Mobile Element Insertions and Associated Structural Variants in Longitudinal Breast Cancer Samples

Cody J. Steely[1*], Kristi L. Russell[1], Julie E. Feusier[1], Yi Qiao[1,2], Gabor Marth[1,2], and Lynn B. Jorde[1]

[1] Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA

[2] USTAR Center for Genetic Discovery, Salt Lake City, Utah 84112, USA

*Corresponding author: Cody Steely, Department of Human Genetics, University of Utah, 15 N. 2030 E. Rm 5100, Salt Lake City, UT 84112, USA. Email: cody.steely@utah.edu

Running Title: Mobile elements in longitudinal breast cancer

Keywords: Retrotransposons, Mobile Elements, Cancer, Structural Variation, Transposable Elements

Abstract

While mobile elements are largely inactive in healthy somatic tissues, increased activity has been found in cancer tissues, with significant variation among different cancer types. In addition to insertion events, mobile elements have also been found to mediate many structural variation events in the genome. Here, to better understand the timing and impact of mobile element insertions and mobile element-mediated structural variants in cancer, we examined their activity in longitudinal samples of four metastatic breast cancer patients. With whole-genome sequencing data from multiple timepoints through tumor treatment and progression, we used mobile element detection software followed by visual confirmation of the insertions. From this analysis we identified 11 mobile element insertions or mobile element-mediated structural variants, and found that the majority (nine of the eleven) of these occurred early in tumor progression. Two of the identified insertions were SVA elements, which have not been examined in previous cancer studies. Most of the variants appear to impact intergenic regions; however, we identified a mobile element-mediated translocation in *MAP2K4* and a mobile element-mediated deletion in *YTHDF2* that likely inactivate reported tumor suppressor genes. *MAP2K4* is part of the *JNK* signaling pathway, influencing cell growth and proliferation. The high variant allele frequency of this translocation and the important function of *MAP2K4* indicate that this mobile element-mediated translocation is likely a driver mutation. Overall, using a unique longitudinal dataset, we find that most variants are likely passenger mutations in the four patients we examined, but some variants impact tumor progression.

Introduction

Mobile elements, or transposable elements, are segments of DNA that are capable of mobilizing from one genomic location to another. These elements compose a significant portion

of the human genome, with estimates ranging from nearly 50% up to 66% (Lander et al. 2001; de Koning et al. 2011). In humans, retrotransposons represent a class of active mobile elements, inserting new elements through a "copy and paste" mechanism (Boeke et al. 1985). While most of these elements have become transcriptionally inactive over time (Brouha et al. 2003; Beck et al. 2010), three mobile element families remain active in humans including LINE-1 (L1), *Alu* elements, and SVA. These three mobile element families compose nearly one-third of the human genome (Lander et al. 2001; Xing et al. 2013). With only a small fraction of these elements retaining the ability to retrotranspose, recent work has shown that germline mobile element insertions in humans are quite rare (Feusier et al. 2019). The majority of these insertions seem to have no adverse impact; however, some insertions have been found to cause disease (Hancks and Kazazian 2016). Mobile element activity in human disease became a subject of interest after two hemophilia A patients were found to have *de novo* L1 insertions in the *F8* gene (Kazazian et al. 1988). Since this initial discovery, mobile element insertions have been found to be associated with more than 130 disease cases (Kazazian and Moran 2017).

While somatic insertion events generally do not appear to occur in most healthy tissues, previous studies show some level of somatic activity in certain diseases or healthy tissues. For healthy tissues, many studies have detected somatic mosaicism of L1 elements in neurons (Muotri et al. 2005; Coufal et al. 2009; Richardson et al. 2014; reviewed in Erwin et al. 2014). This increase in activity may be due to slight changes in methylation at L1 loci (Coufal et al. 2009). The disease that most commonly shows somatic mobile element activity is cancer. Recently, multiple studies have noted varying degrees of activity of mobile elements in numerous cancer tissues (Lee et al. 2012; Helman et al. 2014; Tubio et al. 2014; Doucet et al. 2015; Doucet-O'Hare et al. 2016; reviewed in Burns 2017). The level of activity has been found

to be quite variable in patients with the same type of cancer, and also shows a high degree of variability among cancer types (Helman et al. 2014; Tubio et al. 2014). Regardless of cancer type, roughly half of all tumors have at least one somatic L1 insertion (Tubio et al. 2014).

The impact of mobile elements on the cancer genome is not limited only to somatic insertions, but also includes the structural variation (SV) events that they mediate (reviewed in Cordaux and Batzer 2009). Notably, mobile elements mediate roughly 10% of all SV events over 100 base pairs long found in the human genome (Xing et al. 2009). Mobile elements have been found to mediate SV events that can lead to cancer development (Mauillon et al. 1996; Hsieh et al. 2005). For example, approximately 42% of the *BRCA1* gene is composed of *Alu* elements (Welcsh and King 2001), making this gene a target for non-allelic homologous recombination, specifically *Alu*-*Alu* mediated events (Petrij-Bosch et al. 1997; Puget et al. 1997; Rohlfs et al. 2000; Peixoto et al. 2013). L1 transduction events have also been examined in a variety of cancers (Tubio et al. 2014).

Multiple studies have shown that L1s, in particular, demonstrate increased retrotransposition activity in cancer tissues. However, the timing of these insertions or mobile element-mediated SVs has not been thoroughly investigated. Additionally, L1s and the structural variants that they mediate have been the focus of most previous studies, with little focus on the impact of *Alu* and SVA insertions or the SVs they mediate. In this study, we utilize longitudinally sampled breast cancer tissues to better understand the timing and significance of increased mobile element activity on the cancer genome.

Results

To analyze mobile element activity during tumor progression, we used longitudinal whole-genome sequencing (WGS) data from four metastatic breast cancer patients. Tumor samples were obtained from these patients over multiple years (2-15 years) with varying WGS and bulk RNA-seq timepoints (2-6) available for our analyses. Germline DNA was sequenced from the blood of each patient. The WGS timepoints for each patient, and a summary of treatment information over time, are shown in Figure 1. More detailed treatment information for each patient can be found in the supplementary information of Brady *et al.* 2017. All four patients were estrogen receptor (ER) +. Patients 1 and 2 were human epidermal growth factor receptor 2 (HER2) +, while Patients 3 and 4 were HER2-.

