Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2020 doi:10.20944/preprints202012.0669.v1

Why do centromeres evolve so fast: BIR replication, hypermutation, transposition, and molecular-drive

William R. Rice*

*Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, CA 93106, USA, email: rice@lifesci.ucsb.edu

Key Words: Centromeres, Evolution, Break-Induce Repair, Transposition, Tandem repeat array

Centromeres are among the fastest evolving genomic regions in a diverse array of organisms. The evolutionary process driving this rapid evolution has not been unambiguously established. Here I integrate diverse information to motivate a model in which centromeres evolve rapidly because of their intrinsic molecular phenotype: they tightly bind centromeric proteins throughout the cell cycle. DNA-bound proteins have been shown to cause stalling and collapse of DNA replication forks in many genomic regions, including centromeres. Collapsed replication forks generate one-sided double strand breaks (DSBs) that are repaired by the Break-Induced Repair (BIR) pathway. Here I show why this repair is expected to generate tandem repeat structure and three key features at centromeres: i) increased nucleotide substitution mutation rates, ii) out-of-register re-initiation of replication that leads to indels spanning one or more repeat units, and iii) elevated rates of large and small transpositions within centromeres and between genomic regions. These phenotypes lead to: i) a rapid rate of nucleotide substitutions within a clade of centromeric sequences, ii) continual turnover of monomers within centromeres that fosters molecular-drift and molecular-drive, and iii) recurrent quantum leaps in centromere sequence due to the formation of mosaic monomers and new sequences transposed into non-homologous centromeres. These features are plausibly the major reason centromeres evolve so rapidly. I also speculate on how the DNA sequence of centromeres might perpetually coevolve with the protein sequence of histone CENH3 –the major epigenetic mark of centromeres.

By the beginning of the 21st century, sequence analysis of the centromeres from many different species clearly indicated that they evolve unusually rapidly (Henikoff et al. 2001). Some proteins within the kinetochore that binds this DNA also evolve rapidly (Drinnenberg et al. 2016). The most generally accepted hypothesis for this fast evolution is that intragenomic conflict due to centromere drive (a form of female meiotic drive) causes perpetual antagonistic convolution (Henikoff et al. 2001; Burt and Trivers 2006; Malik and Henikoff 2009; Rosin and Mellone 2017). In the model developed here, **B**reak-**I**nduced **R**epair (BIR) of collapsed DNA replication forks and its down-stream consequences (especially hyper-mutation, transposition and molecular-drive), are the major causative features for the exceptionally rapid evolution at centromeres.

To illustrate the extent and form of rapid centromere evolution, I compare human-chimp nucleotide sequence divergence at centromeres to that of the surrounding chromosomal arms. I first use published data to quantify: i) the local (1 Mb intervals) and chromosome-wide (mean of intervals) sequence divergence between chimps and the human reference genome along the two arms of human chromosome 7 (red dots and horizontal lines in Figure 1A; from Mikkelsen et al. 2005 and Marques-Bonet et al 2009), and ii) this same measure of divergence among different human genomes (blue dots in Figure 1A; from Marques-Bonet et al 2009). I then estimate the minimum centromeric sequence divergence for this chromosome between human and chimp orthologs, as described in Box-1, Figure 1A and the next paragraph.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2020 doi:10.20944/preprints202012.0669.v1

Human centromeres are composed of very long (average 2-3 Mb; Willard 1991) tandem repeat arrays of **H**igher **O**rder **R**epeats (HORs; iterations of groups of two or more monomers) with low sequence variation among HOR units within an array (Willard 1991; Schueler et al. 2001; Schueler and Sullivan 2006). Because the consensus sequence of the centromeric HOR of chromosome 7 is known for humans (Waye et al 1987; Rice 2019A) but unknown for chimps, I compared (Box-1 and Figure 1A) the consensus sequence of the human HOR to the closest matching alpha satellite sequence found during the chimp genome project (Mikkelsen et al. 2005). This analysis indicates that the centromere on human chromosome 7 and its chimp otholog have diverged by a lower limit of 12.3%: a value far in excess of the maximal sequence divergence seen along the chromosomal arms (Figure 1A). Average human-chimp sequence divergence is 1.2% along the chromosomal arms of chromosome 7 (Mikkelsen et al. 2005). Combining measures, the relative rate of chimp-human nucleotide divergence (centromere / arms) is minimally 12.3% / 1.2% = 10.25 times faster for the centromere than the average for the chromosomal arms.

To obtain a more accurate measure of chimp and human centromeric divergence, I next focused on human chromosome 5 (Box-1; Figure 1B). I chose this chromosome because: i) the consensus sequences of the centromere of this chromosome and its chimp ortholog have both been determined (Puechberty et al 1999; Haaf and Willard 1997; Rice 2019A; Rice unpublished), and also ii) the average and range of chimp-human sequence divergence has been measured along the arms of these orthologs (Mikkelsen et al. 2005). At chromosome 5, the estimated sequence divergence between human and chimp centromeres was 39.3%

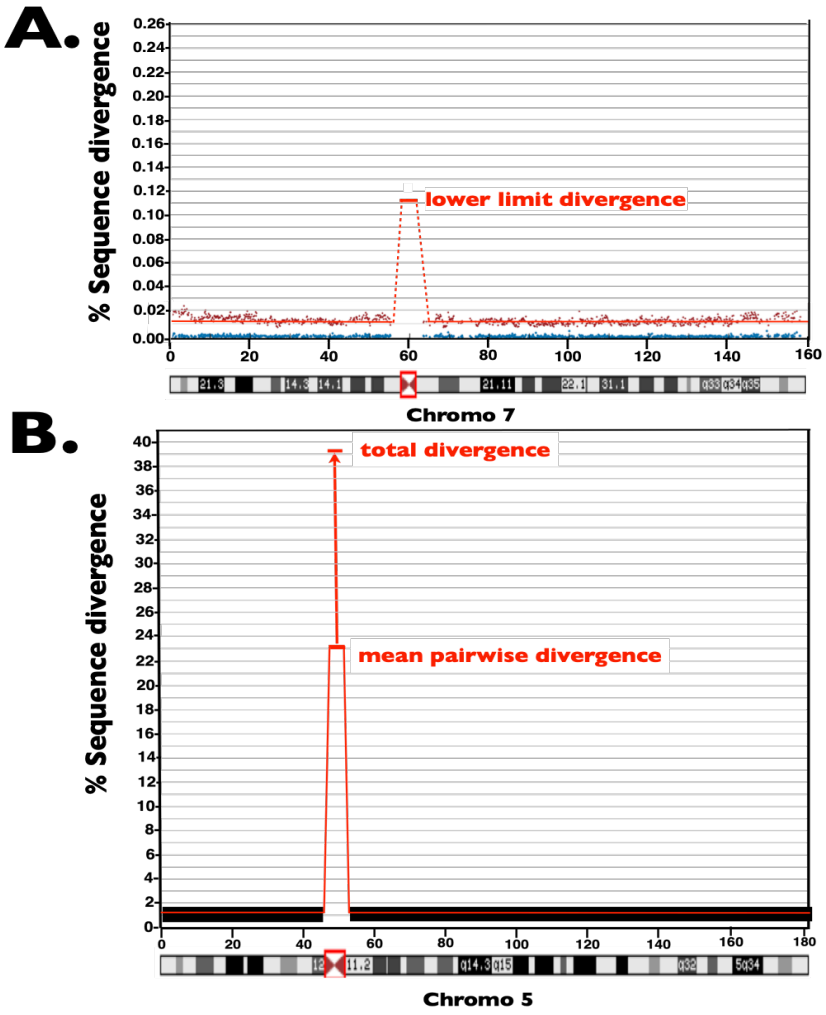


Figure 1. A. Deviations from the human reference sequence among 100 kb intervals along the arms of chromosome 7. The schematic of the chromosome (along the X-axis) is taken from the UCSC genome browser web page and the red square denotes the centromere within the sequencing gap between the assembled arm sequences. Red dots depict deviations for chimp vs. human, blue dots depict deviations among humans, and the horizontal red lines along the bottom of the graph are the arms-wide average of 1 Mb regions (data are from Mikkelsen et al. 2005 and Marques-Bonet et al 2009). Above the centromere, I show the lower limit for the sequence divergence between the human and chimp centromeres (see Box-1 for details). **B.** Chimp sequence deviations from the human reference sequence among 1 Mb intervals along the arms of chromosome 7. Narrow, horizontal red lines on either side of the sequencing gap depict the average deviation from the human reference sequence and the wide horizontal black lines depicts the total range of deviations about the mean; from Mikkelsen et al. 2005). The schematic of the chromosome is taken from the UCSC genome browser web page and the red square denotes the centromere within the sequencing gap between the assembled arm sequences. The first red line above the centromere depicts the average pair-wise sequence divergence between the chimp and human centromeric monomers of the same type: an incomplete measure of total divergence between centromeres. The upper red line red line above the centromere depicts a more complete measure of sequence divergence between the chimp and human centromeres: labeled 'total divergence' (see Box-1 for details).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2020 doi:10.20944/preprints202012.0669.v1

Box-1. Rapid evolution at centromeres is illustrated by the sequence divergence between chimps and humans.

