Genome-Scale Mapping of Human LINE-1 ORF2p Endonuclease Activity

Syndi Koltz,[1] Hanna Kodama,[2,3] Mel Davey,[1] Xu Tan,[1] Mike Kaiser,[1] Martin S. Taylor,[2] Michael Gallagher,[1] John F. Thompson[1]
[1]Nabsys 2.0 LLC, Providence, RI; [2] Department of Pathology and Laboratory Medicine, Brown University, Providence, RI; [3] Department of Molecular Biology, Cellular Biology, and Biochemistry, Brown University, Providence, RI

Abstract

The LINE-1 retrotransposon is a virus-like DNA parasite that has contributed at least a third of the human genome and is the only remaining active human protein-coding transposon. LINE-1 attempts to "copy and paste" itself in the genome through an RNA intermediate via retrotransposition, a process mediated by sequential activities of the encoded ORF2 protein (ORF2p) endonuclease (EN) and reverse transcriptase (RT) domains. This insertional mutagenesis can cause sporadic genetic disease, and ~5% of humans have a new LINE-1–mediated insertion not found in either parent. Understanding this process is critical because LINE-1 contributes to the pathologies of common diseases including cancer by causing inflammation and DNA damage. DNA damage is EN-dependent, and the number of γH2AX damage foci far exceeds completed retrotransposition events in model systems, suggesting DNA damage may occur outside of retrotransposition. Canonical LINE-1 insertions occur preferentially at 5'-TTTTT↓AA consensus motifs, but it is unknown whether this preference comes from site-specific EN DNA nicking or by sequence requirements in the subsequent RT step. What sequences EN cuts in the human genome are not known because only completed events can be recovered from cells due to technical limitations in purifying ORF2p and in mapping nicks. To address these knowledge gaps, we purified ORF2p EN to homogeneity from E. coli as well as an inactive mutant as a control and assayed their cleavage patterns on Lambda, *E. coli*, and human DNAs using electronic genome mapping (EGM), a technique that can directly measure nick positions on a genome-wide scale.

Uncertainties regarding both the nick site sequence and how efficient the process is make designing an assay challenging. We modified our standard analytical methods to minimize false positives so that low level nicking could be detected. With all DNAs tested, frequency of nicks increased in an EN-concentration manner with mutant EN having a lower nicking rate. Mapping of low frequency nicks on the E. coli genome yielded novel sequences that appear to be nicked by ORF2p EN. Conditions are being optimized so that the same type of analysis can be carried out on the human genome. These results reveal a powerful new method to directly measure endonuclease nicking at genome scale. Future applications of this technology may also be used to directly assay other DNA modifications and damage at scale.

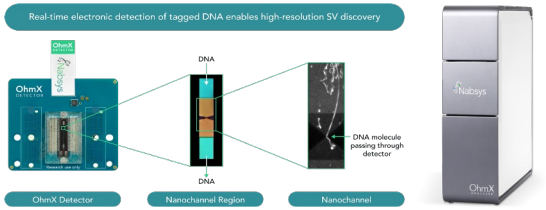


Figure 1. EGM detector and analyzer. EGM is enabled by the novel, solid-state nanodetectors developed by Nabsys. The OhmX™ Detector houses 256 parallel nanochannels, each with its own electronic sensor. Nabsys has developed DNA tagging chemistry that provides a high signal-to-noise. The high signal and spatial resolution allows tags to be closely spaced on the high molecular weight DNA molecules. This yields optimal tag density such that structural variations as small as 300 bp can be routinely identified and mapped genome-wide.

Figure 2. EGM workflow overview. The EGM workflow consists of the following steps: (1) high molecular weight (HMW) DNA isolation, (2) genome preparation using the OhmX Velocity Optimized Labeling and Tagging (VOLT™) Kit, (3) sample loading, (4) electronic detection of tagged molecules using the OhmX Analyzer, and (5) data analysis.