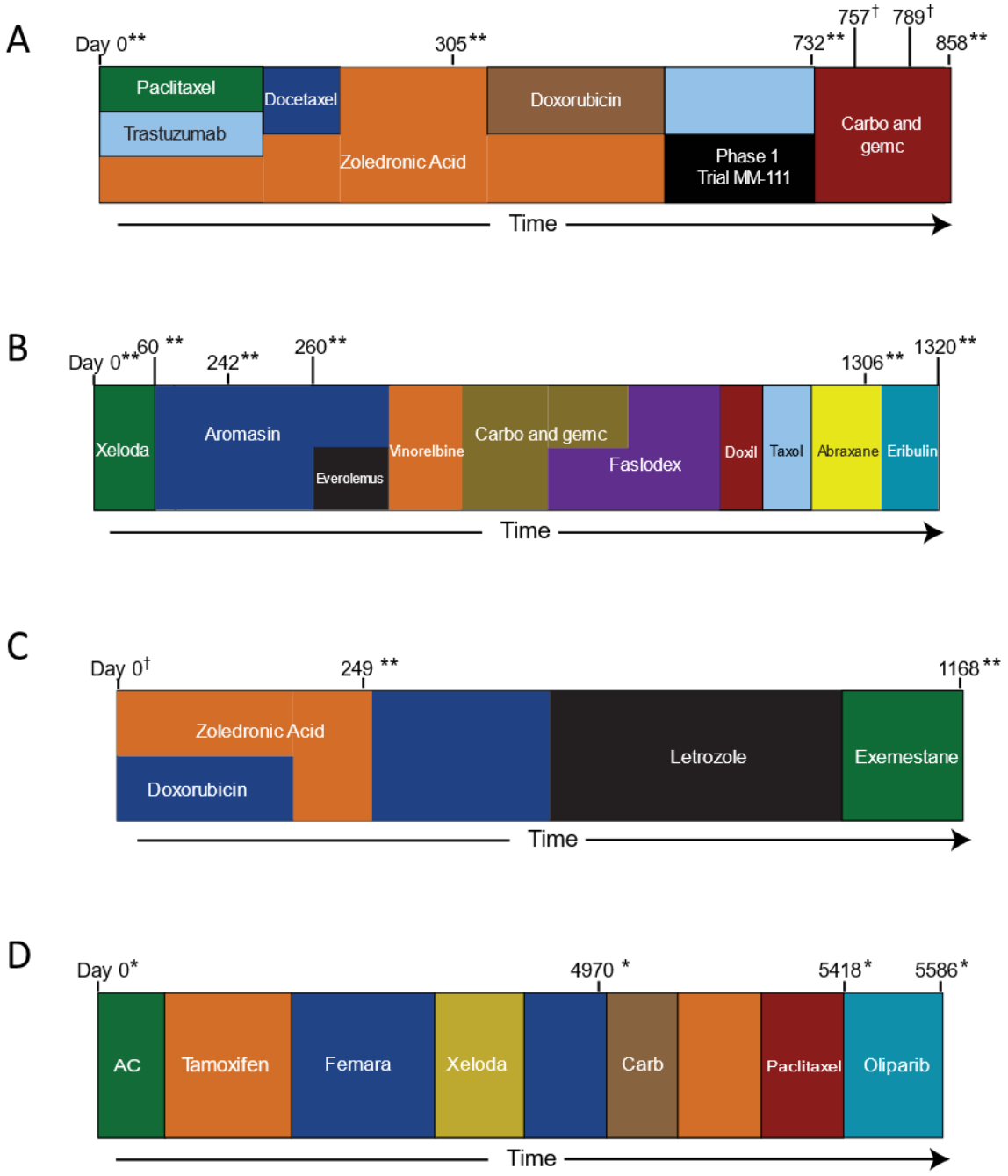

Figure 1. Outline of patient treatment information. Parts A-D correspond with patients 1-4, respectively. Each medication or treatment is identified by a colored box. The length of the box approximates the duration of each treatment. If multiple treatments overlapped, the boxes are stacked through the timepoints for which they overlapped. The timepoint of each sampling event is shown on the top of the box in days after the first sampling timepoint. *= DNA, **= DNA and RNA, †= RNA

We analyzed each of the four patients for mobile element insertions utilizing three tools (MELT, TranSurVeyor, and RUFUS, further described in the Methods section). After identifying and filtering mobile element insertions and SVs (see Methods), we generated a list of potential insertion sites. These potential insertion sites were compared to the matched germline sample for each patient to ensure that the insertion identified was a somatic event. Timepoints for each patient were analyzed individually and compared to the matched germline sample. Following our identification and filtering steps, we were able to identify mobile elements and associated variants at both very high VAF (100%) and very low VAF (~5%).

Collectively, we identified seven mobile element insertions and four structural variants that were mediated by mobile elements in tumor samples. These variants all appear to be somatic mutations acquired during tumorigenesis or during tumor progression because these variants are not present in the corresponding germline samples. The variants that we have identified in each of the four patients are summarized in Table 1. We find that the majority of insertions and variants associated with mobile elements (nine of the eleven) occur before the first sampling timepoint, because these were already present in the first and then all subsequent samples. The two variants that were not present in the first sampling timepoint were both SVA insertions. We also find high variability in the number of insertions and variants that were identified in each of these patients, with three of the patients (Patients 1, 2, and 4) showing multiple insertions and variants, and Patient 3 showing only a single mobile element mediated structural variant. The use of multiple tools and visual inspection for the identified insertions excluded a number of false positive calls. To prevent false negative calls for variants with low frequency, the recommended filters for each tool were not used. Instead, calls made for each timepoint were compared to the matched germline sample. With this filtering, only the mobile elements unique to the tumor

sample remained. After filtering, MELT identified 17,704 potential mobile element variants, and TranSurVeyor identified 69,388 potential mobile element variants. Counts for the images produced by RUFUS were not included here as RUFUS is designed to detect many types of variants, not exclusively mobile elements.

Most of the identified insertions and variants were found in intergenic regions and are unlikely to impact gene expression as they do not overlap known regulatory or ultra-conserved regions. However, we identified multiple insertions that appeared to impact intronic regions: an SVA insertion in *CCT4,* which was identified in Patient 1, and an L1 insertion in *ANKRPD26P*, which we identified in Patient 2. We also identified two *Alu*-mediated somatic structural variation events that impact exons in two genes: an *Alu-Alu* mediated deletion in *YTHDF2* and an *Alu*-mediated translocation impacting *MAP2K4*. Both of these *Alu*-mediated structural variants occurred in Patient 1 and occurred before the first sampling timepoint, but were not present in the germline sequencing data.

Table 1. Mobile element insertions and mobile element mediated structural variants identified in four longitudinal breast cancer patients. *Alu-Alu* mediated deletions are abbreviated as AAMD. L1-L1 mediated deletions are abbreviated as L1MD. Insertion or variant sites found within genes (exonic or intronic) are indicated in the "Genes affected" column.

