Centromere formation remodels chromatin fibre structure

Catherine Naughton[1], Covadonga Huidobro[1], Claudia R. Catacchio[1,2], Adam Buckle[1], Graeme R. Grimes, Ryu-Suke Nozawa[1], Stefania Purgato[1,3], Mariano Rocchi[2], Nick Gilbert[1]*

[1]MRC Human Genetics Unit, The University of Edinburgh, Crewe Rd, Edinburgh, EH4 2XU [2]Department of Biology, University of Bari, Via Orabona 4, 70125 Bari, Italy
[3]Department of Pharmacy and Biotechnology, University of Bologna, 40126 Bologna, Italy

*Nick.Gilbert@ed.ac.uk

Abstract

Human centromeric chromatin is assembled from CENP-A nucleosomes[1] and repetitive α-satellite DNA sequences[2] and provides a foundation for kinetochore assembly in mitosis. Biophysical experiments led to a hypothesis that the repetitive DNA sequences form a highly folded chromatin scaffold necessary for function, but this idea was revised when fully functional evolutionary new centromeres (ENCs) or neocentromeres were found to form on non-repetitive DNA. To understand if centromeres have a special chromatin structure we have genetically isolated a single human chromosome harbouring a neocentromere and investigated its organisation. The centromere core is enriched in RNA pol II, active epigenetic marks and remodelled by transcription to form a negatively supercoiled 'open' chromatin domain. In contrast, centromerisation causes a spreading of repressive epigenetic marks to flanking regions, delimited by H3K27me3 polycomb boundaries and divergent genes. The flanking domain is partially remodelled to form 'compact' chromatin, with characteristics similar to satellite-containing pericentromeric chromatin, but exhibits low level genomic instability. We provide a model for centromere chromatin structure and suggest that open chromatin provides a foundation for a stable kinetochore whilst pericentromeric heterochromatin generates surrounding mechanical rigidity.

To understand how new centromeres are accommodated in chromosomes and to investigate the biophysical properties of centromeric chromatin required to form a stable kinetochore we have used a previously identified neocentromere at 3q24 as a model system[3]. This neocentromere has been propagated across multiple generations, indicating it is stable through the germline, and is located in the vicinity of two genes but within a relatively gene poor segment of the genome. As the parental cells are heterozygous for the neocentromere at 3q24 the chromosome harbouring the neocentromere, Neo3, was genetically isolated from the normal counterpart in a human-hamster hybrid cell line (HybNeo3; Fig 1A) and compared to a human-hamster hybrid cell line, GM10253A, which has a single normal human chromosome 3, termed HSA3. No genetic changes were apparent in Neo3 and there was no evidence for repetitive DNA at 3q24 by deep sequencing, whilst as reported for other neocentromeres[4–6] α-satellite persisted at the original centromere location (Fig. 1A).

The position of the neocentromere was confirmed by using DNA fluorescence *in situ* hybridization (FISH) with probes to 3q24 (Extended Data Fig. 1A-C) and high-resolution mapped using ChIP with antibodies to CENP-A and CENP-C in the parental cells (Extended Data Fig. 2A), revealing a centromere core domain of 130 kb, similar to other synthetically derived neocentromeres[7]. In the derivatized human-hamster HybNeo3 cell line the centromere had drifted ≈30kb away from the telomere and had spread to encompass a ≈190kb domain. Centromere drift is apparent in horse and fission yeast[8,9], and may represent a natural event controlled in part by the constitutive centromere-associated network (CCAN) and buffered by repetitive satellite DNA[10].

Epigenetic remodelling at a human neocentromere

We set out to investigate how the chromatin fibre is remodelled in response to centromere formation and reasoned there could be two distinct possibilities (i) neocentromeres form at a genomic location that already has the features required for centromere function or (ii) neocentromeres have the capacity to remodel the local epigenetic environment. To discriminate between these two scenarios we examined the epigenetic repertoire of 3q24 using ChIP for active (H3K27ac, H3K4me2, H4K20me1) and repressive (H3K9me3, H3K9me2) epigenetic marks. A small block of GC-rich DNA in the vicinity of the neocentromere amplified aberrantly and was blacklisted (Extended Data Fig. 2B).

The canonical 3q24 locus on HSA3 was decorated with active marks (Fig. 1B) coincident with actively transcribed genes in a euchromatic region, whilst repressive heterochromatic marks were absent. In contrast, after neocentromerisation, a large 5 Mb heterochromatin domain marked by H3K9me2/3 formed around the centromere on Neo3 (Fig. 1B, yellow box). Focal active marks in the vicinity of genes were absent and there was a significant loss of H3K27ac, a marker of CBP/P300 activity[11]. The upstream and downstream pericentromeric regions had the epigenetic hallmarks of heterochromatin consistent with the idea that centromeres are remodelled into a repressive state, even in the absence of repetitive DNA, demonstrating that special DNA sequences are not required for heterochromatin formation. At the centromeric core (Fig. 1B-C, blue box), coincident with CENP-C binding, the chromatin was remodelled to a state distinct from the flanking pericentromeric domains, devoid of heterochromatin marks and enriched for H4K20me1. This data suggests that instead of neocentromeres adopting the local epigenetic landscape they can remodel the local chromatin environment[7] to form distinct centromeric and pericentromeric domains (Fig. 1B, blue and yellow boxes, respectively).

To understand the molecular basis for the distinct centromeric and pericentromeric domains we examined patterns of transcription. By RT-qPCR active genes at 3q24 on HSA3 were all silenced upon centromere formation as far as the distal DIPK2A and HLTF genes (Extended Data Fig. 3), suggestive of a spreading activity emanating from the centromere core domain. RNA sequencing was used to further explore the landscape at 3q24 and showed transcriptional repression over a 5 Mb pericentromeric domain on Neo3 (Fig. 1D). Recent data has indicated that (neo)centromeres are transcriptionally active in mitosis[7,12] and in interphase in model organisms[13–15]. ChIP for RNA polymerase II showed it was absent from the pericentromeric domain in Neo3, but statistically significant levels of polymerase were apparent at the centromeric core in both interphase (Extended Data Fig. 4A) and metaphase (Extended Data Fig. 4B) cells. However, no transcripts were detected within the centromeric core domain on Neo3 even after exosome knockdown (Extended Data Fig. 4C). Similarly, very deep sequencing for either short RNA transcripts, long RNA transcripts or nascent transcripts could not identify specific RNA species. This led us to speculate that transcription was at a very low level and dispersed across the centromeric core domain with multiple transcription initiation sites, but sufficient to remodel the local chromatin landscape.

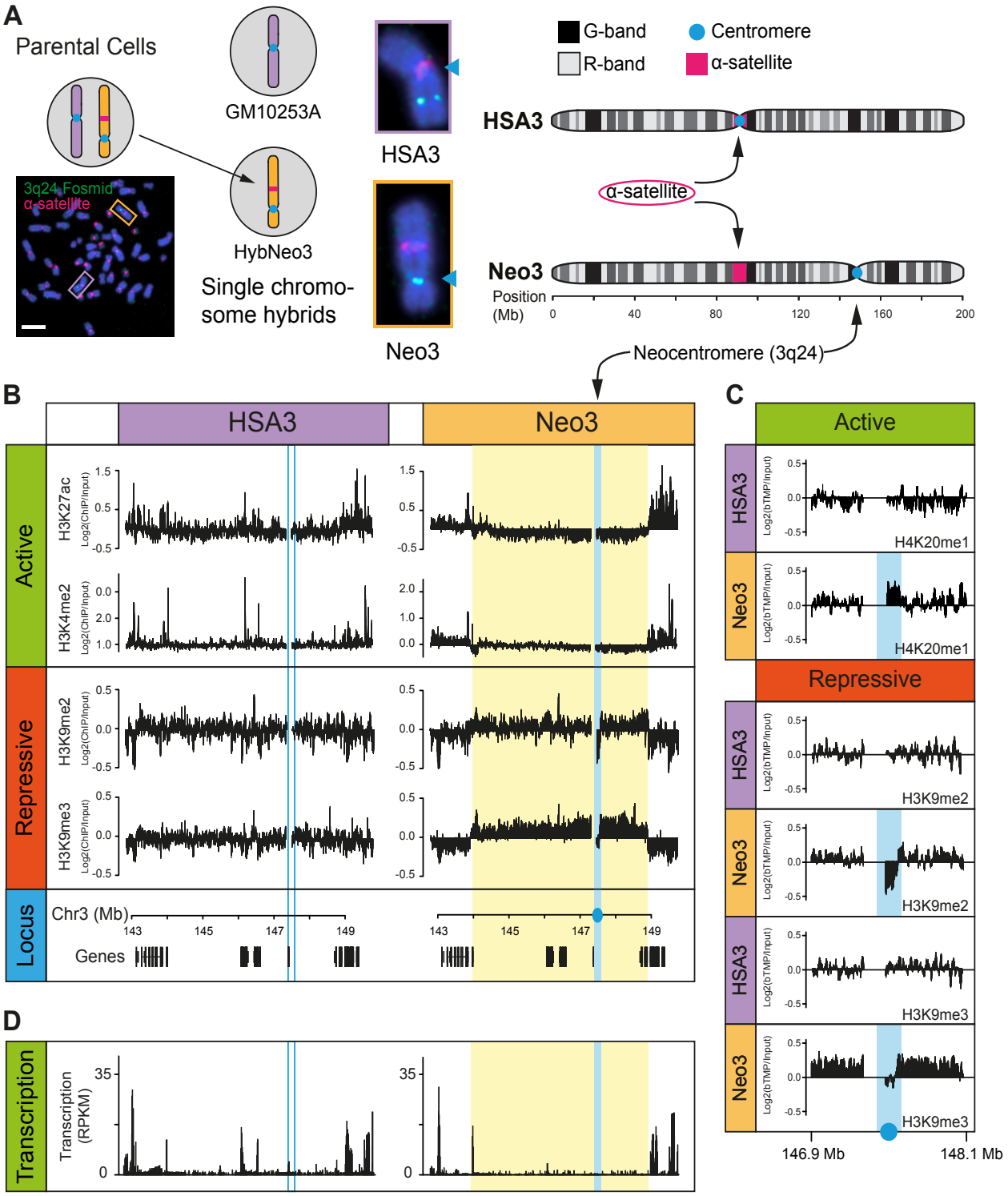


Fig.1. Epigenetic remodelling after centromere formation. (A) Schematic detailing experimental model system. Left; human lymphoblastoid cells (Parental) harbouring a canonical human chromosome 3 (HSA3; purple frame) and a chromosome 3 with a neocentromere located at 3q24 (Neo3; orange frame) were genetically manipulated to isolate Neo3 in a human-hamster hybrid (HybNeo3) for comparison to a human-hamster hybrid (GM10253A) with a single HSA3 chromosome. Middle, DNA FISH with α-satellite specific probe (red) and a fosmid probe (green) for 3q24. The primary constriction (arrow) coincides with the α-satellite array in HSA3 but is located at 3q24 in Neo3. Chromosomes were counterstained with DAPI. Bar is 5 µm. Right, chromosome 3 ideogram to indicate centromere and α-satellite locations on HSA3 and Neo3. **(B)** Distribution of active (H3K4me2, H3K27ac) and repressive (H3K9me3, H3K9me2) epigenetic marks measured by ChIP-chip at the 3q24 neocentromere region on HSA3 and Neo3 chromosomes. Neocentromere is marked in solid blue, equivalent position in HSA3 is marked in open blue box and remodelled pericentromeric heterochromatin domain is marked in yellow. Bottom, position of genes at 3q24 locus. **(C)** Detailed view of active (H4K20me1) and repressive (H3K9me2/3) epigenetic marks surrounding the neocentromere. **(D)** Normalised transcription across HSA3 and Neo3 analysed by RNA-seq.

Distinct pericentromeric domain boundaries

Centromerisation triggered epigenetic remodelling to form a repressive pericentromeric domain (Fig. 1B). To characterise how the domain was delimited we examined facultative epigenetic marks across the locus and identified strong H3K27me3 enrichment at the telomeric end of the domain and a weak enrichment for H3K36me3 at the other end (Fig. 2A). H3K27me3 is a mark indicative of polycomb activity, whilst H3K36me3 is added by Set2 at active genes but is also reported as a facultative mark at heterochromatin[16]. These features suggest that the repressive pericentromere activity spreads until it reaches these boundaries, but raises the question as to what defines them? CTCF is an abundant protein associated with GC-rich DNA sequences and provides boundary activity for marking chromatin interaction domains measured by 3C techniques[17,18]. Although CTCF was present throughout the pericentromeric domain there were strong peaks located at both ends (Fig. 2B). More strikingly and consistent with a recent study in budding yeast[19] the pericentromeric domain was also delimited by the first convergently transcribed gene encountered moving away from the centromere (Fig. 2B). These results reveal a pronounced two domain structure at centromeres with a core and pericentromeric domain flanked by boundary sites defined by convergent genes and CTCF binding.

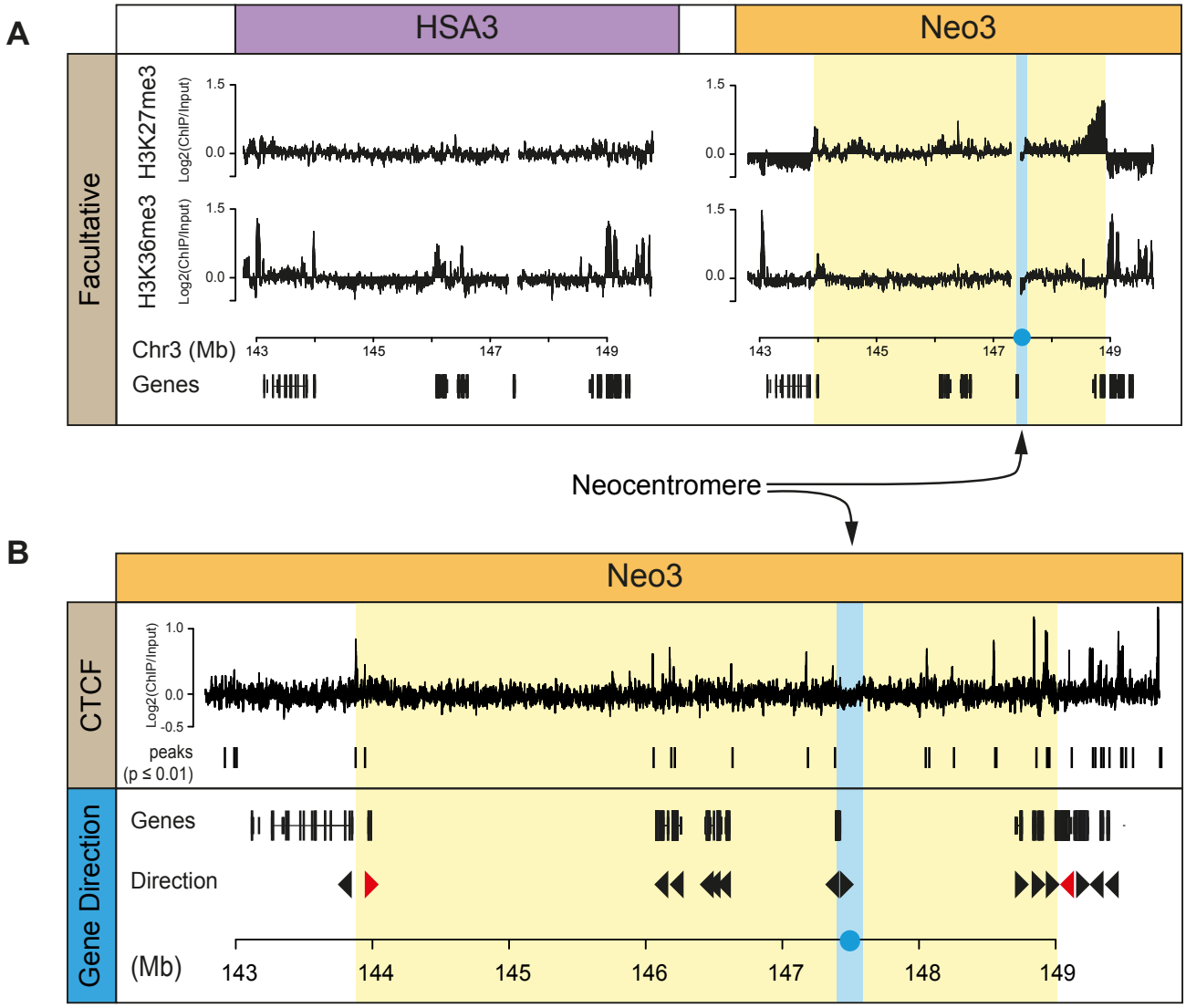


Fig.2. Facultative heterochromatin and convergent genes mark transcriptionally silenced pericentromeric domain. (A) Distribution of facultative heterochromatin (H3K27me3 and H3K36me3) marks measured by ChIP-chip on HSA3 and Neo3 chromosomes. **(B)** Top, ChIP-chip showing distribution of CTCF at pericentromeric boundaries on Neo3 chromosomes. Bottom, schematic showing gene orientation at 3q24, convergent gene boundary shown in red. Vertical blue line corresponds to the neocentromere and repressive pericentromeric heterochromatin domain marked in yellow.

