Slotkin *Mobile DNA* (2018) 9:15
https://doi.org/10.1186/s13100-018-0120-9

Mobile DNA

COMMENTARY Open Access

The case for not masking away repetitive DNA

R. Keith Slotkin

Abstract

In the course of analyzing whole-genome data, it is common practice to mask or filter out repetitive regions of a genome, such as transposable elements and endogenous retroviruses, in order to focus only on genes and thus simplify the results. This *Commentary* is a plea from one member of the *Mobile DNA* community to all gene-centric researchers: *please do not ignore the repetitive fraction of the genome*. Please stop narrowing your findings by only analyzing a minority of the genome, and instead broaden your analyses to include the rich biology of repetitive and mobile DNA. In this article, I present four arguments supporting a case for retaining repetitive DNA in your genome-wide analysis.

Keywords: Repeat mask, Filter, Transposable element, Repetitive DNA, Transposon, Genomics

Main text

From the time of their discovery by McClintock, transposable elements (TEs) have had a history of being ignored [1]. Termed "*controlling elements*" by McClintock for their roles in developmental regulation, TEs were first disregarded, then relegated as an oddity, and are now largely overlooked by most researchers. Today, most biologists believe mobile DNA represents filler (or even worse: "junk") between exons or genes. In single-locus experiments, this bias is propagated by alignment tools such as BLAST that mask low complexity and repetitive regions as a default option. In genome-wide experiments, even though data from all regions should be captured, this blindness to repetitive DNA is exacerbated due to the use of standard bioinformatic pipelines that restrict the analysis to genes. "Whole-genome" datasets are routinely restricted to a concentrated analysis of 22-51% of the human genome corresponding to the non-repetitive portion defined by *RepeatMasker* [2]. For example, in RNA-seq experiments researchers commonly only consider reads aligned to defined genic exons. This leaves 78-49% of the genome unanalyzed in favor of a simplified evaluation of gene expression.

Correspondence: Slotkin.2@osu.edu; kSlotkin@danforthcenter.org
Department of Molecular Genetics and Center for Applied Plant Sciences, The Ohio State University, 500 Aronoff Laboratory, 318 West 12th Ave, Columbus, Ohio 43210, USA

This *Commentary* is not written for the existing readership of *Mobile DNA*, who already appreciate the many roles of TEs and endogenous retroviruses (ERVs). In fact, many of us in the *Mobile DNA* field have enriched our careers by making discoveries from reanalyzing publicly available data in which the original authors did not pay attention to the repetitive fraction of their data. For example, one of the most important discoveries from my own career came from reanalyzing available microarray data for TE expression [3]. Instead, it is my hope that this article is found by researchers when they perform web searches on how to filter out repetitive DNA. This *Commentary* is written by one *Mobile DNA* researcher on behalf of the field as an appeal to the general scientific community: By focusing on only a fraction of the genome, only a fraction of discoveries can be made. Below are four distinct arguments to convince the reader to retain the repetitive portion of the genome in their experiments to perform a true "genome-wide" analysis.

Repetitive DNA critically influences gene and genome regulation

The influence of TEs and ERVs on gene expression is no longer just an oddity of a few examples of meta-stable phenotypes or epialleles. Mobile DNA is now understood to have *bona fide* genome-wide gene regulatory abilities through a number of different mechanisms, including *cis* effects on neighboring genes, and *trans*-effects

at a distance. Repetitive and mobile DNA provide a rich source of gene regulation, evolutionary flexibility and epigenetic catalysts. Although too many examples exist to fit into this short *Commentary* article, I point the reader to thorough review articles on this topic (reviewed in [4, 5]). The dynamic mobile nature of TEs and ERVs, in addition to their interaction with a number of host proteins, provide the cell with a DNA-protein interaction module that if transposed near a gene, may negatively or positively affect the regulation of that gene. If positive and selected for, this module may be evolutionarily coopted as a new enhancer element. TEs and ERVs are now known to regulate single genes as well as entire gene regulatory networks [6, 7]. Therefore, if a researcher were performing a ChIP-seq experiment to identify where a transcription factor binds (for example), or a chromosome confirmation capture (4C/HiC) experiment to identify enhancer-promoter interactions, they would potentially miss important enhancers if repetitive DNA was filtered from their analysis.

