Molecular Therapy
Methods & Clinical Development
Commentary

In Silico Restriction Enzyme Digests to Minimize Mapping Bias in Genomic Sequencing

Jason Roszik,[1,2] György Fenyőfalvi,[3] László Halász,[3,4] Zsolt Karányi,[3] and Lóránt Székvölgyi[3,4]

http://dx.doi.org/10.1016/j.omtm.2017.06.003

A commonly used genome fragmentation method in next generation sequencing, restriction endonuclease (RE) digestion, may severely compromise genomic mapping resolution and prevent the functional annotation of certain chromosomal regions unless REs are applied in correct combinations to sample all genomic regions with an equal probability.

Genome fragmentation by REs is routinely used in multiple genomic mapping technologies, including RNA-DNA hybrid (R-loop) immunoprecipitation sequencing (DRIP-seq),[1,2] chromosome conformation capture (4C/5C, Hi-C),[3] reduced-representation bisulfite sequencing (RRBS),[4] and restriction site associated marker (RAD) genotyping.[5] The performance of these approaches depends on (1) the length distribution of the restriction fragments (determining the spatial resolution of the assay) and (2) the randomness of RE digestion (ensuring that all genomic regions are sampled with an equal probability).[6] Therefore, selecting the proper combination of REs for genome fragmentation is of crucial importance to obtain representative next-generation sequencing (NGS) libraries and to assign clear biological functions to the mapped regions.

Using the DRIP-seq technique, we have recently shown that this technology contains inherent biases related to RE digestion that might prevent functional annotation of a significant fraction of R-loops.[2] R-loops, nucleic acid structures that are composed of an RNA-DNA hybrid and a single-stranded DNA, are involved in multiple cellular processes and may also mediate genomic instability in a pathological context. The DRIP method uses an anti-RNA-DNA hybrid antibody to capture RNA-DNA hybrids associated with RE-fragmented DNA or chromatin, followed by fragment mapping to the genome. The main reason for the over-representation of lengthy DRIP fragments may be that the distribution of restriction enzyme cutting sites is not random in the human genome.[7] This bias is especially enhanced over the first exons. The over-representation of first exons in RE-fragmented samples may also be an issue in other species and sequencing methods. For instance, the mouse genome also contains long intronic sequences that may cause similar biases. Similar to the DRIP method, suboptimal RE fragment size distribution and first exon bias might affect the outcome and interpretation of other frequently used genomic technologies (e.g., all C-based methods [4C, 5C, and Hi-C]), potentially introducing false-positive spatial contacts that fall proximal to open reading frames (ORFs), especially to first exons. Finally, the estimation of the evolutionary conservation of R-loop binding sites between species that reflect the sequence homology/divergence of exonic DRIP fragments,[8] but precisely located R-loop binding sites, is potentially also problematic.

Superimposed on the RE bias, multiple other genome characteristics can affect the efficacy of RE digestion. DNA methylation is present in higher organisms, and the majority of REs do not cut at methylated cytosines. Furthermore, most REs do not cut DNA-RNA hybrids that are prevalent over the chromosomes (constituting 5%–8% of the eukaryote genome). Restriction enzyme accessibility is also limited by the chromatin (nucleosome) structure that inherently prefers the cutting of linker DNA sequences.

The randomness and uniformity of restriction fragment length distributions can be tested for any combination of REs using in silico restriction endonuclease digests (Figure 1), and RE cocktails with theoretically justified cutting parameters can be selected for use in experiments. We recommend using the DECIPHER R package, which is available in Bioconductor.[9] To predict the expected DNA fragments, the "digestDNA" function can be used to perform in silico restriction digestion of given DNA sequences. Issues related to CpG methylation can be experimentally addressed by methylation-insensitive REs that cleave methylated DNA. RNase H1 digestion of nucleic acid preps can also be applied to remove RNA from DNA-RNA hybrids. Furthermore, short treatment of live cells with chromatin decompaction agents (e.g., HDAC inhibitors) may provide increased RE accessibility in experiments involving in situ RE fragmentation (e.g., Hi-C). Collectively, the above recommendations can help identify RE cocktails and experimental conditions that result in proper DNA fragment size distributions and optimal resolution in genomic sequencing technologies.

[1]Department of Melanoma Medical Oncology, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030, USA; [2]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030, USA; [3]MTA-DE Momentum, Genome Architecture and Recombination Research Group, Research Centre for Molecular Medicine, University of Debrecen, Nagyerdei krt 98, Debrecen 4032, Hungary; [4]Department of Biochemistry and Molecular Biology, University of Debrecen, Egyetem sq. 1, Debrecen 4032, Hungary

Correspondence: Lóránt Székvölgyi, PhD, Department of Biochemistry and Molecular Biology, University of Debrecen, Nagyerdei krt. 98, Debrecen 4032, Hungary.
E-mail: lorantsz@med.unideb.hu

Please cite this article in press as: Roszik et al., In Silico Restriction Enzyme Digests to Minimize Mapping Bias in Genomic Sequencing, Molecular Therapy: Methods & Clinical Development (2017), http://dx.doi.org/10.1016/j.omtm.2017.06.003