To contrast the rate of chimp-human sequence divergence of centromeres compared to other regions of the genome, nucleotide divergence between humans and chimps was compared between centromeres and the arms of chromosomes. In Figure 1A, the percent sequence divergence of the two arms of chromosome 7 (at 100 kb intervals) is shown between the human reference genome and: i) a single chimp genome (red dots), and ii) different human genomes (blue dots) (data from Marques-Bonet et al 2009). I also display the average of human-chimp divergences (of 1 Mb intervals) across both chromosomal arms (1.2%, shown by the horizontal red lines above the arms; data from Mikkelsen et al. 2005). This graph illustrates the level of variation in divergences about the mean value at small, non-centromeric regions of the genome. The centromere (red square in the chromosome image) resides within the sequencing gap between the chromosomal arms, where divergence values have not been reported.

The human centromeric sequence (i.e., the consensus HOR sequence) for chromosome 7 has been determined with high accuracy (Waye et al 1987; Rice 2019A), but the corresponding centromeric sequence for the chimp ortholog is not known. To obtain a minimum estimate for sequence divergence at the centromere, I blasted the consensus sequence for the centromeric higher-order repeat (HOR) of human chromosome 7 against a large set of archived chimp sequences from the NCBI web site (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROG_DEF=blastn&BLAST_PROG_DEF=megaBlast&BLAST_SPEC=OGP__9598__12467). The archived chimp sequences were the data set Clint_PTRv2, which contains all assembled chromosomes plus unplaced and unlocalized scaffolds from the reference assembly in Annotation Release 105.

I found a cluster of 488 hits with the closest matching sequences that had a mean divergence from the human HOR of 11.31%. However, this value represents a lower bound for the true divergence between the orthologous chimp and human consensus centromeric HOR sequences because an *in situ* hybridization study by Archidiacono et al. (1995) found that the closest matching chimp centromere to human chromosome 7 is not located on the chimp ortholog. To better estimate the true divergence between chimp and human othologous centromeres, next I focused on human chromosome 5, where the centromeric chimp (Haaf and Willard 1997; Rice unpublished) and human (Puechberty et al 1999; Rice 2019A) consensus HOR sequences are known with high accuracy. The mean, variance and range of human-chimp divergence along the arms of chromosome 7 are highly similar to chromosome 5 (Mikkelsen et al. 2005), as is the level of SNP variation among humans (Shen et al. 2013). The consensus sequence for human chromosome 5 is a short HOR containing a two-monomer dimer: one 'b-box' monomer (containing a 17 bp b-box sequence that binds **CEN**tromere **P**rotein **B** [CENP-B]) and one 'no-b-box' monomer that lacks the b-box sequence. The chimp ortholog is a 5 monomer HOR containing two b-box monomers and three no-b-box monomers. In humans, b-box and no-b-box monomers have substantially different consensus sequences (~16% diverged) with most differences outside the b-box region (see Supplemental Figures S13 and S16 in Rice 2019A).

I characterized chimp-human sequence divergence in two ways. First, I determined the average pairwise percent divergence between the human b-box monomer and the two chimp b-box monomers. I next made this same pair-wise calculation between the human no-b-box monomer and the three chimp no-b-box monomers. These two pairwise divergences averaged 23.35% (Figure 2B). Average pairwise divergence, however, does not reflect the fact that the chimp HOR has three additional monomers and that the chimp monomers are highly diverged from each other (Supplemental Figure S1). To construct a measure that better captures this multi-monomer divergence, I calculated total divergence = the percent of nucleotide positions that differ between: i) the human b-box monomer and either of the two chimp b-box monomers, or ii) the human no-b-box monomer and any of the three chimp no-b-box monomers. This total divergence was 39.3% (Figure 1B). In sum, centromeric sequence divergence is 39.3 / 1.2 = 32.7 times more than the average at non-centromeric regions of chromosome 7. These data indicate that human-chimp sequence divergence at the centromere is far in excess of the range of values found along chromosomal arms (broad horizontal black line in Figure 1B) and more than an order of magnitude higher than the average across the chromosomal arms of chromosome 5 (narrow horizontal red lines at the base of the graph in Figure 1B).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2020 doi:10.20944/preprints202012.0669.v1

(Box-1, Figure 1B), a value far in excess of the range of divergence values found along the orthologs' arms (wide black lines in Figure 1B) and substantially more than an order of magnitude greater that the mean divergence found between the ortholog's arms (narrow red lines, Figure 1B). What causes this extreme divergence at centromeres?

At a functional level, centromeres are the DNA regions that recruit –and tightly bind– the large and complex group of proteins that make up the kinetochore. In most organisms, centromeres are composed of long tandem repeat arrays –which in humans can be as large as 8 Mb (Miga et al. 2014). DNA-kinetochore attachment is achieved by a network of kinetochore proteins (the CCAN = **C**onstitutive **C**entromere-**A**ssociated **N**etwork), the basal members of which bind the centromeric DNA throughout the cell cycle (Musacchio and Desai 2019). Bound protein is well established to cause stalling of DNA replication forks (Mirkin and Mirkin 2007; Beuzer et al 2014), and stalled forks are prone to collapse to form a one-sided double-strand break (DSB). These one-sided DSBs are repaired by the BIR pathway (Costantino et al. 2014).

There is empirical evidence that kinetochore proteins bound to DNA in both the point centromeres of budding yeast (Greenfeder and Newlon 1992) and the regional centromeres of *Canida albicans* (Mitra et al. 2014) cause fork-stalling/collapse, rather than it being caused by DNA secondary structure. There is also evidence for elevated levels of fork-stalling/collapse at human centromeres (Crosetto et al. 2013; Aze et al. 2016).

One expected downstream effect of an elevated rate of fork-stalling/collapse at centromeres is recurrent duplications and deletions (indels) within their tandem repeat arrays that causes a continual turnover of monomers (Figure 2, yellow box). Because centromeric tandem repeat arrays have numerous regions of nearby, flanking homology, BIR repair of collapsed replication forks within a centromere would be expected to re-initiate DNA replication at: i) the same location where the collapse occurred (in-register –with no deletion nor duplication), ii) a downstream location (already replicated) with a sequence matching the break point (out-of-register –leading to a duplication of one or more repeat units), or iii) an upstream location (not previously replicated) with a sequence matching the break point (out-of-register –leading to a deletion of one or more repeat units). These alternatives are illustrated in Supplemental Figure S3 in Rice (2019B). Studies of fork-collapse-induced BIR at rDNA repeat arrays in yeast have shown that downstream out-of-register BIR is more frequent than up-stream repair, causing, on average, a net increase in repeat array length in response to recurrent fork collapses –but only when cohesin-binding of sister chromatids was low (reviewed in Kobayashi 2014). This bias toward array expansion may be the result of differences in chromatin structure in the regions upstream and downstream of replication forks (Poot et al. 2005; also see explanation in Rice 2019B).