In addition to their function in gene regulation, TEs and ERVs can act as developmental and tissue stage-specific markers (such as in early mammalian development [8, 9]), as sources of long non-coding RNAs [10] and as sensors of cellular stress [11]. TE-derived proteins are also involved in centromere function [12, 13], and TEs can substitute as telomeres in the absence of telomerase [14]. Importantly, TEs and ERVs are both associated with (and could be the primary cause of) several diseases [15–17]. Therefore, the message is clear: Repetitive DNA such as TEs and ERVs should not be excluded from genome-wide analyses because they play key roles in gene regulation and evolution.

Technology has advanced to better assay the repetitive portion of the genome

Because of their limited sequenced read length, deep sequencing technologies such as Illumina SBS have fundamentally superior performance for single-copy genic regions compared to repetitive DNA. This problem is most severe for small RNA biology or when only a sequence tag can be obtained, as a large fraction of the reads cannot be mapped to a unique position in the genome because only 15-30 nucleotides of sequence information is available. However, even with limited sequence length, approaches have been developed to successfully assay the repetitive fraction of a genome with short reads (for example, [18]).

Many researchers simply and completely filter out repetitive DNA from their analysis, even though read mapping technologies have evolved or have been specifically designed to efficiently assay repetitive DNA from whole genome data. These tools include MELT, Retro-Seq, EpiTEome and McClintock for TE insertion site identification [19–22], T-lex to identify presence/absence of TE copies [23], Clari-TE to resolve nested TE structure [24], TEtoolkit/TEtranscripts for TE enrichment analysis from experiments such as RNA-seq and ChIP-seq [25, 26], and many others. Which tool to use is specific to the desired type of analysis and biological question. Nevertheless, several overarching guidelines should be considered when mapping sequenced reads to repetitive DNA. First, the location in the genome where the read matches *best* should be the only one reported; however multiple *best*-matching regions may exist (i.e. multi-mapping reads). A conservative approach is to report only the reads that perfectly and uniquely match once in the genome (for example, my lab used this method for MethylC-seq of TEs in [27]). This approach should be used if the researcher wants to be absolutely sure that a particular locus is generating the reads. The problem with this approach is that it disregards a substantially large proportion of reads, especially in highly repetitive (TE-rich) genomes. Rather than having to re-map reads using different approaches based on the number of mismatches (mapping only perfect matches, then 1-mismatch, then 2, …), one can perform an intensive mapping of all possibilities at once, and use downstream tools like *NGSUtils* [28] to parse their desired level of matching preciseness from the output file. A second approach is to randomly divide the multi-mapping reads to their *best* positions in the genome (for example, my lab used this method mapping small RNA reads in [29]), although this will dilute the analysis of individual repetitive elements across their entire repeat families. This approach should be used in TE-rich genomes, and when analysis of TE families is performed but resolution down to individual elements is not necessary. A third approach is to hierarchically distribute multi-mapping reads guided by the evidence of where the uniquely mapping reads cluster [30] (for example, this method was used when mapping small RNA reads in [31]). The idea behind this approach is that we already know that some individual TE elements generate reads (by using the uniquely mapping reads), so it is most likely that these same TEs are also the producers of the multi-mapping reads. This approach should be used when TE regulation requires drilling down beyond the repeat family level to individual elements. The drawback to this approach is it has the potential to falsely assign one or more reads to a locus that did not produce them. Lastly an approach that has been utilized for years within the *Mobile DNA* community, reads can be mapped to single elements or databases of 'consensus' sequences that are created from alignments of many individual repetitive genomic sequences [32] (for example, my lab used this method in Figure 1-2 of [33]). This approach is useful when a family-level (rather than an individual element) view is required in high detail. All of the above mentioned

approaches (with the strong exception of not interrogating the repetitive fraction of the genome at all) are acceptable to the *Mobile DNA* community, given that the researcher is transparent about which approach was used.

