SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
BIOINFORMATICS
BIOPOLYMERS IN VIVO

Received
19 September 2013

Accepted
13 March 2014

Published
31 March 2014

Correspondence and requests for materials should be addressed to S.L.G. (grok@eimb.ru)

Non-random DNA fragmentation in next-generation sequencing

Maria S. Poptsova[1], Irina A. Il'icheva[2], Dmitry Yu. Nechipurenko[1], Larisa A. Panchenko[3], Mingian V. Khodikov[2], Nina Y. Oparina[2], Robert V. Polozov[4], Yury D. Nechipurenko[1,2] & Sergei L. Grokhovsky[2]

[1]Department of Physics, Moscow State University, Moscow, Russia, [2]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia, [3]Department of Biology, Moscow State University, Moscow, Russia, [4]Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, Puschino, Moscow Region, Russia.

Next Generation Sequencing (NGS) technology is based on cutting DNA into small fragments, and their massive parallel sequencing. The multiple overlapping segments termed "reads" are assembled into a contiguous sequence. To reduce sequencing errors, every genome region should be sequenced several dozen times. This sequencing approach is based on the assumption that genomic DNA breaks are random and sequence-independent. However, previously we showed that for the sonicated restriction DNA fragments the rates of double-stranded breaks depend on the nucleotide sequence. In this work we analyzed genomic reads from NGS data and discovered that fragmentation methods based on the action of the hydrodynamic forces on DNA, produce similar bias. Consideration of this non-random DNA fragmentation may allow one to unravel what factors and to what extent influence the non-uniform coverage of various genomic regions.

The genome of a living organism resembles a bookshelf filled with books – chromosomes containing texts made up of letters – nucleotides. The early methods of deciphering biological sequences were based on the precise excision of a particular DNA fragment and its accurate reading. An alternative and, as it initially seemed, incoherent, sequencing method was proposed back in 1979[1], whereby the multiple copies of a whole genome DNA were to be broken up into small fragments, which were sequenced, and then these sequences (termed "reads") were assembled into a continuous text based on the overlapping ends. Nevertheless, the explosive development of automated sequencers and advances in computational power determined the present-time dominance of this method, called random shotgun sequencing. Modern sequencing machines are capable of reading hundreds of millions of reads per day, where each read consists of tens or hundreds of nucleotides.

The first step of DNA sequencing in the NGS technology is DNA fragmentation. Samples of purified DNA are sheared into short fragments, using either mechanical methods (e.g., ultrasonication shearing and nebulization) or enzymatic digestion[2]. The fragmented DNA is ligated at both blunt ends of each fragment with specific adaptors, which serve as primer-binding sites for amplification. Then the adaptor-ligated DNA fragments are size-selected through agarose gel electrophoresis or with paramagnetic beads; at this step the ligation duplicates are removed. Subsequently DNA fragments are melted, and the single-stranded DNAs are immobilized either on planar solid surfaces of a flow cell (Illumina sequencers), or on the surface of micron-scale beads (454-Roche and SOLiD sequencers), or on ionized spheres (Ion Torrent sequencers)[3]. Template amplification is performed by PCR on solid surface, or by emulsion PCR into separate microreactors, beads or spheres within sequencers. Finally, sequencing is achieved by detecting the emission of light or hydrogen ions from every dot on the solid surface or spheres, during enzymatic attachment of complimentary nucleotides to the clusters of identical single-stranded DNA fragments[4].

The required level of resolution for an NGS experiment is achieved by providing sufficient coverage, which generally refers to the average number of reads that align to each base within the sample DNA. Every DNA region must be represented multiple times in different read frames or, in other words, the sequences of the fragments (and thus the reads) must overlap. In today's NGS protocols purified DNA is obtained from many cells, and DNA shearing is performed on multiple genome copies providing a sufficient number of overlapping fragments. In single-cell sequencing in order to generate the sufficient number of overlapping reads, DNA is PCR-amplified prior to fragmentation[5]. In both approaches, for unambiguous determination of the whole genome sequence the overall length of the sequenced reads has to exceed the genome size by a dozen of times[6].