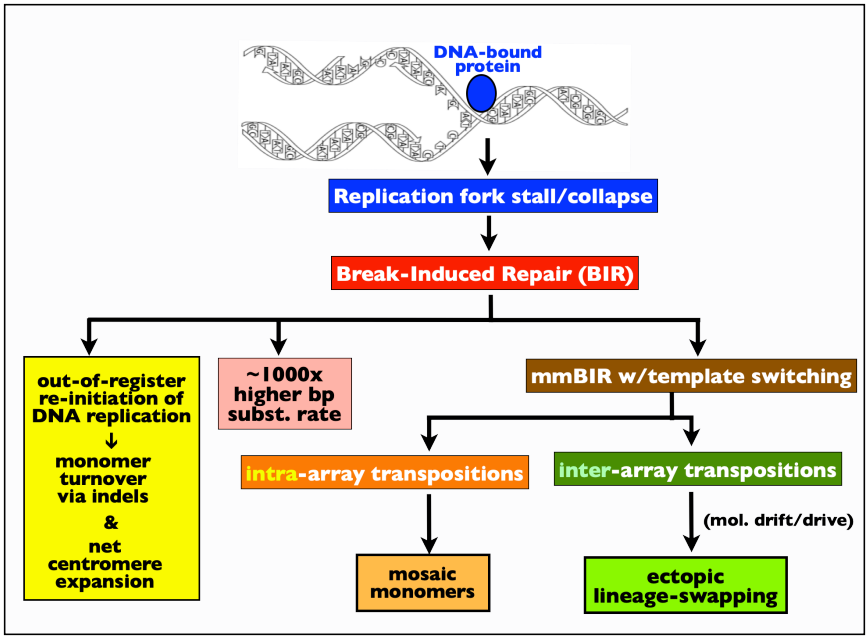


Figure 2. A diagram illustrating the chain of events connecting protein bound to DNA, the fork-stalling and fork-collapse that it generates, BIR repair of collapsed replication forks, and the downstream consequences of BIR repair.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2020 doi:10.20944/preprints202012.0669.v1

An additional factor that is expected to cause monomer turnover within centromeres is recurrent deletions when two-sided DSBs are repaired by the **S**ingle-**S**trand **A**nnealing (SSA) repair pathway (Ozenberger et al. 1991; see also Supplemental Figure S1 in Rice 2019B). Tsouroula et al. (2016) demonstrated that the SSA pathway is used (along with other pathways) during repair of two-sided DSBs at the centromeres of the laboratory mouse (*Mus musculus domesticus)*.

A second expected downstream consequence of increased levels of fork-stalling/collapse at centromeres is a substantially elevated nucleotide-substitution mutation rate (bp substitutions in Figure 2, pink box). BIR replication forks that form after fork-collapse use a unusual combination of DNA polymerase subunits, and this configuration is associated with highly elevated nucleotide-substitution mutations: estimated in yeast to be ~1,000-fold higher compared to normal S-phase replication forks (Sakofsky et al. 2012). The elevated use of BIR after fork-collapse at centromeric repeat arrays would be expected to cause homologous centromeres in related species to diverge in sequence at an elevated rate –as is observed between humans and chimps on their X chromosomes (Rice 2019B).

A third expected downstream effect of expanded levels of fork-stalling/collapse at centromeres is an increased rate of formation of new mosaic monomers that are composed of pieces of extant –usually nearby– monomers (Figure 2, orange box). A substantial proportion of BIR events involve template switching (possibly as high as 20%; Smith et al. 2007). BIR is sometimes mediated by low levels of homology (mmBIR = **m**inimal-homology-**m**ediated **BIR**; Zhang et al. 2009; Hastings et al. 2009A,B). The process of mmBIR with template-switching (mmBIR/templ-switch) is plausibly the major factor leading to **C**opy **N**umber **V**ariation (CNV) within genomes (Hastings et al. 2009A,B). It has the potential to generate new, mosaic monomers when the template switch occurs within the body of a monomer and exchanges part of one monomer with the sequence of another monomer. The resulting mosaic monomer represents a small quantum leap in HOR sequence when the recombined monomers have substantially different sequences –as is common among the monomers that make up the centromeric repeats of humans (see Supplementary Table S1 of Rice 2019A for a complete list of the HOR sequences at all of the active centromeric repeats from a single human genome).

Evidence for mosaic monomers can be found in humans, where monomers cluster into two major groups: those containing a 17 bp CENP-B-binding b-box at their 5' end (b-box monomers), and those lacking this feature (no-b-box monomers) (see Box-1, and also Supplemental Figures S13 and S16 in Rice 2019A). Some monomers are found at human centromeres in which the 5' end contains a b-box sequence while the majority of the remaining monomer has a sequence that strongly clusters with the no-b-box monomers (for examples, see in Rice 2019A Figure 4 and Supplemental Figure S14). These b-box/no-b-box mosaic monomers are consistent with the process of quantum leaps in monomer sequence due to ectopic recombination via mmBIR/templ-switch.

A fourth downstream effect of expanded levels of fork-stalling/collapse at centromeres is 'ectopic lineage swapping' in which one or more monomers from one chromosome's centromere are transposed into the centromere of another non-homologous chromosome, and eventually become the new centromeric repeat sequence (Figure 2, green boxes). Less commonly, the transposed DNA may originate from a non-centromeric location. This transposition process is a larger-scale extension of the mmBIR/templ-switch process described in the previous two paragraphs concerning mosaic monomers except that: i) the template switching is usually between different chromosomes, and ii) the transposed DNA segment is large enough to span one or more monomers. In the human genome there is extensive evidence for transposition between chromosomes into the centromeres of chromosomes 1 and 2 (see Supplemental Figures S4 and S18B in Rice 2019A). There is also evidence for large-scale transpositions from the the high levels of sequence divergence among flanking alpha satellite arrays found on many human chromosomes (see Figure 8 in Rice 2019A).

To become a new centromeric tandem repeat array, transposed DNA must fortuitously already contain a tandem duplication or else a new tandem duplication must be formed *de novo*. As

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2020 doi:10.20944/preprints202012.0669.v1

described earlier, substantial evidence indicates that repair of collapsed replication forks via mmBIR is a major mechanism generating tandem CNVs (Hastings et al. 2009B; Hsiao et al. 2015) and the elevated incidence of mmBIR at centromeres would feasibly provide a pathway to duplicate newly transposed sequences. Elevated rates of BIR at centromeres (due to DNA-bound CCAN proteins) may explain the observation that centromeres of most species are composed of tandem repeat arrays because: i) mmBIR would be expected to generate an initial tandem duplication at an initially non-repetitive centromeric sequence and ii) out-of-register BIR would next be expected to expand this minimal tandem repeat array into the long arrays typically observed at centromeres.

A new and small subarray of a tandemly duplicated transposition that is embedded within a centromere would not lead to centromere-wide sequence change unless the subarray expands to form a much longer tandem repeat subarray –and ultimately becomes the predominant repeated sequence there. As described earlier, BIR-induced duplications and deletions of monomers –and also deletion of monomers from SSA repair of two-sided DSBs– are expected to generate continual turnover of monomers within centromeric repeat arrays. This turnover process produces an opportunity for analogs of genetic drift and natural selection to operate within repeat arrays, i.e. 'molecular-drift' and 'molecular-drive,' respectively (Dover 1982).

The process of molecular-drift occurs when: i) there is heterogeneity in the sequence of repeat units (single monomers or HORs) within a tandem repeat array, and ii) indels cause repeat units to continually turnover, and iii) the proportion of different repeat units (with different sequences) changes over time due to random differences in their deletion and duplication rates. This process can lead to 'molecular-fixation' (all repeat units within an array share –or are descended from when mutated– the same sequence) by chance alone during the continual stochastic turnover of constituent repeat units. In this way, a new and small subarray could spread to an entire centromere.

Molecular fixation can also be driven deterministically by molecular-drive when repeat units with different sequences differ in their duplications/deletions ratio: the repeat unit with the highest ratio is expected to eventually predominate within the array. So if a new and small subarray (initiated by a transposition event) had a higher duplications/deletions ratio, it could expand to the entire centromere by molecular drive.

To recap, the increased prevalence of mmBIR at centromeric repeat arrays is expected to: i) increase the rate of transpositions into centromeric repeat arrays, and ii) foster tandem duplications of transposed DNA segments when not already in tandem repeat form. BIR after fork-collapse and SSA repair of DSBs will generate turnover of repeat units at centromeres and thereby create the opportunity for newly transposed tandem repeats to replace the old centromeric sequence via molecular-drift or molecular-drive. This process represents a form of 'ectopic lineage-swapping' in which a transposed, unrelated sequence replaces an existing centromeric sequence: leading to a 'quantum leap' in centromere sequence that generates highly elevated sequence divergence among othologous centromeres of closely related species.

To examine this ectopic lineage-swapping process in a more detailed molecular context, I next focus on humans and the laboratory mouse, where the structure and function of centromeric tandem repeat arrays has been extensively studied. The influence of molecular-drift and molecular-drive on sequence evolution within centromeric repeat arrays is expected to be influenced by the ability of the repeat unit to: i) expand via tandem duplications, and ii) recruit histone CENH3 –which is the epigenetic mark required to be a functional centromere that assembles a kinetochore (Bodor et al. 2014). Consider the case where transposition via mmBIR generates a small, new tandem-repeat subarray (green rectangle in Figure 3) within an extant, much larger, centromeric repeat array (blue rectangles in Figure 3). The fate of the subarray is expected to be influenced by its sequence and its position within the array.