Patient	Locus	Variant	Region	Genes affected	Present before first tumor timepoint
Patient 1	1:186,755,400	Alu	Intergenic		Yes
	1:29,089,030	AAMD (23kb)	Exon	*YTHDF2*	Yes
	2:62,111,770	SVA	Intronic	*CCT4*	No
	17:11,974,350	Alu-mediated translocation	Exon	*MAP2K4*	Yes
Patient 2	2:15,168,000	Alu	Intergenic		Yes
	8:99,227,400	L1	Intergenic		Yes
	13:19,866,150	L1	Intronic	*ANKRPD26P3*	Yes
Patient 3	6:100,589,720	Alu-mediated deletion	Intergenic		Yes
Patient 4	10:67,431,020	Alu	Intergenic		Yes
	11:81,149,160	L1MD	Intergenic		Yes
	19:23,124,500	SVA	Intergenic		No

To remove contamination from non-tumor cells, we estimated the tumor purity (or cancer cell fraction) of each timepoint. The purity estimates, as well as estimates for genome-wide copy number for each timepoint in each patient, are included in Table 2. There are many changes in genomic copy number, with genome-wide estimates ranging from 1.86, just below normal cellular copy number, up to 4.12, more than twice that seen in most normal cells. In these patients, tumor purity ranged from 30.91% to 94.5%, with a mean value of 76.25%. The only formalin-fixed, paraffin-embedded (FFPE) sample in the dataset had the lowest tumor purity value.

Table 2. Estimates of tumor purity and average genome-wide copy number for each of the samples at each timepoint.

Patient	**Sampling timepoint**	**Estimated Tumor Purity (%)**	**Genomic Copy Number**
Patient 1	Day 0	94.5	1.88
	Day 305	90.75	1.87
	Day 732	79.21	1.86
	Day 858	60.76	1.86
Patient 2	Day 0	89.62	3.9
	Day 60	88.51	3.98
	Day 242	79.75	3.86
	Day 290	87.32	3.96
	Day 1305	78.97	3.73
	Day 1320	72.15	3.73
Patient 3	Day 249	85.51	2.09
	Day 1168	81.82	1.95
Patient 4	Day 0	30.91	4.12
	Day 2970	85.91	1.98
	Day 5418	74.78	2.01
	Day 5586	39.58	2.03

We calculated the variant allele frequencies (VAF) for each of the mobile element insertions or SVs identified after adjusting for tumor purity (Figure 2). The VAFs calculated from the adjusted data are slightly higher, but very similar to the VAFs calculated from the unadjusted VAFs (Figure S1). Of the four insertions and SVs identified in Patient 1 (Figure 2A), the *Alu-Alu* mediated deletion in *YTHDF2* and the translocation in *MAP2K4* were present at very high frequency (>80% for the deletion; 100% for the translocation) at the first timepoint, with decreased frequency over time. The other two identified events were insertions (an *Alu* and SVA) and were present at much lower frequencies. The SVA identified in this patient was not present at the first two sampling timepoints but is present by the third and remained in the tumor cells in the fourth sampling timepoint. Each of these events show a slight decrease over sampling timepoints. We identified three somatic insertions in Patient 2 (Figure 2B), two L1 insertions, and one *Alu* insertion. The L1 insertion on chromosome 13 was present at a higher frequency than the other two insertions found in this patient. The frequency of these variants closely resembles the frequency of the mobile element insertions seen in Patient 1, but does not reach the high frequency seen in the translocation or deletion of Patient 1.

Of the four individuals sampled, Patient 3 (Figure 2C) had the fewest variants, a single TE-mediated deletion, which appears to be an *Alu-Alu* mediated deletion. This deletion was present at a high frequency in the first sampling timepoint (present in about 75% of reads), but decreased to approximately 50% by the second sampling timepoint. In Patient 4 (Figure 2D), we identified two insertions, an *Alu* and SVA, and a single mobile element-mediated SV, a L1-L1 mediated deletion (L1MD). The first sequenced sampling timepoint for Patient 4 has been excluded in the VAF plot as it was a FFPE sample, decreasing our ability to accurately calculate a VAF for this timepoint. The SVA insertion identified in this patient was not present for the first

sampling timepoint, but was present for the next two sampling timepoints. By the final sampling timepoint the SVA was no longer observable in the sequencing data. The L1MD trends upward in VAF before sharply falling at the final timepoint. The *Alu* insertion in this patient appears to maintain a steady VAF of ~20%.

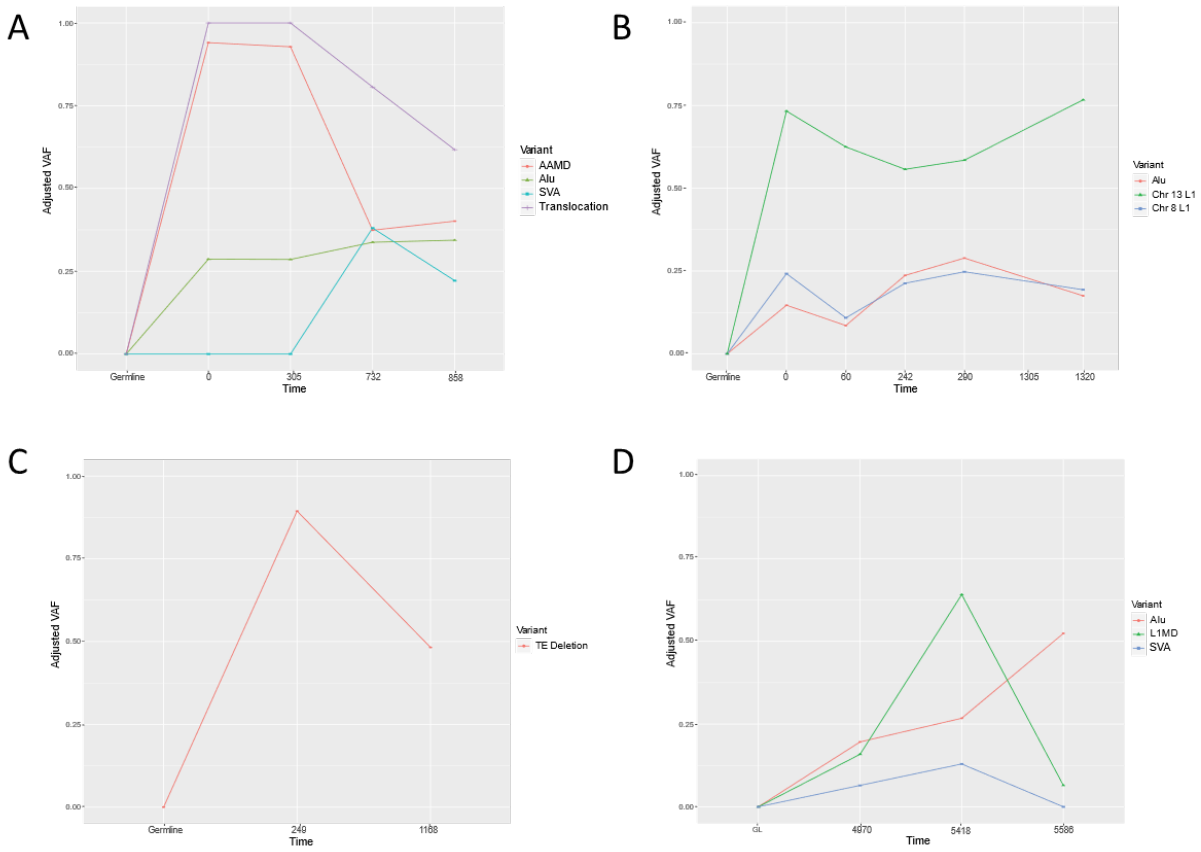

Figure 2. Adjusted variant allele frequency (VAF) for each of the 4 patients. Parts A-D correspond to patients 1-4, respectively. The VAF for each patient has been adjusted to better reflect the percentage of cells in the sample that appear to be from the tumor. VAF is shown on the Y axis with each sampling timepoint for a particular patient shown on the X axis.