A major stumbling block in shotgun sequencing is genomic repeats, the length of which might significantly exceed the size of reads. The repeats are abundant in genomes of eukaryotes[7]. Transposable elements, for example, make up for 45% of the human genome. Among them, Alu, a retrotransposon of ~300 bp in length, is the most common repetitive element with more than a million copies, which comprise 10% of the human DNA[8]. In addition, eukaryotic DNA consists of very large arrays of short, repeated sequences (satellite DNA) near the centromeric region (they can be several megabases long without an interruption) and long repeated DNA sequences in the telomeres (at the ends of linear eukaryotic chromosomes). The presence of repeats may lead to systematic sequencing errors in regions containing many copies of repetitive elements[9–11]. The precise identification of the number of repetitive elements is particularly important for medical research due to existence of polymorphic copy number variants between individuals[12]. Moreover, changes in copy number were found in different tissues of the same organism and in cancer cells[13]. Lately it has been shown that the correct calling of copy number variants is crucial for studying spatial dynamics of genome replication[14].

The development of modern sequencing technologies requires novel methods for precise identification of copy number variants or the number of repetitive elements in the genomes. However the existing methods are hindered by sequencing bias, which leads to over- or undersampling of certain regions of the genome, lowering their resolution and undermining the researcher's ability to accurately identify the mutations and duplicated regions[15]. To avoid sequencing ambiguity produced by the modern sequencers the genome coverage must be high and even[16].

The bias in NGS data has been extensively observed[17–23], but there is no agreement with respect to the sources of the observed bias. Thus, Benjamini and Speed[17] reported the regularities in the GC bias patterns, and found that GC content influences fragment count the most. Since both GC-rich and AT-rich fragments were underrepresented in the sequencing results, the authors hypothesized that the most important cause of the GC bias was PCR. Hansen *et al.*[24] noticed the existence of bias for PCR with random hexamer priming. The authors also reported bias in the first position of reads, but did not determine the source of this bias[24]. Van Heesch, S. *et al.*[25] found that the bias is tissue-specific and related to specific chromatin characteristics. Out of four tissues sampled, brain tissue showed the lowest variation in NGS read coverage, while more homogeneous tissues like blood and liver exhibited the largest bias in read coverage. Auerbach *et al.*[26] showed that in addition to non-uniform read coverage, the read mapping procedure could generate regional bias. Heavy amplification in single-cell sequencing induces bias, which can lead to uneven coverage[5].

We propose to investigate the bias that originated at the fragmentation stage of NGS sequencing procedures. For the first time we venture to show that the bias in NGS reads strongly correlates with the bias produced by sonication of pure restriction DNA fragments, which was shown to be sequence-specific[27]. Hence we report and analyze bias produced by three DNA shearing methods, sonication, nebulization and Covaris, and demonstrate that the instances of bias in different fragmentation methods highly correlate with each other and are common for all of the hydrodynamic DNA shearing methods.

Results

In 2006 we discovered that double-strand breaks resulting from sonication at 22 and 44 kHz of double-stranded DNA fragments with the known nucleotide sequences occurring preferentially in 5′-CpG-3′ dinucleotides[28]. The strand was broken between C and G so that the phosphate group was at the 5′ side of G in the products. The cleavage rate proved to be dependent on the sequences flanking the cleavage site. Subsequent statistical investigations of the sequence-specific influence on cleavage intensities of individual phosphodiester bonds were performed on a data set, which consisted of approximately 20 thousand bands of high-resolution sequencing gels[27,29]. A data set of cleavage rates for all possible di- and tetranucleotides was obtained. This testified to a remarkable enhancement of the cleavage rates of phosphodiester bonds after deoxycytidine, which diminished in the following row of dinucleotides: d(CpG) > d(CpA) ~ d(CpT) > d(CpC). Thus, sonication of DNA restriction fragments leads to the sequence-dependent distribution of the positions of double-stranded DNA breaks. Since DNA fragmentation is the first stage of the contemporary NGS technology, we decided to test if similar regularities were observed when fragmentation was done with ultrasound.

In different NGS platforms three physical shearing methods are now commonly used: nebulization, sonication and adaptive focused acoustics technology (Covaris). These methods produce DNA fragments with heterogeneous ends containing 5′- or 3′-overhangs. To produce blunt ends for primers ligation, T4 DNA polymerase is commonly used, which has single-strand exonuclease activity that removes 3′-overhangs, and DNA polymerase activity that fills in 5′-overhangs. DNA treatment with DNA polymerase could change the initial cleavage points of the double-stranded DNA from the 3′-end, but the cleavage point from the 5′-end remains intact.

