Genome Biology & Evolution, July 18, 2016

Supplementary Files

Centromeres of the yeast *Komagataella phaffii* (*Pichia pastoris)* have a simple inverted-repeat structure

Aisling Y. Coughlan[1], Sara J. Hanson[1], Kevin P. Byrne[1] and Kenneth H. Wolfe*,[1]

[1]UCD Conway Institute, School of Medicine, University College Dublin, Dublin 4, Ireland

*Corresponding author: E-mail: kenneth.wolfe@ucd.ie

Supplementary Figure legends

Figure S1. Construction of 3xHA-tagged *CSE4.* (A) Multiple sequence alignment of Cse4 proteins. The red line shows the site of HA tag insertion previously used for ChIP in *S. cerevisiae* (Scer; Stoler et al. 1995), *Naumovozyma castellii* (Ncas; Kobayashi et al. 2015), *Ogataea polymorpha* (Opol; Hanson et al. 2014) and *K. phaffii* (Kpha; this study). (B) Strategy for tagging *K. phaffii CSE4*. The synthetic DNA fragment, *KanMX* marker and *CSE4* downstream DNA were joined by fusion PCR. The entire cassette was then transformed into *K. phaffii* CBS12964 to replace the endogenous *CSE4* gene by homologous recombination. (C) Sequence of the synthetic DNA. The 3xHA tag is in lowercase.

Figure S2. Dot matrix plot of the four *K. phaffii* centromeres compared to each other. Regions of approximately 10 kb around each centromere were concatenated and compared. The arrows indicate a small similarity between *CEN1* and *CEN4*. The plot was constructed with Dotter (Sonnhammer and Durbin 1995), with the Greyramp parameters set to 40 (min.) / 100 (max.).

Figure S3. Mating-type switching does not induce recombination at centromeric IRs. GS115 S-1 to S-4 are four *MAT*a clones induced by mating-type switching of a GS115 *MAT*alpha strain. Centromere orientation-specific PCR was carried out with primers A-D for each centromere as in Figure 4.

Figure S4. Dot matrix plot of the seven *C. tropicalis* centromeres compared to each other. Parameters are identical to Figure S2. *C. tropicalis* sequence data is from Butler et al. (2009) with centromeres identified by Chatterjee et al. (2016).

Figure S5. Dot matrix plot of the three *Sch. pombe* centromeres compared to each other. Red lines indicate points of concatenation between chromosomes. The plot was constructed with Dotter (Sonnhammer and Durbin 1995), with the Greyramp parameters set to 100 (min.) / 150 (max.) to accommodate the larger scale of this figure compared to Figures S2 and S4. The diagram was constructed using the reference genome sequence of *Sch. pombe* which lacks some copies of *otr* units as indicated by the word GAP on the Y-axis (see Fig. 1 of Wood et al. 2002). The *dg* and *dh* components of the *otr* are marked; *dh* has a different orientation on *cen1* compared to *cen2* and *cen3*. The *tm* region of similarity between *cc1* and *cc3* is also marked.

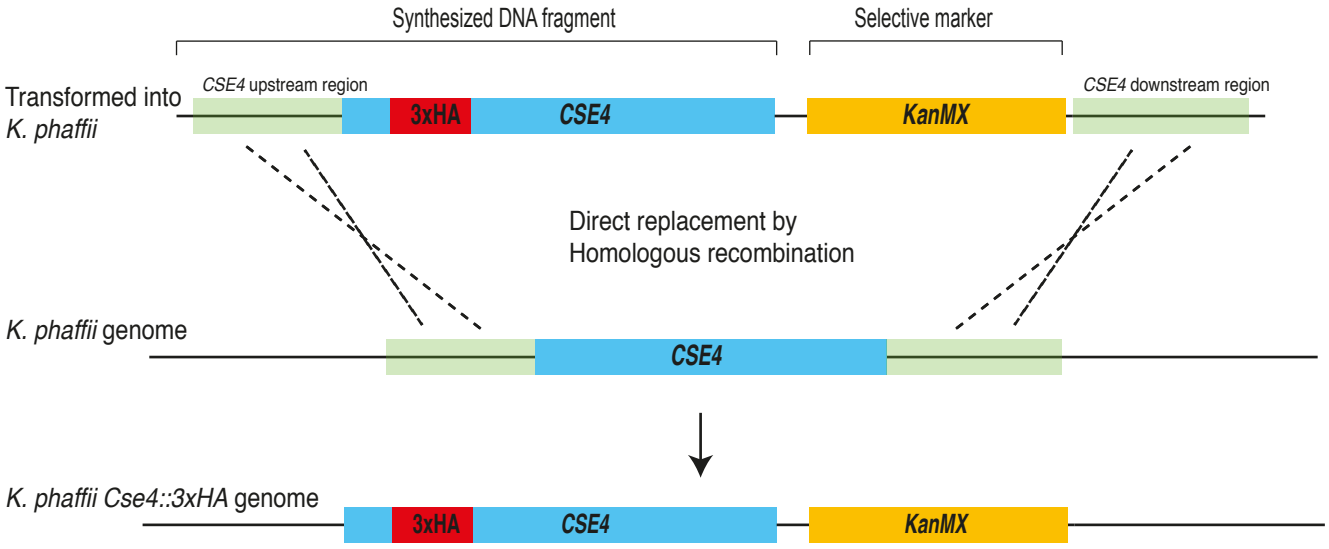
Table S1. Sequences of primers used for centromere orientation-specific PCR.