In contrast to software advances, the key enabling technology that has enhanced the read mappability (and therefore coverage) of the repetitive fraction of the genome is improved read length. On the widely-used Illumina SBS platform, the standard read length has transitioned from 50 to 150 nucleotides, and when coupled with a paired-end sequencing and mapping approach, coverage of the non-genic portion of the genome has greatly improved. The significantly longer reads of PacBio and Nanopore sequencing approaches altogether remove the mapping ambiguity that has previously limited the genome-wide analysis of repetitive DNA. For many experiments, extending the length of the sequence read is already worth the investment to enable unique alignment to the genome, and the future will see these technologies used routinely. Long read lengths also enable improved split-read mapping approaches to identify non-reference genome positions of mobile TEs [21, 34]. Therefore, the read length and mapping approaches described here now exist to at least partially, if not all together, cover the repetitive fraction of the genome, and thus interrogating repetitive DNA should be a necessary component of any "genome-wide" analysis.

Many technologies have originated from mobile DNA or defenses against them

Basic research performed on mobile DNA aims to identify how elements transpose, the mechanisms responsible for repressing activity, and the short-term and evolutionary consequences of transposition/insertion. Medical applied research also aims to identify how TEs and ERVs contribute to cancer, neurological disease and other disorders. In addition to these areas of focus, the investigation of mobile DNA has significantly contributed to the production of new research tools. Mobile DNA biology has enabled improved mutagenesis and gene tagging, transgenesis, and analysis of gene function by gene trapping, just to name a few [35]. In vitro TE systems have been leveraged to generate protocols to assay chromatin accessibility (ATAC-seq) [36] as well as efficient deep sequencing library production (tagmentation) [37]. Even the key enabling technology of this decade, CRISPR-Cas genome editing, is now understood to be based upon mobile DNA [38]. Therefore, paying attention to the analysis of repetitive DNA, even on the basic research level, may pay dividends towards the development of a key technological breakthrough that can propel multiple disciplines forward.

Maximize the data output from your effort (and cost) input

It takes a substantial amount of effort to execute (and fund) research science. Once the funds are successfully obtained, and the experiments are performed, it simply does not make sense to limit the analysis to a fraction of the obtained data. Confining an analysis to only part of the genome significantly restricts the likelihood and scope of the ensuing discoveries. Two recent examples of TE/ERV analyses done well both represent cases where considering repetitive elements answered questions that would otherwise be unexplainable: 1) The structural variation map of 2504 genomes that specifically identified *Alu*, L1 and SVA mobile element insertions across human populations [39] and 2) a ChIP-seq analysis discovered that ERVs have spread interferon-inducible enhancer elements and shaped the evolution of innate immunity [7].

Many published datasets exist that can be downloaded and for the first time analyzed for TE or ERV dynamics. The availability of this only partially analyzed data affords great opportunity to my colleagues in the *Mobile DNA* field; however, the analysis of this data for repetitive DNA dynamics should have been part of the original study. We, the *Mobile DNA* Community, request that you please include repetitive DNA in your genome-wide analyses. If you do not have the tools or the interest, contact a *Mobile DNA* community member in your field to collaborate. Because, after all, if you don't analyze the "whole genome" from your genome-wide data, we will.

Abbreviations

ATAC-seq: Assay for transposase-accessible chromatin using sequencing; ERV: Endogenous retrovirus; HiC/4C: Chromosome confirmation capture experiments; SBS: Sequence by synthesis; TE: Transposable element

Acknowledgements

The author thanks Kaushik Panda and Andrea McCue for editing this manuscript.

Funding

The Slotkin laboratory is funded by NSF grants MCB-1252370, MCB-1608392, and IOS-1340050.

Authors' contributions

RKS conceptualized and wrote this commentary. The author read and approved the final manuscript.

Ethics approval and consent to participate

Not Applicable

Competing interests

The author declares that he/she has no competing interests.

