Supplementary file 1

1. The spread of cell-free DNA sequencing data

In order to investigate the composition of the DNA that is released by 143B cells after 24 hours of incubation, the cfDNA present in the culture medium was isolated and sequenced. Before proceeding with analyses, the spread of the distribution of the cfDNA sequencing data was evaluated. To eliminate potential sequencing artefacts, contigs with coverage less than 20 were excluded from analyses. Contig coverage was displayed as a box plot, with whiskers set at the 10^{th} and 90^{th} percentiles (Supplementary Fig. 1a). This graph shows that the data is skewed significantly to the right. Subsequently, the ROUT method was used to identify outliers, with Q set to 0.1% (the strictest threshold for defining outliers). Of the 4362 contigs that were analyzed, 549 contigs (all with coverage greater than 108) were identified as outliers. Although the latter represents only ~12.5% of the sequences, it constitutes ~53.5% of the data in terms of coverage. Furthermore, in order to discern between the two populations more clearly, the contigs with coverage greater than 1000 (~16% of the data, which skewed the distribution significantly) was omitted and the box plot was redrawn (Supplementary Fig. 1b). Alternatively, this data is presented as a scatter plot (Supplementary Fig. 1c). This clearly shows the presence of at least two cfDNA populations, namely a large number of contigs with a relatively low coverage, and a small number of contigs with a very high coverage.

2. Masking and representation of repetitive elements

To screen for repetitive elements (REs) and regions of low complexity, we used RepeatMasker Open software (4.0). To characterize the REs of the entire cfDNA population and to account for a non-normal distribution (as indicated in Supplementary Fig. 1), three data subsets were investigated, namely: (a) all contigs with coverage greater than 20, (b) contigs with coverage between 20 and 100, and (c) contigs with coverage greater than 100. To determine the representation of REs in each of these data sets, overlapping annotations were first omitted from the output files generated by RepeatMasker (~0.26% of the data). Subsequently, the number of bases masked by each RE in each query sequence was calculated and then multiplied by the coverage of the query sequence with which it aligned. The total number of bases masked by each RE was then added together. To determine the representation of each RE in the cfDNA population of each subset, the total number of bases masked by each RE was divided by the total number of bases in the entire cfDNA population. The total number of bases in the cfDNA population of each subset was calculated by first multiplying the length of each contig (that was submitted for RE screening) with its corresponding coverage value, and then adding it together. The representation of REs was normalized in terms of contig coverage based on the direct relationship between coverage and the number of bases (or copy number) as implied by the Lander/Waterman equation [1]. Furthermore, since the different REs are not equally represented in the human genome, RE representation was expressed as ratios of the percentage of cfDNA masked divided by the expected fraction of the human genome masked by the respective REs (masking values were obtained from RepeatMasker and a Cell SnapShot [2]).

In all subsets investigated only a very small portion of the cfDNA population consists of unique regions, while the amount of REs notably exceeds any value predicted for the human genome [3–5] (Fig. 1). Therefore, the DNA released by 143B cells after 24 hours of incubation is comprised mainly of REs. Furthermore, when taking into consideration the entire cfDNA population (Subset a), LINEs, SINEs, satellites, and simple repeats (mini satellites) make up the majority of the cfDNA population, and is overrepresented compared to the human genome (Fig. 1a). Very interestingly, satellites and simple repeats are significantly overrepresented, while LTR elements and DNA elements are underrepresented. As illustrated in Subset c (Fig. 1c), which depicts only the contigs with a coverage greater than 100, satellites, mini satellites and LINE elements are significantly overrepresented in the cfDNA population as a result of a small number of sequences that have a very high coverage. Therefore, when these contigs (which significantly skew the data) are taken out of consideration (Subset b) (Fig. 1b), it becomes clear that, regardless of the overall masking of each repeat class, specific elements in each class are significantly overrepresented, while others are significantly underrepresented or occur at levels comparable to the human genome: (i) regarding SINEs, *Alu*s are *2.7-fold* overrepresented, while MIRs are *7-fold* underrepresented, (ii) regarding LINEs, L1 is *1.5-fold* overrepresented, while L2 and L3/CR1 is *14.5-fold* and *28.5-fold* underrepresented, respectively, (iii) regarding LTR elements, ERV (K) class II and MaLR elements are *2.4-fold* and *1.8-fold* overrepresented, respectively, while ERV class I and ERV (L) class III elements occur at levels comparable to the human genome, (iv) regarding DNA elements, TcMar-Tigger elements are *1.7-fold* overrepresented, while hAT-Charlie elements are *3-fold* underrepresented.