Centromere	Primer name	Sequence (5'-3')
CEN1	PpCEN1_A1	GGTAATATCCAGCAGTCAGACCC
CEN1	PpCEN1_B1	CTTTGCACCAATTAGCGCATAGC
CEN1	PpCEN1_C1	CCGGCGACAGTATCAATCACTTC
CEN1	PpCEN1_D1	GTCTTTCAGAGAGGAGCAATGC
CEN2	PpCEN2_A1	GATCCGGACTCTTTACAAAAAGC
CEN2	PpCEN2_B1	TGGTCGCATGGCCAACTG
CEN2	PpCEN2_C1	GCGCTGCACTGTTCACATAG
CEN2	PpCEN2_D1	CGATCTCCGTTGATACTCCCAAC
CEN3	PpCEN3_A1	GCTCCGTCAGCTTGAATAAGCC
CEN3	PpCEN3_B1	AATTATGCTAGGGAGAGCTTGC
CEN3	PpCEN3_C1	GGAGGCAGACACGCTTACC
CEN3	PpCEN3_D1	CGACAAGTGGTACACCAGTCAG
CEN4	PpCEN4_A1	CACCTTAACGAGAAGCCGAG
CEN4	PpCEN4_B1	GCAACTTTGGTCCTGAGGTCCTG
CEN4	PpCEN4_C1	GGGCAAACAGCATCCAGC
CEN4	PpCEN4_D1	GTACCCTTTGAAGAGACCACACC

A

```
Scer            MSSKQ-----------QWVSSAIQSDSSGRSLSNVNRLAGD---------QQSINDRALS
Ncas            MSSRKFVEQGHAQNSSSHLFSNILDNDSG-SLSNINRLTLDPDNTEDLLQQEVINERALS
Kpha            ----------------------------------MARLSI----------AGNRLNRLGS
Opol            ----------------------------------MARLN--------------KHKPVS
                                                   : **               .  *

Scer            LLQRTRATKNLFPRREERRR-YESSKSDLDIET------DYEDQAGNLEIETENEEEAEM
Ncas            LLQRTRERRNLLHRFEDKRRYYNQGQDDGDLESVASSHYRSNDVGGNFQFFDQEEDEDEE
Kpha            ISPKT-------PRSAVSNSASGSARTHTGTP-----------------------------R
Opol            TTPKT-------PKMGSSTSFKNSPASIKKTPS---------------------------K
                   .*        .         .

Scer            ETEVPAPVRT--HSYALDRYV------RQKRREKQRKQSLK-RVE-----------KKYT
Ncas            GNAIDDDYGTLDQSNIIDRHQERRKGNKHSRQERHHQRELKQRVEKIRTQRQVGNTKKFT
Kpha            GNKTP-IRNVITSSTGISRNQ------PGDPLPI-AKK------------------YRYK
Opol            SSETPKVKAGSTLSSGLARNQ------PGDPTEIKTQK------------------KRFK
                 .           *  : *          .      :.                  .:.

Scer            PSELALYEIRKYQRSTDLLISKIPFARLVKEVTDEFTTKDQDLRWQSMAIMALQEASEAY
Ncas            PSSLALYEIRKYQRSSELLISKIPFTKLVKEVTDEFTVEDQQLHWQSMAIVALQEASEAY
Kpha            PGTLALREIKRYQKSTDLLLRKLPFARLVREIAQENFIQGEMFQWQSVAILALQEAAEAF
Opol            PGTVALREIRRFQKSTELLIRKLPFARLVREIVQDEF-GTSTYRWQSVAVLALQEAVEAY
                *. :** **..:*.*::**: *:**:.**.*:.::     .  .***:*::***** **:

Scer            LVGLLEHTNLLALHAKRITIMKKDMQLARRIRGQFI
Ncas            LVGLLEHANLLAIHAKRITLMKKDIQLARRIRGQFI
Kpha            LVGLLEDTNLCALHAKRVTIMQKDIQLARRIRGDGI
Opol            LVHLFEDTNLCALHAKRVTIMQKDIHLARRLRGD--
                ** *:* :** *:****:*:*:**::****:**:
```