Centromeric repeat arrays in mice and humans are divided into two functional regions: i) a contiguous centric core that recruits CENH3 at many locations and thereby constitutes the active centromere that assembles a kinetochore (region

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2020 doi:10.20944/preprints202012.0669.v1

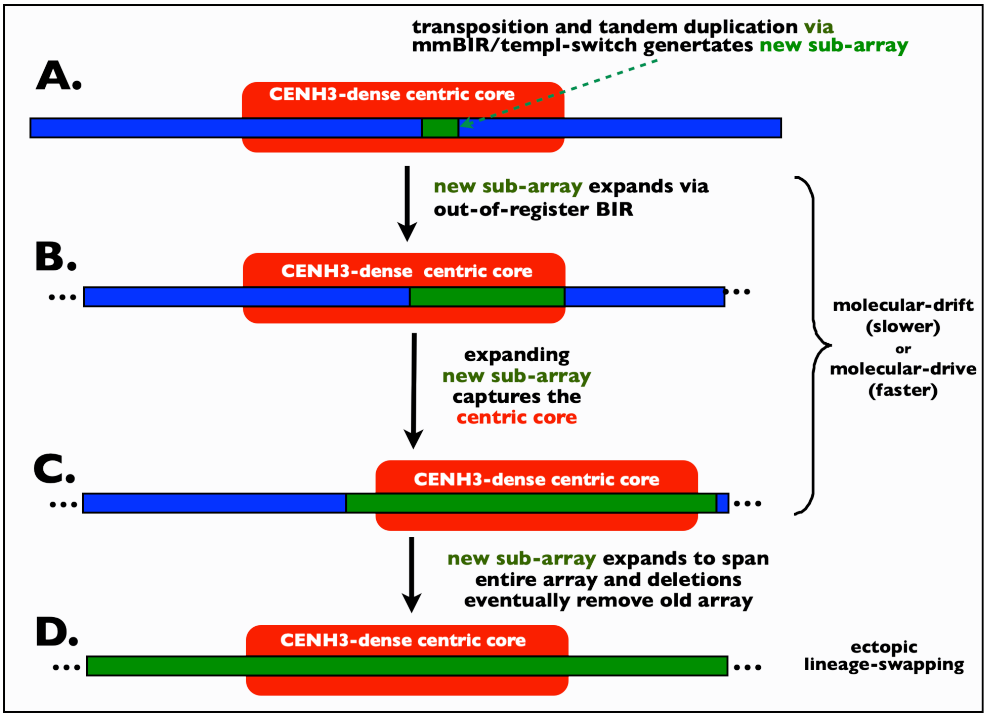


Figure 3. The hypothesized process of ectopic lineage swapping at centromeres in mammals. **A.** A new subarray (green bar) originating from a different genomic location is transposed into the centric core region (red oval background) of an extant centromeric repeat array (blue bar) via mmBIR with template-switching. The two regions outside the centric core (no red oval background) are the pericentric flanks. If the new subarray is not already a tandem repeat, this structure is generated via mmBIR. **B.** Because the subarray is within the CENH3-rich centric core of the resident centromeric repeat array (which as a high density of protein-bound to DNA and a low density of cohesin), it expands via duplications exceeding deletions during out-of-register re-initiation of BIR replication after the collapse of replication forks. **C.** By random chance (molecular-drift) or because the new subarray has a molecular-drive advantage, i) the new subarray retains the CENH3 epigenetic mark as it expands and eventually grows to span the entire centric core, and ii) the extant array is pushed completely out of the centric core and is restricted to the pericentric flanks. **D.** Over time, the old array eroded away by recurrent deletion pressure and only the new subarray is found at the centromere region.

within the red oval background in Figure 3), and ii) the remainder of the array, i.e., the pericentric flanks (Ross et al. 2016; Iwata-Otsubo et al. 2017). Out-of-register, BIR-induced expansion and contraction of the array is expected only (or at least predominantly) within the centric core because only this region i) binds the kinetochore proteins that cause fork-stalling/collapse, and ii) recruit low levels of cohesin (Supplementary Figure S2; see also Rice 2019B). Assuming that the new subarray has a sequence that generates a molecular-drive advantage, and that it resides within the centric core, it will only persist and achieve molecular-fixation (Figure 3D) when it retains (by chance or its molecular phenotype) the CENH3 epigenetic mark as it expands (Figure 3B,C).

To recap, transposition that generates a new tandem-repeat subarray within an extant centromeric array will only lead to replacement of the extant array by the newly transposed sequence when the new subarray resides within the centric core and, i) molecular-drift leads to the new subarray expanding to large size while fortuitously retaining the CENH3 epigenetic mark, ii) molecular-drive favors the new subarray because it has an expansion advantage (i.e., a faster tandem duplication rate of monomers or HORs) that indirectly leads to its retention of the CENH3 epigenetic mark as it expands (described more fully in Supplemental Figure S3), or iii) molecular-drive favors the new subarray because it more strongly recruits/retains CENH3 and thereby pulls the centric core off the larger, extant array as it expands (described more fully in Supplemental Figure S4). In Rice (2019B), I

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2020 doi:10.20944/preprints202012.0669.v1

provide a more general description of the operation of molecular-drive within centromeres, and in Supplemental Figures S2-S4 I summarize a molecular model from this paper for the operation of molecular-drive at centromeres in humans.

In situ hybridization studies demonstrate that centromeric sequencers at all, or nearly all, human and chimp autosomes have diverged to such an extent that probes for the human centromeres no longer bind their chimp orthologs, and sometimes bind non-orthologs (Archidiacono et al. 1995). This finding indicates that, within the human-chimp clade, repeat units (HORs) of centromeres are rapidly (< 5-6 x 10^6 years; De Manuel et al. 2016) replaced by distantly related sequences –and sequencing studies of centromeres at six chimp autosomes support this conclusion (Jorgensen et al. 1992; Haaf and Willard 1997; Warburton et al. 1996). The rapid replacement of all (or nearly all) autosomal centromeric HORs (by distantly related sequences) during divergence between humans and their closest living relative is consistent with the conclusion that the observed centromere evolution is more plausibly driven by deterministic molecular-drive rather than the much slower, stochastic process of molecular-drift.

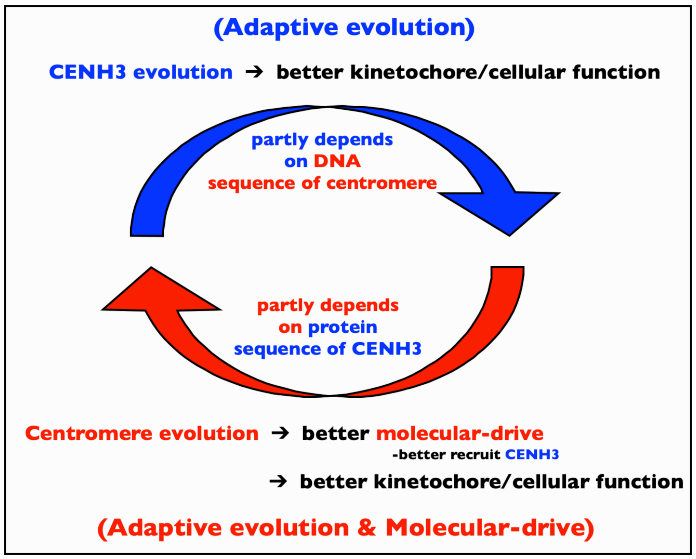


Figure 4. Positive feedback may drive 'push-away' coevolution between the DNA sequence of the centromere and the protein sequence of the histone CENH3. **Top:** The protein sequence of CENH3 is selected to promote kinetochore and cellular functioning. Part of this adaptive process is assumed to depend on the DNA sequence of the centromeric repeat unit. **Bottom:** The DNA sequence of the centromeric repeat unit is also selected to promote kinetochore and cellular functioning, but it is also selected for competitive ability in the context of molecular-drive between alternative sequences of the centromeric repeat unit, and much of this competitive ability is expected to depend on the ability to preferentially recruit CENH3. Part of this molecular-drive selection is assumed to depend on the protein sequence of CENH3. The interdependence of selection on the protein sequence of CENH3 and the DNA sequence of the centromeric repeat unit will feasibly cause perpetual coevolution between them.