After identifying the mobile element insertions and the SVs that they mediate in these four patients (Table 1), we analyzed the impact of these variants on the cancer genome. Because both variants that impacted exons were found at high frequency in Patient 1, we examined the

genomic copy number present at these loci. Specific genome-wide copy number estimates for the first timepoint in the genome of Patient 1 are shown in Figure 3. Both the *Alu-Alu* mediated deletion and the *Alu* mediated translocation appear to have been reduced to only a single copy, contributing to the high frequency for both.

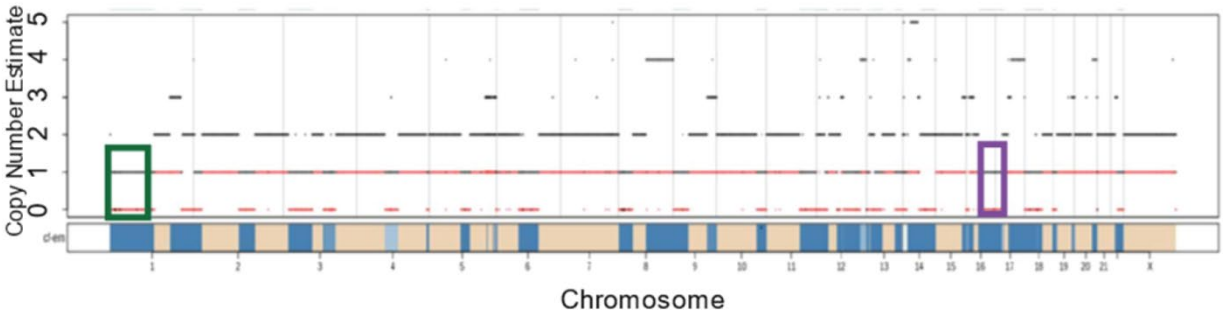

Figure 3. Copy number estimates from FACETS for Patient 1. The approximate location of the deletion in the *YTHDF2* gene is shown in the green box. The approximate location of the breakpoint for the *MAP2K4* translocation that appears to be mediated by *Alu* elements are shown in the purple box on chromosome 17. Both events occurred in regions that have a decreased copy number. The black lines in the figure represent the total copy number, while the red lines show the minor copy number for each segment.

Due to decreased copy number at the *MAP2K4* locus, and the translocation interrupting the one remaining copy of *MAP2K4* (a depiction of this translocation is shown in Figure 4A), we validated absence of the complete transcript and protein in the cancer cells of Patient 1. Creating a *de novo* transcript assembly of RNA-seq data for each timepoint in Patient 1 (see Methods), we find only truncated transcripts of *MAP2K4.* For further validation of lack of protein production by *MAP2K4* in the tumor cells of Patient 1, we performed a Western blot using HEK293T cells as a control and tumor cells taken from an ascites sample from Patient 1 (Figure 4B). While the control cells show clear production of MAP2K4, there appears to be no MAP2K4 being produced in the tumor cells of Patient 1.

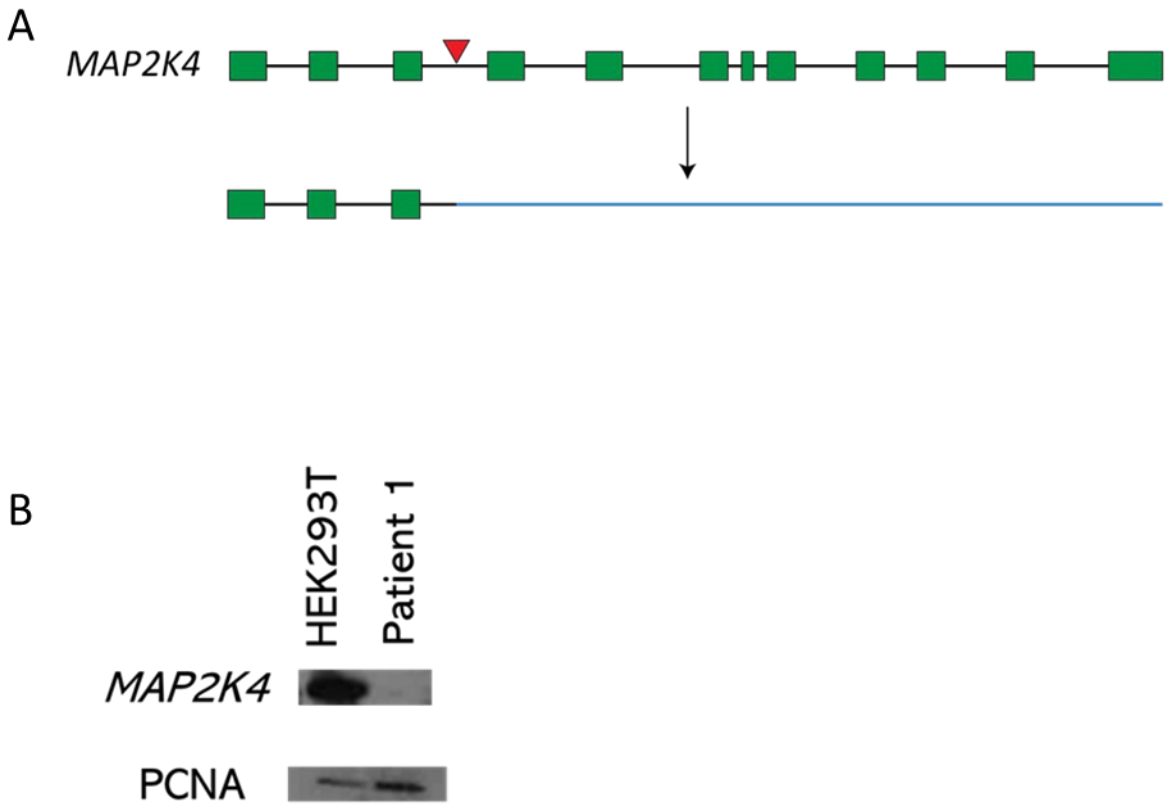

Figure 4. Translocation within *MAPK24* gene abolishes protein in Patient 1. A) Schematic of the primary transcript of *MAP2K4*. The exons of the transcript are shown as green boxes and the red triangle above *MAP2K4* depicts the approximate location of the translocation. The translocation results in a truncation of *MAP2K4* (shown in the bottom portion of this figure), leaving only the first three exons remaining on chromosome 17, while the blue line in the lower portion of the figure represents translocated DNA from chromosome 22. B) Western blot of MAP2K4 in control HEK293T cells and ascites-derived cells from Patient 1. The control cells show clear expression of *MAP2K4*, while the cells taken from Patient 1 show no production of the protein. Proliferating Cell Nuclear Antigen (PCNA) blotting was performed as a loading control.

