scientific reports

OPEN

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], Sean V. Tavtigian[3], Gabor Marth[1,2] & Lynn B. Jorde[1,2]

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 associated structural variants in cancer, we examined their activity in longitudinal samples of four metastatic breast cancer patients. We identified 11 mobile element insertions or associated structural variants and found that the majority of these occurred early in tumor progression. Most of the variants impact intergenic regions; however, we identified a translocation interrupting *MAP2K4* involving *Alu* elements and a deletion in *YTHDF2* involving mobile elements that likely inactivate reported tumor suppressor genes. The high variant allele fraction of the translocation, the loss of the other copy of *MAP2K4*, the recurrent loss-of-function mutations found in this gene in other cancers, and the important function of *MAP2K4* indicate that this translocation is potentially 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.

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 to 66%[1,2]. In humans, retrotransposons represent a class of active mobile elements, inserting new elements through a "copy and paste" mechanism[3]. While most of these elements have become transcriptionally inactive over time[4,5], 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[2,6]. 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[7]. The majority of these insertions seem to have no adverse impact; however, some insertions have been found to cause disease[8]. 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[9]. Since this initial discovery, mobile element insertions have been found to be associated with more than 130 disease cases[10].

Somatic mobile element insertions have not been identified in many healthy tissues, though detection of these low-frequency events is difficult without sufficient coverage. An exception to this is in neurons, where somatic mosaicism of L1 insertions have been identified[11–14]. This increase in activity may be due to slight changes in methylation at L1 loci[11]. Recently, multiple studies have noted varying degrees of activity of mobile elements in numerous cancer tissues[15–22] with some analyses including metastatic samples[20,23,24]. 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[17,20]. Regardless of cancer type, roughly half of all tumors have at least one somatic L1 insertion. Breast cancer patients with L1 activity were found to often have a single L1 insertion, with a small number of patients showing up to five insertions[20].

[1]Department of Human Genetics, University of Utah School of Medicine, 15 N. 2030 E. Rm 5100, Salt Lake City, UT 84112, USA. [2]Utah Center for Genetic Discovery, Salt Lake City, UT 84112, USA. [3]Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT 84112, USA. ✉email: cody.steely@utah.edu

Patient	Locus	Variant	Region	Genes affected	Present before first tumor timepoint
Patient 1	1:186,755,400	*Alu* (non-classical insertion)	Intergenic		Yes
	1:29,089,030	*Alu-Alu* associated deletion (23 kb)	Exon	*YTHDF2*	Yes
	2:62,111,770	SVA insertion	Intronic	*CCT4*	No
	17:11,974,341/22:48,343,831	*Alu*-associated translocation	Exon	*MAP2K4*	Yes
Patient 2	2:15,168,000	*Alu* (non-classical insertion)	Intergenic		Yes
	8:99,227,400	L1 insertion	Exon	*NIPAL2*	Yes
	13:19,866,150	L1 insertion	Intronic	*ANKRPD26P3*	Yes
Patient 3	6:105,643,750	*Alu-Alu* associated deletion	Intergenic		Yes
Patient 4	10:67,431,020	*Alu* insertion	Intergenic		Yes
	11:81,149,160	L1-associated deletion	Intergenic		Yes
	19:23,124,500	SVA insertion	Intergenic		No

Table 1. Mobile element insertions and structural variants involving mobile elements identified in four longitudinal breast cancer patients. Insertion or variant sites found within genes (exonic or intronic) are indicated in the "Genes affected" column. Insertion events described as "non-classical insertions" do not show any hallmarks of L1-mediated insertion events.

The impact of mobile elements on the cancer genome is not limited to somatic insertions but also includes the structural variation (SV) events associated with existing mobile elements (reviewed in[25]). Notably, mobile elements mediate roughly 10% of all SV events larger than 100 base pairs in the human genome[26]. Mobile elements have been found to mediate SV events that can lead to cancer development[27,28]. For example, approximately 42% of the *BRCA1* gene is composed of *Alu* elements[29], making this gene a target for non-allelic homologous recombination, specifically *Alu-Alu* associated events[30–33]. Additional mobile element events have been classified in breast cancer as well[34,35]. L1 transduction events have also been examined in a variety of cancers[20].