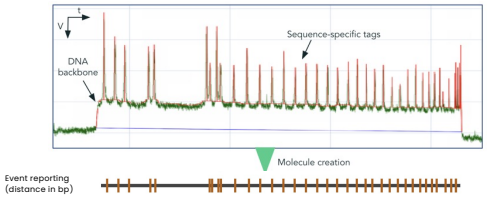
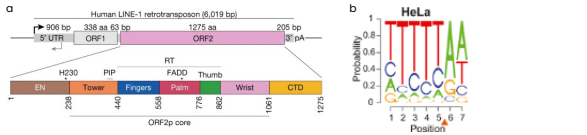


Figure 3. Electronic profile for tagged DNA molecules over 150 kb. The baseline voltage is shown on the left. When the DNA molecules enter the nanochannel, a change in voltage is observed. Whenever a site-specific tag on the DNA is also present, the voltage changes again. The time between tags is converted to distance in base pairs.

LINE-1 Background



A) Baldwin, E. T. et al. Structures, functions and adaptations of the human LINE-1 ORF2 protein. Nature 626, 194–206 (2024)
B) Flasch, D. A. et al. Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication. Cell 177, 837–851 e828 (2019)

Figure 4. Human LINE-1 encodes a retrotransposition-facilitating machinery. (a) LINE-1 ORF2p is a 1275 amino acid multi-domain protein with EN and RT function. We focused on the EN domain of the ORF2 protein. (b) The commonly cited consensus motif (5'-TTTTT↓AA) was derived from genomic insertion sites, which inherently selects for sequences that support reverse-transcription priming. The insertion sites may not be the same sites that LINE-1 EN is binding or nicking.

LINE-1 EN Nicks at Consensus Site in Lambda

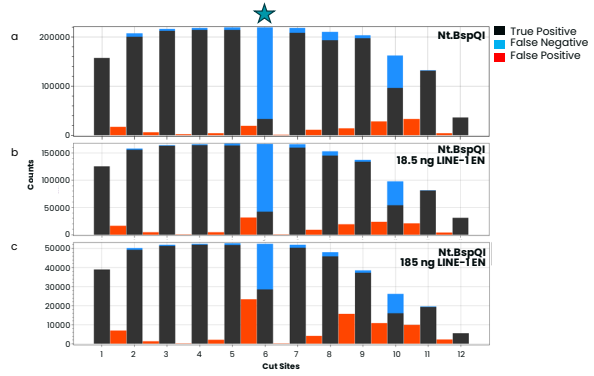
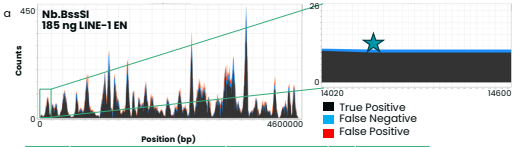


Figure 5. Tag coverage on Lambda DNA with LINE-1 EN. (a) An example tag coverage plot for a run of Lambda DNA nicked with Nt.BspQI is shown with the 10 BspQI sites and the 2 LINE-1 EN consensus sites (6 and 11). This depicts where the tags are located and at which sites there are missing tags. There is a low false negative rate on the BspQI sites and a low false positive rate. Further, site 6 has minimal tagging, as this is a LINE-1 EN consensus site and no LINE-1 EN was included in the experiment. (b, c) Upon addition of LINE-1 EN, there is a dose-dependent signal at site 6 (starred), where increasing LINE-1 EN amount increases nicking at site 6 (indicated by more true positives at the site). Red bars denote nicks not present in the reference, comprising both true false positives and LINE-1 EN nicks that are not the consensus motif.

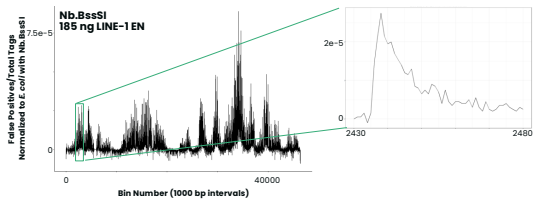
LINE-1 EN Nicks at Consensus Motif in *E. coli*



LINE-1 EN	Nick Sites Included in Reference	%Filtered Remapped	%FP	%FN (> 500 bp)
-	BssSI	63.9	11.59	5.93
+	BssSI	1.2	22.28	8.59
+	BssSI, TTTAAAAA	2.9	19.29	15.25

Figure 6. Tag coverage and remapping data with LINE-1 EN in *E. coli*. Validating the findings in Lambda (Figure 5) in *E. coli*. (a) Tag coverage plot for a run of *E. coli* DNA with Nb.BssSI and LINE-1 EN. The zoom in is on a region that contains a LINE-1 EN consensus site (indicated by the star). (b) When we remap the DNA molecules to the reference in samples that contain Nb.BssSI and LINE-1 EN, we get substantially fewer molecules to align and these molecules have a higher percent false positives (comprising both true false positives and LINE-1 EN nicks) than samples run with just Nb.BssSI. When we add the LINE-1 EN consensus motif to the reference, the remapping statistics improve, indicating that this is likely a real LINE-1 EN nicking site. However, the false positives are still high, indicating there are likely other sequences that LINE-1 EN nicks.