Mobile elements can be divided into subfamilies, classified by diagnostic mutations. As different subfamilies may have different expression patterns in cancer, we examined these subfamilies in our patients. Three of the patients in this study had longitudinal bulk RNA-seq data available for further analysis (Figure 1). There was no germline RNA-seq data available for these patients, but using SQuIRE we were able to analyze mobile element expression on both subfamily (examining all mobile element transcripts from one mobile element subfamily) and locus-based (specific mobile element loci) levels throughout tumor progression. In Patient 1 and Patient 3, we saw no subfamily-level changes across timepoints. However, in Patient 2, while we did not see a subfamily level change for *Alu,* L1, or SVA, we saw significant changes in LTR and HERV subfamilies (Wald test, adjusted for multiple testing with Benjamini and Hochberg (Love et al 2014); adjusted p-value < 0.05). A complete table of subfamilies that reached significance in Patient 2 can be found in Table S1. At the single-locus level, by comparing the four earliest timepoints of Patient 2 to the latter two timepoints (1016 days; nearly 3 years between the fourth timepoint and the fifth), we see significant expression changes (adjusted p-value < 0.05), both increases and decreases, for 337 loci. 125 of these loci are *Alu* elements, 123 are L1s, and 16 are SVA insertions. The remaining loci largely belong to older mobile elements and LTR families. The complete list of mobile element loci that show significantly different expression over time are shown fully in Table S2.

The 264 L1s, *Alu* elements, and SVAs that were found to be differentially expressed between the first set of timepoints and second set of timepoints in Patient 2 were intersected with a list of regulatory regions in mammary epithelial tissue. Of these 264 mobile elements, 72 overlapped with a total of 87 regulatory regions. These 87 regulatory regions include a number of CCCTC-binding factor (CTCF) binding sites, open chromatin regions, promoter flanking

regions, as well as several enhancers, promoters, and transcription factor binding sites. These overlaps are shown in Supplemental Table S3.

Discussion

Using a trio of mobile element and *de novo* variant detection tools, we identified mobile element insertions and variants in longitudinal WGS data from four breast cancer patients (Figure 1). We also find that the visual validation step (IGV or other similar tools) used here reduces the risk of false positive calls and significantly improves the overall accuracy of variant calls (similar to Feusier et al. 2019). This is particularly important for cancer genomes, because most mobile element detection software is not designed for the complexity of a cancer genome. This increased complexity makes the visual inspection step necessary, as there were thousands of potential mobile element insertions suggested by these programs that were not unique to the somatic cells, or were the result of mis-mapping in low complexity or mobile element-rich regions. However, the calls from these programs that passed visual inspection spanned both high variant allele frequency and very low variant allele frequency (Figure 2), showing that the insertions or variants did not have to be present at a particularly high frequency to be detected.

The VAFs for each event change at each timepoint, reflecting the change in subclone frequency shown previously (Brady et al. 2017). Large changes in VAF shown in the current study generally correlate with bottleneck events or large shifts in subclone frequency found previously (Brady et al. 2017). The insertions that share similar VAF may be from the same tumor subclone (see the *Alu* insertion and L1 insertion on chromosome 8 in Patient 2) (Figure 2B), and, after adjusting for tumor purity, those that still show a sharp decline in VAF are likely part of a tumor subclone that showed decreased frequency. The insertions and variants that are present at very high frequency (near 100%) are likely mutations that occurred early in cancer

development. In cancers with a greater number of these events, mobile elements may be valuable markers for identifying tumor subclones, similar to their role in population genetics and evolutionary studies. Using these markers in conjunction with SNPs may increase the level of resolution possible for tumor subclones.

From the summary of the insertions and mobile element mediated SVs identified in four patients (Table 1), we show activity for not only L1s, which we expect based on previous literature (Lee et al. 2012; Tubio et al. 2014; Doucet-O'Hare et al. 2016), but also activity from *Alu* and SVA elements. Most previous studies have largely focused on L1s and the structural variants that they mediate (Helman et al. 2014; Tubio et al. 2014), though others have identified a small number of *Alu* insertions (Lee et al. 2012). SVA has not been a target for most previous studies, but recent work also supports post-zygotic insertions in normal tissue for this mobile element family (Feusier et al. 2019). We also see variation among patients, with most individuals showing multiple insertions or variants but one showing only a single event that appears to be *Alu-Alu* mediated.

The majority of the variants identified in this study appear to occur early in our samples (Table 1, Figures 2 and 3) and likely early in tumor progression or development. As the bulk of these events occur early, it would suggest that mobile elements are likely more active early in cancer, and could play a role in tumorigenesis. This supports previous work done in metastatic tumors from multiple patients showing that most of insertions that occurred in the primary tumor were reflected in metastases (Ewing et al. 2015). In addition to the early insertions and SVs we identified, we find two insertions, both SVAs, that insert later in tumor progression. It is unclear if there is something unique about SVA insertions that would cause them to be more active later in cancer, or if this is just a result of our small sample size. Future studies with a larger sample

size of patients should attempt to identify these insertions to determine if the trend of insertions at later timepoints remains.

The early SV events identified in Patient 1 both appear to be *Alu*-mediated and to affect the coding sequence. The affected genes (*MAP2K4* and *YTHDF2*), have both been implicated in cancer development (Su et al. 2002; Ahn et al. 2011; Koboldt et al. 2012; Xue et al. 2018; Zhong et al. 2019). The deletion we uncovered in *YTHDF2* removes the final exon that would be present in the primary transcript. Though there have been reports of this gene acting as a tumor suppressor, it is not listed in the Catalogue of Somatic Mutations in Cancer (COSMIC) (Tate et al. 2018), and further work likely needs to be completed to validate its role in cancer. The translocation that we identified in Patient 1 interrupted *MAP2K4*, a gene with far more supporting evidence that suggests it plays a role in catalyzing tumor development or metastasis (Ahn et al. 2011; Pavese et al. 2014). The breakpoint on chromosome 17 for this translocation occurs after the third exon of *MAP2K4* (Figure 4A), with a number of mobile elements present in this region. The breakpoint on chromosome 22 disrupts an *Alu* element, and it is likely that this translocation was mediated by an *Alu*-*Alu* interaction, or by the interaction of other repetitive elements in the region surrounding the breakpoints. There have also been previous examples of *Alu*-mediated translocations and recombinations in cancer (Onno et al. 1992; reviewed in Elliott et al. 2005). *MAP2K4* is listed in COSMIC as potentially having a role in both tumor suppression and as an oncogene, depending on its expression level.