First of all we took thirteen samples of human DNA-sequencing data from 1000 Genomes Project[30], one sample of *E.coli* DNA-sequencing data from the Sequence Read Archives (SRA) at NCBI, one sample of *Drosophila mauritiana* genome from University of Veterinary Medicine, Vienna[31], and one sample from *Arabadiposis thaliana* genome from 1001 Genomes project (*see* Supplementary Table 1 for details). Raw reads were aligned to the reference genomes and frequencies of all 4 nucleotides, 16 dinucleotides and 256 tetranucleotides were calculated for each sample at the positions of DNA breaks (see Methods and Supplementary Tables 2–4). The process of read library creation leaves intact the positions prior to the cleavage, which corresponds to the 5′-ends of the reads, so the alignment of the reads to the reference genomes provides the required information about the sequence-specificity of fragmentation.

For the reads fragmented with ultrasound, a typical dependence of mononucleotide frequencies for the region of 200 bp (+/−100 bp) around break point is presented in Fig. 1. One can clearly see the enrichment of C at the position 100, which corresponds to a position prior to the cleavage in the reference genomic sequence adjoined to the 5′-end of the read. Similarly, bias around the break point for all di- and selected NCGN tetranucleotides is presented in Fig. 2 and Fig. 3, correspondingly. Dependence for all 16 NCGN tetranucleotides is presented in Supplementary Fig. 1. In all these figures one may observe bias around breakpoints, which was previously reported by Lazarovici *et al.*[32]. We also observed similar bias of mono-, di- and tetranucleotide frequencies for the region around break point for reads fragmented with nebulization and Covaris (data is not shown). Here we present a comparison of different fragmentation methods for different genomes and sequencing centers with our experimental data at the level of mono-, di- and tetranucleotides.

We begin the exploratory studies looking for structure in these data with the Pearson's correlation matrix for dinucleotide and tetranucleotide relative frequencies centered at the 5′-ends of the reads. At the level of mononucleotides the relative cleavage rates of raw sequencing data from several independent laboratories, where fragmentation was done with ultrasound, and the relative cleavage rates of our experimental data are shown in Fig. 4.

For dinucleotides, all seven analyzed samples from several independent laboratories, where fragmentation was done with ultrasound, show significant correlation with our data (Pearson r = 0.816–0.951). The best correlation (r = 0.961) is observed between our data set and the average data sets of seven ultrasound samples (see Supplementary Table 5).

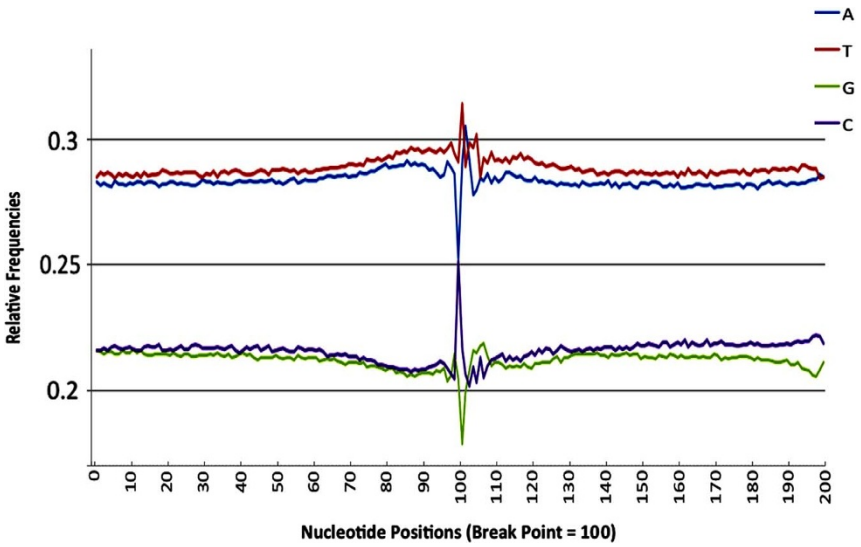

Figure 1 | **Mononucleotide frequencies for the region of 200 bp (+/−100 bp) around break point for sonication method.**

Similarly, for tetranucleotides all seven ultrasound samples show significant correlation with our data (r = 0.669–0.866). The best correlation (r = 0.872) is observed between our data set and the average data sets of seven ultrasound samples (see Supplementary Table 6).

NGS data from experiments that employed other methods of DNA fragmentation, that is, nebulization and Covaris, were taken from 1000 Genomes Project, 1001 Genomes Project and from the University of Veterinary Medicine, Vienna. Remarkably, our ultrasonic cleavage rates showed high correlation with the average cleavage rates for nebulization and Covaris both for dinucleotides (r = 0.972 and r = 0.925; Supplementary Table 5) and tetranucleotides (r = 0.902 and r = 0.824; Fig. 5, Supplementary Table 6).