2.1. Analysis of 143B genomic DNA

As a control, genomic DNA isolated from 143B cells was sequenced, subject to repeat masking, and compared to 143B cfDNA (Supplementary Fig. 2). For this analysis, a coverage range of 5-100 was selected. 143B genomic DNA shows almost normal levels of REs (˜59% of sequences are masked) and normal levels of non-repetitive DNA (gene sequences). In contrast, it is clear that REs are significantly overrepresented in cfDNA (~82% of sequences are masked), while cfDNA contains virtually no unique gene sequences. Interestingly, in 143B genomic DNA, *Alus* and ERV (K) class II elements were found at notably higher proportions, while the other ERVs as well as TcMar-Tigger and satDNA is only slightly elevated. It is possible that these higher proportions may be due to sequencing bias. A couple of research groups have also demonstrated higher than expected levels of *Alus* [6,7]. Alternatively, the higher levels of these elements may be a result of active retrotransposons, and/or gene amplification processes and chromosomal instability, which is a common phenomenon in many cancer types. The fact that cfDNA contains virtually no unique gene sequences, in contrast to 143B genomic DNA, suggests that cfDNA does not represent merely fragmented genomic DNA, but is actively released from specific parts of the genome, possibly as a result of chromosomal instability. This is further discussed in Section 3.4

3. Local alignment analyses and annotation

After repeat masking 250 contigs were, at random, selected for BLAST analyses. Since most of the sequences with a coverage greater than 20 were either completely masked or to a very small degree unmasked, the above selection ensured an accurate representation (Only 395 000 bases of the contigs were unmasked. Of the 250 contigs that were selected, 113 094 bases were unmasked, thus constituting ~30% of the unique bases). Of these sequences, 20% did not align with the human genome, while 36% aligned with a section of a gene. Interestingly, nearly one third of the sequences originate from the centromeres, notwithstanding the already established overrepresentation of satellites and simple repeats (as illustrated in Fig. 1). In addition, 11% of these sequences aligned with one unidentified position in the genome (Supplementary Fig. 3a). Furthermore, annotation of the cfDNA sequences that aligned with part of one gene revealed that less than 10% originate from protein coding regions, while more than 90% aligned with non-coding regions (Supplementary Fig. 3b). We can, therefore, conclude that the cfDNA that is released by 143B cells into the culture medium after 24 hours of incubation is primarily composed of non-coding DNA.

4. Assessing potential sequencing bias and procedural errors

Since REs are typically longer than the reads produced by NGS, they generally have substantially deeper coverage than unique regions [8]. Thus, it is important to evaluate potential coverage bias (deviation from the uniform distribution of reads across the genome). To do this, contigs were first grouped into different coverage ranges at increasing increments in the CLC genomics workbench, after which the FASTA sequences were exported. Each of these datasets was then subjected to repeat analysis and the level of RE masking was plotted against its corresponding coverage. This revealed that there is no correlation between coverage and the percentage of sequences occupied with repeats ($R^2 = 0.01$; $p = 0.72$) (Supplementary Fig. 4a).

Another potential cause of coverage bias is GC-content, since GC-rich and GC-poor regions are typically prone to low coverage [9]. As discussed in Section 2.2, there is considerable variation in total coverage and masking between the different RE populations. In addition, as discussed in Section 2.3, there is considerable variation in the coverage and masking of the contigs that constitute each of the different RE populations. Therefore, the correlation between GC-content and the level of masking by each of the different RE populations was evaluated. To do this, the average GC-content of the contigs that constitute each RE population was plotted against its observed masking value vs. expected masking ratio, in ascending order. This showed a weak negative correlation ($R^2 = 0.43$; $p = 0.022$) (Supplementary Fig. 4b). However, when excluding satDNA (which has the highest coverage and lowest GC-content), there is no longer a statistically significant correlation ($R^2 = 0.18$; $p = 0.20$) (Supplementary Fig. 4c). Furthermore, when examining the average GC-content of each of the different RE populations (Supplementary Fig. 4d), it is clear that none of these values is considered to be GC-rich or GC-poor, and should not cause coverage bias. Indeed, the observed GC-values of each RE population correlate with its expected value in the human genome [10]. Lastly, to account for structural differences between contigs with high- and low coverage, the average GC-content of the overrepresented RE subfamilies was compared to the average GC-content of its corresponding family, which showed that there are no significant differences between any of these groups (Supplementary Fig. 4e). Thus, cfDNA sequences obtained by our methodology are not biased towards GC-rich regions.