We find decreased copy number along multiple regions of chromosome 17, including the region that contains *MAP2K4* (Figure 3). With only a single copy of this gene remaining, the translocation renders it inactive, leading to the lack of production of the protein that we see from the Western blot (Figure 4B). Further analysis of the RNA-seq data for this patient showed that

there were no complete transcripts for *MAP2K4,* with no evidence for fusion transcripts being produced as a result of the translocation. *MAP2K4* is found in the *JNK* pathway, which is responsible for numerous cellular functions (Johnson and Lapadat 2002). Previous work in breast cancer tissue has identified multiple mutations along this pathway (including *MAP2K4* and *MAP2K7*) (Koboldt et al. 2012) and suggests that this mutation could act as a driver by altering function of the *JNK* pathway. Brady *et al.* first identified this structural variant in *MAP2K4* in this patient, but by analyzing mobile elements in these patients we have been able to better classify the cause of the mutation.

Our analysis of RNA-seq data in these patients showed very few subfamily-level expression changes for mobile elements through time. While we did not have control RNA-seq data from the germline, we were able to examine the course of expression change through tumor progression. Of the three patients with RNA-seq data, only one patient showed a subfamily-level change in expression. This was linked to an increase in HERV activity, which previous studies have reported for HERV-K in various cancers (Wallace et al. 2014; Ma et al. 2016). We were also able to examine the expression of specific mobile element loci in which there were informative reads. Here, using the RNA-seq data that we had for Patient 2, which spanned multiple years, we showed that many loci did change expression levels over time. The majority of these elements were in non-coding regions, though some were found in regulatory regions. Previous work has shown that changes in mobile element methylation and expression can impact nearby gene expression (Xie et al. 2013; Jang et al. 2019). This shows that mobile elements, even at the locus level, can change expression throughout cancer progression, potentially altering the expression of one or several genes, or changing which mobile elements are capable of retrotransposition. These expression changes are likely tied to methylation changes due to the

high CpG content of mobile elements (Yoder et al. 1997). We were only able to demonstrate expression changes in Patient 2, the patient that had the most RNA-seq timepoints over the longest course of time. Future studies should examine locus-level expression over time, with germline control comparisons to determine how these expression levels compare with normal cells over longer sampling timepoints.

To our knowledge, our study is the first to examine mobile element activity in longitudinal cancer samples. We find that most mobile element insertions and the structural variation events they mediate appear to occur early in tumor development, and most of these early events appear to be passenger mutations. In addition to these passenger mutations, we find mobile element mediated SV events that disrupt the coding sequence of known (*MAP2K4)* and suspected (*YTHDF2)* driver genes in breast cancer. Though our sample size is small, we identified a number of L1, *Alu,* and SVA insertions occurring during tumor progression. As most studies attempt to identify only L1 insertions and L1-mediated structural variants, we may be underestimating the impact that mobile elements have on mediating driver mutations in cancer. Future studies should examine other types of cancer for the patterns and impact of mobile element insertions that we found to determine which cancers have an increase in *Alu* and SVA activity, as others have done with L1 insertions. Improving our understanding of mobile element insertions and structural variation events in cancer could enhance our ability to identify tumor subclones and our understanding of the mutational landscape in cancer.

Methods

Sequencing data

Information regarding the acquisition of patient sequencing data, as well as quality control and alignment information, can be found in Brady *et al.* 2017. Blood-derived DNA was sequenced and used as the germline DNA sample. Tumor samples were sequenced from ascites and pleural fluid surrounding the breast tumor. DNA sequencing data had previous been aligned to hg19. Data from the Brady *et al.* 2017 publication are available with controlled access on the European Genome-phenome Archive (EGA) under accession EGAS00001002436.

Mobile element insertions and variants identification

We used the Mobile Element Locator Tool (MELT) (Version 2.1.5) (Gardner et al. 2017), RUFUS (https://github.com/jandrewrfarrell/RUFUS), where possible, and TranSurVeyor (Rajaby and Sung 2018) for the identification of mobile elements and related structural variants in our longitudinal breast cancer samples. Each of these three tools incorporates different methods for identifying variants. Insertions and variants identified by these programs were filtered to include only those that appeared absent in the germline sequencing data. To ensure that the potential mobile element insertions or mobile element mediated structural variants were, in fact, mobile element sequences, we used both BLAT (Kent 2002) and RepeatMasker (Smit 2013-2015). Visual inspection was performed to ensure that the sequences matched a known mobile element sequence using the Integrated Genome Viewer (IGV, Version 2.4.13) (Robinson et al. 2011). This step helped to prevent the inclusion of reads that had simply mis-mapped to an incorrect region of the genome. Reads that appeared to have discordant and/or split reads that mapped to a mobile element were included for further validation. Where possible, we used classic hallmarks of retrotransposition events (target site duplications and poly A tails) to ensure that we had identified an insertion. We validated candidate structural variation events that appeared to be mediated by mobile elements with Lumpy (Version 0.2.13) (Layer et al. 2014),

and we used IGV validation to identify signs of mobile element involvement. BEDTools (Quinlan and Hall 2010) intersect was used to determine if the identified variants overlapped with ultraconserved non-coding regions or regulatory elements. Ultraconserved elements and regulatory blocks locations were obtained from UCNEbase (https://ccg.epfl.ch/UCNEbase/) (Dimitrieva and Bucher 2013). The UCSC Genome Browser was manually checked to ensure no other regulatory elements overlapped the insertion site.

Variant allele frequency

Variant allele frequency was calculated by counting the number of reads showing evidence of the insertion or variant in IGV and dividing this by the total number of reads at the breakpoint of the insertion or variant. The variant allele frequency was then adjusted by multiplying the total number of reads at the breakpoint by the tumor purity (or cancer cell fraction) value. These values were calculated using FACETS (Shen and Seshan 2016), which utilizes SNPs in germline and tumor samples, as well as copy number changes, to provide an estimate of the proportion of tumor cells in the sample. FACETS was also used to determine the location of copy number changes throughout the genome, and particularly at the loci at which we identified variants.

RNA-seq analysis and mobile element expression

Mobile element expression was measured on a subfamily-specific level and a locus-specific level using SQuIRE (Version 0.9.9.9a-beta) (Yang et al. 2019). We compared the first timepoint for each patient to later timepoints to understand how expression patterns changed throughout cancer progression for both subfamily-level and locus-level expression. Additionally, where possible, we compared multiple early sampling timepoints with later sampling timepoints

to determine if there was a change during cancer progression. SQuIRE was run according to the documentation provided. Briefly, RNA-seq data were aligned to hg38 using STAR (Version 2.5.3 a) (Dobin et al. 2013); counts of gene expression and mobile element expression were then generated to quantify expression. DESeq2 (Version 1.16.1) (Love et al. 2014) was then used to generate calls for differentially expressed subfamilies and loci. For the locus-specific expression changes at mobile element loci, we intersected these loci with mammary epithelial tissue regulatory regions from Ensembl (Aken et al. 2016) using BEDTools.