Multiple studies have shown that L1s, in particular, demonstrate increased retrotransposition activity in cancer tissues. However, the timing of these insertions or mobile element associated SVs has not been thoroughly investigated. Additionally, L1s and the structural variants that they mediate have been the focus of most previous studies, though *Alu* and SVA insertions or the SVs they mediate have been targeted in some studies[19,20,22]. 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 2–15 years with varying WGS and bulk RNA-seq timepoints (2–6) available for our analyses. Many of the tumor samples were collected from ascites and the surrounding pleural fluid. Tumor cells in ascites are descended from the primary tumor, though they may contain additional mutations and are likely to be polyclonal. Using ascites to analyze tumors may lead to improved characterization of the tumor and of the clonal changes that occur[36,37]. 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 Supplemental Figure 1. More detailed treatment information for each patient can be found in the supplementary information of Brady et al. 2017[38]. 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 − .

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" section), 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 identified insertion 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 Variant Allele Fraction (VAF) (100%) and very low VAF (~ 5%). While the three tools used in this study generally identified the same insertion events, MELT and RUFUS excelled at identifying *Alu* elements, and TranSurVeyor found an additional L1 insertion that was not identified by the other tools.

Collectively, we identified seven mobile element insertions (either classical or non-classical) and four structural variants involving mobile elements (Table 1). Supplemental Table 1 shows which tools detected each variant. 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. IGV images with schematics are shown for three of the events that we identified in Fig. 1, with schematics of the remaining events included as Supplemental Figures 2, 3, and 4. 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. Both SVA insertions identified were present at low frequency in these patients (IGV images shown in Supplemental Figure 5). We also find high variability in the number of insertions and variants

