. Each symbol corresponds to a partial (for $p_{\cdot}$ — partial HORs) or a complete (for $c_{\cdot}$ — canonical HORs) traversal through the cycle in the monomer-graph that corresponds to the HOR ABCDE. The transitions between the consecutive symbols correspond to edges in this graph. For example, the transition from the partial HOR p_{AB} to the partial HOR p_{DB} corresponds to an edge connecting the monomer B to the monomer D in the monomer-graph. We argue that such an approach to describe the structure of a centromere is more natural than the traditional approaches. Indeed, one does not need to rigorously define "structural variants" of HORs and select a first monomer. However, in case the choice of the first monomer is dictated by prior evolutionary knowledge, the traditional approach might be preferred.

Supplementary Note 8: Generating the nucleotide consensus of a HOR

Given a pair of consecutive monomers M_i and M_{i+1} in a t-monomer (cyclic) HOR $H = M_1, ..., M_t$, HORmon identifies all pairs formed by the consecutive M_i-block and M_{i+1}-block in the centromere (we assume that $M_{t+1} = M_1$). It further extracts all nucleotide strings formed by these pairs of monomer-blocks, constructs their multiple alignment, and computes its consensus C_i. Ideally, the resulting strings C_i and C_{i+1} should perfectly overlap over the monomer M_{i+1}. However, in practice, a short suffix of C_i and a short prefix of C_{i+1} may suffer from a somewhat lower nucleotide accuracy due to artifacts of multiple alignment. To overcome this issue, HORmon constructs an

overlap alignment of the suffix of C_i and prefix of C_{i+1} using the fast Edlib tool (Šošić and Šikić, 2017). We denote the ending (starting) coordinate in C_i (C_{i+1}) of the *longest match* in the constructed alignment as $right_i$ ($left_{i+1}$). HORmon concatenates the prefix of C_i ending at position $right_i$ with the suffix of C_{i+1} starting at position $left_{i+1}$ resulting in a more accurate circular nucleotide consensus of *H*. Since *M*-consensus for each monomer *M* is constructed as a multiple alignment of all *M*-blocks, its short prefix and suffix can also suffer from a lower nucleotide accuracy. Thus, concatenation of M_1-consensus, M_2-consensus, …, M_t-consensus does not necessarily coincide with the nucleotide consensus of that HOR *H*. HORmon improves the *M*-consensus for each monomer *M* by launching StringDecomposer on monomers M_1 ,..., M_t and the constructed consensus of *H*. This partitioning ensures that the concatenate of all *M*-consensus coincides with the nucleotide sequence of *H*.

Supplementary Note 9: Running time and memory footprint

All benchmarking was done on a server node with 16 Intel Xeon X7560 2.27 GHz CPUs. We launched HORmon individually on each human centromere (Table S9.1). Since CentromereArchitect is the most time consuming stage of the HORmon pipeline, we benchmarked it separately. Running HORmon on all human centromeres (as presented in the main text) simultaneously is a more time intensive task. Note that since we planned to run HORmon only once for each centromere, its time and memory optimization was not the top priority. We plan to speed-up HORmon in preparation for analysis of multiple human genomes that are now being generated by the HPR Consortium.

CenID	**Time CentromereArchitect (hh:mm:ss)**	**Time HORmon (w/o CentromereArchitect; hh:mm:ss)**	**Memory HORmon (Gb)**
1	1:37:25	47:10	5.8
2	19:21	07:10	2.9

3	09:29	06:14	5.7
4	40:30	26:12	6.4
5	20:44	08:55	2.9
6	21:59	16:48	4.6
7	01:07:27	23:20	3.7
8	12:48	08:05	3.1
9	29:25	11:46	6.1
10	29:32	12:41	8.6
11	22:20	12:39	3.1
12	36:25	09:49	5.0
13	18:43	22:07	3.3
14	28:22	12:28	3.6
15	07:04	04:24	4.1
16	28:10	09:02	5.7
17	01:14:07	35:31	11.4
18	01:00:39	41:35	5.7
19	02:28:12	01:02:59	6.1
20	15:24	11:06	3.9
21	01:03	01:35	2.5

22	24:36	12:03	3.9
X	18:24	12:29	4.6

Table S9.1. HORmon running time and memory footprint. HORmon was launched on each human centromere separately. The first column corresponds to the centromere ID. The second column shows the running time of CentromereArchitect,the most time consuming stage of the HORmon pipeline. The third column shows the running time of the remaining stages of HORmon. The last column shows the memory footprint of the entire HORmon pipeline.