Trinity (Version 2.8.5) (Grabherr et al. 2011) was run on bulk RNA-seq data from the first timepoint of Patient 1 to create a library of *de novo* transcripts. The resulting transcripts were aligned to the transcript of *MAP2K4* using BLAT. Transcripts that aligned to *MAP2K4* were further examined to determine which portions of the *MAP2K4* transcript were being produced.

DNA extraction and PCR validation

For validation of our detection methods, PCR amplification was run on a potential L1 insertion in Patient 2, for whom we had blood-derived germline DNA and were able to extract DNA from cancer cells (ascites). DNA extraction was performed using Qiagen DNeasy Blood and Tissue Kit (50) (Cat No./ID: 69504). PCR amplifications of 25 ng of germline DNA and 25 ng of tumor DNA were performed in 25-µL reactions using Phusion Hot Start Flex DNA polymerase. Initial denaturation was performed for 30 seconds at 98°C, with 40 cycles of denaturation for 10 seconds at 98°C, the optimal annealing temperature of 60°C for 30 seconds, followed by a 2-minute extension at 72°C, and a final extension for 5 minutes at 72°C. The reaction was performed with a negative control (water), the tumor DNA, and the matched germline DNA. Amplicons were run on a 2% agarose gel with ethidium bromide for

approximately 90 minutes at 100 V. The gel was imaged using a Fotodyne Analyst Investigator Eclipse imager. The primer set used for this reaction is listed in Supplemental Table S4, and the corresponding gel image is shown in Supplemental Figure S2.

Immunoprecipitation and Immunoblotting

Patient ascites samples were grown in Human Breast Epithelial Cell Culture Complete Media (Celprogen M36056-01S), and HEK293T cells were grown in DMEM with 1% pen/strep and 10% heat inactivated fetal bovine serum. Magnetic beads were crosslinked to antibody (CST # 13110 PCNA 1:50 dilution, CST #9152 MKK4, 1:50 dilution) using Pierce crosslink magnetic IP kit (ThermoFisher #88805). Cells were lysed with ThermoFisher Co-IP lysis buffer. Crosslinked beads were added to 1000µg of protein lysate and incubated overnight at 4°C. Immunoprecipitation was performed according to Pierce™ Classic Magnetic IP/Co-IP Kit protocol (ThermoFisher #88804). Immunoprecipitated protein was removed from the antibody/bead complex using low-pH elution. Twenty-eight ug of immunoprecipitated protein was loaded per well of NuPAGE 10% Bis-Tris protein gels. After transfer to PVDF membrane, membranes were blocked with 5% milk:PBST, then incubated with primary antibody (PCNA 1:1000, MKK4 1:500) overnight at 4°C. Blots were washed with PBST and incubated with goat anti-rabbit conjugated to HRP secondary antibody (1:1000) for 45 minutes at room temperature. Blots were then incubated in chemiluminescent substrate enhancer for 5 minutes at room temperature (SuperSignal West Pico PLUS (ThermoFisher # 34577)). Bands were visualized using x-ray film.

Acknowledgments

This work was supported by the NIH R35GM118335 (to LBJ) and the NIH/NRSA T32HG008962 (to CJS) from the NHGRI. We would like to thank the Biorepository and Molecular Pathology resource at the Huntsman Cancer Institute for their assistance in accessing samples from these patients. We would also like to thank members of both the Jorde and Marth labs for helpful feedback and discussion.

References

Ahn Y-H, Yang Y, Gibbons DL, Creighton CJ, Yang F, Wistuba II, Lin W, Thilaganathan N, Alvarez CA, Roybal J et al. 2011. Map2k4 functions as a tumor suppressor in lung adenocarcinoma and inhibits tumor cell invasion by decreasing peroxisome proliferator-activated receptor γ2 expression. *Mol Cell Biol* **31**: 4270-4285.

Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, Billis K, Carvalho-Silva D, Cummins C, Clapham P et al. 2016. Ensembl 2017. *Nucleic Acids Research* **45**: D635-D642.

Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, Badge RM, Moran JV. 2010. LINE-1 retrotransposition activity in human genomes. *Cell* **141**: 1159-1170.

Boeke JD, Garfinkel DJ, Styles CA, Fink GR. 1985. Ty elements transpose through an RNA intermediate. *Cell* **40**: 491-500.

Brady SW, McQuerry JA, Qiao Y, Piccolo SR, Shrestha G, Jenkins DF, Layer RM, Pedersen BS, Miller RH, Esch A et al. 2017. Combating subclonal evolution of resistant cancer phenotypes. *Nature Communications* **8**: 1231.

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH, Jr. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A* **100**: 5280-5285.

Burns KH. 2017. Transposable elements in cancer. *Nat Rev Cancer* **17**: 415-424.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**: 691-703.

Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV, Gage FH. 2009. L1 retrotransposition in human neural progenitor cells. *Nature* **460**: 1127-1131.

de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLOS Genetics* **7**: e1002384.

Dimitrieva S, Bucher P. 2013. UCNEbase--a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic acids research* **41**: D101-D109.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Doucet-O'Hare TT, Sharma R, Rodić N, Anders RA, Burns KH, Kazazian Jr HH. 2016. Somatically Acquired LINE-1 Insertions in Normal Esophagus Undergo Clonal Expansion in Esophageal Squamous Cell Carcinoma. **37**: 942-954.

Doucet O, Hare TT, Rodić N, Sharma R, Darbari I, Abril G, Choi JA, Young Ahn J, Cheng Y, Anders RA et al. 2015. LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. *Proceedings of the National Academy of Sciences* **112**: E4894.

Elliott B, Richardson C, Jasin M. 2005. Chromosomal Translocation Mechanisms at Intronic Alu Elements in Mammalian Cells. *Molecular Cell* **17**: 885-894.

Erwin JA, Marchetto MC, Gage FH. 2014. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat Rev Neurosci* **15**: 497-506.

Ewing AD, Gacita A, Wood LD, Ma F, Xing D, Kim M-S, Manda SS, Abril G, Pereira G, Makohon-Moore A et al. 2015. Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. *Genome research* **25**: 1536-1545.

Feusier J, Watkins WS, Thomas J, Farrell A, Witherspoon DJ, Baird L, Ha H, Xing J, Jorde LB. 2019. Pedigree-based estimation of human mobile element retrotransposition rates. *Genome research* **29**: 1567-1577.

Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, Mills RE, Devine SE. 2017. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res* **27**: 1916-1929.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**: 644-652.

Hancks DC, Kazazian HH, Jr. 2016. Roles for retrotransposon insertions in human disease. *Mob DNA* **7**: 9-9.