Local chromatin fibre remodelling after centromere formation

In gene-rich euchromatin nucleosomal DNA is packaged into chromatin fibres which have a disrupted or 'open' configuration[20], a structure that is particularly conspicuous at transcription start sites[21]. In contrast, the specialised chromatin found at centromeres is formed from alpha satellite[2] and CENP-A containing nucleosomes[1]. Alpha satellite makes up 3% of the human genome and positions nucleosomes precisely *in vivo*[22] and *in vitro*[23]. Sedimentation studies to examine the biophysical properties of satellite containing chromatin indicate that it has the characteristics of a rigid rod-like particle which may enable it to fold into an ordered or crystalline array[24,25]. To establish the biophysical properties of centromeric chromatin (Fig. 3A) soluble chromatin was prepared from nuclei containing HSA3 and Neo3 chromosomes and fractionated by sucrose gradient sedimentation and pulsed-field gel electrophoresis[20] (PFGE) (Extended Data Fig. 5A). Subsequently DNA corresponding to 'open' or disrupted chromatin was isolated from the gel (Extended Data Fig. 5B) and used to map the chromatin fibre structure in the centromeric domain. Chromatin fibres located at the neocentromere were substantially remodelled to have a pronounced 'open' configuration (Fig. 3B) which was restricted to the CENP-C containing core, and is similar to the characteristics observed at transcription start sites[21]. Due to the presence of low-level RNA polymerase in the centromeric domain (Extended Data Fig. 4) we speculated that chromatin remodelling was linked to transcription, as has been observed in model organisms[13–15]. Transcription inhibition (Extended Data Fig. 6) completely abrogated the formation of disrupted chromatin fibres (Fig. 3B) demonstrating that centromeric chromatin is remodelled to have a transcription dependent 'open' structure.

Previous studies have suggested that CENP-A containing chromatin is folded differently and this could be linked to DNA supercoiling[26]. If DNA is twisted in a right-handed direction it becomes over-wound (positive supercoiling) whilst twisting in the opposite direction it adopts an under-wound (negatively supercoiled) configuration[27]. Our earlier work showed that the level of supercoiling is transcription dependent[28] so we hypothesised that low level RNA polymerase II activity could impact the local DNA configuration. Using biotinylated 4,5,8-trimethylpsoralen (bTMP) as a DNA structure probe[28] centromeric chromatin was found to be enriched in negatively supercoiled DNA (Fig. 3C), in a transcription dependent manner. These data further indicate that RNA polymerase is not present as a static component but is engaged in active transcription that remodels DNA and local chromatin fibre structure.

Satellite-containing pericentromeric chromatin fibres found at canonical centromeres have a rigid rod-like structure[25] that may facilitate kinetochore formation and increase the fidelity of chromosome segregation. In contrast, neocentromeres do not have repetitive α-satellite DNA sequences[4–6] to precisely position nucleosomes, so it was important to ask if centromerisation, *per se*, affected nucleosome positioning within the core or flanking DNA sequences. Mono and di-nucleosome fragments prepared from HSA3- and Neo3-containing nuclei using DFF nuclease (Extended Data Fig. 7A-B) were selected using biotinylated baits (Extended Data Fig. 7C) and deep sequenced. After centromerisation the size of the nucleosomal fragments did not change, despite centromeric nucleosomes being enriched in CENP-A (Extended Data Fig. 7D-F) and we were unable to observe any difference in nucleosome positioning (Extended Data Fig. 7G) or periodicity (Extended Data Fig. 7H). Despite no apparent change in the nucleosomal arrangement we speculated that over time pericentromeric chromatin may be remodelled to adopt a more compact configuration (scenario 2; Fig. 3D), analogous to the structure observed at canonical pericentromeres[25]. Consistent with this idea a 250 kb region at the pericentromeric boundary had a compact chromatin fibre structure (Fig. 3D) coincident with H3K36me3 (Fig. 2A) a mark that has previously been observed at constitutive and facultative heterochromatin[16]. This indicates that H3K9me2/3 heterochromatin marks are not sufficient to generate a compact chromatin fibre structure, but pericentromeric chromatin can be remodelled to form a structure analogous to canonical satellite-containing pericentromeric chromatin.

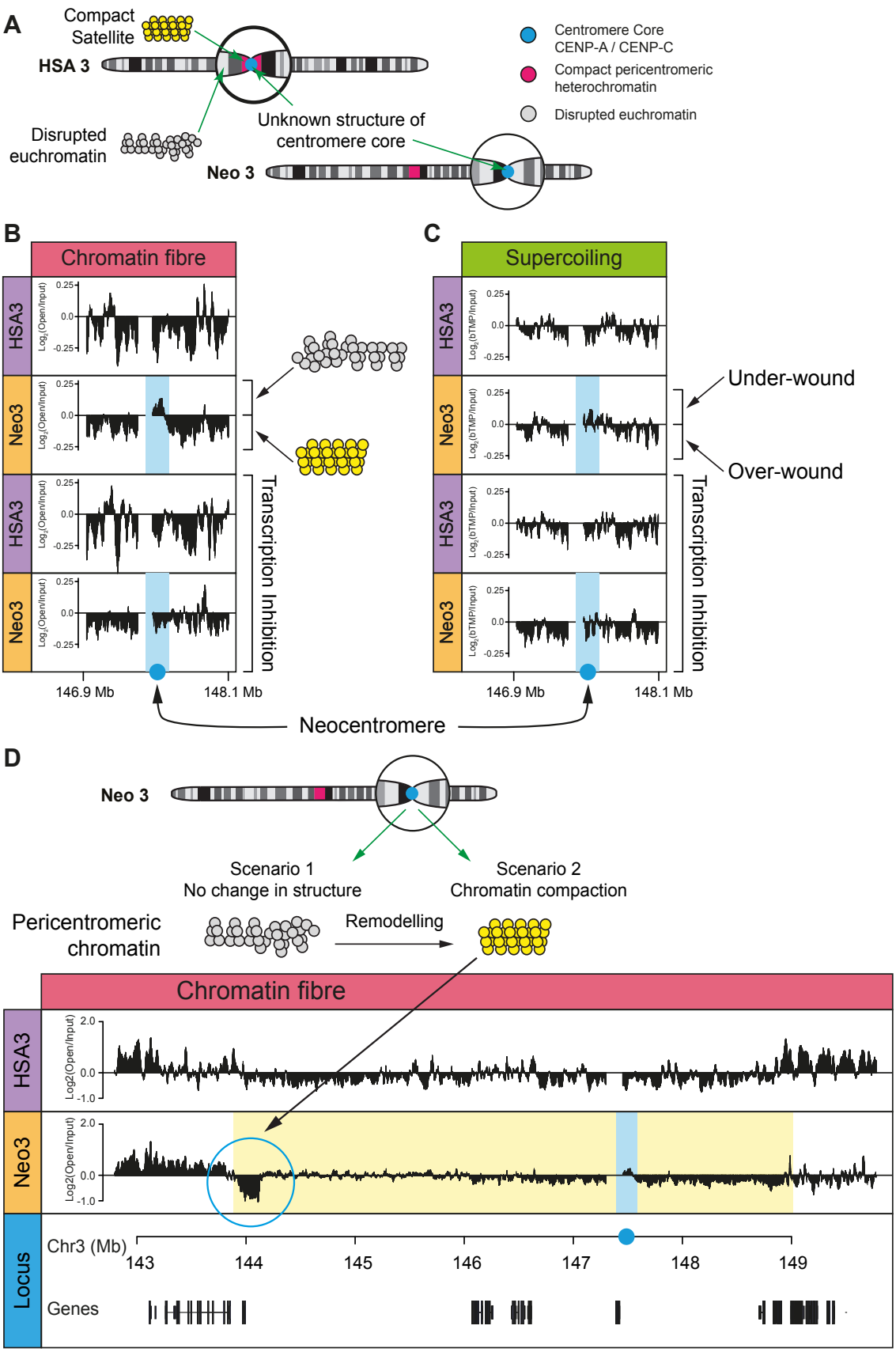


Fig.3. Centromeric chromatin has a transcription dependent underwound and disrupted chromatin fibre structure. (A) Top, schematic indicating the lack of understanding of chromatin structure at canonical centromeric chromatin (HSA3) and after centromerisation (Neo3). **(B)** Distribution of disrupted chromatin fibres across a 1.2 Mb region on HSA3 and Neo3, in the presence or absence of transcription inhibition (5 h α-amanitin treatment). **(C)** Organisation of negative (under-wound) and positive (over-wound) supercoils mapped by bTMP binding before or after transcription inhibition (5 h α-amanitin treatment). **(D)** Top, diagram showing potential scenarios for chromatin remodelling at pericentromere after centromerisation. Bottom, distribution of disrupted chromatin fibres, across the 3q24 locus in the HSA3 and Neo3 chromosomes. Blue bar corresponds to neocentromere and the yellow domain marks the silenced pericentromeric region.

Decompacted large-scale centromeric chromatin

As previous studies indicated an inter-relationship between different levels of chromatin organisation[20,21,28], we speculated that after centromere formation local changes in chromatin structure might be propagated and influence large-scale chromatin compaction[29]. To directly test this hypothesis 3D DNA FISH, with pairs of differentially labelled fosmid probes ≈300 kb apart (Supplementary Table 1), was used to ascertain large-scale chromatin compaction at the core and pericentromeric chromatin domains (Fig. 4A). In the hamster-hybrid cells harbouring HSA3 and Neo3 chromosomes there was no apparent change in compaction in pericentromeric regions but a pronounced decompaction at the centromeric core (Fig. 4B). To ensure this difference was not a consequence of comparing different cell lines the analysis was repeated in the parental cells using CENP-C immuno-FISH to discriminate between the Neo3 and HSA3 chromosomes. This similarly revealed a significant large-scale chromatin decompaction after centromerisation (Extended Data Fig. 7) showing that remodelling occurs at multiple levels of centromere organisation. Centromeric large-scale chromatin structure was also transcription dependent, with both α-amanitin and flavopiridol treatment causing chromatin compaction (Fig. 4C; Extended Data Fig. 8). As bleomycin treatment (introduces nicks) also caused large-scale chromatin to compact (Fig. 4C), this suggested the fibres were under topological strain, consistent with being negatively supercoiled (Fig. 3C). Although we were unable to find evidence for centromeric derived transcripts (Extended Data Fig. 4C) we speculated that transcripts may act locally to impact chromatin structure[30]. Consistent with this idea, RNase H treatment (hydrolyses RNA in the context of a DNA/RNA hybrid) compacted centromeric but not pericentromeric chromatin structure (Fig. 4D) suggesting that a transcription-dependent RNA component stabilises decompacted centromeric chromatin.

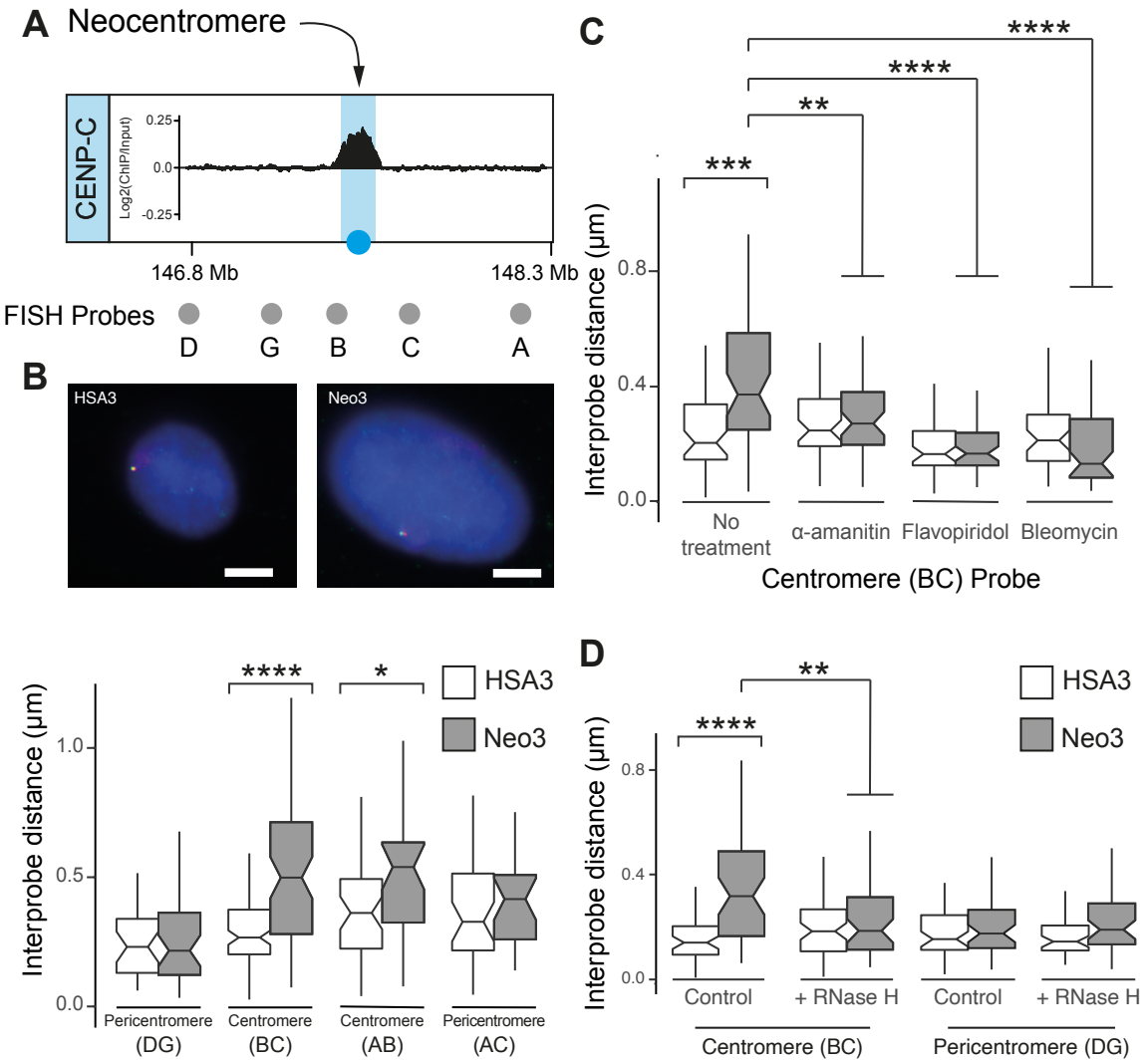


Fig.4. Large-scale neocentromere chromatin fibre decompaction. (A) Diagram showing fosmid FISH probes (grey circles) surrounding the CENPC-marked centromeric core domain (blue). **(B)** Top, representative images of 3D-FISH on HSA3 and Neo3 chromosomes hybridized to probe B (red) and C (green) in single chromosome human-hamster hybrid nuclei, counterstained with DAPI. Bar is 5 µm. Bottom, boxplot showing interprobe distance measurements (µm) for pairs of fosmid probes. Horizontal line, median; whiskers, 1.5 interquartile range; p-values are for a Wilcoxon test; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$. **(C)** Boxplot showing interprobe distance (µm) between the BC (centromere) fosmid probes in HSA3 (white) and Neo3 (grey) chromosomes in GM10253A and HybNeo3 cell lines respectively, after treatment with α-amanitin (5 h), flavopiridol (3 h) or bleomycin (10 min). **(D)** Boxplot showing interprobe distance (µm) for the BC (centromere) and DG (pericentromere) pairs of fosmid probes on HSA3 or Neo3 chromosomes following RNase H treatment.

Inherent genome instability at a neocentromere

Whilst neocentromeres form fully functional kinetochores that are stably propagated[31], they are still associated with higher chromosome mis-segregation rates and mitotic errors[32,33]. To quantify neocentromere stability we examined the chromosome architecture and copy number of the centromeric and pericentromeric domains in cells propagated for different amounts of time (Fig. 5A). At low passage almost all chromosomes had a normal structure (Fig. 5B and Extended Data Fig. 9A) but after approximately 50 passages the pericentromeric region upstream of the neocentromere had undergone break and or fusion events in 70% of cells. Copy number analysis was used to estimate the position of breakage events (Extended Data Fig. 9B) which were predominantly in the vicinity of the centromere. However, a region of pronounced DNA amplification was visible near the DIPK2A gene located proximal to the pericentromere boundary (Extended Data Fig. 9C). FISH of the cell population with the border fosmid indicated this region was amplified and located on other chromosomes but strikingly was rarely visible (2%) after chromosome breakage and fusion events (Extended Data Fig. 9D) indicating that breaks occurred in the pericentromeric chromatin domain upstream of the neocentromere.