If molecular-drive has been operating at the centromeres of eukaryotes for eons, and centromeres have essentially retained the same function in organisms as diverse as humans and fungi, then why would there continue to be molecular-drive for new sequences? Put another way: Why haven't optimal molecular-drive centromeric sequences evolved long ago and then persisted? In humans, I have described evidence for an intransitive competitive hierarchy between different HOR sequences that is expected to drive rapid and perpetual evolution of centromeric HOR sequences (Rice 2019B). However, this intransitivity is dependent on the influence of CENP-B at human centromeres –and most organisms do not recruit this protein to their centromeres. An alternative to competitive intransitivity is a continually changing optimal sequence for 'winning' in molecular-drive at centromeres. One way that such a 'moving target' optimum could occur is simply due to the inevitable and continuous evolution of the genomic background of numerous molecules that influence molecular-drive among centromeric subarrays (Rice 2019B). In addition, there may be positive feedback between the DNA sequence of centromeres and the protein sequence of CENH3. Below I describe this speculation.

Natural selection is expected to cause CENH3 to evolve a protein sequence that reliably recruits this molecule to the centromeres and that increases cellular functioning by assembling a well-operating kinetochore. The DNA sequences of centromeres will also experience natural selection for proper cellular functioning, but these sequences will additionally experience selection in the context of molecular-drive to better compete against other subarrays within a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2020 doi:10.20944/preprints202012.0669.v1

centromeric repeat array. This molecular-drive phenotype would be expected to include the capacity to better retain the CENH3 epigenetic mark as its subarray expands –compared to other subarrays within the same centric core– in order to remain within the centric core and not lose centromeric functioning (Figure S3 and Supplemental Figure S4).

The observation that ectopic neocentromeres have been found to occur at a wide diversity of DNA sequences (Marshall et al. 2008), would intuitively indicate that centromere sequence is relatively unimportant in CENH3 recruitment. However, the findings that: i) neocentromeres recruit reduced levels of CENH3 (Bodor et al. 2014; Fachinetti et al. 2015), ii) centromeric HORs differ in their ability to recruit CENP-A (Bodor et al. 2014; Fachinetti et al. 2015; Aldrup-MacDonald et al. 2016) and iii) different DNA sequences are better at recruiting CENH3 to artificial chromosomes (Masumoto et al 1998; Basu et al. 2005; Molina et al. 2017), supports the conclusion that DNA sequence does influence CENH3 recruitment and hence molecular-drive within tandem repeat arrays. The combination of i) natural selection on the protein sequence of CENH3, and ii) both natural selection and selection via molecular-drive on centromeric repeat sequences, motivates the potential for coevolution between the DNA sequence of centromeres and the protein sequence of CENH3 (Figure 4).

Selection on CENH3 for cellular functioning is expected to depend at least in part on some form of congruence between the protein sequence of CENH3 and the DNA sequence at centromeres (top Figure 4). Similarly, selection on centromere DNA sequence via molecular-drive performance is expected to depend in part on the protein sequence of CENH3 (bottom of Figure 4). As centromeric DNA sequences continually evolve in response to molecular drive, the resulting cumulative change in their DNA sequences would be expected to gradually 'push' the optimal protein sequence of CENH3 away from its current position. Eventually, the cumulative change in optimum would be sufficient to lead to an evolutionary change in the protein sequence of CENH3, which in turn would 'push' the optimal DNA sequence of a centromere away from its current position. Collectively, these interactions would be expected to drive open-ended, 'push-away' coevolution between the protein sequence of CENH3 and the DNA sequence of the centromeres because: i) each sequence type is evolving in response to different selection regimes, and ii) evolution by each type of sequence changes the optimum of the other type of sequence (Figure 4).

Summary

Centromere sequences are expected to evolve rapidly because of a fundamental aspect of their phenotype: they tightly bind centromeric proteins throughout the cell cycle. Bound protein is established to lead to substantially elevated rates of fork-collapse during DNA replication –and subsequent repair of collapse-generated one-sided DSBs via the BIR pathway. BIR in turn is expected to lead to many downstream consequences at centromeres: i) the formation of tandem repeat structure at new centromeres or new centromeric subregions lacking this structure, ii) expansion of centromeric arrays due to an excess of monomer duplications over deletions, iii) perpetual monomer turnover that creates an opportunity for molecular-drift and molecular-drive, iv) an elevated nucleotide-substitution mutation rate, v) small-scale transpositions that create new, mosaic monomer sequences, and vi) larger-scale transpositions that create new subarrays of monomers or HORs within extant centromeres. The increased mutation rate alone would be expected to make centromeres evolve more rapidly than other genomic regions, but this effect is magnified by mmBIR-induced transpositions. These horizontal transfer events, in combination with molecular-drift and especially molecular-drive, are expected to cause monomer sequence to rapidly change due to quantum leaps (large steps) in sequence over time via the formation of mosaic monomers and ectopic lineage swapping. Repeated episodes of molecular-drive may be perpetual because: i) of intransitivity among centromeric sequences, as seems plausible in humans, or more generally, ii) the genetic background of genes coding for molecules that interact with the centromere continually evolve and change the optimal centromeric sequence for molecular-drive. More speculatively, push-away coevolution driven by positive feedback between the protein sequence of CENH3 and the DNA sequence centromeres may also contribute to rapid evolution of centromeric sequences.

Acknowledgments

I thank Kathryn Schoenrock for copy-editing assistance.

References

Aldrup-MacDonald, M. E., Kuo, M. E., Sullivan, L. L., Chew, K., & Sullivan, B. A. (2016). Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. Genome research, 26(10), 1301-1311.

Archidiacono, N., Antonacci, R., Marzella, R., Finelli, P., Lonoce, A., & Rocchi, M. (1995). Comparative mapping of human alphoid sequences in great apes using fluorescence in situ hybridization. Genomics, 25(2), 477-484.

Aze, A., Sannino, V., Soffientini, P., Bachi, A., & Costanzo, V. (2016). Centromeric DNA replication reconstitution reveals DNA loops and ATR checkpoint suppression. Nature cell biology, 18(6), 684-691.

Basu, J., Stromberg, G., Compitello, G., Willard, H. F., & Bokkelen, G. V. (2005). Rapid creation of BAC-based human artificial chromosome vectors by transposition with synthetic alpha-satellite arrays. Nucleic acids research, 33(2), 587-596.

Beuzer, P., Quivy, J. P., & Almouzni, G. (2014). Establishment of a replication fork barrier following induction of DNA binding in mammalian cells. Cell Cycle, 13(10), 1607-1616.

Bodor, D. L., Mata, J. F., Sergeev, M., David, A. F., Salimian, K. J., Panchenko, T., ... & Jansen, L. E. (2014). The quantitative architecture of centromeric chromatin. *Elife*, *3*, e02137.

Burt, A., & Trivers, R. (2006). Genes in Conflict: The Biology of Selfish Genetic Elements (Belknap, Cambridge, MA).

Costantino, L., Sotiriou, S. K., Rantala, J. K., Magin, S., Mladenov, E., Helleday, T., ... & Halazonetis, T. D. (2014). Break-induced replication repair of damaged forks induces genomic duplications in human cells. Science, 343(6166), 88-91.

Crosetto, N., Mitra, A., Silva, M. J., Bienko, M., Dojer, N., Wang, Q., ... & Pasero, P. (2013). Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. Nature methods, 10(4), 361-365.

De Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., ... & Schmidt, J. M. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science, 354(6311), 477-481.

Dover, G. (1982). Molecular drive: a cohesive mode of species evolution. Nature, 299(5879), 111-117.

Drinnenberg, I. A., Henikoff, S., & Malik, H. S. (2016). Evolutionary turnover of kinetochore proteins: a ship of theseus?. Trends in cell biology, 26(7), 498-510.

Fachinetti, D., Han, J. S., McMahon, M. A., Ly, P., Abdullah, A., Wong, A. J., & Cleveland, D. W. (2015). DNA sequence-specific binding of CENP-B enhances the fidelity of human centromere function. Developmental cell, 33(3), 314-327.

Greenfeder, S. A., & Newlon, C. S. (1992). Replication forks pause at yeast centromeres. Molecular and cellular biology, 12(9), 4056-4066.

Haaf, T., & Willard, H. F. (1997). Chromosome-specific α-satellite DNA from the centromere of chimpanzee chromosome 4. Chromosoma, 106(4), 226-232.

Hastings P, Ira G, Lupski JR (2009A) A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS genetics 5, e1000327.

Hastings P, Lupski JR, Rosenberg SM, Ira G (2009B) Mechanisms of change in gene copy number. Nature Reviews Genetics 10, 551-564.