Then, for all analyzed samples, we performed hierarchical cluster analysis and the resulting tree is presented in Supplementary Fig. 2. The choice of clustering algorithm did not affect the shape of the cluster tree. As one can see from the cluster tree, neither fragmentation methods nor species group together. This result indicates that all investigated fragmentation methods have common physicochemical nature coupled with mechanochemical breakage of DNA due to shearing forces, which originate in high-gradient liquid flows[33,34].

Therefore a comparison of the cleavage rates obtained from the studied fragmentation methods with the values resulting from our experiment on DNA restricted fragments, which characterize purely the ultrasonic fragmentation specificity bias, might elucidate another cause of the systematic bias. The bias may originate from various procedures, such as PCR amplification, primer ligation, primary fragment separation in agarose gel or computer processing of sequenced data[2,24,25]. Moreover, the enhancement in cleavage rates observed for d(CpG) dinucleotide might be caused by epigenetic modifications, such as cytosine methylation. According to our preliminary results methylation of cytosine in the 5-position increases the relative cleavage intensity of the d(CpG) dinucleotide, but the stability of this effect and its physical nature require further investigations.

It is evident from the analysis of DNA cleavage rates for genomes of *E. coli*, *Drosophila mauritiana*, *Arabidopsis thaliana*, as well as for

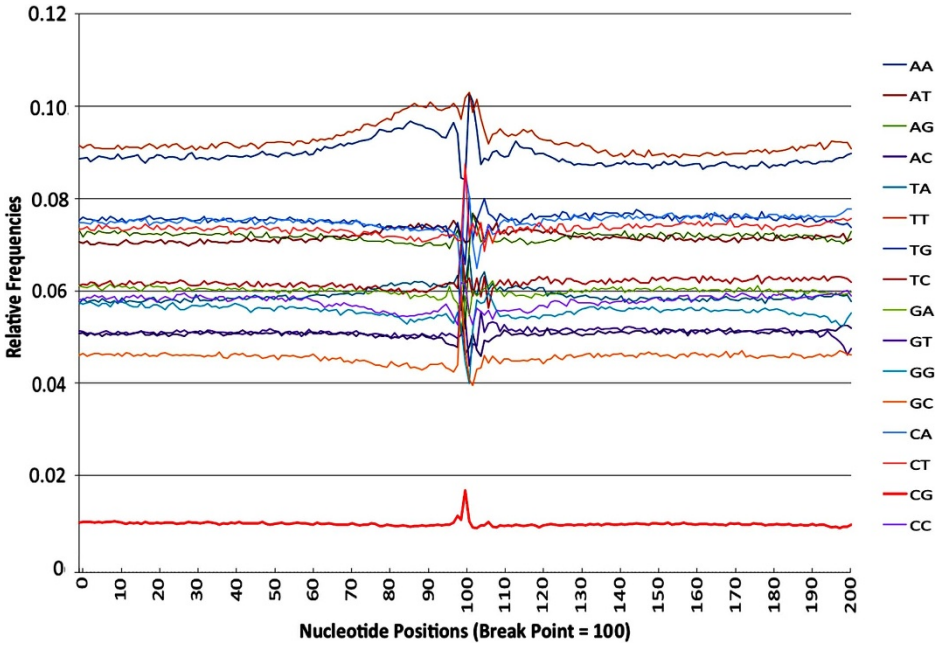

Figure 2 | **Dinucleotide frequencies for the region of 200 bp (+/−100 bp) around break point for sonication method.**