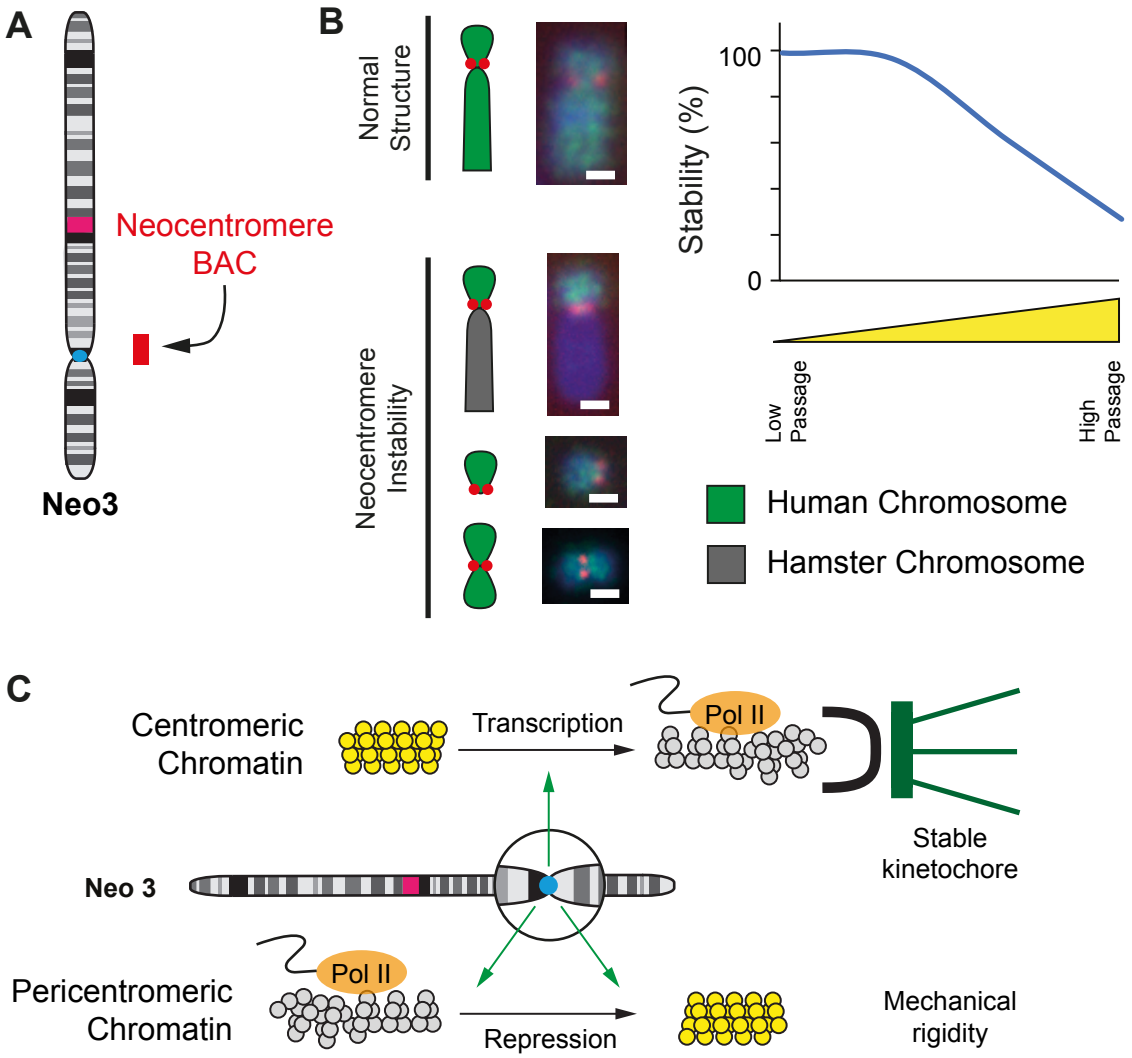


Fig. 5. Neocentromere chromosome instability. (A) Neo3 ideogram and BAC probe used to analyse chromosome stability. **(B)** Left, Representative 2D DNA FISH images of Neo3 chromosome architecture. Neo3 chromosomes were hybridized with a human chromosome 3 paint (green) and neocentromere BAC probe (red) and scored as normal or abnormal (displaying neocentromere instability in the form of breaks, fusions to hamster chromosomes (dark grey) or duplications). Right, graph quantifying loss in chromosome stability as passage number increased. Bar is 2 µm. **(C)** Model showing transcription dependent centromere remodelling to a disrupted euchromatin state (grey) and pericentromeric chromatin repression to form heterochromatin (yellow). We speculate euchromatin provides a suitable foundation for a high-fidelity kinetochore whilst heterochromatin and the accumulation of satellite sequences generates surrounding mechanical rigidity.

Discussion

For efficient and accurate segregation kinetochores assemble onto CENP-A-containing chromatin via CENP-C[34]. After centromere formation centromeric chromatin recruits active RNA polymerase which facilitates nucleosome turnover[13–15] and generates negatively supercoiled DNA. Concomitantly the chromatin fibre is remodelled to have a disrupted or 'open' chromatin fibre structure accompanied by large scale chromatin decompaction which we speculate may be important to for kinetochore assembly in mitosis. Canonical centromeres are typically composed of satellite DNA sequences which periodically position the underlying nucleosomes to assemble a regular and rigid chromatin fibre[25,35] that may withstand the forces exerted by microtubule pulling necessary for chromosome segregation. This two-domain model is then consistent with the idea of a flexible centromere core flanked by rigid pericentromeric chromatin (Fig. 5C).

Although pericentromeric chromatin surrounding a neocentromere is epigenetically remodelled and transcriptionally repressed, presumably by an activity emanating from the centromeric core, only a small region of pericentromeric chromatin had a compact structure, indicating that epigenetic remodelling is insufficient to form a compact chromatin fibre structure. Consistent with immaturity, neocentromeres exhibit a low level of aneuploidy suggesting that evolutionary new centromeres recruit satellite DNA sequences over time to form a stable chromatin platform[25] (Fig. 5C). We therefore suggest that centromere formation is accompanied by significant, transcriptional, epigenetic and chromatin fibre remodelling to form a suitable environment for kinetochore assembly and chromosome segregation in mitosis.

Acknowledgements

We would like to thank all members of our groups for useful discussions and to Alison Pidoux for critical comments on the manuscript. This work was funded by the UK Medical Research Council (MR/J00913X/1; MC_UU_00007/13).

Author Contributions

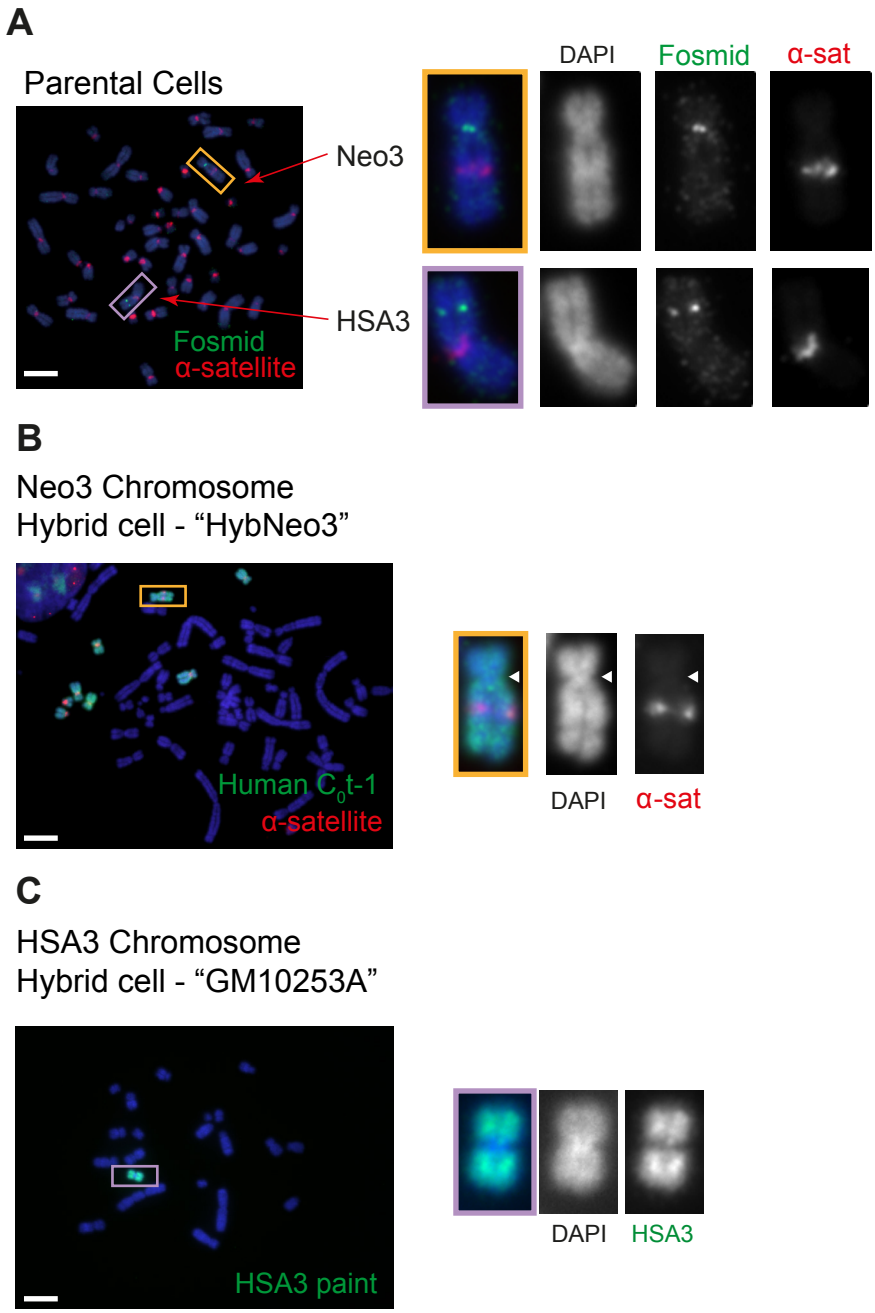
C.N., C.H., C.R.C., A.B., R.-S.N, S.P. undertook experiments; G.R.G and N.G. analysed data; M.R. and N.G. designed experiments; C.N. and N.G. wrote the manuscript with input from all authors.

References

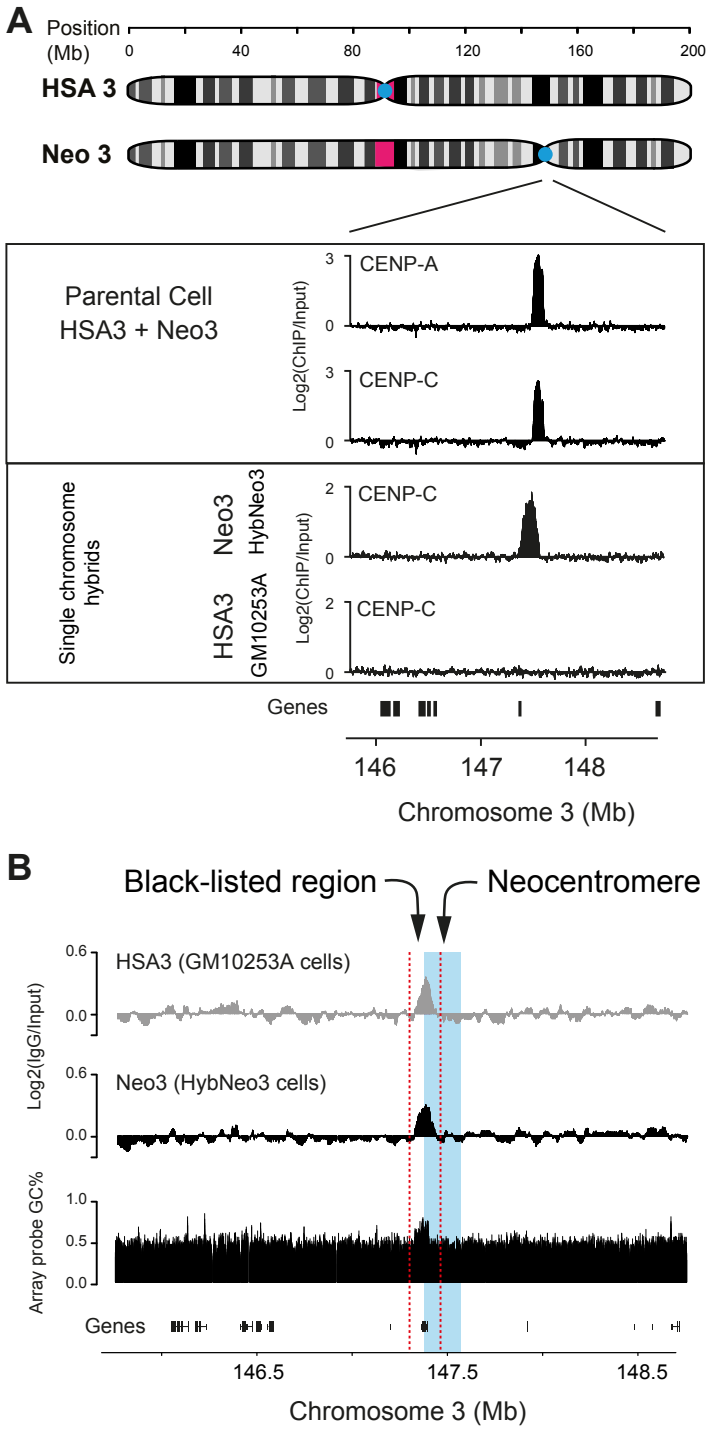
1. Black, B. E. & Cleveland, D. W. Epigenetic centromere propagation and the nature of CENP-A nucleosomes. *Cell* **144**, 471–479 (2011).
2. Willard, H. F. Chromosome-specific organization of human alpha satellite DNA. *American Journal of Human Genetics* **37**, 524–532 (1985).
3. Ventura, M. *et al.* Recurrent sites for new centromere seeding. *Genome Research* **14**, 1696–1703 (2004).
4. Voullaire, L. E., Slater, H. R., Petrovic, V. & Choo, K. H. A. A functional marker centromere with no detectable alpha-satellite, satellite III, or CENP-B protein: Activation of a latent centromere? *American Journal of Human Genetics* **52**, 1153–1163 (1993).
5. Depinet, T. W. *et al.* Characterization of neo-centromeres in marker chromosomes lacking detectable alpha-satellite DNA. *Human molecular genetics* **6**, 1195–1204 (1997).
6. Wandall, A., Tranebjaerg, L. & Tommerup, N. A neocentromere on human chromosome 3 without detectable alpha-satellite DNA forms morphologically normal kinetochores. *Chromosoma* **107**, 359–365 (1998).
7. Murillo-Pineda, M. *et al.* Induction of spontaneous human neocentromere formation and long-term maturation. *Journal of Cell Biology* **220**, (2021).
8. Purgato, S. *et al.* Centromere sliding on a mammalian chromosome. *Chromosoma* **124**, 277–287 (2015).
9. Yao, J. *et al.* Plasticity and epigenetic inheritance of centromere-specific histone H3 (CENP-A)-containing nucleosome positioning in the fission yeast. *The Journal of biological chemistry* **288**, 19184–96 (2013).

10. Hori, T. *et al.* Constitutive centromere-associated network controls centromere drift in vertebrate cells. *Journal of Cell Biology* **216**, 101–113 (2017).
11. Raisner, R. *et al.* Enhancer Activity Requires CBP/P300 Bromodomain-Dependent Histone H3K27 Acetylation. *Cell Reports* **24**, 1722–1729 (2018).
12. Chan, F. L. *et al.* Active transcription and essential role of RNA polymerase II at the centromere during mitosis. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 1979–1984 (2012).
13. Shukla, M. *et al.* Centromere DNA Destabilizes H3 Nucleosomes to Promote CENP-A Deposition during the Cell Cycle. *Current Biology* **28**, 3924–3936 (2018).
14. Singh, P. P. *et al.* Hap2-Ino80-facilitated transcription promotes de novo establishment of CENP-A chromatin. *Genes and Development* **34**, 226–238 (2020).
15. Bobkov, G. O. M., Gilbert, N. & Heun, P. Centromere transcription allows CENP-A to transit from chromatin association to stable incorporation. *Journal of Cell Biology* **217**, 1957–1972 (2018).
16. Chantalat, S. *et al.* Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin. *Genome research* **21**, 1426–1437 (2011).
17. Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome research* **19**, 24–32 (2009).
18. Narendra, V. *et al.* CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science (New York, N.Y.)* **347**, 1017–1021 (2015).
19. Paldi, F. *et al.* Convergent genes shape budding yeast pericentromeres. *Nature* **582**, 119–123 (2020).
20. Gilbert, N. *et al.* Chromatin architecture of the human genome: Gene-rich domains are enriched in open chromatin fibers. *Cell* **118**, 555–566 (2004).
21. Naughton, C., Sproul, D., Hamilton, C. & Gilbert, N. Analysis of Active and Inactive X Chromosome Architecture Reveals the Independent Organization of 30 nm and Large-Scale Chromatin Structures. *Molecular Cell* **40**, 397–409 (2010).
22. Horz, W., Fittler, F. & Zachau, H. G. Sequence specific cleavage of African green monkey alpha-satellite DNA by micrococcal nuclease. *Nucleic acids research* **11**, 4275–4285 (1983).
23. Linxweiler, W. & Horz, W. Reconstitution experiments show that sequence-specific histone-DNA interactions are the basis for nucleosome phasing on mouse satellite DNA. *Cell* **42**, 281–290 (1985).
24. Ghirlando, R. & Felsenfeld, G. Hydrodynamic studies on defined heterochromatin fragments support a 30-nm fiber having six nucleosomes per turn. *Journal of molecular biology* **376**, 1417–1425 (2008).
25. Gilbert, N. & Allan, J. Distinctive higher-order chromatin structure at mammalian centromeres. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 11949–11954 (2001).
26. Furuyama, T. & Henikoff, S. Centromeric Nucleosomes Induce Positive DNA Supercoils. *Cell* **138**, 104–113 (2009).
27. Gilbert, N. & Allan, J. Supercoiling in DNA and chromatin. *Current Opinion in Genetics and Development* **25**, 15–21 (2014).
28. Naughton, C. *et al.* Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nature Structural and Molecular Biology* **20**, 387–395 (2013).
29. Tolomeo, D. *et al.* Epigenetic origin of evolutionary novel centromeres. *Scientific Reports* **7**, 41980–41980 (2017).
30. Maison, C. *et al.* Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. **30**, 329–334 (2002).
31. Saffery, R. *et al.* Human centromeres and neocentromeres show identical distribution patterns of >20 functionally important kinetochore-associated proteins. *Human Molecular Genetics* **9**, 175–185 (2000).
32. Fachinetti, D. *et al.* DNA Sequence-Specific Binding of CENP-B Enhances the Fidelity of Human Centromere Function. *Developmental Cell* **33**, 314–327 (2015).