Supplementary Note 10: Annotation of centromeres in *Arabidopsis thaliana*

Naish et al., 2021 presented the first assembly of *Arabidopsis thaliana* that resolved all five centromeres. Below we use HORmon to provide initial insights into the monomeric and HOR structure of assembled centromeres. Sequences of satellite arrays were extracted from https://github.com/schatzlab/Col-CEN/tree/main/v1.2, file "Col-CEN_v1.2.fasta.gz". As an input, CentromereArchitect (the prerequisite for HORmon) utilized the consensus monomer *Consensus* (180 bp long) that is reported in Naish et al., 2021:

```
AGTATAAGAACTTAAACCGCAACCCGATCTTAAAAGCCTAAGTAGTGTTTCCTTGTTAGA
AGACACAAAGCCAAAGACTCATATGGACTTTGGCTACACCATGAAAGCTTTGAGAAGCA
AGAAGAAGGTTGGTTAGTGTTTTGGAGTCGAATATGACTTGATGTCATGTGTATGATTG
```

Table S10.1 contains the coordinates of centromeres on each chromosome. These coordinates define the longest contiguous segment of a chromosome with all monomer-blocks sharing the same strands and with identity to *Consensus* exceeding *MinBlockIdentity* (the default value *MinBlockIdentity* = 70%, extracted using StringDecomposer, Dvorkina et al., 2020). To simplify the analysis, we ran CentromereArchitect and HORmon on each centromere separately. For consistency of notation with the analysis of human centromeres, we refer to the set of frequent monomers generated by

CentromereArchitect as *MonomersNew*, and to the final set of monomers generated by HORmon as *MonomersFinal*.

Table S10.2 summarizes information about monomers extracted from centromeres using HORmon. Table S10.3 presents the monomer graphs built on the monomer-sets *MonomersNew* and *MonomersFinal.*

The putative HOR in centromere 1 consists of a dimer A1+B1+C1+D1+F1+G1+H1+I1+J1+M1+N1+P1 (refer to as X1 for brevity), K1. However, tandem occurrences of X1 and K1 are quite numerous: X1 (K1) is followed by X1 (K1) in 3765 (833) cases while X1 (K1) is followed by K1 (X1) in 2226 (2225) cases (Table S10.2).

The putative HOR in centromere 2 also consists of a dimer. A very frequent monomer (multiplicity 5060, compared to 941 and 874 for individual monomers within a dimer respectively) is a hybrid of A and B. Monomer graph for *MonomersFinal* of centromere 2 (Table S10.2) shows frequent tandem occurrences of that hybrid monomer — monomer C follows itself 4058 times.

The monomer graph for centromere 3 consists of a single monomer and the monomer graph for centromere 4 and 5 have a complex topology that prevents extraction of HORs. Even though the CE Postulate implies that all monomers within the HOR are different, it is unclear whether the CE Postulate holds for *Arabidopsis thaliana*.