B



C

```
ATGGCGAGGCTTAGTATAGCAGGTAACAGGTTGAATAGATTAGGTTCTATTTCGCCTAAAACGCCTAGATCAGCAGTATC
CAATTCTGCCTCCGGGTCAGCTAGAACCCACACAGGAACACCAtacccatacgatgttcctgactatgcgggctatccgt
atgacgtcccggactatgcaggatcctatccatatgacgttccagattacgctAGAGGTAATAAGACTCCAATACGGAAT
GTAATAACAAGTAGCACAGGAATTAGTAGAAATCAACCCGGAGACCCTCTACCCATTGCAAAGAAATACAGATACAAACC
TGGAACGTTAGCTTTAAGAGAGATTAAAAGATATCAAAAATCCACAGACTTGCTCTTGCGTAAGCTACCCTTTGCTCGAC
TAGTGAGGGAAATAGCCCAGGAGAACTTCATTCAAGGAGAAATGTTTCAATGGCAAAGCGTGGCCATCTTAGCTTTACAA
GAAGCGGCCGAAGCCTTCTTGGTAGGCTTACTGGAAGATACAAATCTGTGCGCCCTCCACGCAAAAAGAGTCACCATAAT
GCAGAAAGACATCCAACTAGCAAGAAGGATAAGAGGTGATGGGATC
```