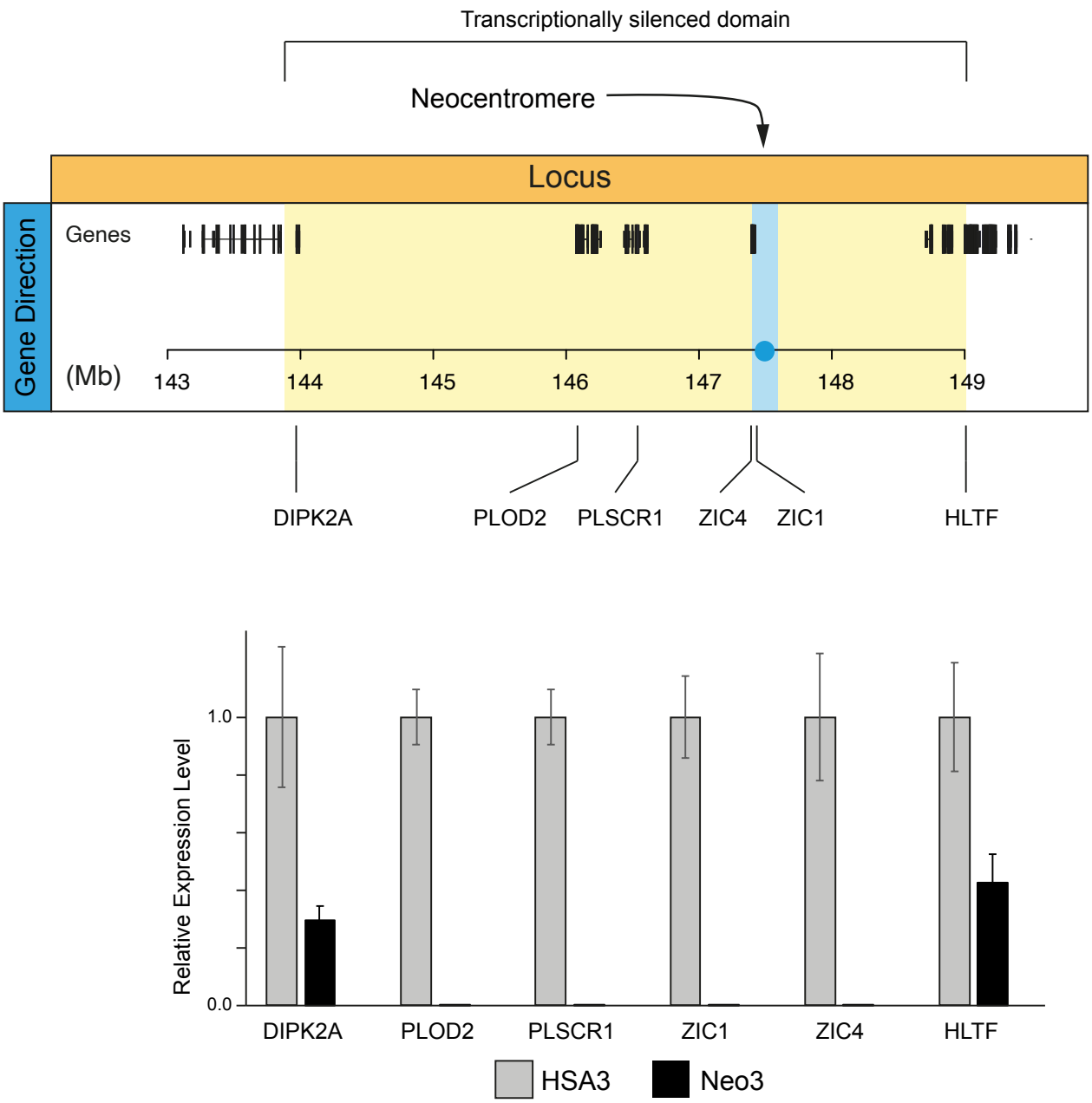
33. Bassett, E. A. *et al.* Epigenetic centromere specification directs aurora B accumulation but is insufficient to efficiently correct mitotic errors. *Journal of Cell Biology* **190**, 177–185 (2010).
34. Carroll, C. W., Milks, K. J. & Straight, A. F. Dual recognition of CENP-A nucleosomes is required for centromere assembly. *Journal of Cell Biology* **189**, 1143–1155 (2010).
35. Kruithof, M. *et al.* Single-molecule force spectroscopy reveals a highly compliant helical folding for the 30-nm chromatin fiber. *Nature structural & molecular biology* **16**, 534–540 (2009).
36. Mitchell, A. R., Gosden, J. R. & Miller, D. A. A cloned sequence, p82H, of the alphoid repeated DNA family found at the centromeres of all human chromosomes. *Chromosoma* **92**, 369–377 (1985).
37. Kimura, H., Hayashi-Takanaka, Y., Goto, Y., Takizawa, N. & Nozaki, N. The organization of histone H3 modifications as revealed by a panel of specific monoclonal antibodies. *Cell structure and function* **33**, 61–73 (2008).
38. Trazzi, S. *et al.* The C-terminal domain of CENP-C displays multiple and critical functions for mammalian centromere formation. *PloS one* **4**, e5832 (2009).
39. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078–2079 (2009).
40. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841–842 (2010).
41. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)* **30**, 923–930 (2014).



Extended Data Fig. 1. Cell line model system used to interrogate centromeric chromatin structure. **(A)** Parental human lymphoblastoid cells with one canonical chromosome 3 (HSA3, purple frame) and one chromosome 3 with the centromere relocated to form a neocentromere at 3q24 (Neo3, orange frame). DNA FISH with α-satellite specific probe (red) and a 3q24 fosmid probe (green). **(B)** The Neo3 chromosome was retained after fusion of the parental cell line with a hamster cell to create a hybrid line called “HybNeo3”. Metaphase spread of HybNeo3 hybridized with human C_0t-1 DNA (green) identifying the seven human chromosomes present in this human/hamster hybrid (4,6,8,11,13,18, X and Neo3 (orange frame)) and a human α-satellite specific FISH probe (red). **(C)** The GM10253A hybrid cell line has a single canonical human chromosome 3 (HSA3). Metaphase spread of GM10253A hybridized with a human chromosome 3 paint (green; purple frame). Nuclei are counterstained with DAPI. Bar is 5 μm.

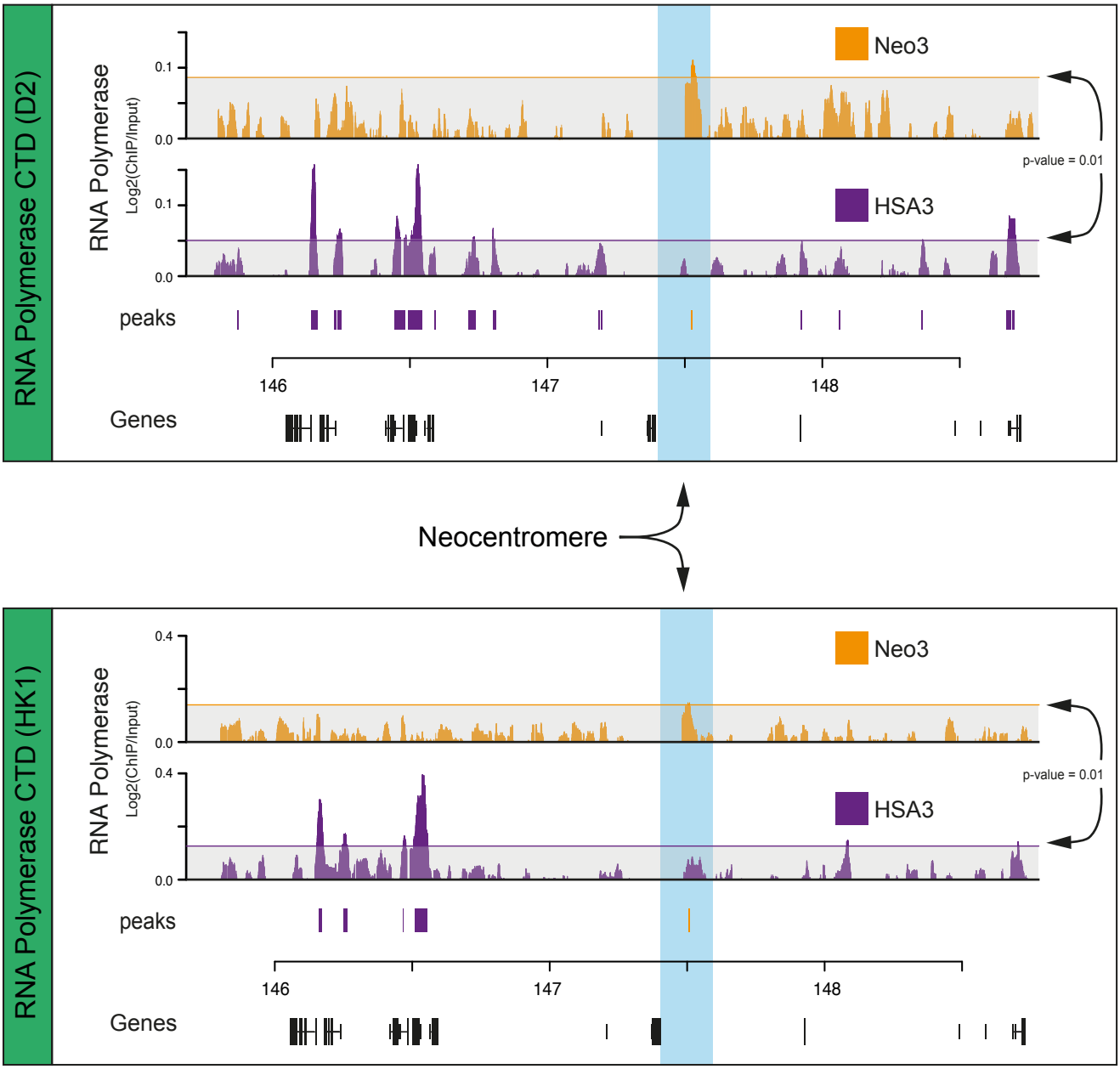


Extended Data Fig. 2. Neocentromere on Neo3 maps to 3q24. (A) Top, ideogram depicting chromosomes HSA3 and Neo3. Bottom, distribution of CENP-A and CENP-C ChIP signal in parental and hybrid cells at 3q24. **(B)** Top, signal for control IgG ChIP-chip and bottom microarray probe GC composition (%). Blacklisted region (chr3: 147324413-147482213; hg38) is marked by red dashed lines. Neocentromere core (Chr3: 147400413-147591023) defined from CENP-C (panel A) is marked in blue.

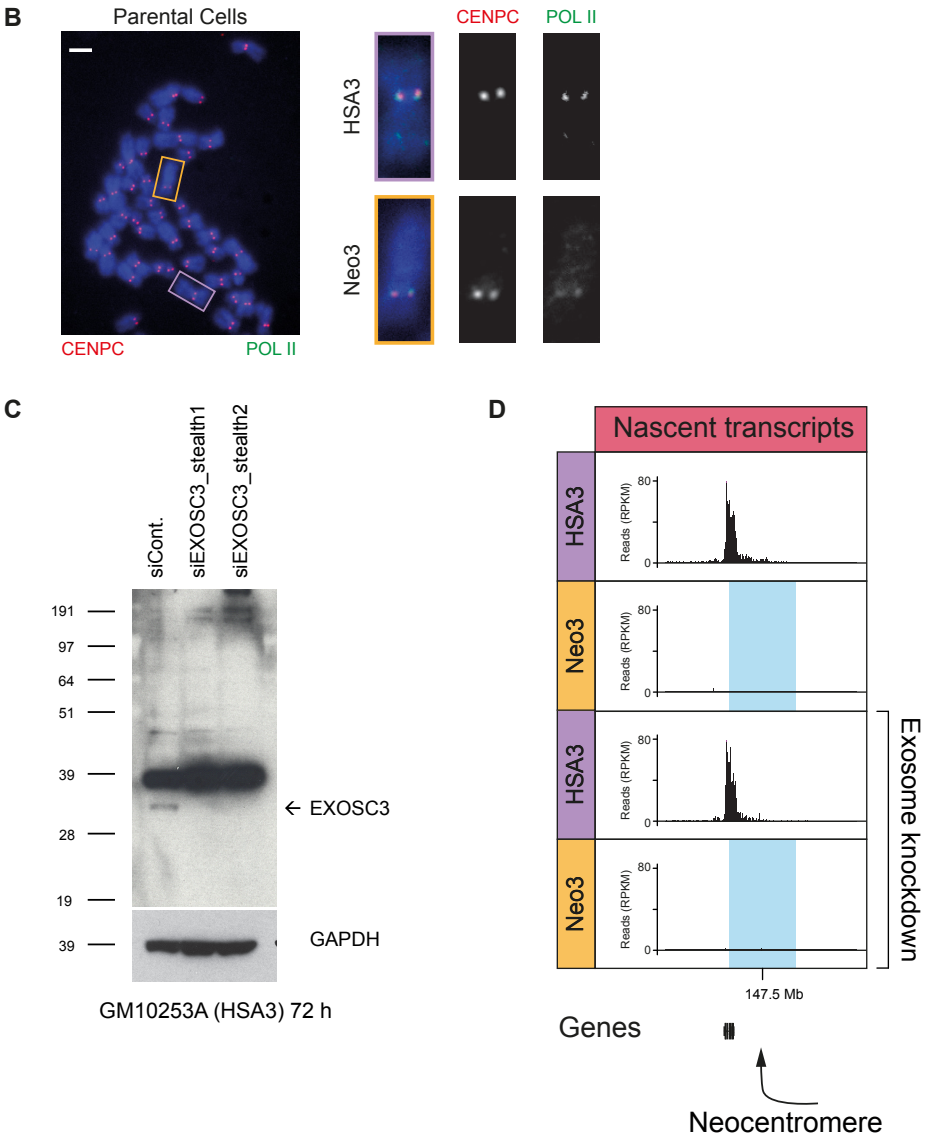


Extended Data Fig. 3. Transcriptional repression at the pericentromeric domain. Top, diagram showing individual genes at 3q24, yellow block corresponds to pericentromere domain, blue is the neocentromere. Bottom, RT-qPCR expression data for genes within and bordering the pericentromeric heterochromatic domain, showing expression from HSA3 and gene silencing on Neo3.

