Genomics

Volume 113, Issue 6, November 2021, Pages 3804-3810

Original Article

Characterization of full-length LINE-1 insertions in 154 genomes

Jessica S. Wong [a, 1], Tanaya Jadhav [a, 1], Eleanor Young [a, 1], Yilin Wang [a], Ming Xiao [a, b]

Show more

https://doi.org/10.1016/j.ygeno.2021.09.011

Abstract

Long interspersed nuclear elements (LINEs) are retrotransposons that contribute to genetic variation in the human genome. LINE-1 elements in larger-scale studies are challenging to identify using sequencing technologies due to cost and scalability. We developed an approach using optical mapping for detection of full-length LINE-1 insertions and 10× sequencing for confirmation. We found 51 true positive full-length LINE-1 insertions, of which 4 are novel insertions, in NA12878. Repeating our analysis on a larger sample set representing 26 populations, we identified 329 full-length LINE-1 elements, of which 123 are novel. 24.8% of these 329 LINE-1 insertions were shared amongst all 5 superpopulations (AFR, AMR, EUR, EAS, SAS). The African superpopulation has a higher percentage of population-specific LINE-1 insertions than any other superpopulation. These data indicate that our approach can provide high-speed, cost-effective, and increased accuracy for LINE-1 detection. These data also provide an insight into variations of LINE-1 elements between different populations.

Introduction

Transposable elements (TEs) of various classes contribute to a large fraction of most eukaryotic genomes, including 45% of human genomes [13]. Active TEs are highly mutagenic and cause recombination and rearrangement within the genome [21]. Long interspersed elements (LINE-1), a subclass of TEs, are retrotransposons (elements that cannot move) that lack long terminal repeats (LTRs), are concentrated in AT-rich regions, and are active in approximately 96 disease-causing insertions [9,10]. Full-length LINE-1 insertions are approximately 6 kb in length, contain an internal polymerase II promoter, and encode two non-overlapping open reading frames (ORFs) [13].

LINE-1 is an important source of structural variation and while the reference genome captures fixed alleles and high-frequency alleles, it does not include common variants [23]. Because of their highly repetitive nature, it is often challenging to capture the accurate number of LINE-1 insertions, their locations in a genome, and their exact sequence [27]. The current human reference hg38 includes 146 full-length LINE-1 elements, but due to variation between LINE-1 elements there may actually be more present [5,17]. Further, variation between individuals suggests this number will likely vary.

A recent study linked active LINE-1 elements to a variety of human diseases, including neurological disorders, genetic disorders, and cancer [30]. Developing comprehensive catalogs of LINE-1 insertions is therefore important for recognizing recurrent insertion patterns, associating insertions with phenotypes, and identifying activity [23].

Targeted advancements toward understanding the complete scope of LINE-1 elements and their variations in human genomes are still underway. A previous study using long-read assemblies indicates the presence of at least 50% more intact LINE-1 elements in a diploid human genome than previous estimates (290 intact LINE-1 elements at 194 loci) [27]. Another recent approach used PacBio long-read sequencing data to identify LINE-1 insertions that were previously absent in short-read studies and genotyped 63% of those insertions in 1000 Genomes Project sequences [31]. These findings indicate that long-read sequencing approaches can detect LINE-1 insertions that are often missed by short-read sequencing approaches.

Short-read sequencing technologies lack the ability to accurately detect long insertions [24]. Despite advances in long-read sequencing technology increasing throughput and accuracy, it is still cost prohibitive to perform whole-genome sequencing to detect intact LINE-1 insertion, especially to the depth needed to accurately identify LINE-1 elements [2]. Optical mapping technology utilizes longer molecules than sequencing, with read lengths averaging 300 kb and ranging up to 1 Mb [29]. It has found wide applications in assisting genome assembly and characterization of complex structural variants [14,16,28]. When the size distribution of insertions from human genomes are plotted, a peak at 6 kb is always observed, which could be mostly due to full-length LINE-1 insertions. But optical mapping lacks precise

sequence information to determine which 6 kb insertions are true LINE-1 elements, thus further confirmation with sequencing reads is necessary.

In this paper, we developed a method identifying intact full-length LINE-1 insertions using optical mapping assemblies in conjunction with sequencing data. In this method, we first selected all of the 6 kb insertions containing mapping motif signature within the insertion and then confirmed these insertions containing LINE-1 sequences with low coverage 10× genomics sequencing reads. After validating the method with known LINE-1 datasets in NA12878 [31], we then applied this method in a 154 sample set across 26 populations. In total, we detected 329 intact full-length LINE-1 insertions, of which 123 are novel. We also assessed the LINE-1 population patterns in this sample set. These results demonstrate that a combination of optical mapping data and 10× sequencing data provides a nearly comprehensive catalog of LINE-1 insertions.