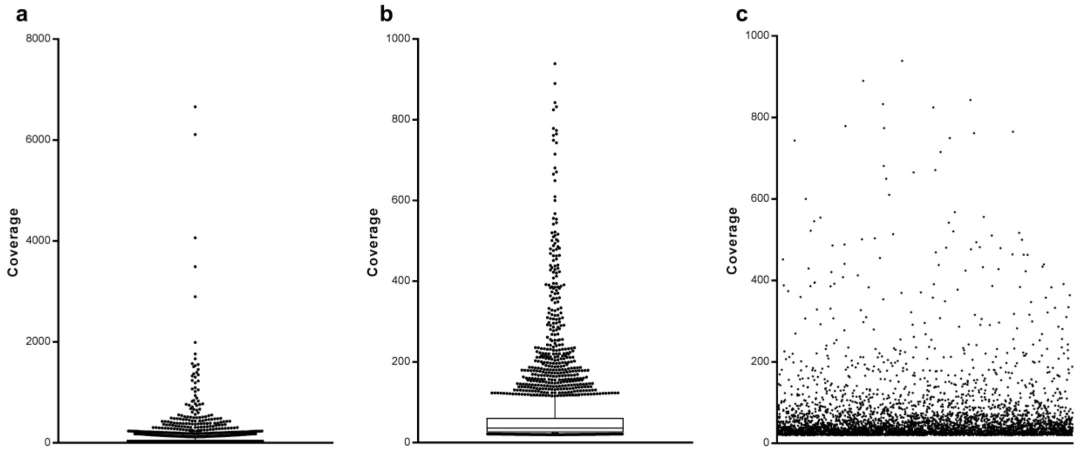
5. Identification of sequences that did not align with the human genome

To investigate the origin of the sequences that did not align with the human genome, a second search against the National Center for Biotechnology Information (NCBI) nucleotide collection (nr/nt) database was performed using the Megablast algorithm (optimized for highly similar sequences). For the sequences that again returned no results, a third search was performed using the Blastn algorithm (optimized for somewhat similar sequences). The top 10 scoring hits of each BLAST query were indexed, after which the hits with the highest maximum score and ID percentage of each query were tabulated. In addition, all binomial names were converted to their non-scientific counterparts (e.g., *Wuchereria bancrofti* = parasitic roundworm). Taken together, the majority of sequences appear to originate either from domesticated cattle, sheep and parasitic roundworms (that typically infect cattle). Thus, we can argue that the presence of these sequences can be ascribed mainly to the presence of these DNA sequences in the fetal bovine serum (FBS) that was used to fortify the growth medium. (Results are summarized in Supplementary file 4).

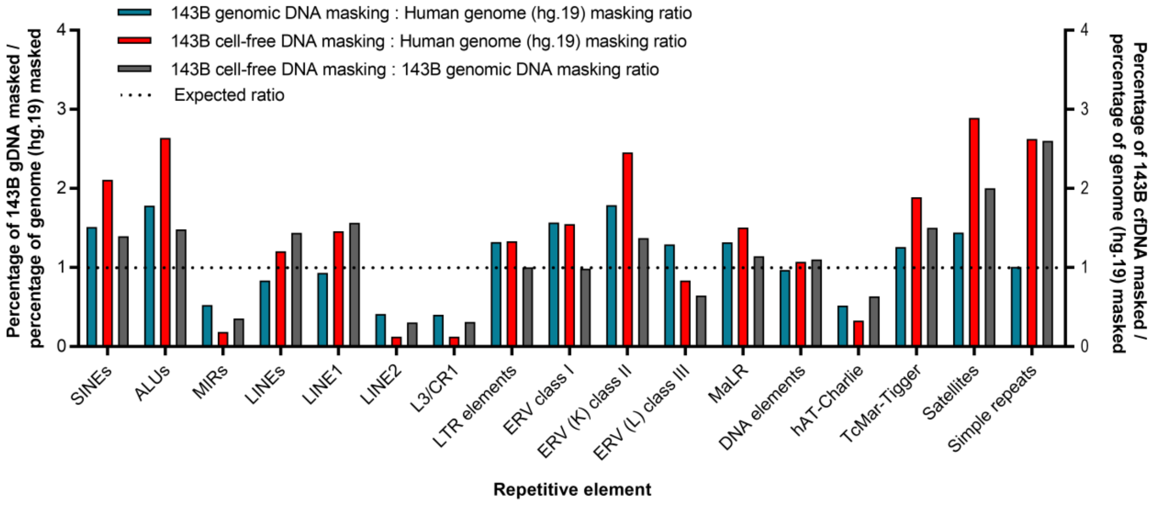
6. Movement of DNA transposons in the human genome