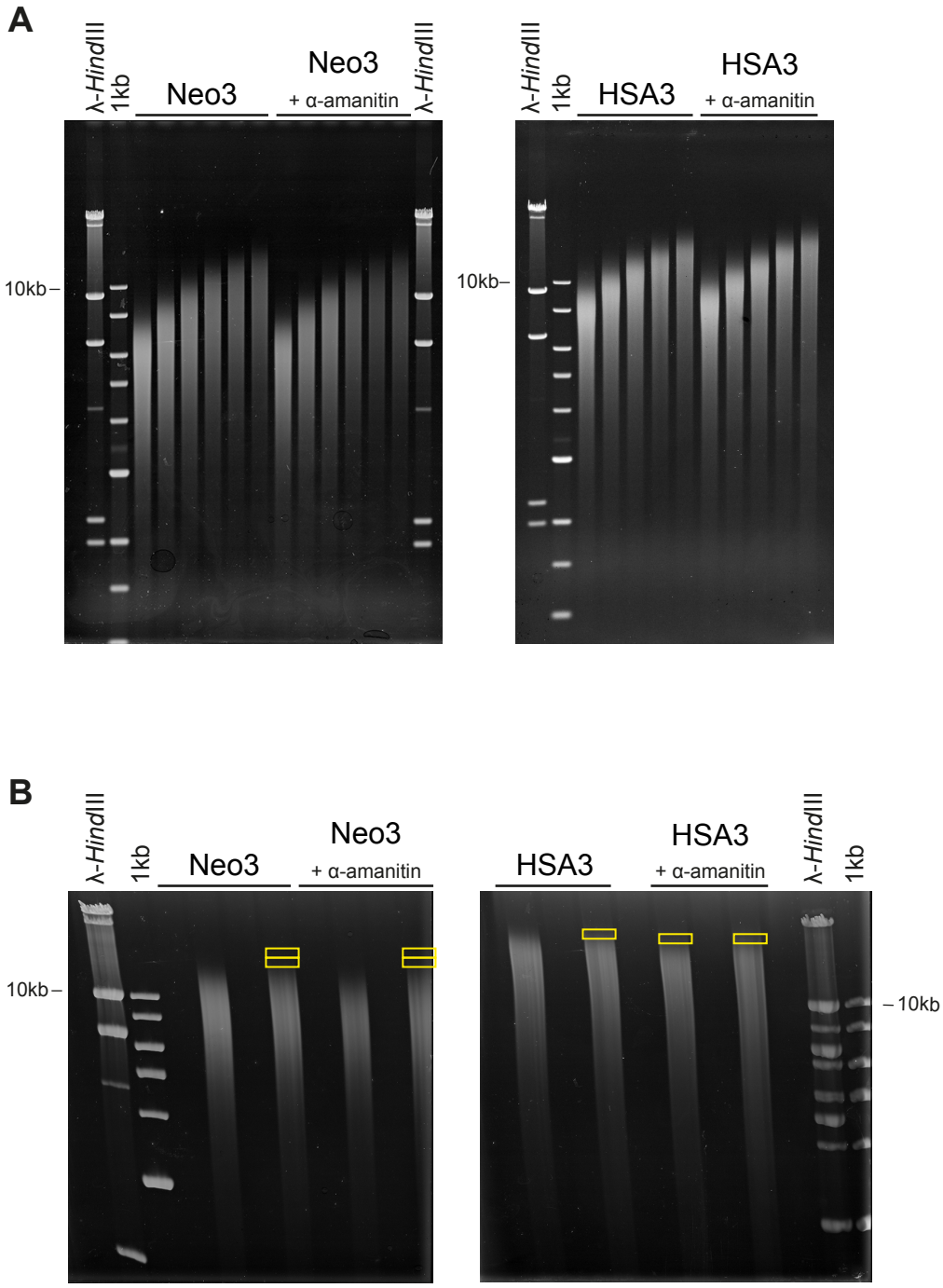
A



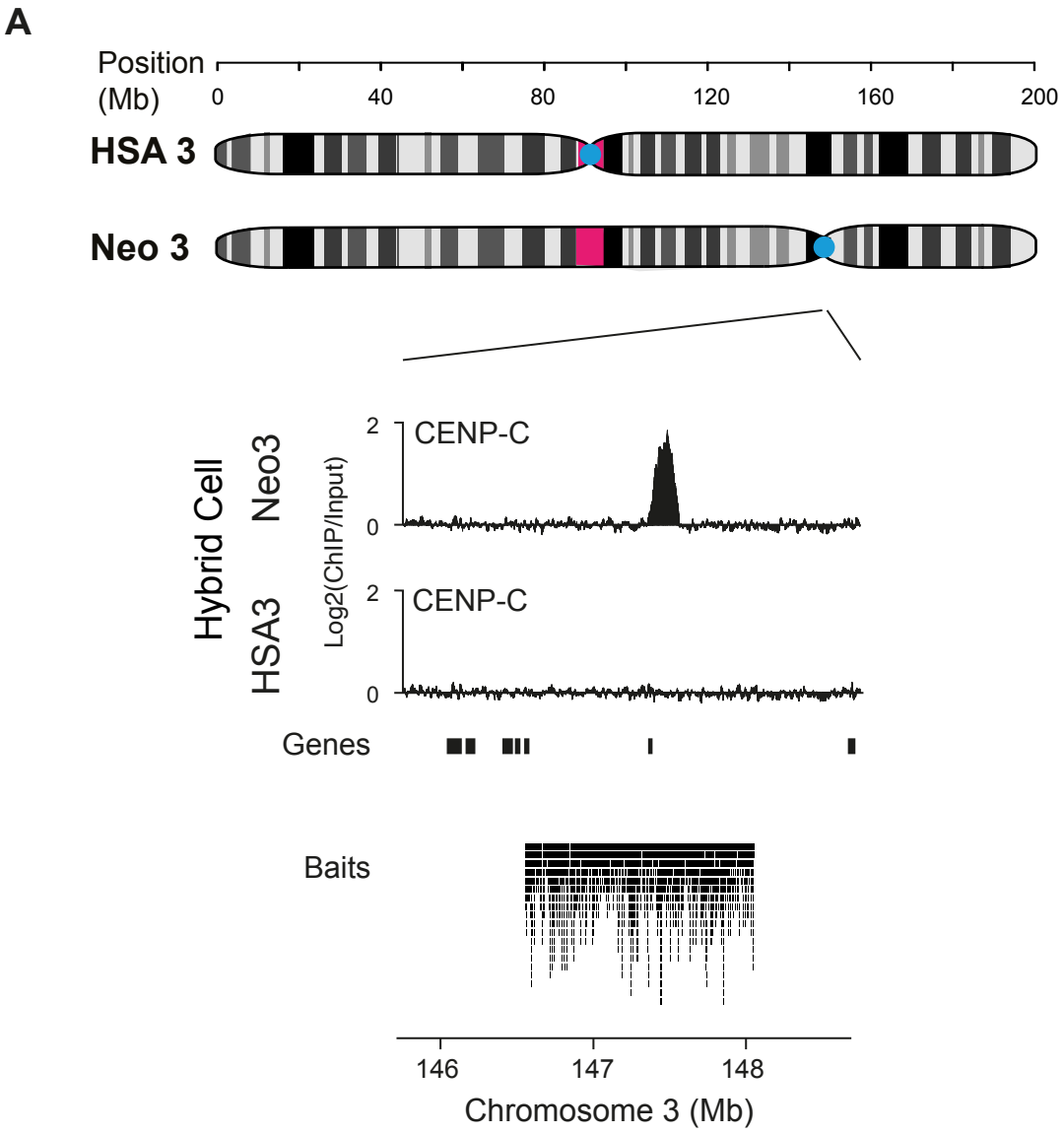
Naughton et al., Extended Data Figure 4



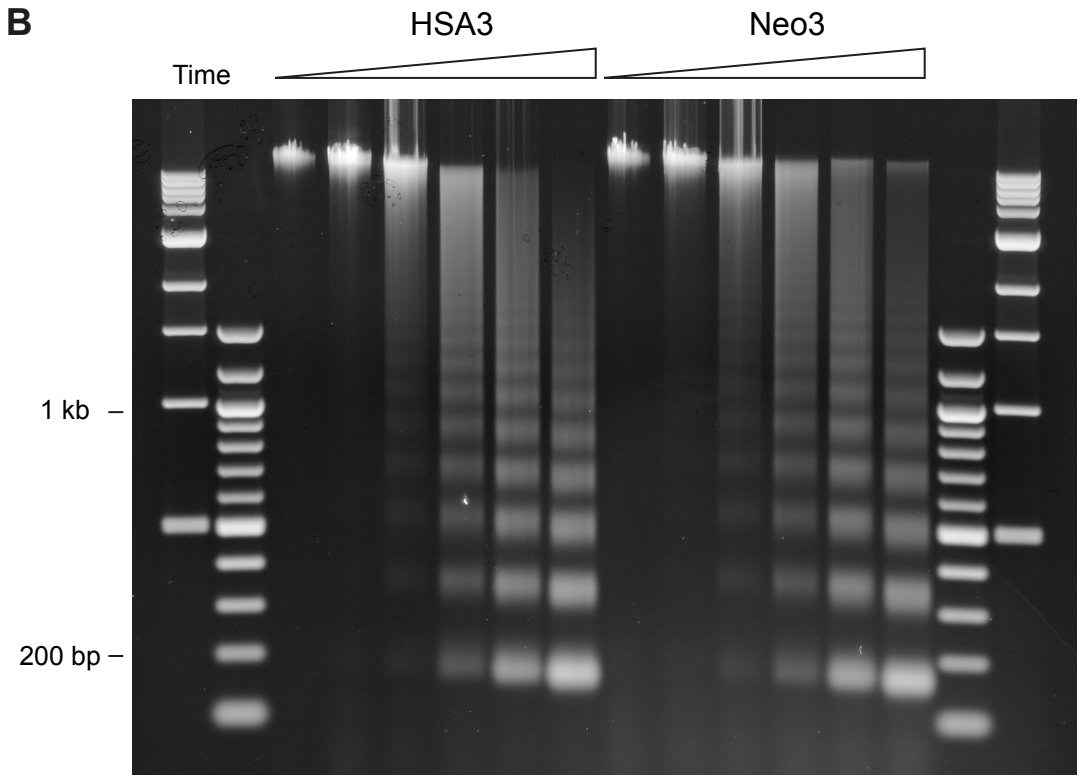
Extended Data Fig. 4. RNA pol II binding at a functioning human centromere. (A) Distribution of RNA pol II binding at 3q24 for Neo3 (orange) and HSA3 (purple) chromosomes. Top, RNA pol II CTD antibody (D2, Diagenode), bottom, RNA pol II CTD antibody from Hiroshi Kimura (HK1). Horizontal line (arrows) corresponds to a sampled background model with p = 0.01. Significant peaks (p<0.001) called using the Ringo package for Neo3 (orange) and HSA3 (purple) are shown below tracks. **(B)** RNA pol II (CTD domain, Diagenode) (green) and CENP-C (gift, W. Earnshaw) (red) immunofluorescence staining on a metaphase spread from parental cells (HSA3, purple; Neo3, orange). Bar is 5 µm. **(C)** Western blot confirming EXOSC3 protein (arrow) knockdown following 72 h RNAi treatment in GM10253A cells. **(D)** Distribution of nascent transcripts (normalised RPKM) in 600 kb window around neocentromere (blue), mapped using TT-seq. Top panels correspond to RNAi control, bottom panels are for exosome RNAi knockdown.

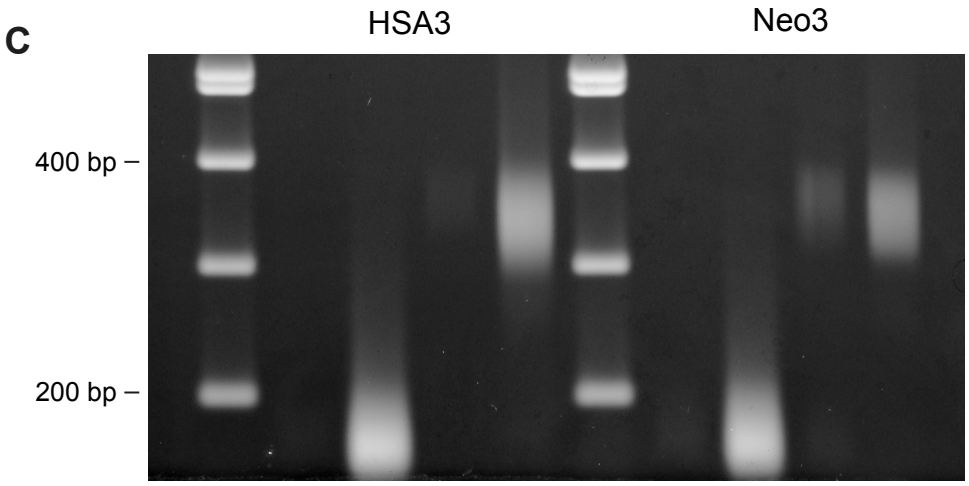


Extended Data Fig. 5. Isolation of "open" chromatin probes. Soluble chromatin isolated from hybrid cell nuclei by MNase digestion was fractionated by size and structure on a sucrose gradient. **(A)** Top, agarose gel electrophoresis of DNA purified from sucrose gradient fractions. **(B)** DNA from gradient fractions was size selected by PFGE. DNA fragments ≈10 kb longer than the bulk of the DNA signal, corresponding to "open" or disrupted chromatin, was purified from gel slices (yellow boxes).