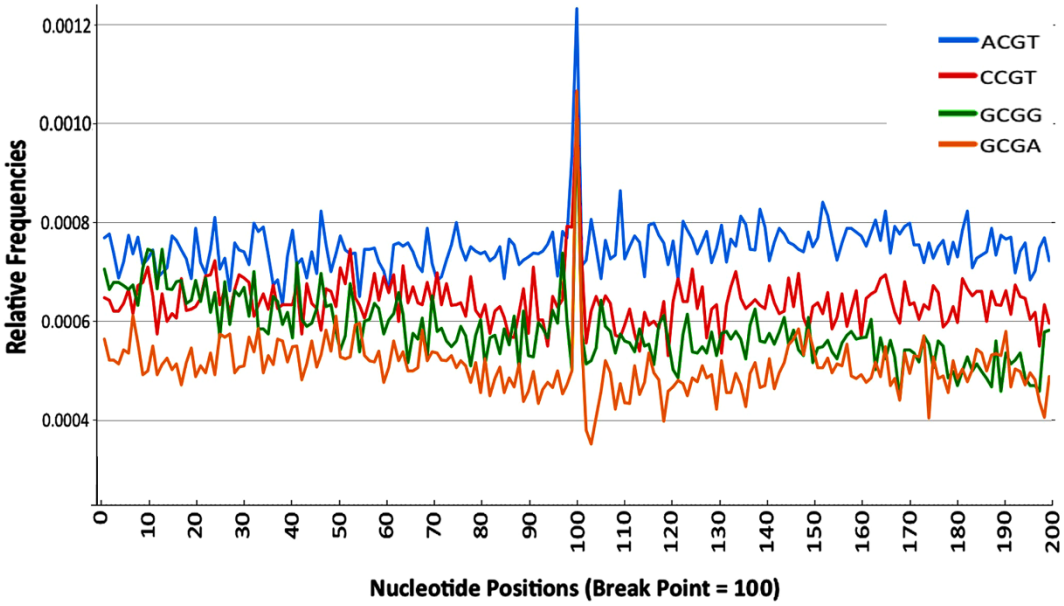

Figure 3 | Four selected tetranucleotide frequencies for the region of 200 bp (+/−100 bp) around break point for sonication method.

the human genome, that cleavage rates of mono-, di- and tetranulceotides do not depend on species. However, in general, this may not be true because different species can have various base modifications.

Discussion

Previously we have shown that for the sonicated restriction DNA fragments the comparative rates of double-stranded breaks depend on the nucleotide sequence[27,28]. Here we analyzed genomic DNA reads from NGS data generated by the ultrasonic shearing methods. After aligning the reads to the reference genomes we found that at the positions of DNA breaks the nucleotide relative frequencies at their 5′-ends are in good agreement with the cleavage rates obtained in our experiments. We analyzed reads from other DNA fragmentation methods – nebulization and Covaris – and discovered that these methods, based on the action of hydrodynamic forces, also produce similar bias.

Since NGS methods commonly use hydrodynamic cleavage DNA and basically imply that this cleavage is non-specific, the observed effect of the sequence specificity of ultrasound DNA cleavage should be taken into account in order to avoid systematic errors during sequence assembly. Thus the sequence-specific fragmentation bias evaluation might lower the uncertainties arising during DNA copy number calculation[15,35].

It is generally accepted that in regions of elevated GC content the number of reads was increased due to the PCR-amplification step[16,36,37]. But in accordance with our results, excessively intense ultrasound treatment of genomic DNA could induce amplified cleavage of GC-rich areas of genome. Further removal of the shortest fragments in agarose gel may actually cause AT-bias. So, an estimation of the effect of the relative cleavage rates is complicated and the proposed bias corrections are not straightforward and require further development.

An analysis of this non-random DNA fragmentation allows one to unravel what factors and to what extent influence the non-uniform coverage of various genomic regions. A selection of some specific conditions and reagents might diminish the bias in the DNA fragmentation, and reduce the amount of repetitive sequencing runs. Moreover, further studies of the comparative DNA cleavage rates of the base modified nucleotides might serve as basis for the development of new methods for identification of the epigenetic patterns.

Recently we demonstrated that addition of particular metallic ions (Ag^+, for example) results in sufficient impairment of the observed ultrasonic fragmentation bias[33]. So, by choosing the proper chemical agents, it is possible to lower the systematic bias associated with the action of high-gradient liquid flows.

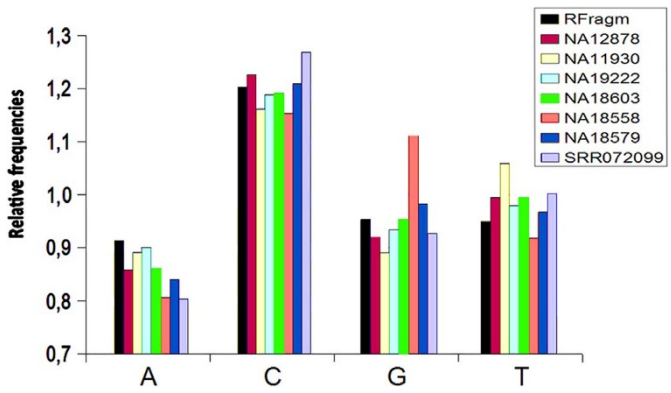

Figure 4 | Comparison of the relative cleavage rates for 4 mononucleotides derived from NGS data fragmented by ultrasound and from our experiments[27].