Figure S1

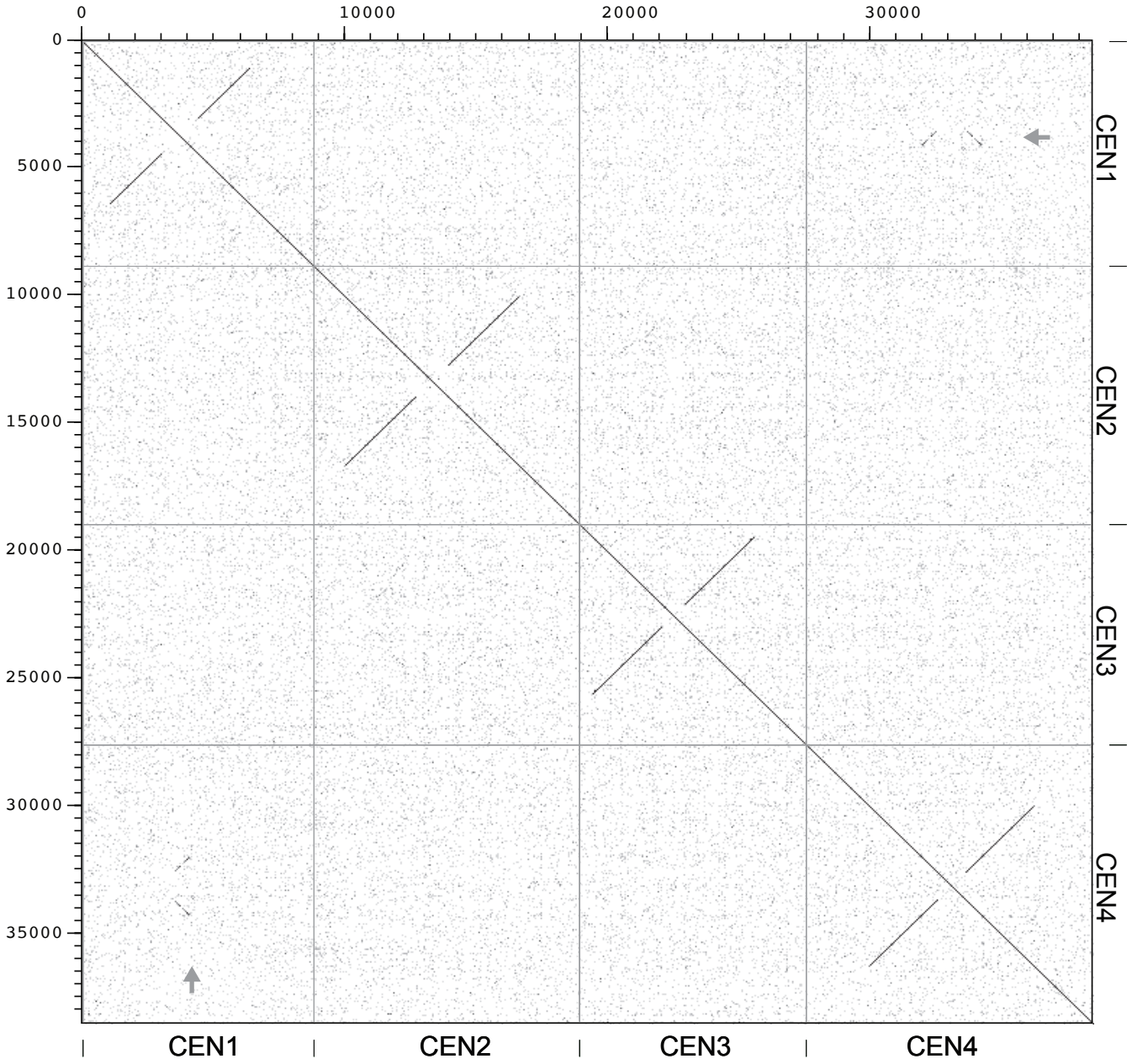

0
10000
20000
30000
0
5000
10000
15000
20000
25000
30000
35000
CEN1
CEN2
CEN3
CEN4
CEN1
CEN2
CEN3
CEN4

Figure S2

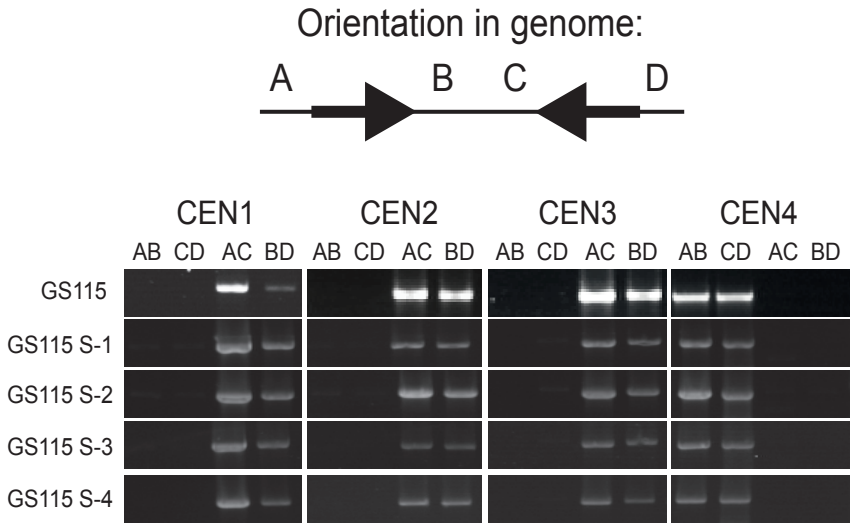

Orientation in genome:
A
B
C
D
CEN1
CEN2
CEN3
CEN4
AB CD AC BD AB CD AC BD AB CD AC BD AB CD AC BD
GS115
GS115 S-1
GS115 S-2
GS115 S-3
GS115 S-4