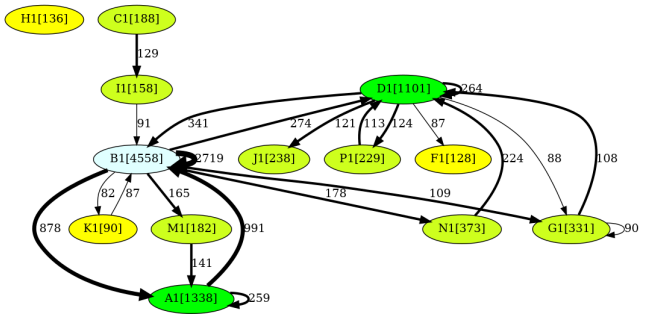
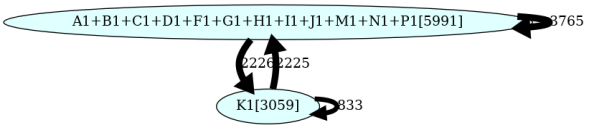
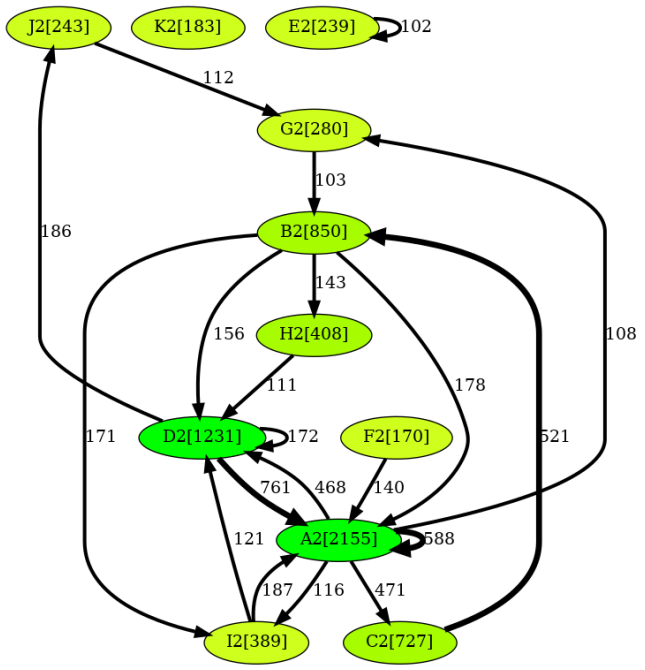
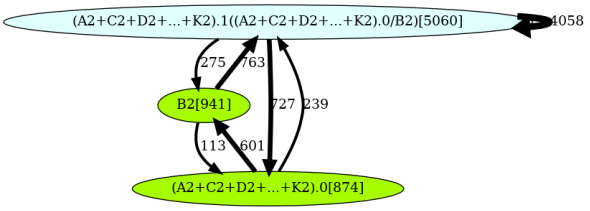
The similarity of monomer blocks to *Consensus* in all centromeres exceeds 92% (for human genome, the average similarly is only 75-80%), suggesting that the default values for some HORmon parameters might not be optimal for *Arabidopsis thaliana* centromeres. However, in absence of ground truth for inference of monomers and HORs for this genome, careful tuning of these parameters remains challenging.

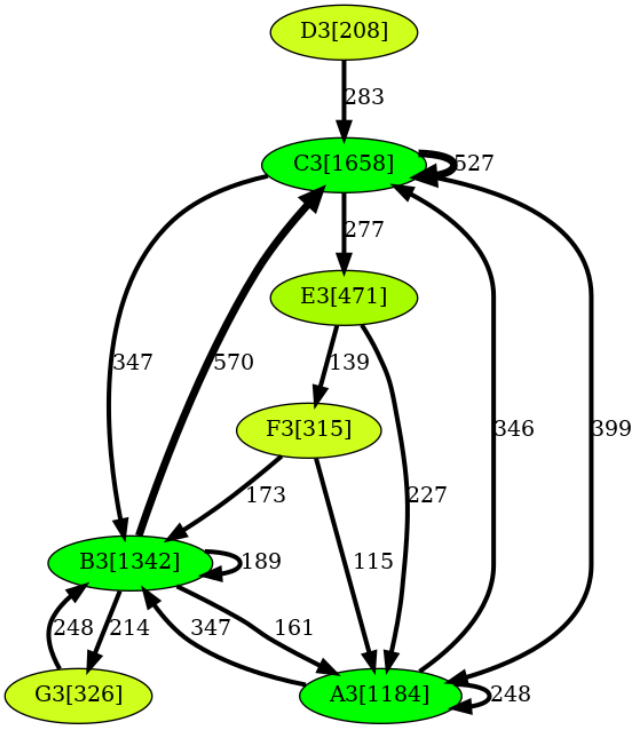
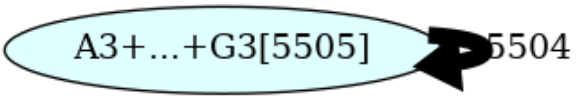
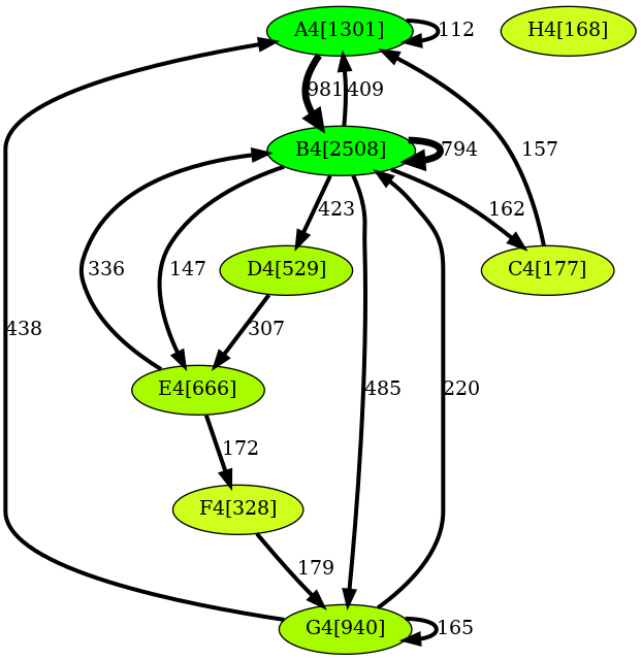
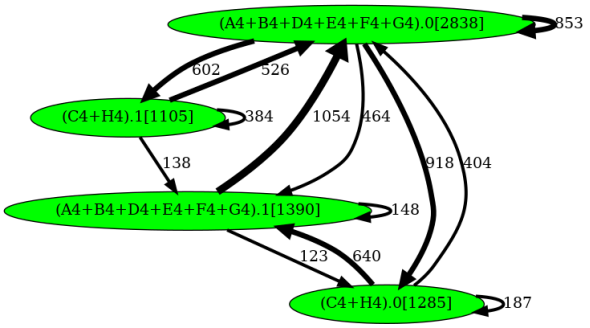
chromosome	start	end
1	15 520 362	17 128 808
2	3 826 809	5 048 992
3	14 755 631	15 733 923
4	4 530 001	5 707 115
5	12 750 675	13 217 483