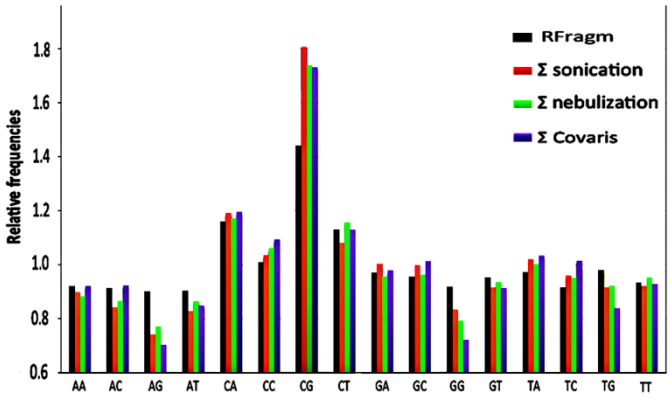

Figure 5 | Comparison of the relative cleavage rates for 16 dinucleotides derived from NGS data fragmented by various methods and from our experiments[27].

The observed sequence dependence of DNA breakage by ultrasound, nebulization and Covaris and the high correlation of DNA relative cleavage rates between all three methods are quite surprising. It seems that sequence-specificity of hydrodynamic shearing reflects local variations in DNA structural dynamics. Hence, the reported DNA cleavage bias may also provide a basis for developing new methods of studying sequence effects on local structural dynamics of the DNA sugar phosphate backbone.

Thus, the reported specificity of DNA fragmentation with ultrasound, nebulization and Covaris is important not only for improvement of DNA sequencing protocols, but may also be interesting for the study of DNA structure and its dependence on sequence context. In tetranucleotides the effect of flanking nucleotides on the cleavage rates of all 16 types of central dinucleotides was also reported as statistically significant. The sequence-dependent ultrasonic cleavage rates of dinucleotides were consistent with the reported data on the intensity of the conformational motion of their 5′-deoxyribose. The sequence specificity of ultrasonic cleavage is the result of sequence-dependent conformational dynamics, and is likely modulated by the intensity of the sugar ring S<=>N interconversion[27,38,39]. Moreover, the enhanced ultrasonic cleavage of dCpG dinucleotide might also reflect its functional role: epigenetic mechanisms based on d(CpG) methylation[32,40] might be the consequence of the unique properties of this dinucleotide. DNA methyltransferase enzyme may recognize the unusual conformational dynamics of the d(CpG) dinucleotide and flip cytosines out of the DNA helix during methylation more efficiently than the other bases.

It is of particular interest that cleavage rates for complementary pairs of dinucleotides are not identical[27]. Moreover, they differ from each other to a varying degree. The most pronounced is the difference between the cleavage rates in two complementary dinucleotides, namely AG/CT and CA/TG. This difference exceeds a quarter of the average level of their cleavage rates. Thus we propose that the sticky ends, which originated after mechanochemical fragmentation of DNA, derive their existence from these differences in cleavage rates of complementary dinucleotides.

Sequence specificity of mechanochemical breakage of DNA due to shearing forces is also important as it helps us to understand the role a nucleotide sequence may play in functional potential of DNA regions. Really, it reveals the diversity of conformational dynamics in both complementary strands, which now can be characterized independently by the relative cleavage rate. Such numerical evaluation may be useful for identifying promoter regions in the genome as well as assessing preferences for nucleosome positioning[41].

Methods

Sequencing data. For analysis of the bias at the 5′ ends of the reads we used publicly available data from various sequencing centers. For human genome we took the available sequencing data from 1000 Genome Project (ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/); for *Arabadiposis thaliana* genome we choose the sequencing data from 1001 Genomes project (http://1001genomes.org/data/), specifically from the JGIHeazlewood2011 project sequenced by the DOE Joint Genome Institute (http://1001genomes.org/data/JGI/JGIHeazlewood2011/releases/2012_05_30/TAIR10/strains/Alc-0/); for *Drosophila mauritiana* genome we used the sequencing data of the group from the University of Veterinary Medicine, Vienna[31] (http://www.popoolation.at/mauritiana_genome/index.html), and for the genome of *Escherichia coli* the raw data was downloaded from the Sequence Read Archives (SRA) at NCBI (http://www.ncbi.nlm.nih.gov/sra) under the number SRR072099. A full list of data replete with accession numbers, fragmentation method and names of the sequencing centers is presented in Supplementary Table 1. For all data except *E. coli*, we used alignment files in .bam format, already prepared by the sequencing centers and available for downloading. For *E. coli* we downloaded the raw sequencing data from SRA archive, and then aligned the raw reads with bwa aligner, version 0.6.2, to the reference genome of *Escherichia coli* K 12 substr MG1655 (accession number NC_000913), downloaded from ftp.ncbi.nlm.nih.gov/genomes/Bacteria.