Figure S3

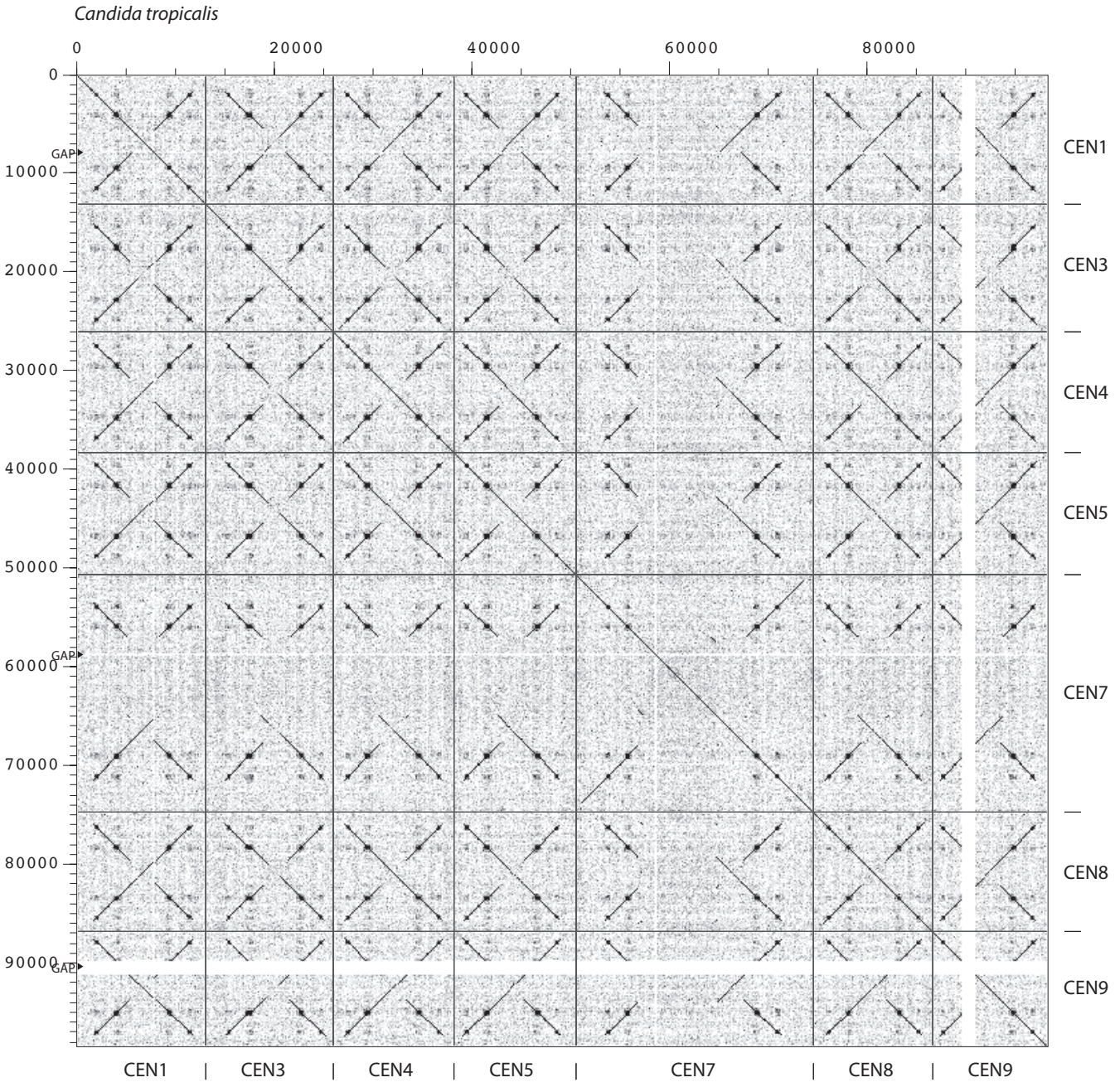

Candida tropicalis
0
20000
40000
60000
80000
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
GAP
GAP
GAP
CEN1
CEN3
CEN4
CEN5
CEN7
CEN8
CEN9
CEN1
CEN3
CEN4
CEN5
CEN7
CEN8
CEN9

