Sensitive Detection of LINE-1 Retrotransposition Across Cancer Uncovers Biological Correlates

Authors: Alexander Solovyov[1]*, Julie Behr[2]*, Eric Banks[2], Jimmy Z. Zhong[2], Enrique Garcia-Rivera[2], Wilson McKerrow[2], Bryan Thornlow[2], Dennis Zaller[2], Menachem Fromer[2], Benjamin Greenbaum[1,3]****

[1] Computational Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[2] ROME Therapeutics, Cambridge, MA
[3] Physiology, Biophysics & Systems Biology, Weill Cornell Medical College, New York, NY, USA

The Mobile DNA Conference: Evolution, Diversity, and Impact
June 5-9, 2022

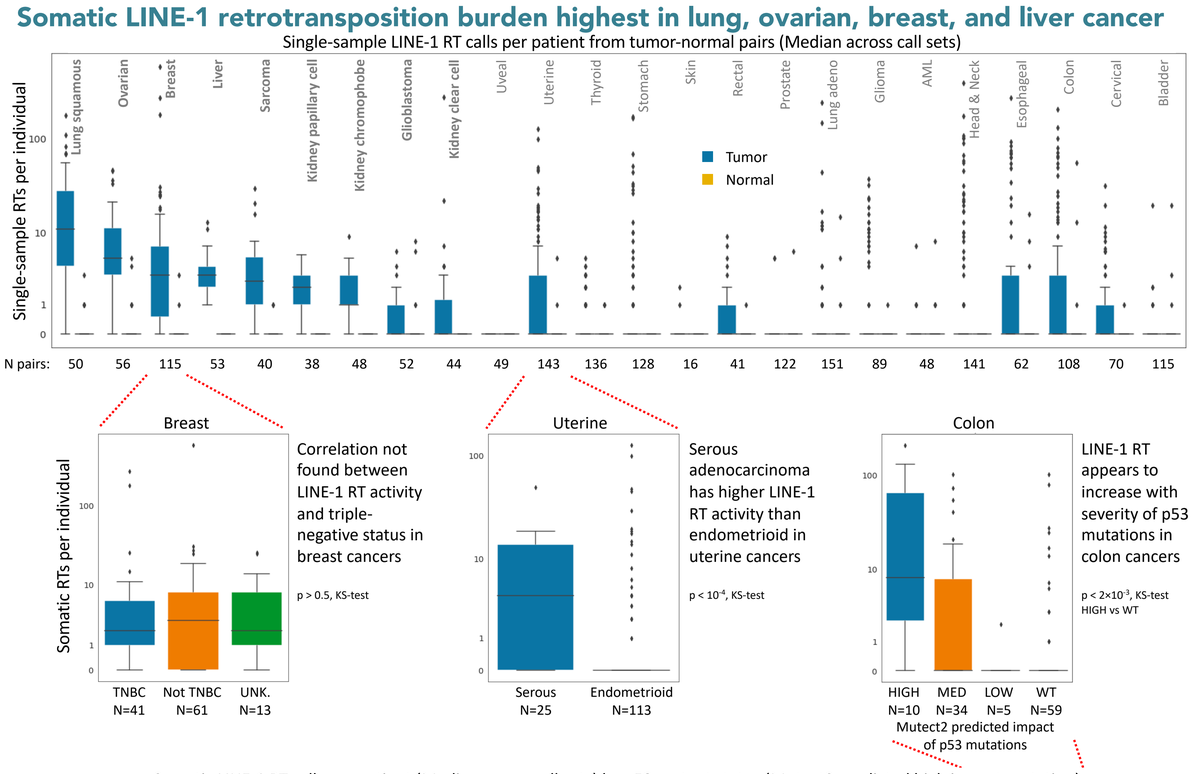
Abstract

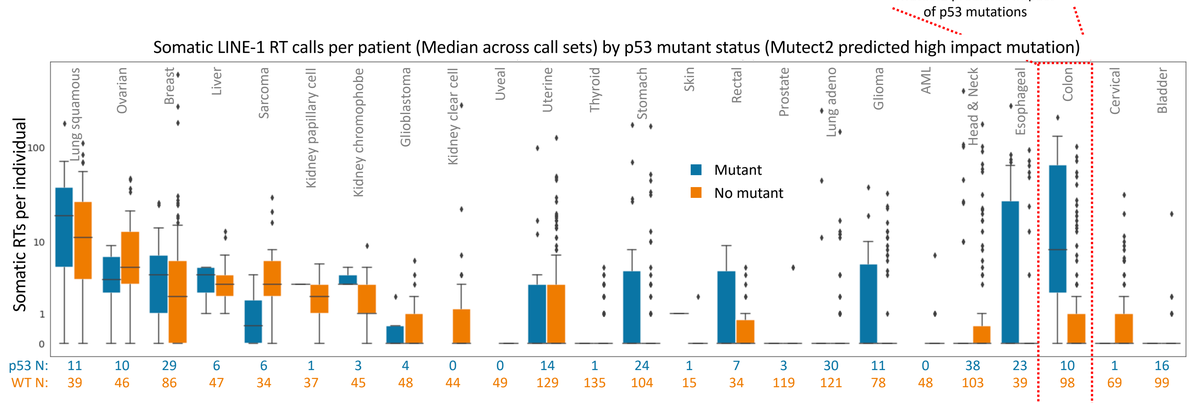
Retrotransposons, including the autonomous element Long Interspersed Nuclear Element 1 (LINE-1), are abundant throughout the human genome but typically repressed in healthy somatic tissues. Increasing evidence suggests activity of LINE-1 in many cancers, including those arising from gastrointestinal tissues, which may contribute to genomic instability. The impact of activated LINE-1 on the immune system may provide a new opportunity for therapeutic intervention, if the mechanisms and corresponding cellular pathways of elevated LINE-1 activity in cancers were better understood.

Detection of LINE-1 retrotransposition from short-read DNA sequencing data continues to be challenging due to the high number of LINE-1 copies in the genome, the length of LINE-1 insertions relative to short reads, and frequent mis-processing of