Henikoff, S., Ahmad, K., & Malik, H. S. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. Science, 293(5532), 1098-1102.

Hsiao M-C, Piotrowski A, Callens T, et al. (2015) Decoding NF1 intragenic copy-number variations. The American Journal of Human Genetics 97, 238-249.

Jørgensen, A. L., Laursen, H. B., Jones, C., & Bak, A. L. (1992). Evolutionarily different alphoid repeat DNA on homologous chromosomes in human and chimpanzee. Proceedings of the National Academy of Sciences, 89(8), 3310-3314.

Kobayashi, T. (2014). Ribosomal RNA gene repeats, their stability and cellular senescence. Proceedings of the Japan Academy, Series B, 90(4), 119-129.

Malik, H. S., & Henikoff, S. (2009). Major evolutionary transitions in centromere complexity. Cell, 138(6), 1067-1082.

Marques-Bonet, T., Ryder, O. A., & Eichler, E. E. (2009). Sequencing primate genomes: what have we learned?. Annual review of genomics and human genetics, 10, 355-386.

Marshall, O. J., Chueh, A. C., Wong, L. H., & Choo, K. A. (2008). Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. The American Journal of Human Genetics, 82(2), 261-282.

Masumoto, H., Ikeno, M., Nakano, M., Okazaki, T., Grimes, B., Cooke, H., & Suzuki, N. (1998). Assay of centromere function using a human artificial chromosome. Chromosoma, 107(6-7), 406-416.

Miga KH, Newton Y, Jain M, et al. (2014) Centromere reference models for human chromosomes X and Y satellite arrays. Genome research 24, 697-707.

Mikkelsen, T., Hillier, L., Eichler, E., Zody, M., Jaffe, D., Yang, S. P., ... & Archidiacono, N. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature, 437(7055), 69-87.

Mirkin, E. V., & Mirkin, S. M. (2007). Replication fork stalling at natural impediments. Microbiology and molecular biology reviews, 71(1), 13-35.

Mitra, S., Gómez-Raja, J., Larriba, G., Dubey, D. D., & Sanyal, K. (2014). Rad51–Rad52 mediated maintenance of centromeric chromatin in Candida albicans. PLoS Genet, 10(4), e1004344.

Molina, O., Kouprina, N., Masumoto, H., Larionov, V., & Earnshaw, W. C. (2017). Using human artificial chromosomes to study centromere assembly and function. Chromosoma, 126(5), 559-575.

Musacchio, A., & Desai, A. (2017). A molecular view of kinetochore assembly and function. Biology, 6(1), 5.

Ozenberger, B. A., & Roeder, G. S. (1991). A unique pathway of double-strand break repair operates in tandemly repeated genes. Molecular and cellular biology, 11(3), 1222-1231.

Poot RA, Bozhenok L, Berg DLvd, Hawkes N, Varga-Weisz PD (2005) Chromatin remodelling by WSTF-ISWI at the replication site: opening a window of opportunity for epigenetic inheritance? Cell Cycle 4, 543-546.

Puechberty, J., Laurent, A. M., Gimenez, S., Billault, A., Brun-Laurent, M. E., Calenda, A., ... & Roizès, G. (1999). Genetic and physical analyses of the centromeric and pericentromeric regions of human chromosome 5: recombination across 5cen. Genomics, 56(3), 274-287.

Rice, W. R. (2019A). A Game of Thrones at Human Centromeres I. Multifarious structure necessitates a new molecular/evolutionary model. BioRxiv, 731430.

Rice, W. R. (2019B). A Game of Thrones at Human Centromeres II. A new molecular/evolutionary model. BioRxiv, 731471.

Rosin, L. F., & Mellone, B. G. (2017). Centromeres drive a hard bargain. Trends in Genetics, 33(2), 101-117.

Ross JE, Woodlief KS, Sullivan BA (2016) Inheritance of the CENP-A chromatin domain is spatially and temporally constrained at human centromeres. Epigenetics & chromatin 9, 20.

Sakofsky, C. J., Ayyar, S., & Malkova, A. (2012). Break-induced replication and genome stability. Biomolecules, 2(4), 483-504.

Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K., & Willard, H. F. (2001). Genomic and genetic definition of a functional human centromere. Science, 294(5540), 109-115.

Schueler MG, Sullivan BA (2006) Structural and functional dynamics of human centromeric chromatin. Annu. Rev. Genomics Hum. Genet. 7, 301-313.

Shen, H., Li, J., Zhang, J., Xu, C., Jiang, Y., Wu, Z., ... & Tian, Q. (2013). Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians. PLoS One, 8(4), e59494.

Smith CE, Llorente B, Symington LS (2007) Template switching during break-induced replication. Nature 447, 102-105.

Tsouroula, K., Furst, A., Rogier, M., Heyer, V., Maglott-Roth, A., Ferrand, A., ... & Soutoglou, E. (2016). Temporal and spatial uncoupling of DNA double strand break repair pathways within mammalian heterochromatin. Molecular cell, 63(2), 293-305.

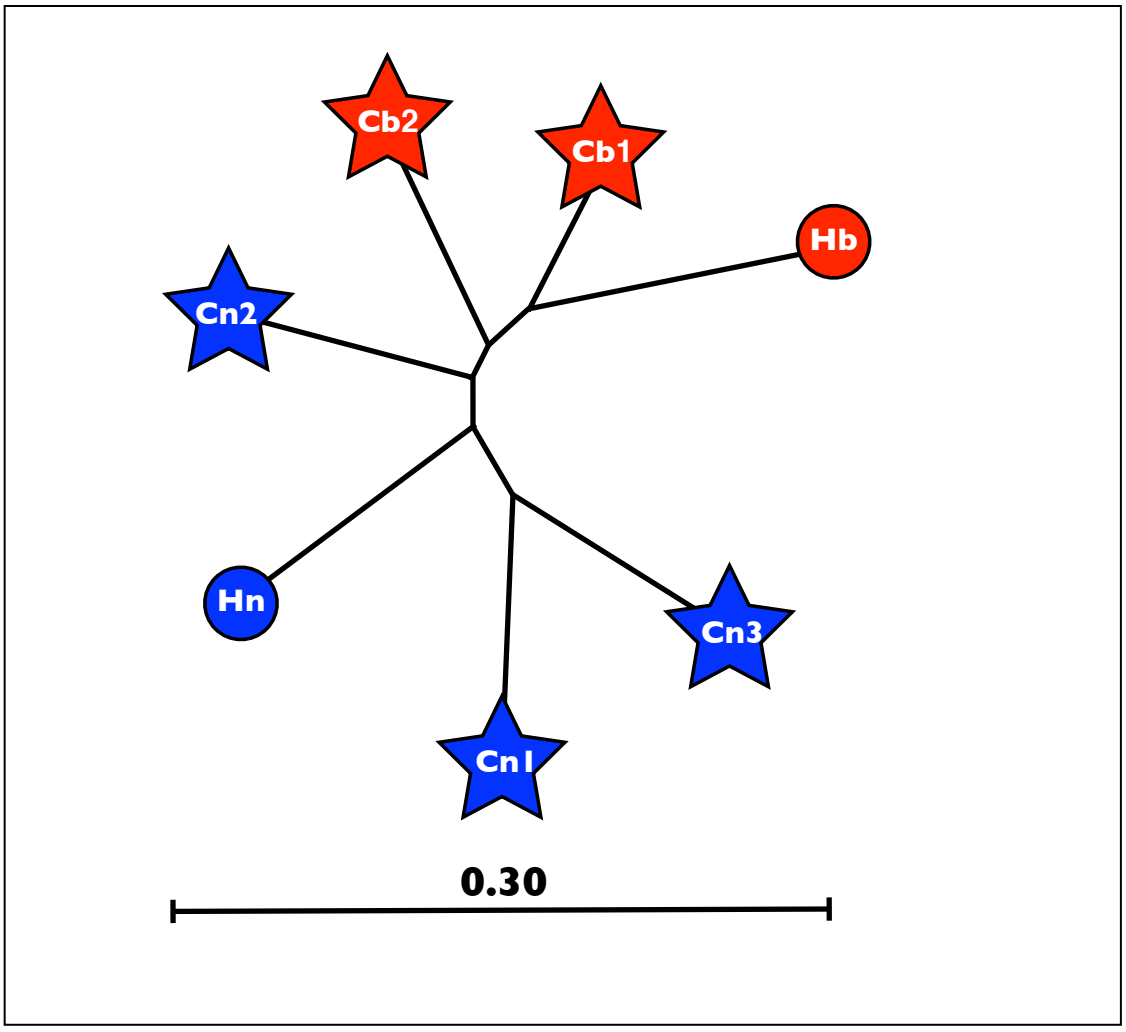
Warburton, P. E., Haaf, T., Gosden, J., Lawson, D., & Willard, H. F. (1996). Characterization of a chromosome-specific chimpanzee alpha satellite subset: evolutionary relationship to subsets on human chromosomes. Genomics, 33(2), 220-228.