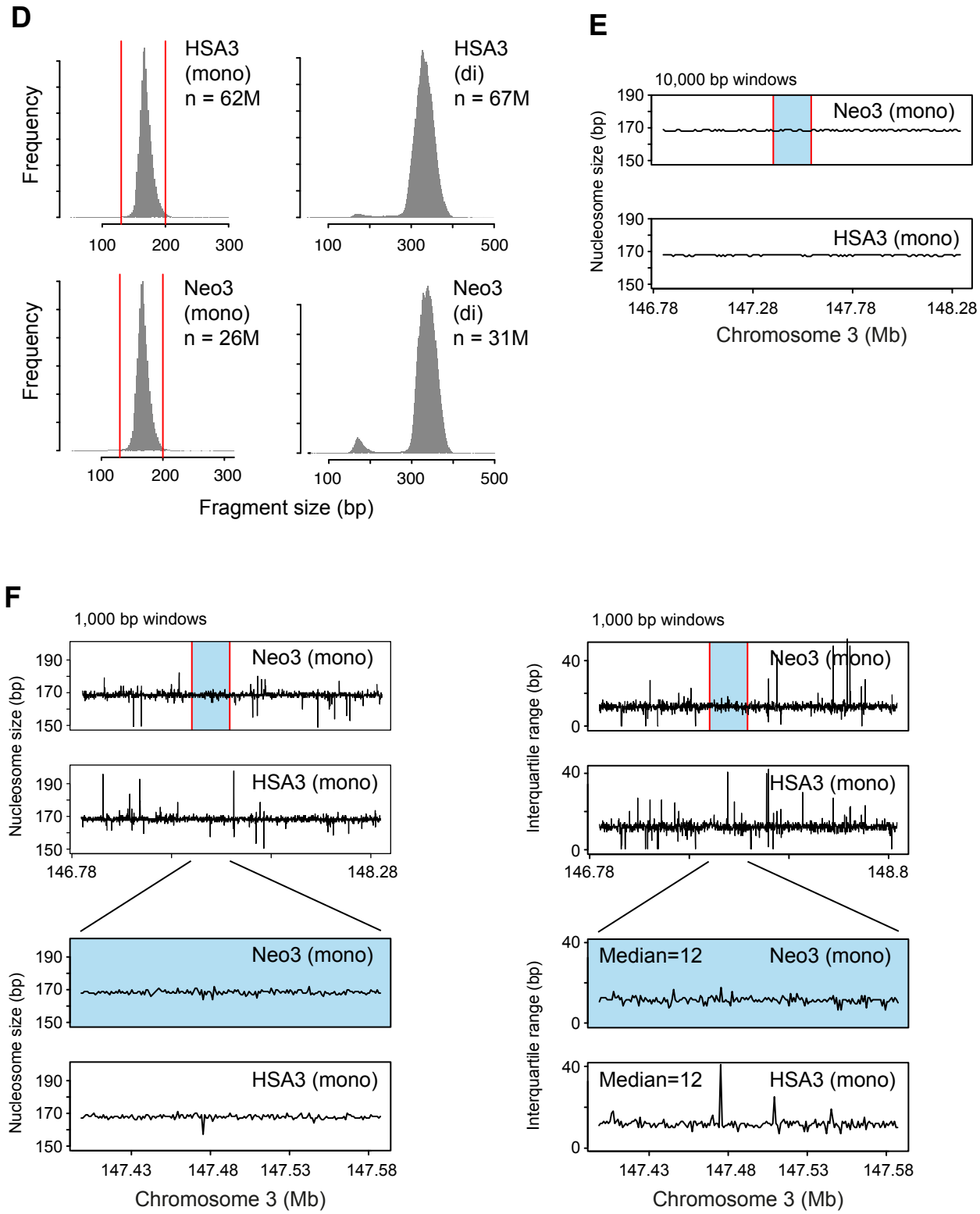
Naughton et al., Extended Data Figure 6



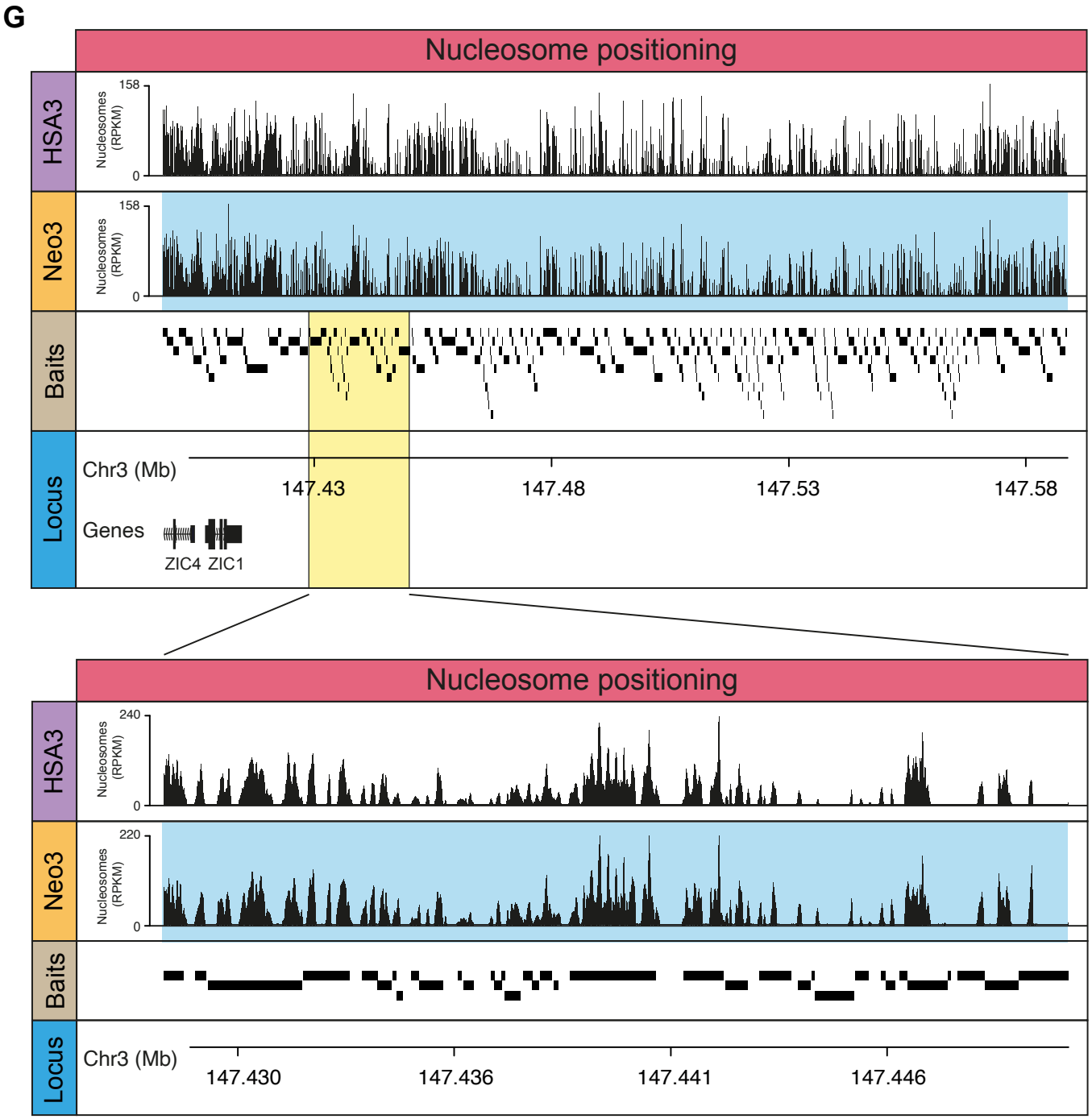


Gel purified fragments

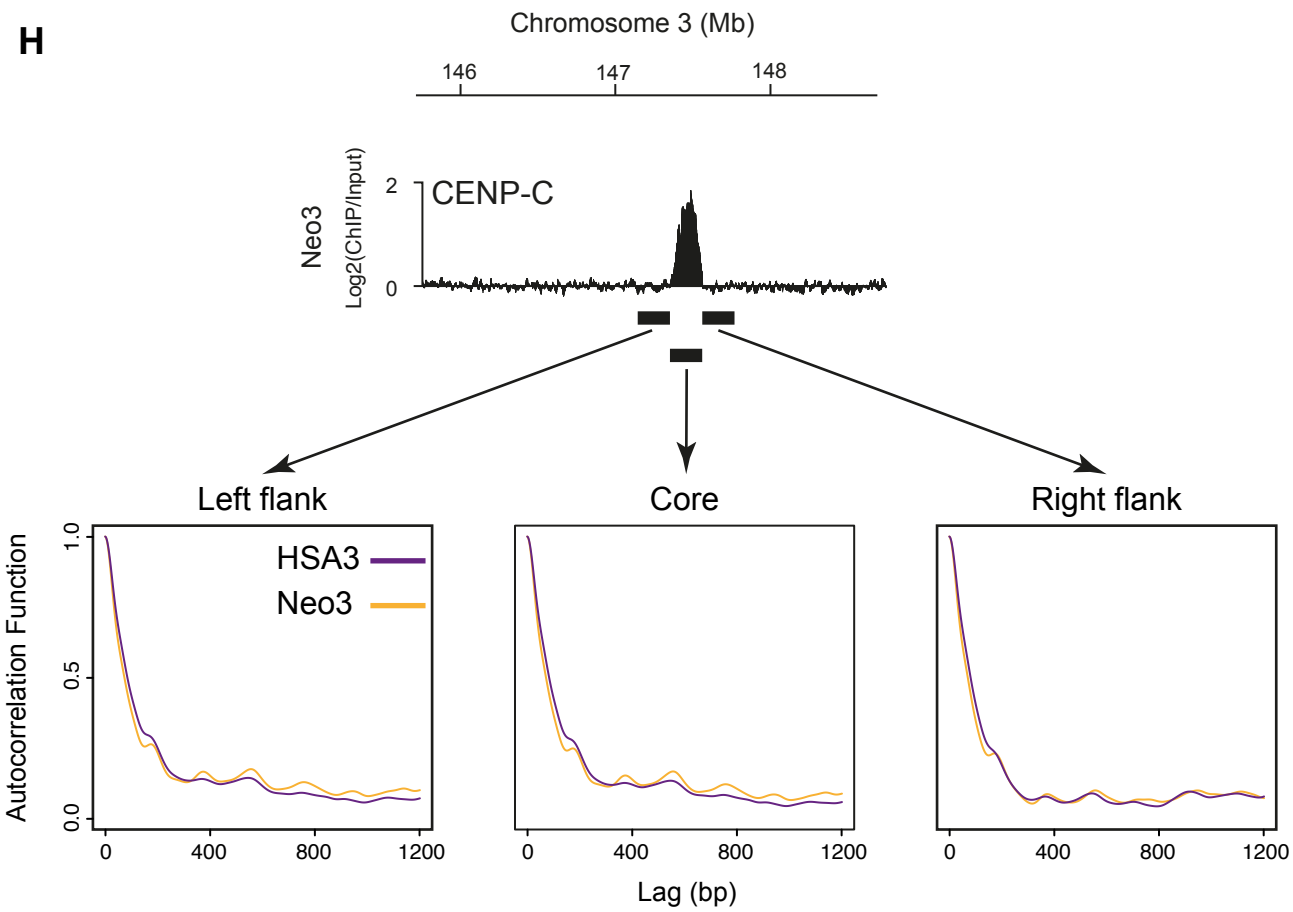
Naughton et al., Extended Data Figure 6



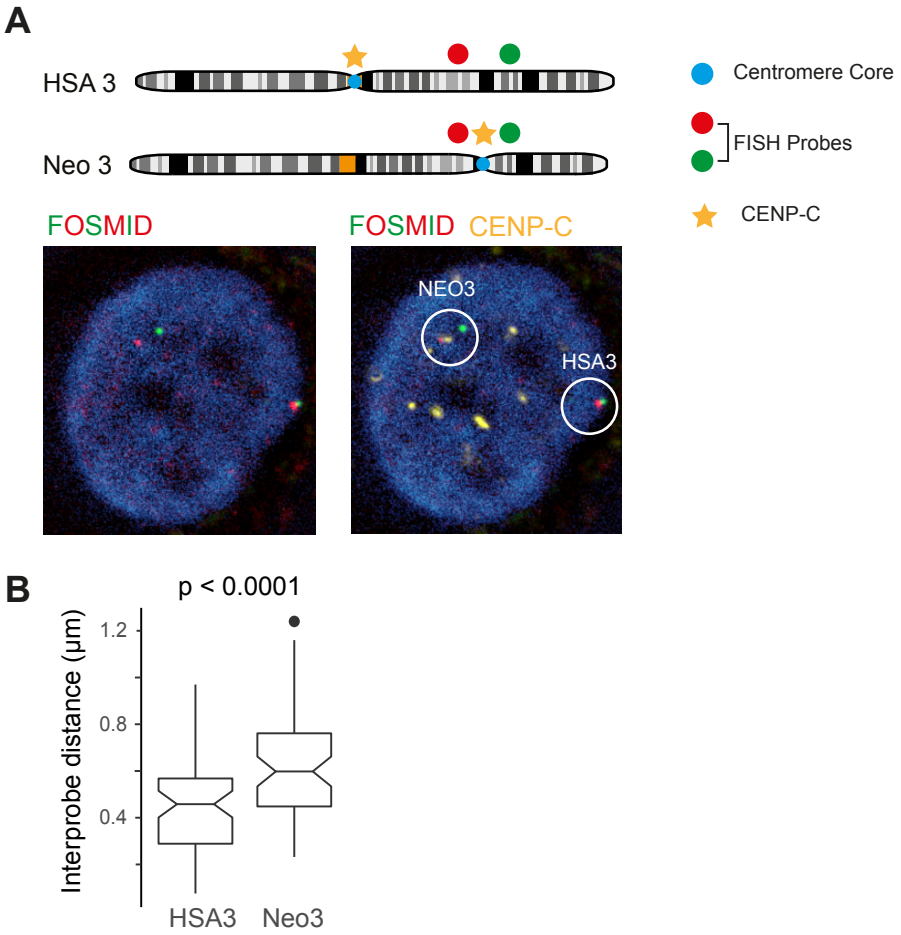
Naughton et al., Extended Data Figure 6



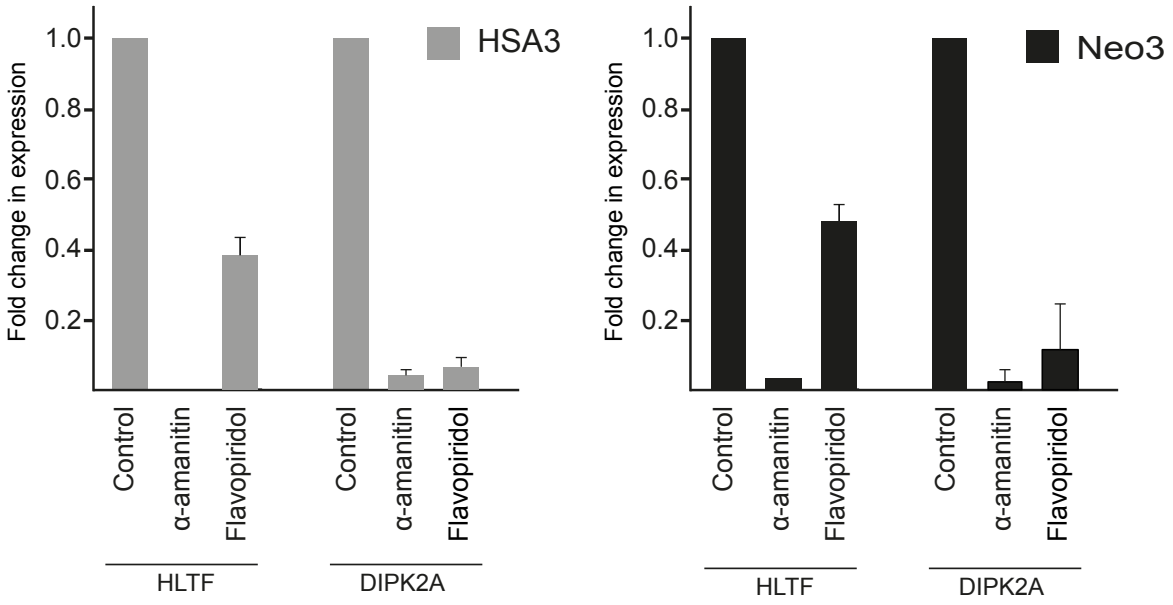
Naughton et al., Extended Data Figure 6



Extended Data Fig. 6. Arrangement of nucleosomes around neocentromere (A) Top, ideogram depicting HSA3 and Neo3 chromosomes with enlargement of 3 Mb region around the neocentromere (marked by CENP-C). Bottom, genomic location of the capture baits used to enrich for 1.5 Mb of neocentromeric region. **(B)** Agarose gel electrophoresis of DFF digested nuclei isolated from HSA3 and Neo3 containing cells. Mono and di-nucleosome fragments were excised, and the DNA extracted by β-agarase treatment. **(C)** Agarose gel electrophoresis of purified mono- and di-nucleosomes fragments used for nucleosome mapping. **(D)** Size distribution of mono and di nucleosomes (bp) isolated from HSA3 and Neo3 cells for 1.5 Mb around the region corresponding to the neocentromere. **(E)** Mono nucleosome size distribution in 10 kb windows across the 1.5 Mb captured domain (neocentromere marked in blue). (**F)** Left, mono nucleosome size (median) in 1 kb windows across the 1.5 Mb captured domain and focussed region covering the neocentromere (marked in blue). Right, variance (interquartile range) in mono nucleosome size in 1 kb windows across the 1.5 Mb captured domain and focussed region covering the neocentromere (marked in blue). **(G)** Nucleosome dyad coverage on the HSA3 and Neo3 chromosome at the neocentromere domain (blue) with enlarged region (below) **(H)** Autocorrelation of nucleosome dyad coverage at left flank, centromere core and right flank for different lag (bp).

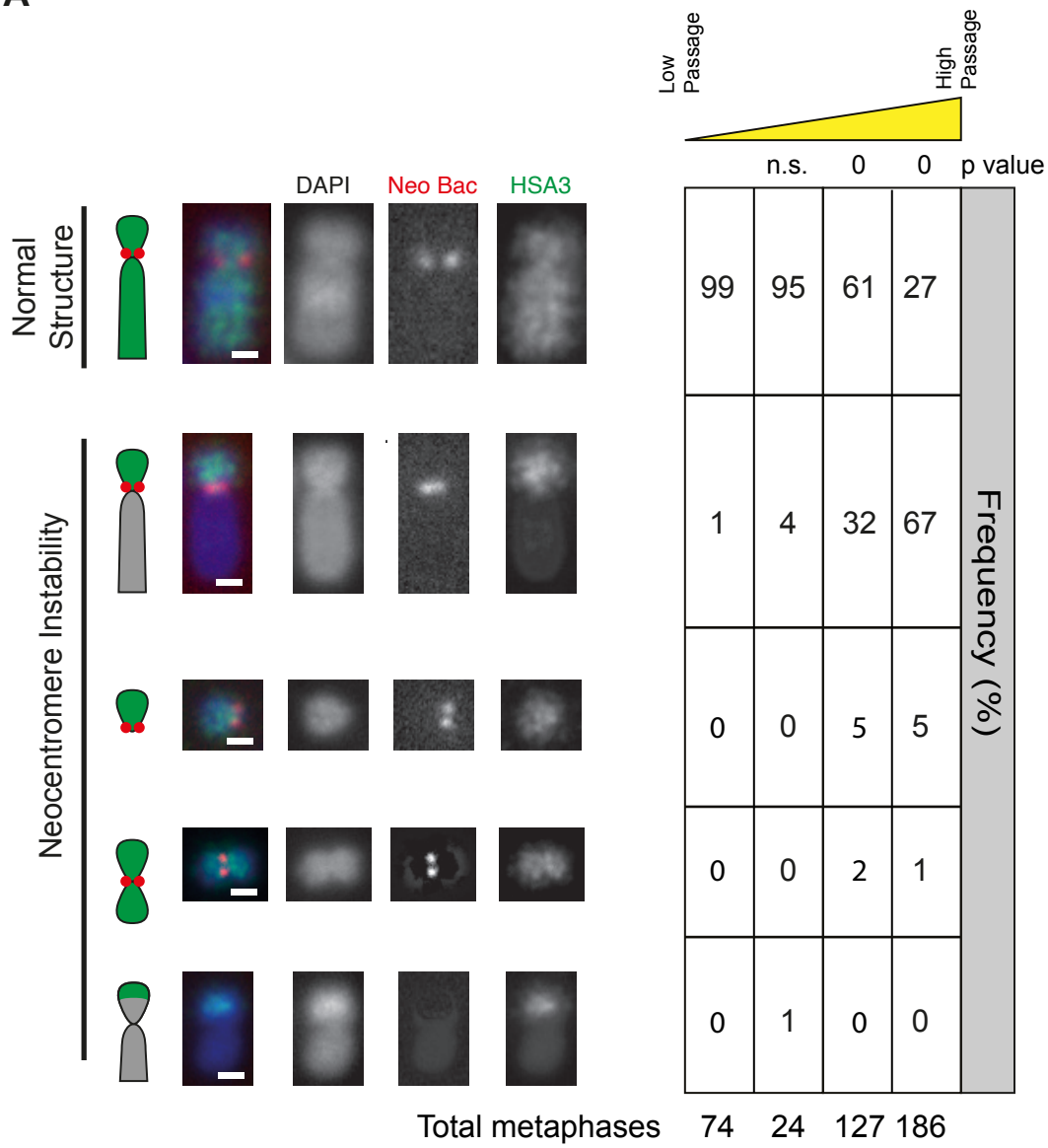


Extended Data Fig. 7. Large scale chromatin is decompacted at the neocentromere in the parental cell line. (A) Top, chromosome 3 ideogram indicating CENP-C immunofluoresence signal (yellow) and the neocentromere specific FISH probes (green and red). Bottom right, representative image of 4 colour 3D immuno-FISH for identifying the chromosome 3 harbouring a neocentromere at 3q24 due to the presence of the CENP-C signal proximal to one pair of fosmid probes. Below left, three colour representation of the same image used for measuring interprobe distance. **(B)** Boxplot showing interprobe distance measurements (µm) between the pair of fosmid probes (B and C, see Fig 4A) for the HSA3 and Neo3 chromosomes in the parental lymphoblastoid cells.

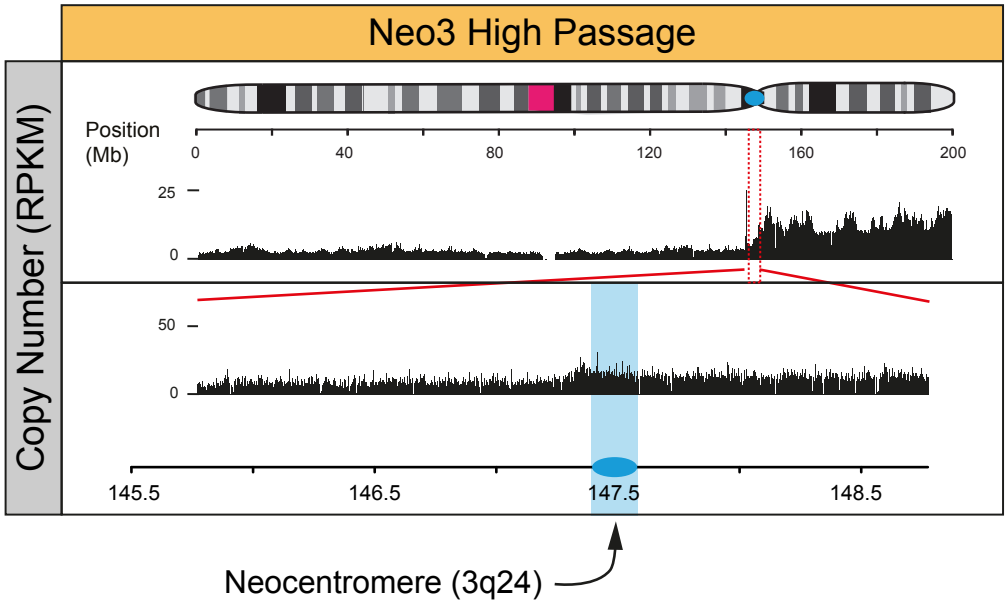


Extended Data Fig. 8. Transcription inhibition. RT-qPCR expression data for genes in the pericentromere region flanking the neocentromere domain in cells carrying the HSA3 and Neo3 chromosomes, following transcription inhibition with α-amanitin (5h) or flavopiridol (3h).