Table S10.1. Coordinates of centromeres on each chromosome in *Arabidopsis thaliana* (Naish et al., 2021).

cen	# monomers reported by Centromere Architect	# monomers in *MonomersNew*	Max/Min # blocks for *MonomersNew*	#merge/split operations	#hybrid decom-s	# monomers in *MonomersFinal*	Max/Min # blocks for *MonomersFinal*
1	16	13	4535/90	11/0	0	2	5994/1084
2	35	11	1943/170	9/0	1	3	5060/874
3	26	7	2173/134	6/0	0	1	5505
4	19	8	2344/131	6/0	0	4	2838/1105
5	28	13	724/34	10/0	1	3	2010/274

Table S10.2. Information about monomers in *Arabidopsis thaliana* extracted by HORmon. The centromere ID (first column), number of monomers reported by CentromereArchitect (second column), number of monomers in the monomer-set *MonomersNew* — the set of frequent monomers generated by CentromereArchitect (third column), maximum / minimum number of blocks for monomers in the monomer-set *MonomersNew* (fourth column), number of merging / splitting operations performed to generate the monomer-set *MonomersFinal* — the final set of monomers generated by HORmon (fifth column), number of hybrid decomposition operations to generate the monomer-set *MonomersFinal* (sixth column), maximum / minimum number of blocks for monomers in the monomer-set *MonomersFinal.*

cen	Monomer graph for *MonomersNew*	Monomer graph for *MonomersFinal*
1		
2		

3	 D3[208] 283 C3[1658] 527 277 E3[471] 347 570 139 F3[315] 346 399 173 227 B3[1342] 189 115 248 214 347 161 G3[326] A3[1184] 248	 A3+...+G3[5505] 5504
4	 A4[1301] 112 H4[168] 981 409 B4[2508] 794 157 423 162 336 147 D4[529] C4[177] 438 307 E4[666] 485 220 172 F4[328] 179 G4[940] 165	 (A4+B4+D4+E4+F4+G4).0[2838] 853 602 526 (C4+H4).1[1105] 384 1054 464 138 918 404 (A4+B4+D4+E4+F4+G4).1[1390] 148 123 640 (C4+H4).0[1285] 187