Waye, J. S., England, S. B., & Willard, H. F. (1987). Genomic organization of alpha satellite DNA on human chromosome 7: evidence for two distinct alphoid domains on a single chromosome. Molecular and cellular biology, 7(1), 349-356.

Willard, H. F. (1991). Evolution of alpha satellite. Current opinion in genetics & development, 1(4), 509-514.

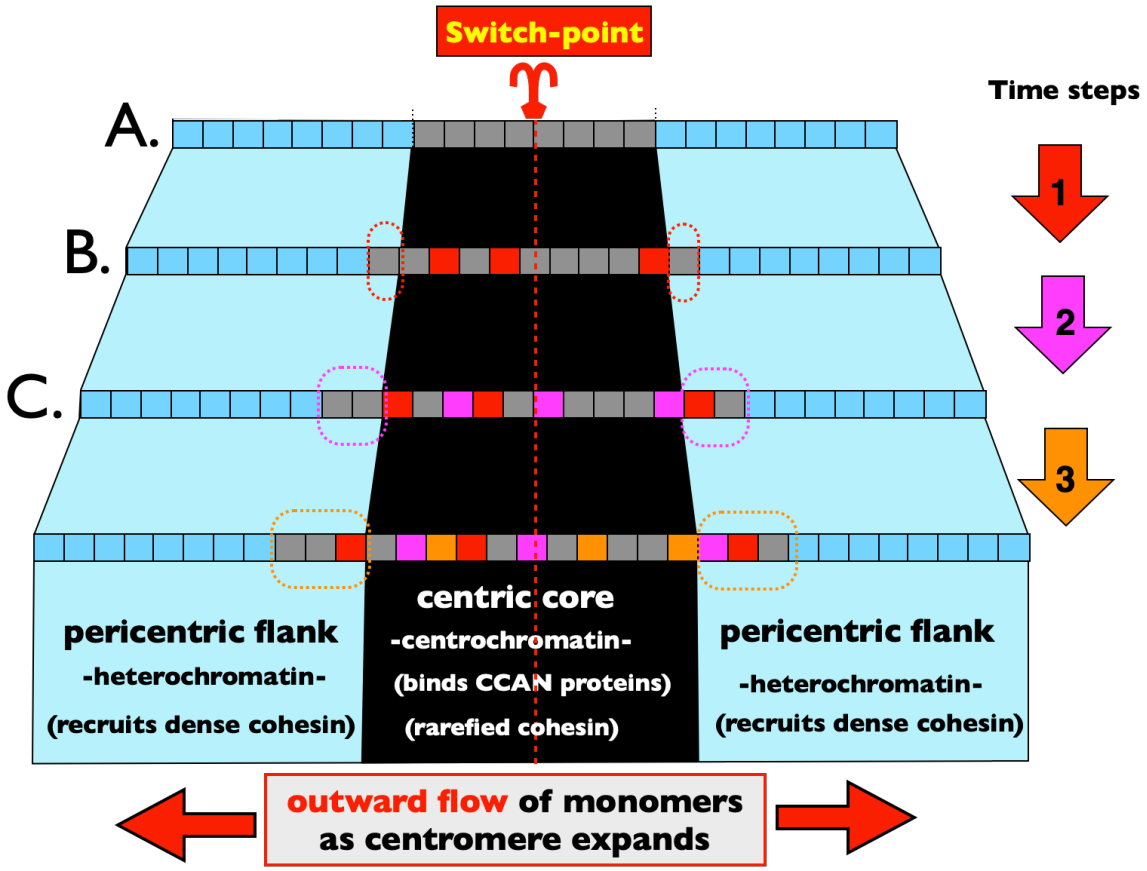
Zhang, F., Khajavi, M., Connolly, A. M., Towne, C. F., Batish, S. D., & Lupski, J. R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nature genetics, 41(7), 849-853.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2020 doi:10.20944/preprints202012.0669.v1



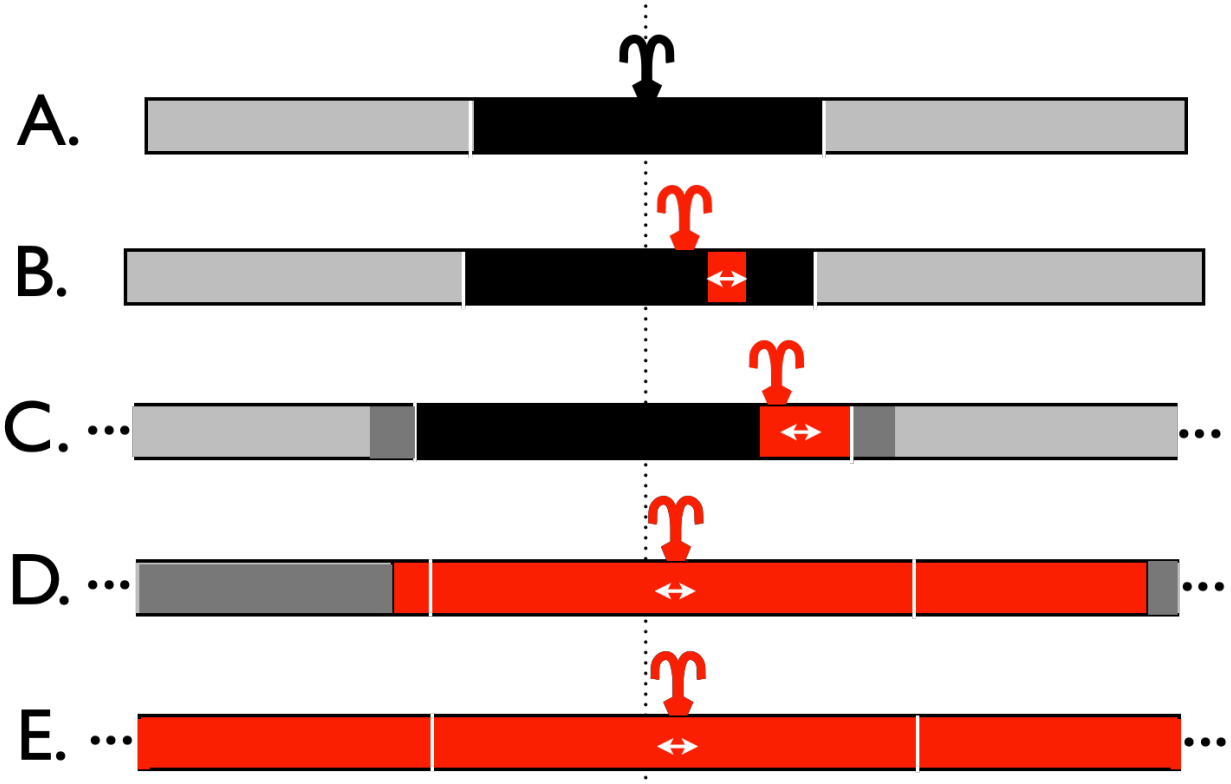
Supplemental Figure S1. A neighbor-joining cluster diagram of: i) the consensus sequence of the two monomers from the human dimeric HOR at chromosome 5 (one b-box monomer and one no-b-box monomer), and ii) the consensus sequences of the two b-box and three no-b-box monomers from the orthologous 5-monomer HOR in the chimp. Note the substantial sequence divergence. Red symbols are for b-box monomers and blue for no-b-box-monomers. Circles depict human monomers and stars depict chimp monomers. Symbol letters: H = human, C =chimp, b = b-box, n = no-b-box, numbers (when present) denote different monomers of the same type within a species.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2020 doi:10.20944/preprints202012.0669.v1