retrotransposition-supporting reads by standard analysis pipelines. While methods have been developed for detecting retrotranspositions[1,2], we set out to build a more sensitive approach by focusing on two key signals when aligning sequencing reads to a reference genome: (a) reads that span an insertion site will contain insertion site sequence as well as the inserted LINE-1 sequence ("clipped reads"); (b) paired-end reads arising from fragments spanning retrotransposed LINE-1 may map with one of the pairs near the insertion site but the other read mapping to LINE-1 elsewhere in the genome ("discordant pairs"). By anchoring our approach in "clipped reads," we often derive more precise genomic breakpoints.

We have evaluated 1,925 independent tumor/normal whole-genome sequencing sample pairs across 25 different tumor types from The Cancer Genome Atlas to identify non-reference LINE-1 insertions and associate somatic retrotransposition (RT) burden with other biological properties, including LINE-1 expression estimates across these tumor types.

Somatic LINE-1 retrotransposition burden highest in lung, ovarian, breast, and liver cancer





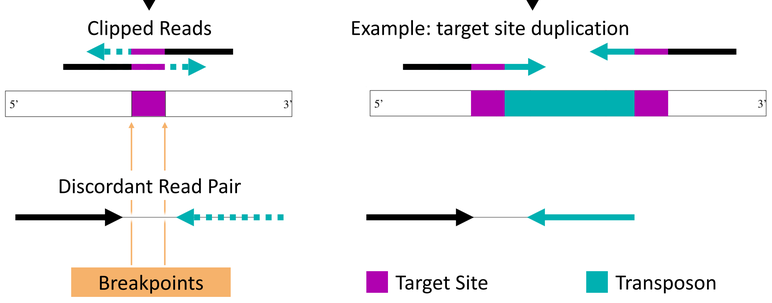
Total ReCall algorithm finds retrotranspositions (RT) by prioritizing clipped reads

What sequencing reads look like when mapped to reference genome:

What the genome looks like:

Clipped (or "split") reads:
- Solid purple line - Sequence present in normal genome
- Dashed green line - Inserted transposon sequence