A



B



Naughton et al., Extended Data Figure 9

Extended Data Fig. 9. Neocentromere associated genome instability (A) Left, representative FISH images of Neo3 metaphase chromosomes hybridised to a human chromosome 3 paint (green) and BAC (red) located at the neocentromere. Chromosome morphology was scored as normal or showing instability: deletions, fusions or duplications. Bar is 2μm. Right, quantification (%) of different chromosome morphologies with increasing passage number (low ~ 10, high ~ 100) over time. P values are for a χ^2 test compared to low passage. **(B)** High passage Neo3 Chromosome copy number (RPKM), with a zoom in of the 3 Mb region around the neocentromere (blue). **(C)** High passage Neo3 Chromosome copy number (RPKM), with a zoom in of the 7 Mb region around the neocentromere (blue). Locations of DNA FISH probes are shown, neocentromere BAC probe (green) and border fosmid (red). **(D)** Left, representative FISH images of Neo3 metaphase chromosomes hybridised to a BAC (green) located at the neocentromere and a fosmid (red) located at the border. Bar is 2μm. Right, chromosome morphology was scored and quantified (%) with increasing passage number over time. Loss of the border probe signal was coincident with fusion of human Neo3 fragment (light grey) to a hamster chromosome (dark grey). Bar is 2μm.

Supplementary Table 1
Fosmid and BAC probes used in this study

FISH Probe	ID	Probe location (HG38)	Genomic Band	Type
G248P81751C5	A	chr3:148152154-148190706	3q24	Fosmid
G248P87211G9	B	chr3:147386613-147424083	3q24	Fosmid
G248P87849H8	C	chr3:147698288-147739058	3q24	Fosmid
G248P8432C9	D	chr3:146792070-146832279	3q24	Fosmid
G248P8026G2	G	chr3:147128170-147169864	3q24	Fosmid
RP11-1068F8	Neocentromere BAC	chr3:147402395-147564508	3q24	BAC
WI2-2366F2	Border Fosmid	chr3:144079778-144118953	3q24	Fosmid
p82H	a-satellite			Plasmid

Fosmids and BACs were obtained from BacPac resources, DNA co-ordinates are hg38

Materials and Methods

Cell lines

The human parental lymphoblastoid cell line was grown in RPMI 1640 with L-Glutamine (Life Technologies) supplemented with 20% FCS, penicillin (100 U.ml^{-1}) and streptomycin (100 µg.ml^{-1}). Human/Hamster hybrid cell lines GM10253A and HybNeo3, harbouring HSA3 and Neo3 respectively, were grown in the same media but with 10% FCS. All cells were maintained at 37 °C in an atmosphere of 5% CO2 and subjected to regular mycoplasma testing. Transcription was blocked by adding α-amanitin (50 µg.ml^{-1}) or flavopiridol (100 µM) to cells for the times indicated.

DNA fluorescence *in situ* hybridization (FISH)

FISH was performed on both metaphase chromosome spreads and interphase nuclei. Metaphase chromosomes were prepared by treating cells with 0.1 µg.ml^{-1} Colcemid (Life Technologies, Cat No 15210-040) for 30 min (hybrid cells) or 4 hr (parental lymphoblastoid cells) prior to harvest to induce mitotic arrest and increase the number of mitotic cells. Cells were recovered by trypsin treatment and washed in PBS. Hypotonic solution, containing 75 mM KCl was added drop wise to a final 5 ml volume. Hypotonic treatment was performed at room temperature for 10 min, after which cells were pelleted by centrifugation at 1200 rpm (250g) for 5 min and fixed three times in 5 ml of a freshly prepared solution of 3:1 ratio (v/v) methanol: acetic acid (MAA). The MAA fixative was added to the cell pellet dropwise with constant agitation. Chromosome preparations were stored at −20°C. To prepare slides with metaphase spreads, metaphase chromosome preparations were dropped onto glass slides. The glass slides were pre-treated in a dilute solution of HCl in Ethanol for at least one hour prior to use. The chromosome preparations were pelleted by centrifugation at 1500 rpm (350g) for 5 min and resuspended in freshly prepared MAA solution until the suspension became cloudy. Two drops of the suspension were dropped onto a pre-treated glass slide from a height of 20 cm and dried at room temperature overnight before staining or hybridization.

For 3D FISH on interphase nuclei hybrid cells were grown overnight on glass slides whilst parental non-adherent lymphoblastoid cells (3 x 10^4 cells) were cytospun onto glass slides at 600 rpm (50g) for 10 min. Slides were rinsed with PBS and fixed in 4% paraformaldehyde (PFA) for 10 min. Slides were then rinsed with PBS and cells were permeabilized for 10 min on ice with PBS supplemented with 0.2% triton. After rinsing, slides were stored in PBS (for immunohistochemistry) or 70% ethanol (for FISH) at 4 °C. For chromatin nicking and RNase H treatment cells were grown on slides overnight, rinsed gently whilst still in the slide tray three times with PBS and then treated with bleomycin (100 µM) in PBS or RNase H (100U.ml^{-1}, NEB MO297S) in PBS supplemented with 0.1% triton, 1mM Ca^{2+} and 1mM Mg^{2+} for 10 min at 37°C. Slides were then rinsed with PBS and PFA fixed as before.

FISH was carried out as described[27] except that MAA fixed metaphase spreads were denatured for 1 min in 70% formamide in 2× SSC, pH 7.5, at 70°C and interphase cells (grown on glass slides or cytospun) were 4% PFA fixed and were denatured for 45 min at 80 °C. Following denaturation, slides were submerged in

ice-cold 70% ethanol for 2 min and then dehydrated through 90% and 100% ethanol for 2 min each at room temperature. Fosmid and BAC clones (BACPAC Genomics) were labelled by nick translation with digoxigen-in-11-dUTP (Roche, #11093070910) or biotin-16-dUTP (Roche, #11093088910) for antibody based detection as previously described or alternatively directly labelled with Green 500 dUTP (Enzo-42845) or red-dUTP (ChromaTide Alexa Fluor 594-5-dUTP C11400). α-satellite probe p82H[35] was labelled by nick translation. For hybridization, 150 ng of labelled probe was combined with 5 µg salmon sperm and 10 µg human C_0t1 DNA (Invitrogen, Cat No 15279011). Two volumes of ethanol were added and the probe mix was collected by centrifugation and dried. Dried probes were resuspended in 10 µl of hybridization buffer containing 50% formamide (v/v), 1% Tween-20 and 10% dextran sulfate (Sigma Aldrich, Cat No D8906-100G) in 2 x SSC. Chromosome 3 paint (XCP3 Green, Metasystems) was supplied already labelled with a green fluorophore and dissolved in hybridization solution and ready to use. Probes were denatured at 70°C for 5 min and re-annealed at 37°C for 15 min and chilled on ice. Probes were pipetted onto slides and hybridization was performed at 37°C overnight. Coverslips were removed and slides were washed four times in 2 x SSC at 45°C for 3 min and four times in 0.1 x SSC at 60°C for 3 min. Slides were blocked in 5% milk in 4 x SSC for 5 min at RT. Detection of biotin label was performed with sequential layers of fluorescein (FITC)-conjugated avidin, biotinylated anti-avidin and a further layer of FITC-avidin. Digoxigenin was detected with sequential layers of Rhodamine-conjugated anti-digoxigenin and Texas-Red (TR) –conjugated anti-sheep IgG. Slides were DAPI stained, mounted in Vectashield (Vector Laboratories, Cat No H-1000) Epifluorescent images were acquired using a Photometrics Coolsnap HQ2 CCD camera on a Zeiss Axioplan II fluorescence microscope with a Plan-neofluar/apochromat 100x objective (Carl Zeiss, Cambridge, UK), a Mercury Halide fluorescent light source (Exfo Excite 120, Excelitas Technologies) and Chroma #83000 triple band pass filter set (Chroma Technology Corp., Rockingham, VT) with the single excitation and emission filters installed in motorised filter wheels (Prior Scientific Instruments, Cambridge, UK). Data was collected using Micromanager software and analyzed using custom scripts in iVision.

For four-color 3D Immuno-FISH immunocytochemistry for CENP-C was performed prior to FISH. Slides stored in PBS were blocked in 5% horse serum then incubated overnight with anti-CENP-C antibody (1:200, gift, W. Earnshaw) before 1 hr incubation with Texas Red labelled anti-rabbit (1:100, Jackson ImmunoResearch Laboratories) secondary antibody. CENP-C signal was fixed with 4% PFA for 45 min followed by denaturation with 50% formamide in 2× SSC, pH 7.5, at 80°C for 45 min. Slides were then dipped briefly in 2 x SSC followed by incubation overnight at 37°C with pairs of labelled fosmid probes. Slides were then washed and processed as above. Epifluorescent images were acquired using a Photometrics Coolsnap HQ2 CCD camera and a Zeiss AxioImager A1 fluorescence microscope with a Plan Apochromat 100x 1.4NA objective, a Nikon Intensilight Mercury based light source (Nikon UK Ltd, Kingston-on- Thames, UK) and a Chroma 89000ET single excitation and emission filters (Chroma Technology Corp., Rockingham, VT) with the excitation and emission filters installed in Sutter motorised filter wheels (Sutter Instrument, Novato, CA). A piezo-electrically driven objective mount (PIFOC model P-721, Physik Instruments GmbH & Co, Karlsruhe) was used to control movement in the z dimension. Hardware control, image capture and analysis were performed using Nikon Nis-Elements software (Nikon UK Ltd, Kingston-on-Thames, UK) and Volocity (Perkinelmer, Inc.). Images were deconvolved using a calculated point spread function with the constrained iterative algorithm of Volocity. Image analysis was carried out using Imaris software that calculate the distance between two fosmid probe signals.
The significance of compaction between pairs of probes was tested using the nonparametric Wilcoxon test for paired samples, $P < 0.05$ was considered significant.
FISH probes are described in Supplementary Table 1.

Immunocytochemistry
Metaphase chromosome spreads derived from parental cells were rinsed in PBS, blocked in 5% horse serum then incubated overnight with anti-CENP-C antibody (1:200, gift from W. Earnshaw) and anti-RNA pol II (1:1000, Abcam Ab24758) antibody. Secondary antibodies were FITC-conjugated anti-mouse and Texas Red-conjugated anti-rabbit antibodies (1:150, Jackson ImmunoResearch Laboratories). Slides were DAPI stained, mounted in Vectashield (Vector Laboratories, Cat No H-1000) and imaged on a Zeiss epifluorescence microscope using a 100x objective.

Chromatin immunoprecipitation
ChIP was done as described[36] except that a Soniprep 150 (Sanyo) was used for sonication. In brief, cells (5-6 x 10^6 in 10 cm dishes) were cross-linked with 10 ml 1% formaldehyde (Sigma) in medium for 5 min at room temperature and then incubated in 10 ml 200mM glycine in medium for 5 min. Cells were rinsed twice with

PBS and incubated with 7 ml lysis buffer (10 mM Tris-HCl (pH 7.5), 10 mM NaCl and 0.5% NP-40) for 10 min at room temperature with mild rotation. This lysis buffer was then aspirated off and cells were scraped into 1 ml lysis buffer and centrifuged at 3000 rpm for 3 min at 4°C. Cell pellet was resuspended in 100 µl SDS lysis buffer (50 mM Tris-HCl (pH 7.5), 10 mM EDTA and 1% SDS) and mixed by pipetting. 400 µl ChIP dilution buffer (50 mM Tris-HCl (pH 7.5), 167 mM NaCl, 1.1% Triton X-100, 0.11% sodium deoxycholate and protease inhibitor cocktail (complete EDTA-free; Roche)) was added before sonication (fifteen times for 20 seconds at 2 µm). After centrifugation at 13,000 rpm for 15 min at 4°C to remove the insoluble material the supernatant was removed to a new 1.5 ml tube and the volume made up to 500 µl with ChIP dilution buffer. 50 µl was removed as input for ChIP and the rest of the sample was incubated with antibody-bound Dynabeads overnight at 4°C with rotation. Dynabeads were prepared in advance by taking 50 µl of beads and washing three times with 500 µl cold RIPA-150 mM NaCl buffer (50 mM Tris-HCl (pH 7.5), 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.11% sodium deoxycholate and protease inhibitor cocktail). Beads were then incubated with 500 µl cold RIPA-150 mM NaCl buffer plus antibody for 2 hr at 4°C with rotation. Beads were then washed three times with 500 µl cold RIPA-150 mM NaCl buffer and were then ready for overnight incubation with the ChIP sample. Beads were washed sequentially with 1 ml cold RIPA- 500 mM NaCl (50 mM Tris-HCl (pH 7.5), 500 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.11% sodium deoxycholate and protease inhibitor cocktail) and twice with 1 ml TE (10 mM Tris-HCl (pH 8.0) and 1 mM EDTA). DNA was eluted by the addition of 200 µl ChIP direct elution buffer (10mM Tris-HCl (pH8.0), 300 mM NaCl, 5mM EDTA and 0.5% SDS) and incubated overnight at 65°C. Samples were then treated with DNase-free RNase (Roche; 5 $\mu g.ml^{-1}$; 37°C; 30 min) and proteinase K (250 $\mu g.ml^{-1}$; 55°C; 1 hr). DNA was extracted with phenol-chloroform-isoamyl alcohol (25:24:1) and ethanol precipitated with carrier (1µl glycogen, Invitrogen) on dry ice for 30 min. Following 70% ethanol wash the DNA pellet was resuspended in 20 µl water and quantified using a NanoDrop. For microarray hybridization, immunoprecipitated DNA was amplified using whole-genome amplification (Sigma).

Magnetic sheep anti-mouse IgG beads (Dynabeads, Invitrogen, 11201D) were used for mouse antibodies and protein G beads were used for rabbit antibodies (Dynabeads, Invitrogen, 10004D). Antibodies used were to CENP-A[37] and CENP-C (gift from W. Earnshaw), H3K27ac (Abcam, Ab4729), H3K4me2 (Millipore, 07-030), H3K9me2 (Millipore, 07-212), H3K9me3 (Abcam, Ab8898), H3K27me3 (Abcam, Ab6002), H3K36me3 (Abcam, Ab9050), CTCF (D31H2) (Cell Signaling Technology, 3418), RNA Polymerase II (Diagenode, C15100055), RNA Polymerase II (gift from H. Kimura). All antibodies were characterized using western blots, and ChIP was optimized using quantitative PCR assays.

Analyzing changes in DNA supercoiling

Biotinylated psoralen (bTMP) uptake was used to analyse DNA supercoiling as previous described[27]. Cells were treated with 500 $\mu g.ml^{-1}$ of bTMP in PBS for 20 min at room temperature in the dark. bTMP was UV cross-linked to DNA at 360 nm for 10 min. DNA was purified from cells using SDS and proteinase K digestion and extracted using phenol-chloroform-isoamyl alcohol (25:24:1). DNA was fragmented by sonication (thirteen times for 30s at 2 µm). Biotin incorporation into DNA was detected by dot blotting using alkaline phosphatase−conjugated avidin as a probe. The bTMP−DNA complex in TE was immunoprecipitated using avidin conjugated to magnetic beads (Dynabeads MyOne Streptavidin Invitrogen, 65001) for 2 h at room temperature and then overnight at 4°C. Beads were washed sequentially for 5 min each at room temperature with TSE I (20 mM Tris-HCl, pH 8.1, 2 mM EDTA, 150 mM NaCl, 1% Triton X-100 and 0.1% SDS), TSE II (20 mM Tris-HCl, pH 8.1, 2 mM EDTA, 500 mM NaCl, 1% Triton X-100 and 0.1% SDS) and buffer III (10 mM Tris-HCl, pH 8.1, 0.25 M LiCl, 1 mM EDTA, 1% NP40 and 1% deoxycholate). Beads were then washed twice with TE buffer for 5 min. To extract DNA and to release psoralen adducts, the samples were boiled for 10 min at 90°C in 50 µl of 95% formamide with 10 mM EDTA. Samples were then made up to 200 µl with water, and the DNA was purified using a Qiagen MinElute PCR purification kit. bTMP bound DNA was amplified using whole-genome amplification (Sigma) prior to microarray hybridization.

Chromatin Fractionation

Disrupted or 'open' chromatin was isolated as described previously[20]. In brief, cell nuclei were digested with micrococcal nuclease and soluble chromatin released overnight followed by fractionation on a 6-40% isokinetic sucrose gradient in 80 mM NaCl, 0.1 mM EDTA, 0.1 mM EGTA and 250 µM PMSF. DNA purified from gradient fractions was analyzed by electrophoresis through 0.7% agarose in 1 × TPE buffer (90 mM Tris-phosphate, 2 mM EDTA) with buffer circulation. Preparative fractionation of DNA from gradient fractions was carried out by pulsed-field gel electrophoresis (PFGE) (CHEF system, Biorad) through 1% low melting point agarose in 0.5 × TBE, at 180 V, for 40 hr, with a 0.1–2 s switching time. Size markers were 1 kb (Promega) and λ-*Hind*III (NEB) DNA ladders. EtBr-stained gels were scanned using a 473 nm laser and a 580 nm

band-pass filter on a Fuji FLA-3000. DNA of ~ 20 kb, corresponding to “open” chromatin, was isolated by β-agarase (NEB) digestion and amplified by whole genome amplification (Sigma) prior to microarray hybridization.

Microarray hybridization, data processing and analysis

Whole-genome amplified DNA (ChIP/bTMP/’open’ chromatin) was labelled and hybridised as previously (NSMB) to custom 180K Agilent microarrays) (7 Mb spanning the neocentromere domain (chr3:142781158-149782213; GRCh38 (hg38). In brief, 500 ng DNA was random prime labeled (ENZO) with Cy3 (Sample DNA) or Cy5 (Input DNA) and purified on a MinElute PCR purification column (Qiagen). Labeled DNA was diluted in hybridization buffer (Agilent) and hybridized to arrays for 24 h at 65°C. Slides were washed according to the manufacturer’s instructions and scanned on a Nimblegen Microarray scanner at 2 micron resolution generating a TIFF file.