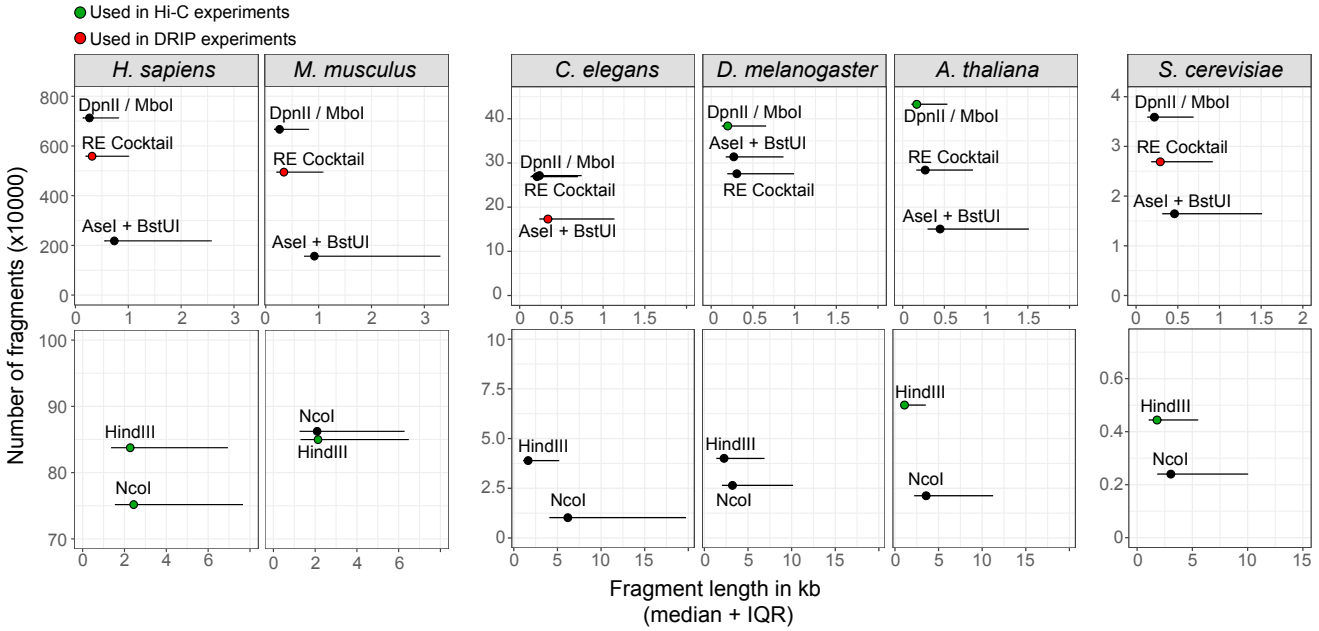

Figure 1. Genome Fragmentation by In Silico Restriction Enzyme Digestion in Species That Were Analyzed by DRIP-seq or HI-C

The absolute number of restriction fragments is shown in terms of the average fragment lengths (mean + interquartile range [IQR]) obtained by the indicated restriction enzymes applied alone or in combination. RE cocktail denotes the HindIII, EcoRI, BsrGI, XbaI, and SspI enzymes.

ACKNOWLEDGMENTS

This work was supported by the Hungarian Academy of Sciences (Lendület programme, LP2015-9/2015) and grants from the National Research, Development and Innovation Office (NKFIH-ERC-HU-117670 and GINOP-2.3.2-15-2016-00024).

REFERENCES

1. Ginno, P.A., Lott, P.L., Christensen, H.C., Korf, I., and Chédin, F. (2012). R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol. Cell *45*, 814–825.

2. Halász, L., Karányi, Z., Boros-Oláh, B., Kuik-Rózsa, T., Sipos, É., Nagy, É., Mosolygó-L, Á., Mázló, A., Rajnavölgyi, É., Halmos, G., and Székvölgyi, L. (2017). RNA-DNA hybrid (R-loop) immunoprecipitation mapping: an analytical workflow to evaluate inherent biases. Genome Res. *27*, 1063–1073.

3. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

4. Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. *33*, 5868–5877.

5. Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A., and Johnson, E.A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. *17*, 240–248.

6. Bystrykh, L.V. (2013). A combinatorial approach to the restriction of a mouse genome. BMC Res. Notes *6*, 284.

7. Hartono, S.R., Korf, I.F., and Chédin, F. (2015). GC skew is a conserved property of unmethylated CpG island promoters across vertebrates. Nucleic Acids Res. *43*, 9729–9741.

8. Sanz, L.A., Hartono, S.R., Lim, Y.W., Steyaert, S., Rajpurkar, A., Ginno, P.A., Xu, X., and Chédin, F. (2016). Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. Mol. Cell *63*, 167–178.

9. Wright, E.S. (2016). Using DECIPHER v2.0 to analyze big biological sequence data in R. R J. *8*, 352–359.