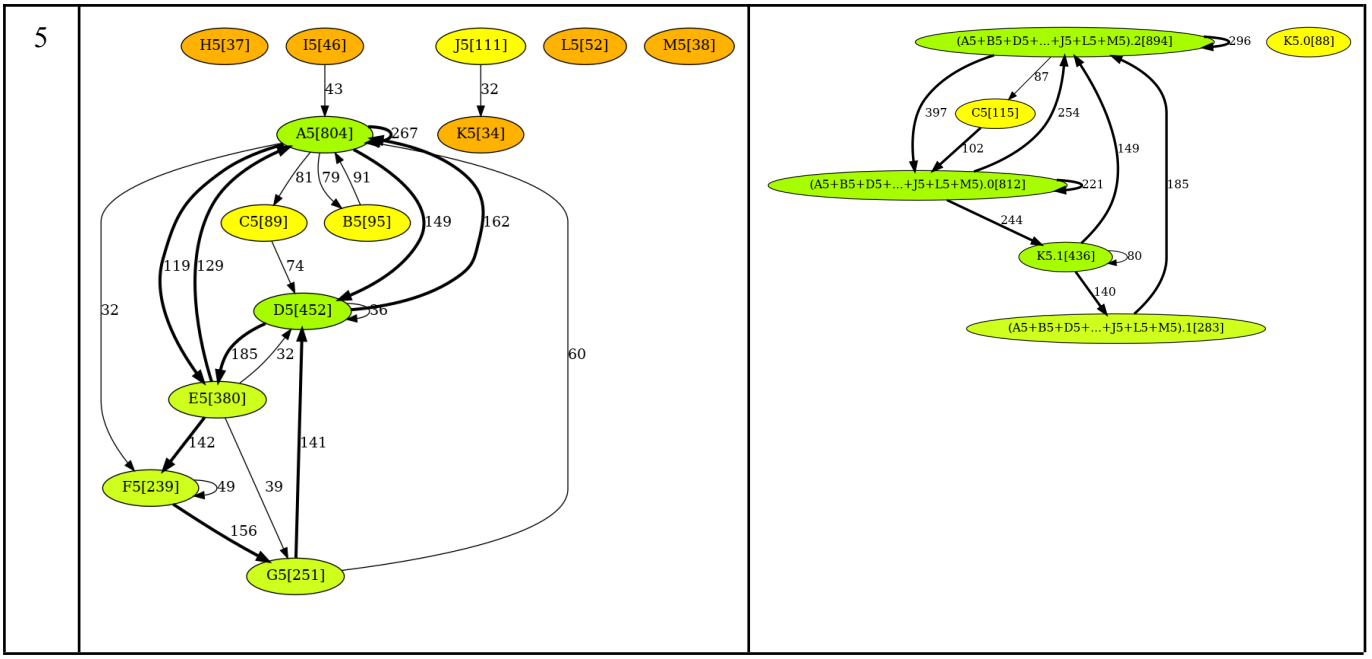


Table S10.3. Monomer Graphs for *Arabidopsis thaliana*. The first column corresponds to the centromere ID. The second (third) column shows monomer graphs built on the monomer-set *MonomersNew* (*MonomersFinal*). The label of each vertex represents the monomer ID and its count in the monocentromere (in parentheses). The rules of monomer naming are described in the Supplementary Note 4. The label of an edge in the monomer-graph represents its multiplicity. The width of an edge (color of a vertex) reflects its multiplicity (count of a monomer).

Supplementary Note 11: Annotation of the centromere on Chromosome 8 in Chimpanzee

Logsdon et al., 2021 generated the first complete assembly of centromeres on Chromosome 8 (cen8) of Chimpanzee, Orangutan, and Macaque. Below we use HORmon to provide initial insights into the monomer and HOR structure of cen8 in Chimpanzee. As an input, CentromereArchitect (the prerequisite for HORmon) utilized the same consensus monomer *Consensus* that is used for analysis of centromeres in the human genome. The start and end coordinates of the centromere (haplotype 1) are 801,093 — 2,232,981, length is 1,431,888 bp. These coordinates define the longest contiguous segment of the chromosome with all monomer-blocks sharing the same strands and with identity to *Consensus* exceeding *MinBlockIdentity* (the default value *MinBlockIdentity* = 70%, extracted using StringDecomposer, Dvorkina et al., 2020).