Spot signal intensity was extracted from the TIFF files using Agilent Feature Extraction software and were pre-processed in R using the RINGO bioconductor package to give the raw Cy5 and Cy3 signal intensities for each spot. Individual Cy5 and Cy3 channels were normalized to each other and between arrays using a variance stabilizing algorithm (for bTMP arrays) and loess, vsn (for ChIP arrays) or nimblegen (“open” chromatin arrays) normalized and scaled, using the standard Bioconductor LIMMA package. All arrays were quality controlled by checking array hybridization patterns, analyzing signal profiles and using MA plots. For data analysis log2(sample/input) data was loaded in to the ZOO package in R and for display the data was smoothed using a rolling median.

RNA extraction and RNA-Seq

Total RNA was extracted from cells using RNeasy mini kit (Qiagen) with on-column DNase I digestion (RNAse-Free DNase Set, Qiagen). For RT-qPCR RNAs were reverse transcribed (Superscript II, Invitro-gen) using random primers and quantified by qPCR (Fast start SYBR green, Roche). Primer sequences are described below. Ribosomal RNA was depleted using Illumina® Ribo-Zero Plus rRNA Depletion Kit (Illumina, 20040526) following the manufacturer’s instructions and libraries for RNA-seq were prepared and indexed using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (NEB #E7645L) and NEBNext Singleplex Oligos for Illumina (NEB #E7335, E7500) following the manufacturer’s instructions. Libraries were sized and quality controlled on a D1000 Tapestation tape (Agilent). Single-end RNA-seq of 50-bp read length was performed on Illumina Hi-Seq 2000 (UMC, Amsterdam). FASTQ sequence files were obtained and the RNA-seq reads were aligned to the human reference genome (hg38) using TopHat v2 and processed with Samtools v1.6[38] and the Bedtools “genome coverage” tool[39]. Aligned BAM files were processed with the Subread v1.5 “feature counts” tool[40] to generate FPKM scores against hg38 RefSeq genes.

Primer name	Sequence	Figure
DIPK2A_Fwd	TTTGCAGTTGGTCCTAGAGATG	Extended Data Fig. 3
DIPK2A _Rev	AAGCCTCCTTATCACAGTCATC	Extended Data Fig. 3
PLOD2_Fwd	CTGAACGACAGCGTTCTCTTC	Extended Data Fig. 3
PLOD2_Rev	CCACCTCCCTGAAAGTCTTCT	Extended Data Fig. 3
PLSCR1_Fwd	GGACAGAGGGTTTACTTTGC	Extended Data Fig. 3
PLSCR1_Rev	GGTCTCTCCAGAGTTATGAC	Extended Data Fig. 3
ZIC1_Fwd	GTCCTACACGCATCCCAGTT	Extended Data Fig. 3
ZIC1_Rev	GTGGAGGATTCGTAGCCAGA	Extended Data Fig. 3
ZIC4_Fwd	AGGCTCAAAGTCAGAAAATG	Extended Data Fig. 3
ZIC4_Rev	TAAGAGTGTTTCGGTAAAGC	Extended Data Fig. 3
HLTF_Fwd	TGAAATGGAACCAGCTGAGG	Extended Data Fig. 3
HLTF_Rev	GTATAAGTCATTTCGCTGTTCCC	Extended Data Fig. 3
DIPK2A _Exon/Intron_Fwd	CTACGCGGAGACCAAGGAC	Extended Data Fig. 8
DIPK2A_Exon/Intron_Rev	CCACTTCTCCCGACTCTGAT	Extended Data Fig. 8
HLFT_Exon/Intron_Fwd	AGCGGTTTCAGATCAGTTGA	Extended Data Fig. 8
HLFT_Exon/Intron_Rev	AGTTGTACCTTGGAGCCTTGA	Extended Data Fig. 8

TT-seq

Nascent RNA was labelled by adding 500 µM 4-thiouridine (4sU)(Sigma, T4509) to cells harbouring HSA3 and Neo3 in T75 flasks and incubating at 37°C for 10 min. Media was aspirated and RNA extraction was performed with TRIzol (Invitrogen) following the manufacturers' instructions. After DNase treatment (Turbo DNase, Thermo Fisher Scientific) RNA concentration and purity were determined using a NanoDrop. RNA (70 µg) was fragmented in 100 µl H_20 to <1.5kb by 20 cycles of 30 seconds on/30 seconds off at high power in a Biorupter plus and RNA size assessed by agarose gel electrophoresis. Fragmented 4sU labelled RNA was biotinylated by adding 140 µl of EZ-Link Biotin-HPDP (1mg.ml^{-1} in dimethylformamide; Pierce, 21341), 70 µl of 10 x biotinylation buffer (100 mM Tris-HCl pH 7.5, 10 mM EDTA) and H_20 to a final volume of 700 µl. This was incubated at room temperature for 1.5 h with rotation. Unincorporated biotin-HPDP was removed by two rounds of chloroform extraction with 2 ml Phase lock gel heavy tubes (Eppendorf). RNA was precipitated with 1/10 volume of 5M NaCl and an equal volume of Isopropanol. This was inverted to mix and incubated at room temperature for 10 min followed by centrifugation at 10,000g for 20 min at room temperature. RNA pellet was washed with 80% EtOH and centrifuged at 13,000 rpm for 10 min at 4°C. RNA was resuspended in 100 µl H_20 and dissolved by heating to 40°C for 10 min with agitation. RNA was then immediately placed on ice and RNA concentration determined by Nanodrop spectrophotometer. Biotinylated 4sU labelled RNAs were then recovered using µMACS Streptavidin MicroBeads (Miltenyi, 130-074-101) and separation on a µMACS Separator. For the concentration of total RNA in µg per sample an equal amount in µl of Streptavidin microbeads was added. This was incubated at room temperature for 15 min with rotation. µMacs columns were equilibrated with 900 µl room temperature washing buffer (100 mM Tris-HCl pH 7.5, 10 mM EDTA, 1 M NaCl, 0.1% Tween20). The RNA/streptavidin bead solution was then applied to the column followed by three washes with 900 µl of washing buffer at 65°C and three washes with 900 µl of washing buffer at room temperature. RNA was eluted with 2 x 100 µl of fresh elution buffer (100 mM dithiothreitol in RNase-free H_20) directly into 2 ml lobind tubes (Eppendorf) containing 700µl Buffer RLT (RNeasy MinElute Cleanup Kit, Qiagen). 500 µl of 100% ethanol was added to the RNA solution, and mixed thoroughly by pipetting before RNA was purified through RNAeasy MinElute Spin Columns. RNA concentration was determined using a Nanodrop and libraries for RNA-seq were prepared and indexed using NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina® (NEB #E7645L) and NEBNext Singleplex Oligos for Illumina (NEB #E7335, E7500) following the manufacturer's instructions. Libraries were sized and quality controlled on a D1000 Tapestation tape (Agilent) and Illumina sequencing (paired-end RNA-seq of 50-bp read length) was performed on NovaSeq S1 (Edinburgh Genomics). FASTQ sequence files were obtained and the RNA-seq reads were aligned to a Human(hg38) /Hamster (GCA_003668045.1) hybrid reference genome using Bowtie2 and processed with Samtools v1.6[38], and the deepTools "bamCoverage" tool[39] with RPKM normalisation.

Exosome RNA interference

For siRNA treatment, cells (GM10253A and HybNeo3; 10%–20% confluent) were transfected with 10 nM Silencer Select Pre-designed siRNA targeting EXOSC3 (Ambion, Life Technologies) using Lipofectamine RNAi MAX (ThermoFisher) 24 h after seeding and again 48 h later. After a further 48 h exosome knockdown was confirmed by western blotting and TTseq was performed. Silencer Select RNA sequence for EXOSC3 were GAGATATATTCAAAGTTGA, part number s83102. The control RNA was Stealth RNAi siRNA Negative Control (ThermoFisher). For western blotting cells were suspended in NuPAGE LDS sample buffer (ThermoFisher) with 10 mM DTT, incubated at 100°C for 5 min and sonicated briefly. Protein samples were resolved on 12% bis-tris gels (ThermoFisher) and transferred to Immobilon-P PVDF 0.45 mm membrane (Merck Millipore) by wet transfer. Membranes were probed with anti-EXOSC3 antibody (Abcam, Ab156683) using standard techniques and detected by enhanced chemiluminescence.

Shallow DNA Sequencing

The hybrid HSA3 and Neo3 cell lines were shallow sequenced to confirm copy number. Genomic DNA was prepared from cells and 500 ng DNA was fragmented using a Covaris sonicator. Genomic DNA libraries were prepared using Illumina TruSeq Nano DNA LT sample prep kit as per manufacturer's instructions and Illumina sequencing (50-bp, single end reads) was performed on Illumina Hiseq 4000 (VUMC Cancer Centre, Amsterdam). FASTQ sequence files were obtained and reads were aligned to the human reference genome (hg19) using BWA and processed with Samtools v1.6[38]. In R the BAM files were loaded into the Bioconductor package QDNAseq for copy number analysis. Human reference genome HG19 was used here as QDNAseq has pre-calculated bin annotations for genome build hg19.

Neocentromere capture DNA sequencing

NimbleGen Sequence Capture technologies were employed for targeted deep sequencing of the neocen-

tromere domain. Capture probes tiling a 1.5Mb domain across the neocentromere were designed using Nimblegen capture design software and sequence capture performed using this custom SeqCap EZ Choice probe pool and SeqCap EZ HE-Oligo Kit A and SeqCap EZ Accessory Kit (Nimblegen) according to manufacturer's instructions. In brief, genomic DNA (gDNA) from the parental lymphoblastoid cell line was fragmented to ~200-500bp with 20 cycles of 30 seconds on/30 seconds off on a Biorupter. 1 µg gDNA was then used to prepare the gDNA sample library using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (NEB #E7645L) following the manufacturer's instructions. The neocentromere domain was captured by hybridising this gDNA library with the biotinylated SeqCap EZ library. 342ng of gDNA library was mixed with 5µg C_0t1 DNA, 1000 pmol of SeqCap HE Universal Oligo 1 and 1000 pmol SeqCap HE Index Oligo, 7.5 µl 2 X Hybridisation Buffer and 3µl Hybridisation Component A. This was vortexed for 10 seconds and centrifuged at maximum speed for 10 seconds before denaturing at 95°C for 10 min. This gDNA/C_0t1/Oligo/Hybridisation cocktail was then combined with the SeqCap EZ library (provided as 4.5µl single-use aliquots in 0.2ml tubes), vortexed for 3 seconds and centrifuged at maximum speed for 10 seconds. Hybridization was then performed on a thermocycler at 47°C and incubated for 70 hours. Each hybridization reaction was then bound to streptavidin beads from SeqCap EZ Pure Capture Bead Kit and washed with SeqCap EZ Hybridization and Wash Kit (Nimblegen), following the manufacture's protocol. Captured libraries were re-amplified using Post LM-PCR oligos (Nimblegen) and Q5 High-Fidelity DNA polymerase (NEB) directly from the beads. A mastermix consisting of 65 µl 2 × NEBNext Ultra II Q5 Master Mix NEB, 50 µl captured library (beads in H_20), 13 µl LM-PCR Oligo mix (Oligo 1 & 2, 2µM final concentration of each)(Nimblegen) was made up to 50 µl with H_20, vortexed to mix and then split into 2 × 65 µl samples for PCR using the following PCR cycling conditions. Initial incubation at 98°C for 30 seconds, 14 cycles of 98°C for 10 seconds, 65°C for 30 seconds and 72°C for 30 seconds. Final incubation of 72°C for 5 min and hold at 4°C. The two PCR reactions were recombined and the captured DNA was purified using 1.8:1 AMPure XP Beads: DNA ratio. Capture efficiency was determined to be between 94 and 117 fold using a Nimblegen Sequence Capture control locus qPCR assay. Neocentromere captured DNA libraries were sized and quality controlled on a D1000 Tapestation tape (Agilent), and paired-end sequenced (50 bp) on an Illumina MiSeq (Edinburgh Genomics).

Nucleosome positioning

Nuclei were extracted from cells carrying HSA3 and Neo3 as described[24] and resuspended in NB-R (85 mM KCl, 10 mM Tris-HCl [pH 7.6], 5.5% (w/v) sucrose, 1.5 mM $CaCl_2$, 3 mM $MgCl_2$, 250 µM PMSF). Nuclei (800 µl at 5A260) were digested with DFF nuclease (PMID: 17626049) for increasing amounts of time (100 µl of digested nuclei were removed to a new tube after 1, 2, 4, 8, 16 and 32 min digestion) at room temperature in the presence of 100 $\mu g.ml^{-1}$ RNaseA. Digestion was stopped by adding EDTA to 10 mM. DNA was purified with SDS/Proteinase K digestion, phenol/chloroform extraction and ethanol precipitation. Agarose gel electrophoresis of DFF digested nuclei confirmed digestion to mono and di-nucleosomes. The 4 min and 8 min samples and the 16 min and 32 min samples were pooled and ran on 2.5% LMP GTG agarose in 1 × Sybr Safe (Thermo Fisher) dye. Mono and di-nucleosome DNA bands were excised from the gel and purified by β-agarase (NEB) digestion followed by phenol/chloroform extraction and ethanol precipitation. 500ng DNA samples (50ng from the 4 and 8 min DFF digestion pool plus 450 ng of the 16 and 32 min DFF digestion pool) were concentrated to 55 µl volume using 2:1 AMPure XP Beads: DNA ratio for the mono nucleosome samples and 1.6:1 AMPure XP Beads: DNA ratio for the di nucleosome samples. Genomic DNA sample libraries were then prepared using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (NEB #E7645L) following the manufacturer's instructions. This included adaptor-ligated DNA size selection of between 100-200 bp for the mono nucleosome samples and 300-400 bp for the di nucleosome samples. DNA library yield was increased by a further round of PCR with Post LM-PCR oligos (Nimblegen) and Q5 High-Fidelity DNA polymerase (NEB). The PCR reaction consisted of 30 µl of the mono or di nucleosomal DNA libraries (150-270 ng), 50 µl 2 × NEBNext Ultra II Q5 Master Mix NEB, 5 µl Post LM-PCR oligo mix (Oligo 1 & 2, 2 µM final concentration of each; Nimblegen) and 15 µl H_20. PCR cycling conditions were an initial incubation at 98°C for 30 seconds, 5 cycles of 98°C for 10 seconds and 65°C for 75 seconds and a final incubation of 65°C for 5 min. DNA was purified using 1.8:1 AMPure XP Beads: DNA ratio and libraries were quantified and sized on a D1000 Tapestation tape (Agilent). Libraries, representative of mono and di nucleosome positions throughout the genome, were pooled in equimolar amounts (150 nM) and then subjected to neocentromere capture (as above) to examine nucleosome positioning across the neocentromere region. 1.25 µg of the mono and di nucleosome library pool was mixed with 5µg C_0t1 DNA, 1000 pmol of SeqCap HE Universal Oligo 1 and 1000 pmol SeqCap HE Index Oligos 14, 16, 18 and 19, 7.5 µl 2 × Hybridisation Buffer and 3µl Hybridisation Component A. This was vortexed for 10 seconds and centrifuged at maximum speed for 10 seconds before denaturing at 95°C for 10 min. This gDNA/C_0t1/Oligo/Hybridisation cocktail was then combined with the SeqCap EZ library (provided as 4.5µl single-use aliquots in 0.2 ml tubes), vortexed for 3 seconds and

centrifuged at maximum speed for 10 seconds. Hybridization was then performed on a thermocycler at 47°C and incubated for 70 hours. Each hybridization reaction was then bound to streptavidin beads from SeqCap EZ Pure Capture Bead Kit and washed with SeqCap EZ Hybridization and Wash Kit (Nimblegen), following the manufacture's protocol. Captured libraries were re-amplified using Post LM-PCR oligos (Nimblegen) and KAPA High-Fidelity DNA polymerase (KAPA Biosystems) directly from the beads. 25 µl 2 × KAPA HiFi Hot-Start ReadyMix, and 5 µl LM-PCR Oligo mix (Oligo 1 & 2, 2µM final concentration of each; Nimblegen) was added to the 20 µl captured library (beads in H_20), vortexed to mix and PCR amplified using the following PCR cycling conditions. Initial incubation at 98°C for 45 seconds, 14 cycles of 98°C for 15 seconds, 60°C for 30 seconds and 72°C for 30 seconds. Final incubation of 72°C for 1 min and hold at 4°C. Captured DNA was purified using 1.8:1 AMPure XP Beads: DNA ratio. Capture efficiency was determined to be between 120 and 240 fold using a Nimblegen Sequence Capture control locus qPCR assay. Neocentromere captured DNA libraries were sized and quality controlled on a D1000 Tapestation tape (Agilent), and paired-end sequenced (50 bp) on an Illumina MiSeq (Edinburgh Genomics).

Nucleosome positioning analysis

Paired end sequencing reads were mapped to hg38 using Bowtie 2 with high quality (mq > 20) and paired reads selected for further analysis. Start and end positions of reads were extracted from bamfiles using bed-tools bamtobed function and analysed in R. The NucleR package was used to calculated nucleosomal dyad positions with 40 bp trimming and coverage, data was formatted in the ZOO package and plotted using the lattice package. The acf function in R was used to calculate nucleosome autocorrelation.