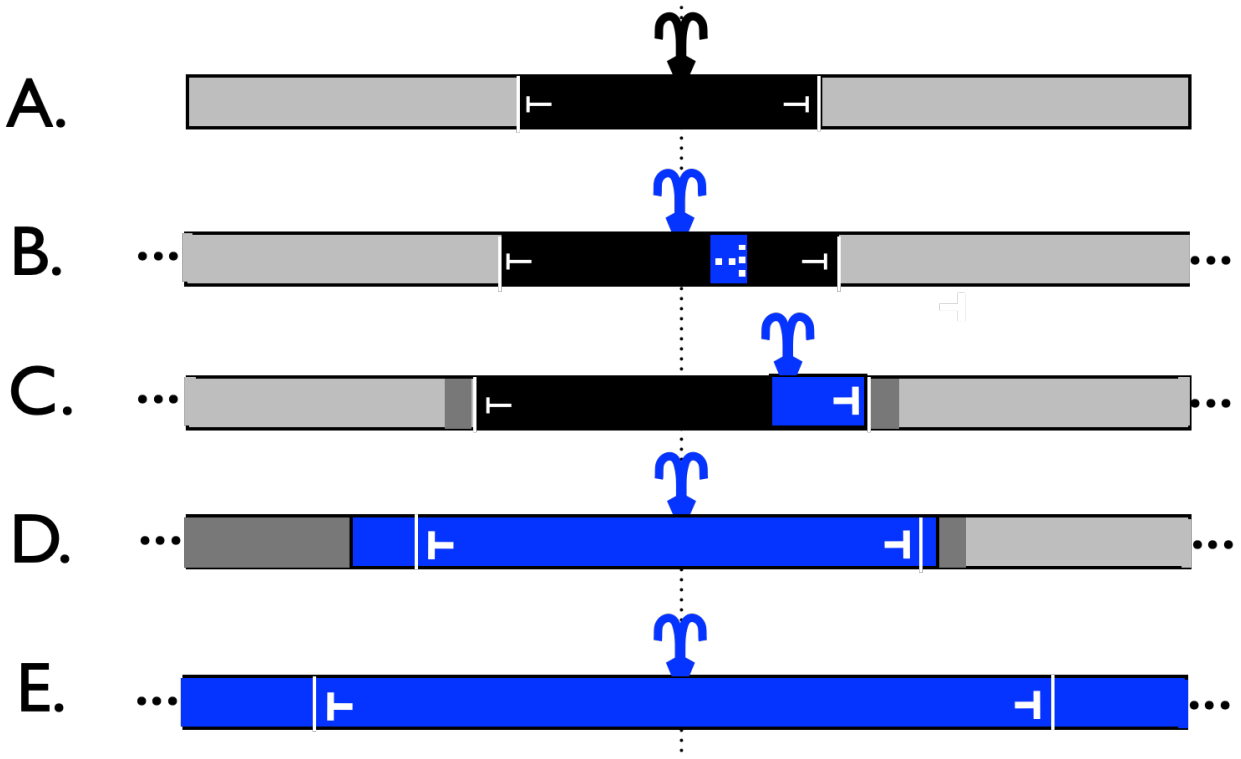
Supplemental Figure S2. The outward flow of monomers within a human centromeric tandem repeat array generates a 'Switch-point' that determines the 'winner' of molecular-drive. Consider a centromere array (DNA) composed of tandemly repeated units (assumed here to be monomers for simplicity) depicted by side-by-side rectangles. **A.** The array is partitioned into a centric core (assembled as centrochromatin) and pericentric flanks (assembled as heterochromatin) (Sullivan et al. 2011). Only the DNA of the centric core has the characteristics needed for expansion: bound protein and sparse cohesin clamps (Kobayashi 2014; Rice 2019). **B-D.** In each time step, three randomly positioned monomers within the centric core tandemly duplicate (colored rectangles) and cause the array to expand. Because the proportionate size of the centric core remains constant (about one third of the total; Sullivan et al. 2011; Ross et al. 2016), the centric/pericentric boundary ingresses as the centric core expands –causing the edges of the centric core to be pushed into the pericentric flanks (colored, dotted ovals). This ingression of the centric/pericentric boundary is plausibly due to dilution of a fixed amount of CENH3 (400 molecules per centric core; Bodor et al 2014) as the core expands (see Supplemental Figure S5 in Rice 2019B). Monomers entering the pericentric flanks from the edges of the centric core as it expands are expected to be eventually removed by recurrent deletion pressure –especially via SSA repair of double strand breaks (Bhargava et al. 2016). Recurrent expansions within the centric core generate a net outward flow of monomers, with the direction of flow reversing at a central location called the 'Switch-point' (red fountain symbol; Rice 2019B). The Switch-point is a key feature in molecular-drive at centromeres because any sequence variant that evolves to span this landmark will be spread bidirectionally and eventually populate the entire centromeric array, i.e., they will 'win' in the context of molecular-drive. The model developed here operates both when the tandem repeat array is composed of repeated monomers or HORs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2020 doi:10.20944/preprints202012.0669.v1



Supplemental Figure S3. A new subarray with a faster lateral expansion rate can 'capture' the Switch-point and 'win' in molecular-drive at human centromeres. A-B. Consider a centromeric tandem repeat array. The centric core [assembled as centrochromatin] of the array is shown in black, the pericentric flanks [assembled as heterochromatin] in grey, and the centric/pericentric boundary as a white line. The switch-point is shown by the fountain symbol. The centromeric repeat array has acquired (via transposition and tandem duplication) a new subarray that has a faster lateral expansion rate (red region with a white double-arrow; see Rice 2019B for molecular examples of this phenotype). The Switch-point can be defined as the position where the the average expansion rate is equal on both sides, assuming that the edges of the centric core are equally permeable to ingression by the pericentric flanks as tandem duplications within the centric core accumulate. The newly established subarray with faster lateral expansion will cause the right half of the centric core to expand faster and thereby move the Switch-point toward the right: the faster the lateral expansion rate of the new subarray, and the larger the size of this subarray, the greater the displacement of the Switch-point toward the new subarray. **C.** If during its expansion and movement toward the right, the new subarray 'captures' (spans) the Switch-point before it is pushed off the side of the centric core, its sequence will spread bidirectionally: causing the new subarray's sequence to eventually spread to the entire centric core (**D.**) and ultimately the entire centromeric array once monomers from the original centromeric array (that have been pushed into the pericentric flanks) are removed via recurrent deletion pressure (**E.**). Note that regions of the original centric core that recently have been pushed into the pericentric flanks are shown in dark grey. The model developed here operates both when the tandem repeat array is composed of repeated monomers or HORs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2020 doi:10.20944/preprints202012.0669.v1



Supplemental Figure S4. A new subarray with a higher resistance to ingression by the pericentric flanks can 'capture' the Switch-point and 'win' in molecular-drive at human centromeres. A-B. Consider a centromeric tandem repeat array. The centric core [assembled as centrochromatin] of the array is shown in black, the pericentric flanks [assembled as heterochromatin] in grey, and the centric/pericentric boundary as a white line. The switch-point is shown by the fountain symbol. The centromeric repeat array has acquired (via transposition and tandem duplication) a new subarray that has higher PHI-resistance (**P**ericentric **H**eterochromatin **I**nvasion **resistance;** blue region with a larger T-shaped arrow; see Rice 2019B for molecular examples of this phenotype) that is only expressed when the subarray resides at the edge of the centric core (dashed T-arrow becomes solid). Higher PHI-resistance reduces the exiting of monomers into the pericentric flanks, so that more monomers will exit on the opposite side with lower PHI-resistance as the centric core expands but remains the same proportionate size. When the edges of the centric core have unequal PHI-resistance, the Switch-point can be defined as the position where a monomer has an equal probability of eventually exiting the centric core on either of its sides. The newly established subarray with higher PHI-resistance will cause fewer monomers to exit on the edge of its side (right side) of the centric core and thereby moves the Switch-point closer to its position (toward the right in the figure): the stronger the PHI-resistance the greater the displacement of the Switch-point toward the new subarray. Initially the new subarray does not express its phenotype because of its central position but it does expand as it gradually moves toward the closest centric/pericentric boundary (on the right in the figure). **C.** When the new subarray reaches the right centric/pericentric boundary, it expresses its stronger PHI-resistance. If the combination of the new subarray's size and stronger PHI-resistance is sufficient, it will 'capture' the Switch-point before it is pushed off the side of the centric core and its sequence will spread bidirectionally: causing the new subarray's sequence to eventually spread to the entire centric core (**D.**) and ultimately the entire centromeric array once monomers from the original centromeric array (that have been pushed into the pericentric flanks) are removed via recurrent deletion pressure (**E.**). Note that regions of the original centric core that recently have been pushed into the pericentric flanks are shown in dark grey. The model developed here operated both when the tandem repeat array is composed of repeated monomers or HORs.