Figure S4

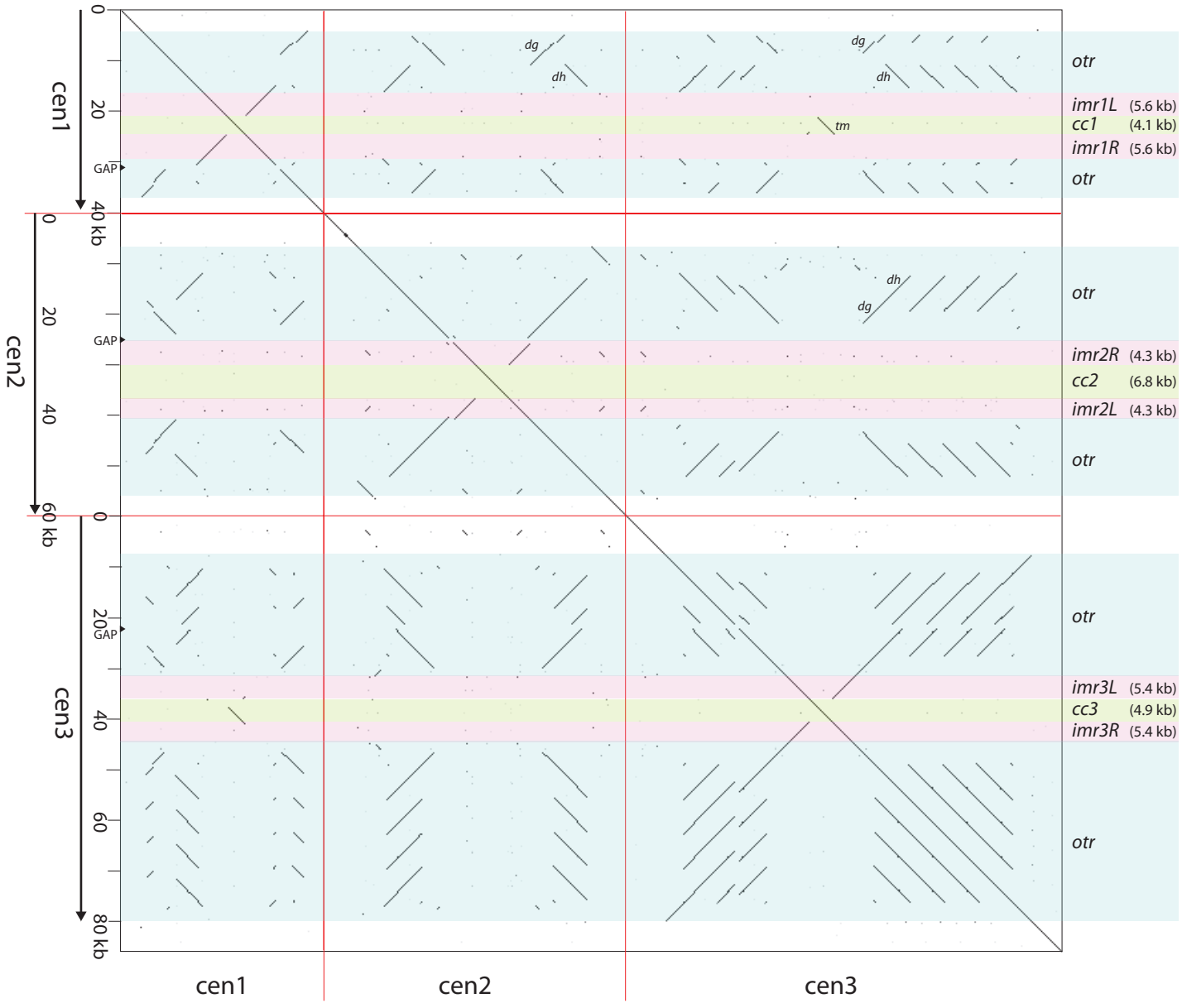

cen1
cen2
cen3
0
20
GAP
40 kb
0
20
40
GAP
60 kb
0
20
GAP
40
60
80 kb
dg
dh
dg
dh
tm
dh
dg
otr
imr1L (5.6 kb)
cc1 (4.1 kb)
imr1R (5.6 kb)
otr
otr
imr2R (4.3 kb)
cc2 (6.8 kb)
imr2L (4.3 kb)
otr
otr
imr3L (5.4 kb)
cc3 (4.9 kb)
imr3R (5.4 kb)
otr
cen1
cen2
cen3

Figure S5