(i) Several cases of their horizontal transfer (HT) among insect species have been documented, which suggest that DNA elements rely heavily on HT for their evolutionary conservation [11,12], (ii) DNA transposons are particularly well adapted for HT, as several *in vitro* studies have shown that transposase is the only protein required for transposition [13], (iii) TcMar-Tigger elements are non-autonomous, i.e., they themselves encode for the transposase protein that is necessary for their mobilization, (iv) Our results show that TcMar-Tigger elements are overrepresented in the DNA that is actively released by cells, and (v) cfDNA fragments have been shown to be readily assimilated by most cells and can be integrated into the genome [14].

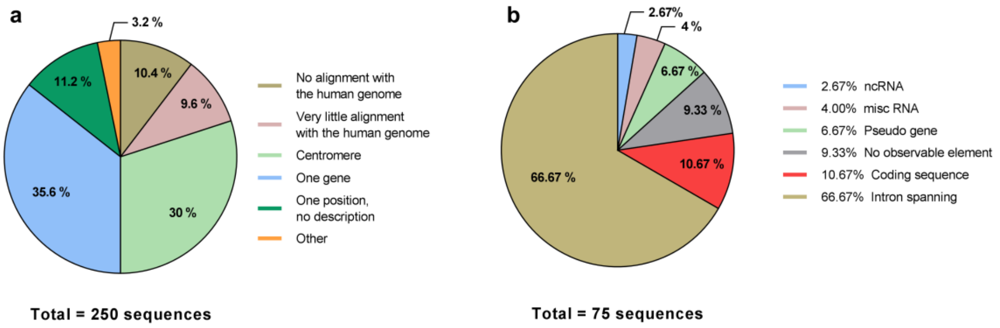
7. Supplementary figures



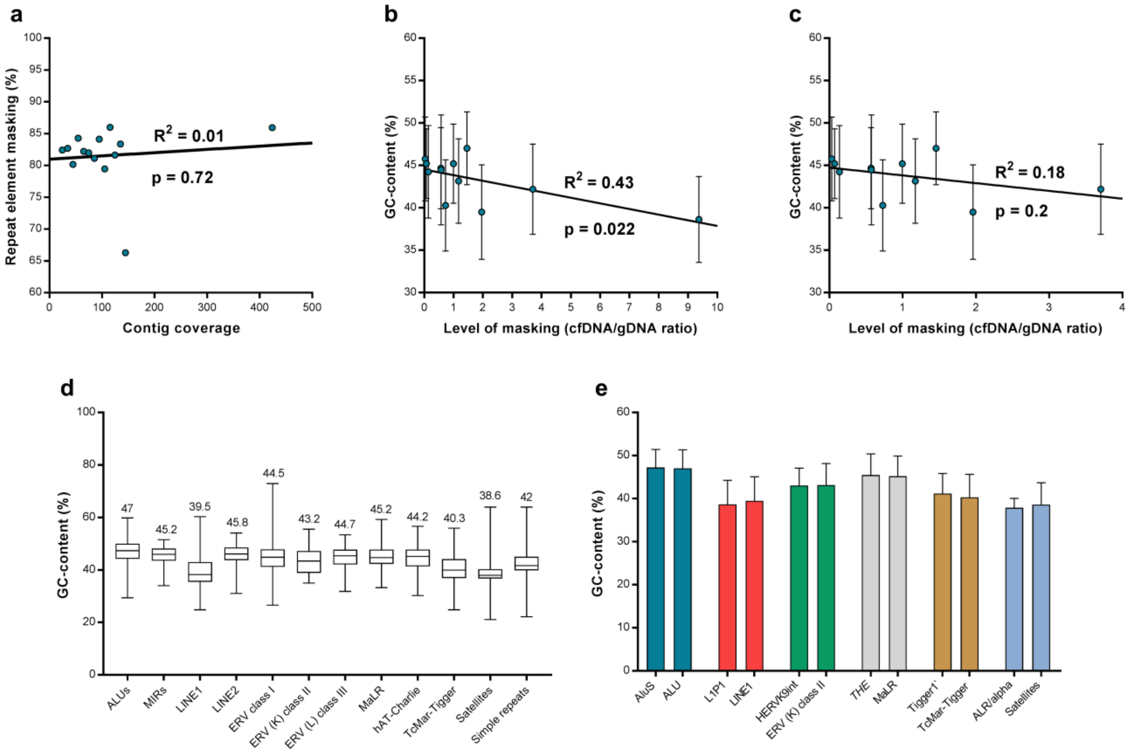
Supplementary Fig. 1. Coverage distribution of cell-free DNA sequences. (a) Coverage distribution of all cfDNA contigs with coverage greater than 20. (b) Box plot illustrating the coverage distribution of the data in a when contigs with coverag e greater than 1000 are omitted. (c) A scatter plot of the data presented in b.