For consistency of notation with the analysis of human centromeres, we refer to the set of 16 frequent monomers generated by CentromereArchitect as *MonomersNew*, and to the final set of 13 monomers generated by HORmon as *MonomersFinal*. Maximum (minimum) multiplicity of monomers in *MonomersNew* is 1363 (117). HORmon conducts 5 merge and 2 split transformations, and no dehybridizations. Maximum (minimum) multiplicity of monomers in *MonomersFinal* is 1403 (168).

Figure S11.1 shows monomer graphs for *MonomersNew* and *MonomersFinal*. HORmon has detected two HORs consisting of monomers J8,K8,L8,(F8+I8+H8+D8)1.1 and M8,N8,O8,P8. One could argue that merging monomer (F8+I8+H8+D8).1.0 with monomer (F8+I8+H8+D8).0 would result in another HOR: A8,B8,C8+E8+G8,(F8+I8+H8+D8).0. Similarly to the section "What is a HOR in cen9?", this argument reflects the difficulty of developing an automated approach to centromere annotation and defining parameters of these approaches that work across all centromeres over many organisms.

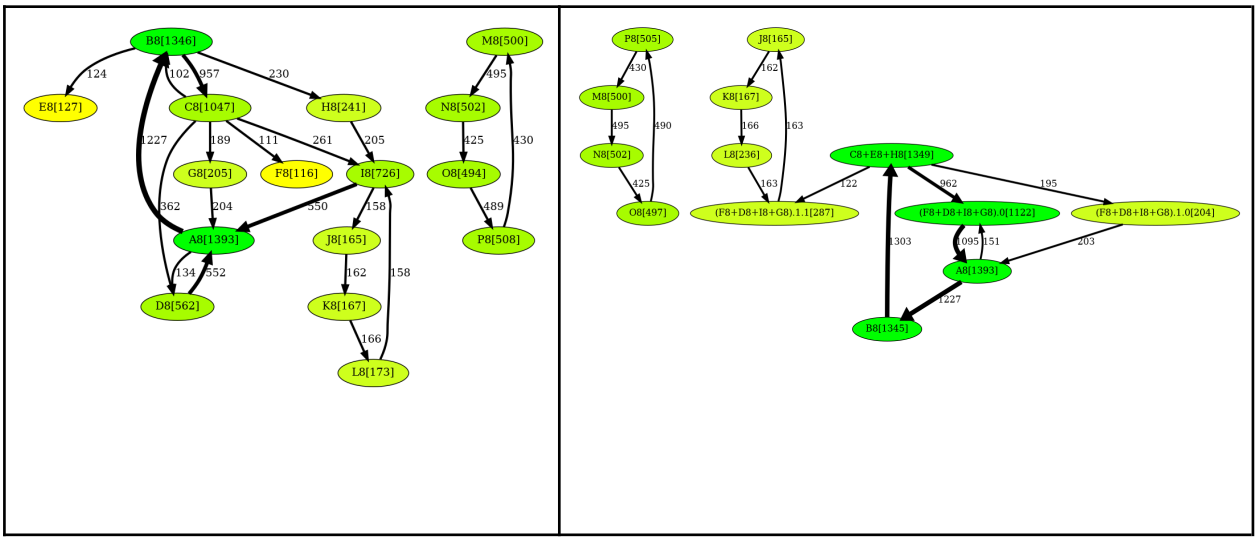


Figure S11.1. Left (right) is the monomer graph built for *MonomersNew* (*MonomersFinal*). The label of each vertex represents the monomer ID and its count in the monocentromere (in parentheses). The rules of monomer naming are described in the Supplementary Note 4. The label of an edge in the monomer-graph represents its multiplicity. The width of an edge (color of a vertex) reflects its multiplicity (count of a monomer).

Supplementary Note 12: The pseudocode of the split-and-merge module of HORmon

```
SplitAndMerge(Centromere, Monomers, minPI, minPosSim, splitValue)
while there are positionally-similar monomers in Monomers (wrt minPI and minPosSim)
   identify the most positionally-similar monomers M' and M'' in Monomers
   identify a new monomer M as the consensus of all M'-blocks and M''-blocks in Centromere
   remove monomers M' and M'' from Monomers
   add monomer M to Monomers
for each breakable monomer M in Monomers (wrt parameter splitValue)
   for each M-candidate-pair (X,Y)
      identify a new monomer M' as the consensus of all M-blocks in triples XMY
      add M' to Monomers
   remove M from Monomers
return Monomers
```