Searching for Novel LINE-1 EN Motifs in *E. coli*



LINE-1 EN	Nick Sites Included in Reference	%Filtered Remapped	%FP	%FN (> 500 bp)	Δ %Filtered Remapped/site	# of Sites
-	BssSI	63.9	11.59	5.93	-	807
+	BssSI	1.2	22.28	8.59	-	807
+	BssSI, TTTAAAAA	2.9	19.29	15.25	0.0051	333
+	BssSI, TTTAAAAA, CTTATCAGGCCT	3.5	19.21	15.48	0.0146	158
+	BssSI, TTTAAAAA, CGGATAAGGCGTTNACGCCGCATCC	3.6	19.09	15.15	0.0197	122
+	BssSI, TTTAAAAA, GCCGCCGCG	4.5	18.96	16.23	0.0244	135

Figure 7. Mining false positives for new LINE-1 EN motifs. (a) False positive (representing LINE-1 EN nick sites) sites in *E. coli* after an OhmX run with Nb.BssSI and LINE-1 EN. The false positives were normalized to the total number of tags observed. To eliminate more of the true false positives, the same analysis was performed with *E. coli* nicked with just Nb.BssSI and these values were subtracted from the values calculated for the *E. coli* sample run with Nb.BssSI and LINE-1 EN. The zoom in shows a false positive peak. The sequences of the peaks were analyzed using MEME Suite to identify motifs. (b) The new motifs were added to the reference for remapping the DNA molecules and the motifs shown improve the remapping statistics. Thus, these motifs are potential new LINE-1 EN motifs.

Patterns in LINE-1 EN Nick Site Location

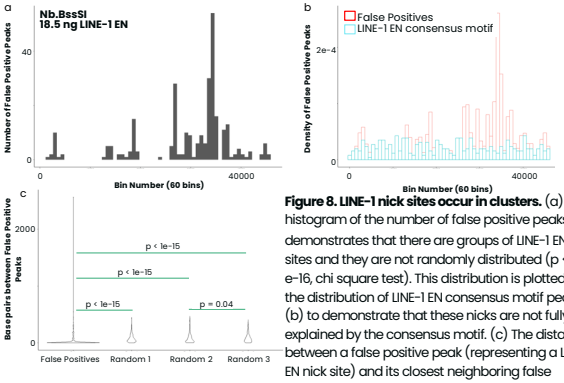


Figure 8. LINE-1 nick sites occur in clusters. (a) A histogram of the number of false positive peaks demonstrates that there are groups of LINE-1 EN nick sites and they are not randomly distributed ($p < 2.2$ e-16, chi square test). This distribution is plotted with the distribution of LINE-1 EN consensus motif peaks (b) to demonstrate that these nicks are not fully explained by the consensus motif. (c) The distance between a false positive peak (representing a LINE-1 EN nick site) and its closest neighboring false positive peak was plotted. Three sets of random integers containing the same number of integers as the number of false positive peaks were generated and the above analysis was repeated. Reported p values are the significant results of the K-S test (differences in distribution) with Bonferroni correction. This shows that the peaks are closer together than expected by chance.

Conclusions

- **Accurate detection:** EGM can be used to detect nicks made by LINE-1 EN
- **Validation:** The LINE-1 EN consensus motif was confirmed as a nicking site
- **Discovery:** Additional potential LINE-1 EN nicking sites were identified and can be validated using alternative methods. Another possibility is that these motifs are in regions that are prone to breaking when a LINE-1 enzyme attaches to the DNA, bending it before nicking it. One possibility for why the LINE-1 EN nick sites are grouped together is that once LINE-1 finds a recognition sequence to attach to the DNA, it can track along the DNA before nicking.
- **Alternative methods:** ChIP-seq, END-seq, and CUT&Run could be used, but EGM is faster and lower-cost.

Learn more at Nabsys.com

© 2026 Nabsys. For Research Use Only. Not for use in diagnostic procedures.