Discordant read pairs:
- One read (black) in the pair maps near the breakpoint
- Another read (green) maps to the transposon

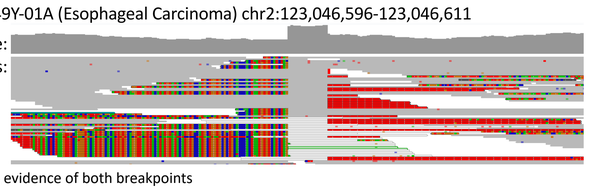


- "First pass": require strong evidence of BOTH left and right breakpoint
- "Second pass": consider calls with weaker evidence of left and/or right breakpoint

Parameters considered for filtering RT calls

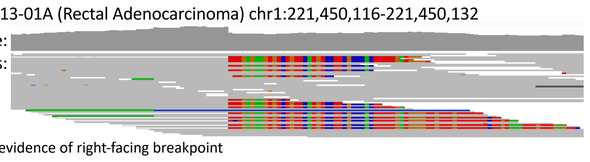
These features, used to identify a "call", can be filtered at varying threshold levels to generate call sets

TCGA-LN-A49Y-01A (Esophageal Carcinoma) chr2:123,046,596-123,046,611



Strong evidence of both breakpoints

TCGA-AF-3913-01A (Rectal Adenocarcinoma) chr1:221,450,116-221,450,132



Weak evidence of right-facing breakpoint

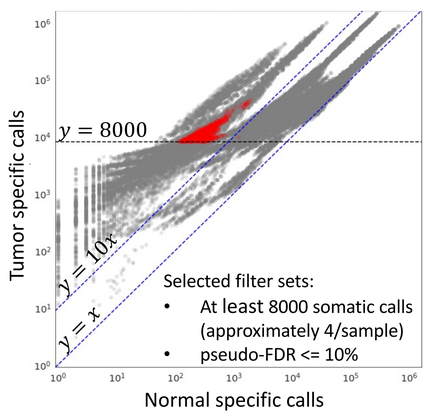
A. Minimum clipped reads at both sides
B. Minimum total insertion-supporting (clipped + any discordant) reads at both sides
C. Minimum LINE insertion-supporting (clipped + discordant aligned to LINE-1) reads at both sides
D. Minimum median read depth (case sample)
E. Require 5' and/or 3' "L1HS" consensus sequences found in the breakpoint signatures
F. Minimum length (bp) of assembled sequence at both sides
G. Maximum coverage ratio noise (case and control samples)
H. Maximum distance (bp) from left to right breakpoint
I. Maximum clusters of clipped reads at either breakpoint
J. Maximum independent samples with the same call (both breakpoints exactly matching)
K. First pass call
L. Are the breakpoints in low complexity sequence
M. Minimum clipped reads at both sides with respect to median coverage

Evaluating Total ReCall Specificity in tumor-normal pairs

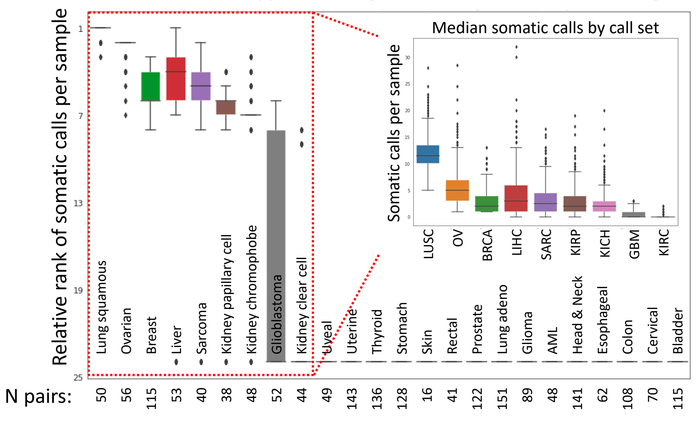
Evaluation of sensitivity/specificity of call sets with respect to curated ground-truth/other methods is ongoing