Helman E, Lawrence MS, Stewart C, Sougnez C, Getz G, Meyerson M. 2014. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res* **24**: 1053-1063.

Hsieh S-Y, Chen W-Y, Yeh T-S, Sheen IS, Huang S-F. 2005. High-frequency Alu-mediated genomic recombination/deletion within the caspase-activated DNase gene in human hepatoma. *Oncogene* **24**: 6584-6589.

Jang HS, Shah NM, Du AY, Dailey ZZ, Pehrsson EC, Godoy PM, Zhang D, Li D, Xing X, Kim S et al. 2019. Transposable elements drive widespread expression of oncogenes in human cancers. *Nature Genetics* **51**: 611-617.

Johnson GL, Lapadat R. 2002. Mitogen-Activated Protein Kinase Pathways Mediated by ERK, JNK, and p38 Protein Kinases. *Science (New York, NY)* **298**: 1911.

Kazazian HH, Jr., Moran JV. 2017. Mobile DNA in Health and Disease. *N Engl J Med* **377**: 361-370.

Kazazian HH, Jr., Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**: 164-166.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656-664.

Koboldt DC Fulton RS McLellan MD Schmidt H Kalicki-Veizer J McMichael JF Fulton LL Dooling DJ Ding L Mardis ER et al. 2012. Comprehensive molecular portraits of human breast tumours. *Nature* **490**: 61-70.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology* **15**: R84.

Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, 3rd, Lohr JG, Harris CC, Ding L, Wilson RK et al. 2012. Landscape of somatic retrotransposition in human cancers. *Science (New York, NY)* **337**: 967-971.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**: 550.

Ma W, Hong Z, Liu H, Chen X, Ding L, Liu Z, Zhou F, Yuan Y. 2016. Human Endogenous Retroviruses-K (HML-2) Expression Is Correlated with Prognosis and Progress of Hepatocellular Carcinoma. *Biomed Res Int* **2016**: 8201642-8201642.

Mauillon JL, Michel P, Limacher J-M, Latouche J-B, Dechelotte P, Charbonnier F, Martin C, Moreau V, Metayer J, Paillot B et al. 1996. Identification of Novel Germline hMLH1

Mutations Including a 22 kb Alu-mediated Deletion in Patients with Familial Colorectal Cancer. *Cancer Research* **56**: 5728.

Muotri AR, Chu VT, Marchetto MCN, Deng W, Moran JV, Gage FH. 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**: 903-910.

Onno M, Nakamura T, Hillova J, Hill M. 1992. Rearrangement of the human tre oncogene by homologous recombination between Alu repeats of nucleotide sequences from two different chromosomes. *Oncogene* **7**: 2519-2523.

Pavese JM, Ogden IM, Voll EA, Huang X, Xu L, Jovanovic B, Bergan RC. 2014. Mitogen-activated protein kinase kinase 4 (MAP2K4) promotes human prostate cancer metastasis. *PLoS One* **9**: e102289-e102289.

Peixoto A, Pinheiro M, Massena L, Santos C, Pinto P, Rocha P, Pinto C, Teixeira MR. 2013. Genomic characterization of two large Alu-mediated rearrangements of the BRCA1 gene. *Journal of Human Genetics* **58**: 78-83.

Petrij-Bosch A, Peelen T, van Vliet M, van Eijk R, Olmer R, Drusedau M, Hogervorst FB, Hageman S, Arts PJ, Ligtenberg MJ et al. 1997. BRCA1 genomic deletions are major founder mutations in Dutch breast cancer patients. *Nature genetics* **17**: 341-345.

Puget N, Torchard D, Serova-Sinilnikova OM, Lynch HT, Feunteun J, Lenoir GM, Mazoyer S. 1997. A 1-kb Alu-mediated Germ-Line Deletion Removing BRCA1 Exon 17. *Cancer Research* **57**: 828.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Rajaby R, Sung W-K. 2018. TranSurVeyor: an improved database-free algorithm for finding non-reference transpositions in high-throughput sequencing data. *Nucleic Acids Research* **46**: e122-e122.

Richardson SR, Morell S, Faulkner GJ. 2014. L1 retrotransposons and somatic mosaicism in the brain. *Annu Rev Genet* **48**: 1-27.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nature Biotechnology* **29**: 24-26.

Rohlfs EM, Puget N, Graham ML, Weber BL, Garber JE, Skrzynia C, Halperin JL, Lenoir GM, Silverman LM, Mazoyer S. 2000. An Alu-mediated 7.1 kb deletion of BRCA1 exons 8 and 9 in breast and ovarian cancer families that results in alternative splicing of exon 10. *Genes, chromosomes & cancer* **28**: 300-307.

Shen R, Seshan VE. 2016. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic acids research* **44**: e131-e131.

Smit A, Hubley, R & Green, P. 2013-2015. RepeatMasker Open-4.0.

Su GH, Song JJ, Repasky EA, Schutte M, Kern SE. 2002. Mutation rate of MAP2K4/MKK4 in breast carcinoma. *Human mutation* **19**: 81.

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E et al. 2018. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* **47**: D941-D947.

Tubio JM, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, Gundem G, Pipinikas CP, Zamora J, Raine K. 2014a. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**: 1251343.

Wallace TA, Downey RF, Seufert CJ, Schetter A, Dorsey TH, Johnson CA, Goldman R, Loffredo CA, Yan P, Sullivan FJ et al. 2014. Elevated HERV-K mRNA expression in PBMC is associated with a prostate cancer diagnosis particularly in older men and smokers. *Carcinogenesis* **35**: 2074-2083.

Welcsh PL, King MC. 2001. BRCA1 and BRCA2 and the genetics of breast and ovarian cancer. *Hum Mol Genet* **10**: 705-713.

Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL et al. 2013. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nature Genetics* **45**: 836-841.

Xing J, Witherspoon DJ, Jorde LB. 2013. Mobile element biology: new possibilities with high-throughput sequencing. *Trends Genet* **29**: 280-289.

Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA et al. 2009. Mobile elements create structural variation: analysis of a complete human genome. *Genome research* **19**: 1516-1526.

Xue Z, Vis DJ, Bruna A, Sustic T, van Wageningen S, Batra AS, Rueda OM, Bosdriesz E, Caldas C, Wessels LFA et al. 2018. MAP3K1 and MAP2K4 mutations are associated with sensitivity to MEK inhibitors in multiple cancer models. *Cell Research* **28**: 719-729.

Yang WR, Ardeljan D, Pacyna CN, Payer LM, Burns KH. 2019. SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Research* **47**: e27-e27.

Yoder JA, Walsh CP, Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends in genetics : TIG* **13**: 335-340.

Zhong L, Liao D, Zhang M, Zeng C, Li X, Zhang R, Ma H, Kang T. 2019. YTHDF2 suppresses cell proliferation and growth via destabilizing the EGFR mRNA in hepatocellular carcinoma. *Cancer Letters* **442**: 252-261.