Publisher's Note

Received: 22 February 2018 Accepted: 23 April 2018
Published online: 01 May 2018

References

1. Comfort NC. From controlling elements to transposons: Barbara McClintock and the Nobel Prize. Trends Genet. 2001;17:475–8.
2. de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 2011;7:e1002384.
3. Slotkin RK, Vaughn M, Borges F, Tanurdzic M, Becker JD, Feijó JA, et al. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell. 2009;136:461–72.
4. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017;18:71–86.
5. McCue AD, Slotkin RK. Transposable element small RNAs as regulators of gene expression. Trends Genet. 2012;28:616–23.
6. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 2014;24:1963–76.
7. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016;351:1083–7.
8. Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature. 2012;487:57–63.
9. Gerdes P, Richardson SR, Mager DL, Faulkner GJ. Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. Genome Biol. 2016;17:1286.
10. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLoS Genet. 2013;9:e1003470.
11. Horváth V, Merenciano M, Gonzalez J. Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response. Trends Genet. 2017;33:832–41.
12. Klein SJ, O'Neill RJ. Transposable elements: genome innovation, chromosome diversity, and centromere conflict. Chromosom Res. 2018; 2012:947089.
13. Mateo L, Gonzalez J. Pogo-like transposases have been repeatedly domesticated into CENP-B-related proteins. Genome Biol Evol. 2014;6:2008–16.
14. George JA, DeBaryshe PG, Traverse KL, Celniker SE, Pardue M-L. Genomic organization of the Drosophila telomere retrotransposable elements. Genome Res. 2006;16:1231–40.
15. Reilly MT, Faulkner GJ, Dubnau J, Ponomarev I, Gage FH. The role of transposable elements in health and diseases of the central nervous system. J Neurosci. 2013;33:17577–86.
16. Wylie A, Jones AE, Abrams JM. p53 in the game of transposons. BioEssays. 2016;38:1111–6.
17. Burns KH. Transposable elements in cancer. Nat Rev Cancer. 2017;17:415–24.
18. Faulkner GJ, Forrest ARR, Chalk AM, Schroder K, Hayashizaki Y, Carninci P, et al. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics. 2008;91:281–8.
19. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. 2017;27:1916–29.
20. Keane TM, Wong K, Adams DJ. RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics. 2013;29:389–90.
21. Daron J, Slotkin RK. EpiTEome: Simultaneous detection of transposable element insertion sites and their DNA methylation levels. Genome Biol. 2017;18:7704.
22. Nelson MG, Linheiro RS, Bergman CM. McClintock: An Integrated Pipeline for Detecting Transposable Element Insertions in Whole-Genome Shotgun Sequencing Data. G3. 2017;7:2763–78.
23. Fiston-Lavier A-S, Carrigan M, Petrov DA, Gonzalez J. T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. Nucleic Acids Res. 2011;39:e36–6.
24. Daron J, Glover N, Pingault L, Theil S, Jamilloux V, Paux E, et al. Organization and evolution of transposable elements along the bread wheat chromosome 3B. Genome Biol. 2014;15:546.
25. Jin Y, Tam OH, Paniagua E, Hammell M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics. 2015;31:3593–9.
26. Jin Y, Hammell M. Analysis of RNA-Seq Data Using TEtranscripts. Methods Mol Biol. 2018;1751:153–67.
27. Panda K, Ji L, Neumann DA, Daron J, Schmitz RJ, Slotkin RK. Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA methylation. Genome Biol. 2016;17:100.
28. Breese MR, Liu Y. NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. Bioinformatics. 2013;29:494–6.
29. Nuthikattu S, McCue AD, Panda K, Fultz D, DeFraia C, Thomas EN, et al. The Initiation of Epigenetic Silencing of Active Transposable Elements Is Triggered by RDR6 and 21-22 Nucleotide Small Interfering RNAs. Plant Physiol. 2013;162:116–31.
30. Johnson NR, Yeoh JM, Coruh C, Axtell MJ. Improved Placement of Multi-Mapping Small RNAs. G3. 2016;6:2103–11.
31. Martinez G, Wolff P, Wang Z, Moreno-Romero J, Lez JS–GX, Conze LL. et al, Paternal easiRNAs regulate parental genome dosage in Arabidopsis. Nat Genet. 2018;50:193–98.
32. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.
33. McCue AD, Panda K, Nuthikattu S, Choudury SG, Thomas EN, Slotkin RK. ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the establishment of DNA methylation. EMBO J. 2015;34:20–35.
34. Ewing AD. Transposable element detection from whole genome sequence data. Mob DNA. 2015;6:164.
35. Ivics Z, Li MA, Mátés L, Boeke JD, Nagy A, Bradley A, et al. Transposon-mediated genome manipulation in vertebrates. Nat Methods. 2009;6:415–22.
36. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol. 2015; 109:21.29.1–9.
37. Picelli S, Björklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. 2014;24:2033–40.
38. Koonin EV, Makarova KS. Mobile Genetic Elements and Evolution of CRISPR-Cas Systems: All the Way There and Back. Genome Biol Evol. 2017;9:2812–25.
39. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526:75–81.