Mono, di and tetra-nucleotides relative frequence calculations at the 5′-ends of the reads. To obtain the relative frequencies at the 5′-ends of the reads we selected reads that were 100% aligned, without a single mismatch, to the positive strand of the reference genome (bam FLAG = 99) and composed datasets consisting of 500 000 fully mapped reads. Prior to the analysis we tested if the size of a dataset influences the relative frequence calculations. For that purpose we extracted, for one dataset from MPIMG (NA12878), all reads fully aligned to the positive strand from the entire genome, and obtained a set of 31 000 000 reads. A comparison of the relative frequencies of mono-, di- or tetranucleotide occurrence at the 5′-ends of the reads, obtained from a set of 500 000 and a set of 31 000 000 reads, showed a very high Pearson's correlation, r = 0.999, proving that the size of a dataset does not significantly affect the results. The analysis of relative frequencies of mono-, di- and tetranucleotide occurrence at the 5′-ends of the reads for all other samples was confined to the sets composed of 500 000 reads.

The values of mono-, di- and tetranucleotide frequencies at positions of the 5′-ends of the reads were calculated as the number of occurrence of a given mono-, di- or tetranucleotide nucleotide divided by the size of the data set (here, 500 000). The relative frequencies were calculated as the ratio between the numbers of mono-, di- or tetranucleotides at the positions of the cleavage normalized to the average frequencies of mono-, di- or tetranucleotides from two regions covered: −10 ÷ −20 and the +10 ÷ +20 bp positions around the break point.

For estimation of the frequency of mononucleotides, we counted the number of nucleotides right before the 5′-end of the read. For an estimation of the frequency of dinucleotides, the first position was taken right before the 5′- end of the read, and the second position coincided with the beginning of the read. For an estimation of the frequency of tetranucleotides, two first positions were taken right before the 5′- end of the read, and the third and the fourth positions corresponded to the first and the second nucleotides of the read.

Correlation analysis. Pearson's correlation matrices of relative cleavage rates of the reads obtained in different experiments by the same or different fragmentation methods for all 16 dinucleotides and all 256 tetranucleotides are given in Supplementary Tabes 5 and 6. The correlations, which are significant at $p < 0.05$, are marked in red.

Cluster analysis. Hierarchical cluster analysis was performed for all analyzed samples using the Ward's Method and the $(1 - r)$ distance, where r is the Pearson r for relative frequencies of tetranucleotides centered at the 5′-ends of the reads (Supplementary Table 6).

1. Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* **6**, 2601–2610 (1979).
2. Knierim, E., Lucke, B., Schwarz, J. M., Schuelke, M. & Seelow, D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PLoS One* **6**, e28240 (2011).
3. Loman, N. J. *et al.* High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol* **10**, 599–606 (2012).
4. Metzker, M. L. Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31–46 (2010).
5. Nawy, T. Single-cell epigenetics. *Nat Methods* **10**, 1060 (2013).
6. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135–1145 (2008).
7. Feschotte, C. & Pritham, E. J. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* **41**, 331–368 (2007).
8. Kramerov, D. A. & Vassetzky, N. S. SINEs. *Wiley Interdiscip Rev RNA* **2**, 772–786 (2011).
9. Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
10. Pinto, D. *et al.* Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol* **29**, 512–520 (2011).
11. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**, 36–46 (2012).
12. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
13. Liang, W. S. *et al.* Long insert whole genome sequencing for copy number variant and translocation detection. *Nucleic Acids Res* (2013).
14. Muller, C. A. *et al.* The dynamics of genome replication using deep sequencing. *Nucleic Acids Res* **42**, e3 (2014).
15. Medvedev, P., Fiume, M., Dzamba, M., Smith, T. & Brudno, M. Detecting copy number variation with mated short reads. *Genome Res* **20**, 1613–1622 (2010).
16. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G + C)-biased genomes. *Nat Methods* **6**, 291–295 (2009).
17. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* **40**, e72 (2012).
18. Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y. & Hwang, C. C. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS One* **8**, e62856 (2013).
19. Schwartz, S., Oren, R. & Ast, G. Detection and removal of biases in the analysis of next-generation sequencing reads. *PLoS One* **6**, e16685 (2011).
20. Taub, M. A., Corrada Bravo, H. & Irizarry, R. A. Overcoming bias and systematic errors in next generation sequencing data. *Genome Med* **2**, 87 (2010).
21. Guo, Y. *et al.* The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* **13**, 666 (2012).