Preliminary assessment of specificity was performed by treating paired samples as replicates (true positive calls identified in the normal sample should also be in found in the paired tumor):

pseudo-false discovery rate (FDR): $\frac{normal\ specific\ calls}{all\ normal\ based\ calls}$

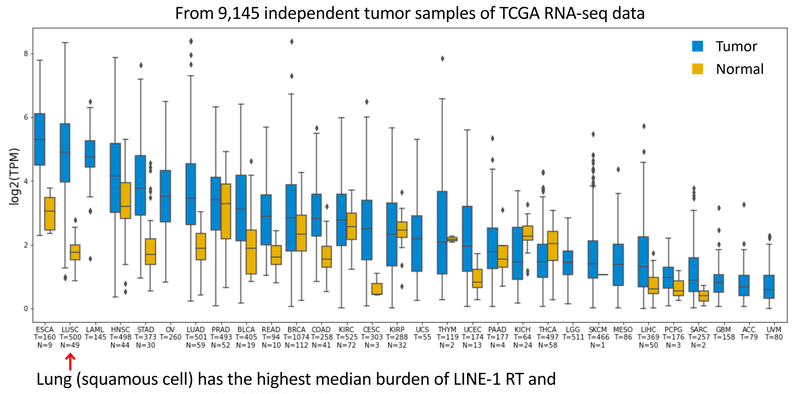


Relative tumor type rankings are robust across filterings

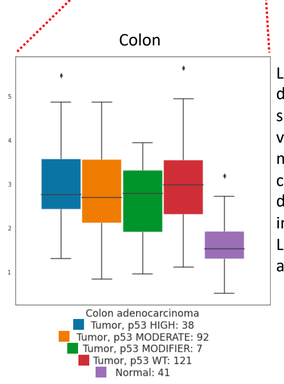


LINE-1 RNA expression[3,4] differs from RT activity

Estimated total LINE-1 expression (in log2(transcripts per million)) per sample
From 9,145 independent tumor samples of TCGA RNA-seq data



Lung (squamous cell) has the highest median burden of LINE-1 RT and the second highest median LINE-1 RNA levels



LINE-1 RNA does not significantly vary with p53 mutations in colon cancers, despite their impact on LINE-1 RT activity

Conclusions

- Total ReCall ranking of TCGA tumor types by highest to lowest median somatic burden of LINE-1 RT per sample: Lung squamous cell, Ovarian, Breast, Liver, Sarcoma, Kidney papillary cell, Kidney Chromophobe, Glioblastoma, Kidney clear cell
  All other TCGA tumor types have an estimated median somatic LINE-1 RT burden per sample of 0.
- Lung squamous cell carcinoma also highly expresses LINE-1 at the RNA level.
- In Uterine cancer, LINE-1 RT only seems to be active in serous adenocarcinoma, and not in endometrioid adenocarcinoma.
- In Colon cancer, p53 mutation status does not influence LINE-1 RNA expression, but does seem to influence LINE-1 RT.

References

[1] Rodriguez-Martin, B., Alvarez, E.G., Baez-Ortega, A., Zamora, J., Supek, F., Demeulemeester, J., Santamarina, M., Ju, Y.S., Temes, J., Garcia-Souto, D. and Detering, H., 2020. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nature genetics, 52(3), pp.306-319.
[2] Chu, C., Borges-Monroy, R., Viswanadham, V.V., Lee, S., Li, H., Lee, E.A. and Park, P.J., 2021. Comprehensive identification of transposable element insertions using multiple sequencing technologies. Nature Communications, 12(1), pp.1-12.
[3] McKerrow, W. and Fenyö, D., 2020. L1EM: a tool for accurate locus specific LINE-1 RNA quantification. Bioinformatics, 36(4), pp.1167-1173.
[4] McKerrow, W., Wang, X., Mendez-Dorantes, C., Mita, P., Cao, S., Grivainis, M., Ding, L., LaCava, J., Burns, K.H., Boeke, J.D. and Fenyö, D., 2022. LINE-1 expression in cancer correlates with p53 mutation, copy number alteration, and S phase checkpoint. Proceedings of the National Academy of Sciences, 119(8), p.e2115999119.