Section snippets

Results

Optical mapping mostly relies on mapping the short sequence motifs across the whole genome [4,6,12,26]. The sequence motifs GCTCTTC (Nt.BspQ1) and CTTAAG (DLE-1) are available commercially on Bionano Genomics platforms. The patterns of these motifs with LINE-1 signature could be used to confirm LINE-1 insertions. We first investigated the signature of these motifs with 300 LINE-1 references collected from [7,20], and [31] (Fig. 1A) [7,20,31]. The vast majority of LINE-1 references contain 3…

Discussion

LINE-1 insertions contribute considerably to genetic variation found in the human genome [27]. Previous analyses revealed that LINE-1 families have frequently recruited novel 5'UTRs in the human lineage and have considerable variation in copy numbers. This indicates that LINE-1 may have large differences in replicative success [11], and active LINE-1 elements may give rise to new, active progeny at a faster rate than they are inactivated [5]. Observing population patterns may suggest…

Data collection

This dataset has been collected and analyzed by Levy-Sakin et al. The dataset includes 154 samples from 26 different populations. High-molecular-weight DNA for 52 samples was

extracted, nicked, and labeled using both DLE-1 and Nt.BspQ1 enzymes, and the remainder 102 were labeled using only Nt.BspQ1 enzyme. All 154 Nt.BspQ1 samples underwent imaging using the Bionano Genomics Irys System to generate single-molecule maps. The 52 DLE-1 genomes underwent the Saphyr system. These 52 samples also...

Declaration of Competing Interest

The authors declare no competing interests....

Acknowledgements

This research is supported in part by grants from NIH (R01HG005946)....

References (33)

C. Aston *et al.*
Optical mapping and its potential for large-scale sequencing projects
Trends Biotechnol. (1999)

C.R. Beck *et al.*
LINE-1 Retrotransposition activity in human genomes
Cell (2010)

D.C. Hancks *et al.*
Active human retrotransposons: variation and disease
Curr. Opin. Genet. Dev. (2012)

Y. Yuan *et al.*
Advances in optical mapping for genomic research
Comput. Struct. Biotechnol. J. (2020)

X. Zhang *et al.*
New understanding of the relevant role of LINE-1 retrotransposition in human disease and immune modulation
Front. Cell Dev. Biol. (2020)

Mark J.P. Chaisson *et al.*
Multi-platform discovery of haplotype-resolved structural variation in human genomes
Nature Communications (2019)

Eugene J. Gardner *et al.*
The Mobile Element Locator Tool (MELT): population-scale mobile element discovery

and biology
Genome research (2017)

W. Zhou *et al.*
Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology
Nucleic Acids Res. (2019)

E. Young *et al.*
Comprehensive analysis of human subtelomeres by whole genome mapping
PLoS Genet. (2020)

L. Yang *et al.*
Characterization of LINE-1 transposons in a human genome at allelic resolution
bioRxiv (2019)

View more references

Cited by (1)

Methodological and Biological Factors Influencing Global DNA Methylation Results Measured by LINE-1 Pyrosequencing Assay in Colorectal Tissue and Liquid Biopsy Samples

2022, International Journal of Molecular Sciences

Recommended articles (6)

Research article

Genome-wide DNA arrays profiling unravels the genetic structure of Iranian sheep and pattern of admixture with worldwide coarse-wool sheep breeds

Genomics, Volume 113, Issue 6, 2021, pp. 3501-3511

Show abstract

Research article

Evaluation of deep learning approaches for modeling transcription factor sequence specificity

Genomics, Volume 113, Issue 6, 2021, pp. 3774-3781

Show abstract

Research article

Clustering genomic organization of sea cucumber miRNAs impacts their evolution and expression

Genomics, Volume 113, Issue 6, 2021, pp. 3544-3555

Show abstract

Research article

The *Clausena lansium* (Wampee) genome reveal new insights into the carbazole alkaloids biosynthesis pathway

Genomics, Volume 113, Issue 6, 2021, pp. 3696-3704

Show abstract

Research article

Up-regulated microRNA-132 reduces the cognition-damaging effect of sevoflurane on Alzheimer's disease rats by inhibiting FOXA1

Genomics, Volume 113, Issue 6, 2021, pp. 3644-3652

Show abstract

Research article

Identification and expression analysis of miRNAs in germination and seedling growth of Tibetan hulless barley

Genomics, Volume 113, Issue 6, 2021, pp. 3735-3749

Show abstract

---

[1] Equal contribution

View full text