22. Le, H. S., Schulz, M. H., McCauley, B. M., Hinman, V. F. & Bar-Joseph, Z. Probabilistic error correction for RNA sequencing. *Nucleic Acids Res* **41**, e109 (2013).
23. Cheung, M. S., Down, T. A., Latorre, I. & Ahringer, J. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Res* **39**, e103 (2011).
24. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* **38**, e131 (2010).
25. van Heesch, S. *et al.* Systematic biases in DNA copy number originate from isolation procedures. *Genome Biol* **14**, R33 (2013).
26. Auerbach, R. K. *et al.* Mapping accessible chromatin regions using Sono-Seq. *Proc Natl Acad Sci U S A* **106**, 14926–14931 (2009).
27. Grokhovsky, S. L. *et al.* Sequence-specific ultrasonic cleavage of DNA. *Biophys J* **100**, 117–125 (2011).
28. Grokhovsky, S. L. Specificity of DNA Cleavage by Ultrasound. *Molecular Biology (in Russan)* **40**, 276–283 (2006).
29. Grokhovsky, S. L. *et al.* in *Gel Electrophoresis - Principles and Basics* (ed Dr. Sameh Magdeldin) (InTech, 2012).
30. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
31. Nolte, V., Pandey, R. V., Kofler, R. & Schlotterer, C. Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in Drosophila mauritiana. *Genome Res* **23**, 99–110 (2013).
32. Lazarovici, A. *et al.* Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc Natl Acad Sci U S A* **110**, 6376–6381 (2013).
33. Grokhovsky, S. L. *et al.* in *Ultrasonics: Theory, Techniques and Practical Applications* (ed Hanako Ayabito & Mitsuko Katsukawa) 1–24 (Nova Science Publishers, 2013).
34. Lentz, Y. K., Anchordoquy, T. J. & Lengsfeld, C. S. DNA acts as a nucleation site for transient cavitation in the ultrasonic nebulizer. *J Pharm Sci* **95**, 607–619 (2006).
35. Kim, S., Medvedev, P., Paton, T. A. & Bafna, V. Reprever: resolving low-copy duplicated sequences using template driven assembly. *Nucleic Acids Res* **41**, e128 (2013).
36. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**, e105 (2008).
37. Mutter, G. L. & Boynton, K. A. PCR bias in amplification of androgen receptor alleles, a trinucleotide repeat marker used in clonality studies. *Nucleic Acids Res* **23**, 1411–1418 (1995).
38. Il'icheva, I. A., Nechipurenko, D. Y. & Grokhovsky, S. L. Ultrasonic cleavage of nicked DNA. *J Biomol Struct Dyn* **27**, 391–398 (2009).
39. Nechipurenko, Y. D. *et al.* Characteristics of ultrasonic cleavage of DNA. *Journal of Structural Chemistry* **50**, 1007–1013 (2009).
40. Doerfler, W. & Böhm, P. *DNA Methylation: Basic Mechanisms*. (Springer-Verlag, 2006).
41. Parker, S. C., Hansen, L., Abaan, H. O., Tullius, T. D. & Margulies, E. H. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* **324**, 389–392 (2009).

Acknowledgments

We thank Georgy Gursky and Mikhail Thamm for useful discussions. This work has been carried out with the financial support of the Program of the Presidium of the Russian Academy of Sciences for Molecular and Cellular Biology and the Russian Foundation for Basic Research (Grant No. 14-04-01269).

Author contributions

The authors have made the following declarations about their contributions: S.L.G., I.A.I. and Y.D.N. – – conceived and designed the article; M.S.P., L.A.P. and M.V.K. – performed the experiments; D.Y.N., M.S.P., L.A.P., N.Y.O. and R.V.P. – analyzed the data; all authors reviewed the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Poptsova, M.S. *et al.* Non-random DNA fragmentation in next-generation sequencing. *Sci. Rep.* **4**, 4532; DOI:10.1038/srep04532 (2014).