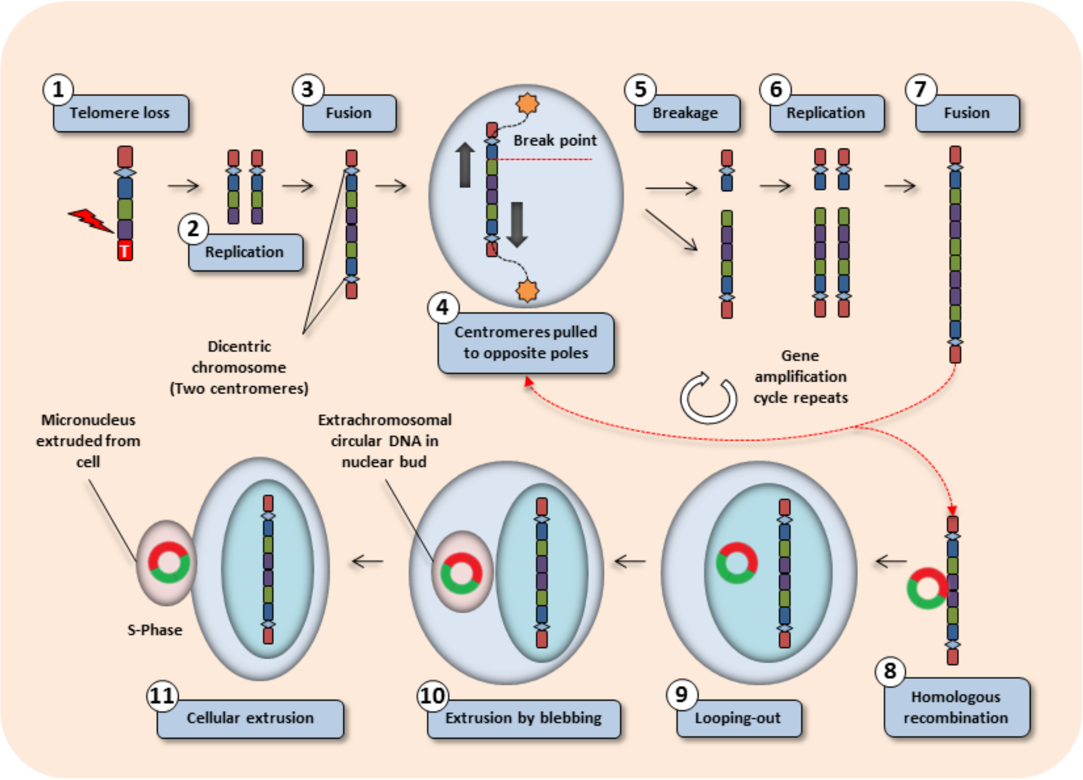
Supplementary Fig. 2. 143B cell-free DNA vs 143B genomic DNA masking. Representation of repetitive elements in both cell-free DNA and genomic DNA isolated from 143B cells. Bars illustrate the representation of each repetitive element type, expressed as observed/expected ratios (i.e., the percentage of cfDNA and gDNA masked by each repetitive element was divided by the percentage of the human genome masked by the corresponding element). Masking of 143B genomic DNA is plotted on the left y-axis (blue), while 143B cell-free DNA is plotted on the right y-axis (red). The dashed line denotes the expected ratio of masking for all repeats. Thus, values that extend above this line indicate an overrepresentation of the repetitive element, while values below this line indicate an underrepresentation of the repetitive element.