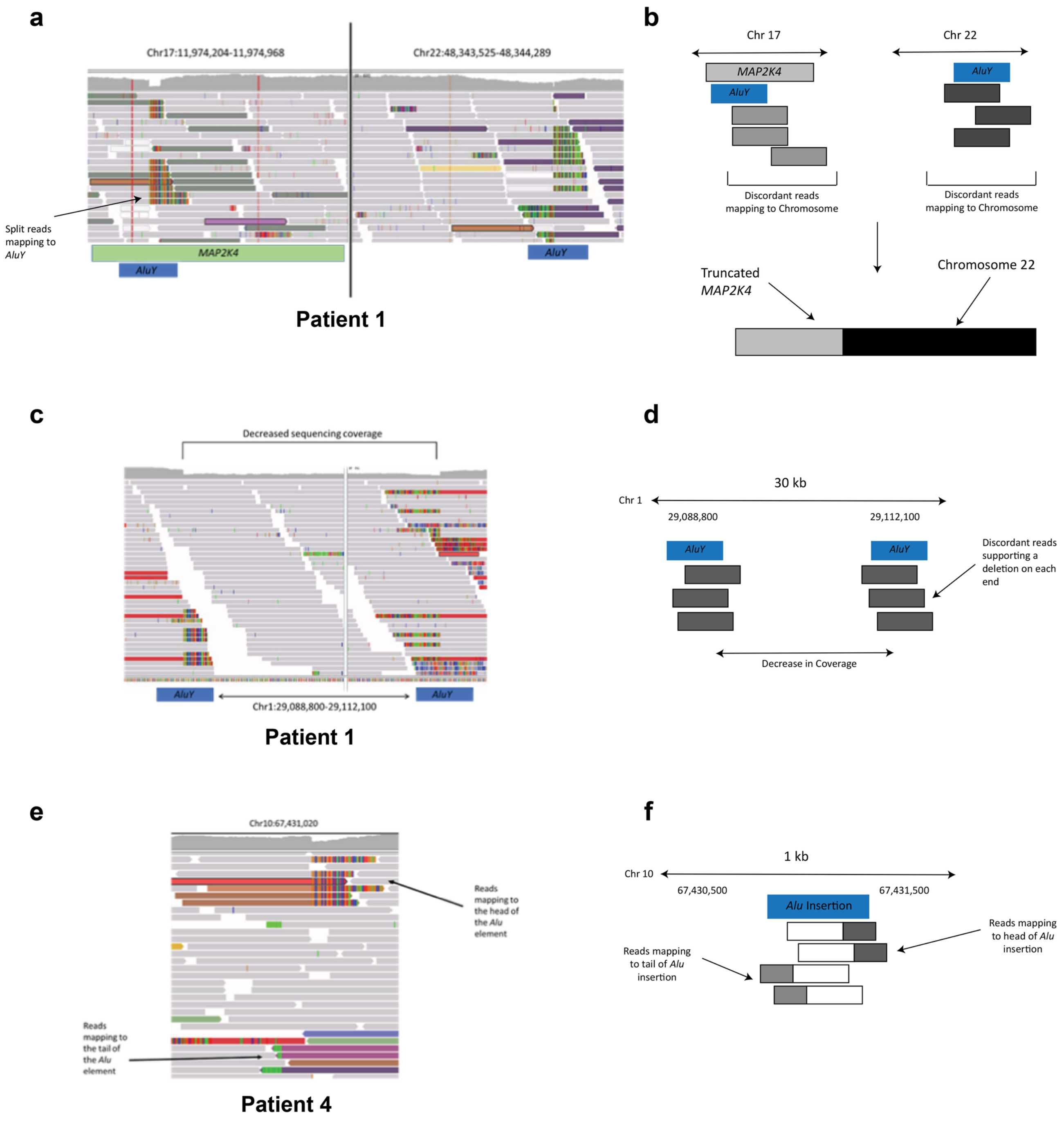
a
Chr17:11,974,204-11,974,968
Chr22:48,343,525-48,344,289
Split reads mapping to AluY
MAP2K4
AluY
AluY
Patient 1
b
Chr 17
Chr 22
MAP2K4
AluY
AluY
Discordant reads mapping to Chromosome
Discordant reads mapping to Chromosome
Truncated MAP2K4
Chromosome 22
c
Decreased sequencing coverage
AluY
Chr1:29,088,800-29,112,100
AluY
Patient 1
d
30 kb
Chr 1
29,088,800
29,112,100
AluY
AluY
Discordant reads supporting a deletion on each end
Decrease in Coverage
e
Chr10:67,431,020
Reads mapping to the head of the Alu element
Reads mapping to the tail of the Alu element
Patient 4
f
1 kb
Chr 10
67,430,500
67,431,500
Alu Insertion
Reads mapping to head of Alu insertion
Reads mapping to tail of Alu insertion

Figure 1. IGV images and schematics of some of the identified insertions and structural variants. (**A**) Image of the translocation between Chromosomes 17 and 22 in Patient 1 that interrupts *MAP2K4*. Discordant reads on each chromosome map to the other end of the translocation. The split reads on Chromosome 17 map to a non-reference *Alu* element. The *Alu* element suspected to be involved in the translocation on Chromosome 17 (from the split reads) is shown as a blue box, and the reference *Alu* involved with the translocation on Chromosome 22 is also shown as a blue box. (**B**) A schematic showing the translocation event (not to scale). The *Alu* elements associated with the translocation are shown as blue boxes. The discordant reads are shown as gray or purple boxes. A schematic of the translocation is shown below the arrow. (**C**) Deletion involving *Alu* elements in Patient 1. *Alu* elements involved in the deletion are shown as blue boxes below the image. (**D**) A schematic of the deletion involving two *Alu* elements. Each *Alu* element is shown in blue, with the discordant reads shown in red. (**E**) Somatic *Alu* element insertion in Patient 4. The reads that map to the head and tail of the element are labeled by arrows. (**F**) A schematic of the *Alu* insertion on Chromosome 10. The approximate location of the insertion (blue box) is shown with the split and discordant reads mapping to the head of the *Alu* shown in white and red boxes, while the split and discordant reads mapping to the tail of the *Alu* are shown in white and green boxes.