Supplementary Fig. 3. Local alignment analysis and annotation. (a) The position of masked cfDNA sequences in the human genome inferred from local alignment analyses. (b) Annotation of masked sequences originating from one gene.



Supplementary Fig. 4. Evaluation of coverage bias. (a) Correlation between the level of repeat masking and contig coverage. All contigs with coverage >20 were divided into a sliding window of coverage size with a 10 unit step up until 150 coverage, after which all contigs were grouped. The average coverage of each group was then plotted against its corresponding level of repeat masking. (b) Scatterplot showing the GC-content of each RE population against its level of masking. Elements are sorted from lowest to highest cfDNA/gDNA masking ratios (as determined in section 2.2). (c) Scatterplot showing the GC-Content of each RE population against its level of masking when the satDNA group is omitted. In each graph, the solid black line indicates the linear regression line and the (R2) values indicate the regression coefficient, where values close to 1 denote a perfect correlation and values close to 0 denote no correlation. Statistical significance is indicated by (p), where a p-value of less than 0.05 indicates a statistically significant result. (d) Box plots (min/max) illustrating the GC-content distribution of the contigs that comprise each RE population. The average GC-content of each RE population is indicated above the maximum whisker. (e) Comparison of the GC-content of overrepresented RE subfamilies with is corresponding RE class. Error bars indicate standard deviation.


1 Telomere loss
T
2 Replication
3 Fusion
Dicentric chromosome (Two centromeres)
Break point
4 Centromeres pulled to opposite poles
5 Breakage
6 Replication
7 Fusion
Gene amplification cycle repeats
8 Homologous recombination
9 Looping-out
Extrachromosomal circular DNA in nuclear bud
10 Extrusion by blebbing
Micronucleus extruded from cell
S-Phase
11 Cellular extrusion

Supplementary Fig. 5. Breakage-fusion-bridge cycle.

8. References

1. Illumina. Estimating Sequencing Coverage.

2. Mandal PK, Kazazian HH. SnapShot: Vertebrate Transposons. *Cell*. Epub ahead of print 2008. DOI: 10.1016/j.cell.2008.09.028.

3. E. S. Lander et al. Initial sequencing and analysis of the human genome. 409.

4. de Koning APJ, Gu W, Castoe TA, et al. Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genet*. Epub ahead of print 2011. DOI: 10.1371/journal.pgen.1002384.

5. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*. Epub ahead of print 2012. DOI: 10.1038/nrg3174.

6. Beck J, Urnovitz HB, Riggert J, et al. Profile of the Circulating DNA in apparently healthy individuals. *Clin Chem*. Epub ahead of print 2009. DOI: 10.1373/clinchem.2008.113597.

7. Stroun M, Lyautey J, Lederrey C, et al. Alu Repeat Sequences Are Present in Increased Proportions Compared to a Unique Gene in Plasma/Serum DNA Evidence for a Preferential Release from Viable Cells?

8. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics*. Epub ahead of print 2012. DOI: 10.1038/nrg3117.

9. Ross MG, Russ C, Costello M, et al. Characterizing and measuring bias in sequence data. *Genome Biol*. Epub ahead of print 2013. DOI: 10.1186/gb-2013-14-5-r51.

10. Medstrand P, Van De Lagemaat LN, Mager DL. Retroelement Distributions in the Human Genome: Variations Associated With Age and Proximity to Genes.

11. De Almeida LM, Carareto CMA. Multiple events of horizontal transfer of the Minos transposable element between Drosophila species. *Mol Phylogenet Evol*. Epub ahead of print 2005. DOI: 10.1016/j.ympev.2004.11.026.

12. Robertson KD. DNA methylation and chromatin – unraveling the tangled web. *Oncogene* 2002; 21: 5361–5379.

13. Craig JM, Earle E, Canham P, et al. Analysis of mammalian proteins involved in chromatin modification reveals new metaphase centromeric proteins and distinct chromosomal distribution patterns. *Hum Mol Genet*. Epub ahead of print 2003. DOI: 10.1093/hmg/ddg330.

14. Mittra I, Khare NK, Raghuram GV, et al. Circulating nucleic acids damage DNA of healthy cells by integrating into their genomes. *J Biosci*. Epub ahead of print 2015. DOI: 10.1007/s12038-015-9508-6.