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 structural variant associated with mobile elements.

The use of multiple tools and visual inspection of the identified insertions excluded a large number of false positive calls. In total, MELT identified 8825 potential mobile element variants, and TranSurVeyor identified 69,394 potential mobile element variants. For MELT, we analyzed every polymorphic variant (regardless of genotype) to determine if it was unique to the tumor sample. For TranSurVeyor, we did not use the built-in filtering mechanism to ensure that low frequency variants would be included for further examination. Following this filtering, we retained only those variants that passed visual inspection (see "Methods" section). These 11 variants are shown in Table 1. 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, an L1 insertion in *NIPAL2* in Patient 2*,* and an L1 insertion in *ANKRPD26P*, which we identified in Patient 2. We also identified two *Alu*-associated somatic structural variation events that impact exons in two genes: an *Alu*-*Alu* associated deletion in *YTHDF2* (with ~ 90% homology between the two *Alu* elements, excluding the poly(A) tail) and a translocation involving *Alu* elements that impacted *MAP2K4*. Both of these *Alu*-associated structural variants occurred in Patient 1 and occurred before the first sampling timepoint but were not present in the germline sequencing data.

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 Supplemental 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.

We calculated the VAF for each of the mobile element insertions or SVs identified after adjusting for tumor purity (Fig. 2). The VAFs calculated from the adjusted data are slightly higher, but very similar to the VAFs calculated from the unadjusted VAFs (Supplemental Figure 6). Of the four insertions and SVs identified in Patient 1 (Fig. 2A), the deletion in *YTHDF2* involving *Alu* elements, 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 was 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 (Fig. 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 (Fig. 2C) had the fewest variants, a single deletion involving *Alu* elements (~ 98% homology between the short sequences involved). This deletion was present at a high frequency in the first sampling timepoint (about 75% of reads), but decreased to approximately 50% by the second sampling timepoint. In Patient 4 (Fig. 2D), we identified two insertions, an *Alu* and SVA, and a single mobile element associated SV, a L1–L1 associated deletion. 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 L1–L1 associated deletion trends upward in VAF before sharply falling at the final timepoint. The *Alu* insertion in this patient maintains a steady VAF of ~ 20%.

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 in these regions. Specific genome-wide copy number estimates for the first timepoint in the genome of Patient 1 are shown in Fig. 3. Both the deletion involving *Alu* elements and the *Alu* associated translocation appear to have been reduced to only a single copy, contributing to the high frequency for both.

Due to decreased copy number at the *MAP2K4* locus, and the translocation interrupting the one remaining copy of *MAP2K4* (shown in Fig. 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" section), we find only truncated and hybrid transcripts of *MAP2K4* (Fig. 4A; Sequences for these transcripts are shown in Supplemental Table 3. To confirm that there was evidence of this translocation in the RNA-seq reads, we manually reviewed reads in the region surrounding the translocation. Multiple RNA-seq reads mapped to either Chromosome 17 or 22, and had a mate that mapped to the other chromosome, with some split reads surrounding the breakpoint. To further validate a lack of protein production by *MAP2K4* in the tumor cells of Patient 1, we performed a Western blot using MCF7 cells as a control and tumor cells taken from an ascites sample from Patient 1 (Fig. 4B; uncropped image is shown as Supplemental Figure 7). While the control cells show clear production of MAP2K4, there appears to be a decrease in MAP2K4 in the tumor cells of Patient 1. The remaining production of MAP2K4 may be due to non-tumor cells in the sample. Further, mapping the transcripts from the RNA-seq data in Patient 1 (see "Methods" section), we find an average of only 2.92 TPM that map to *MAP2K4*. The number of transcripts aligning to *MAP2K4* in Patient 1 is much lower than

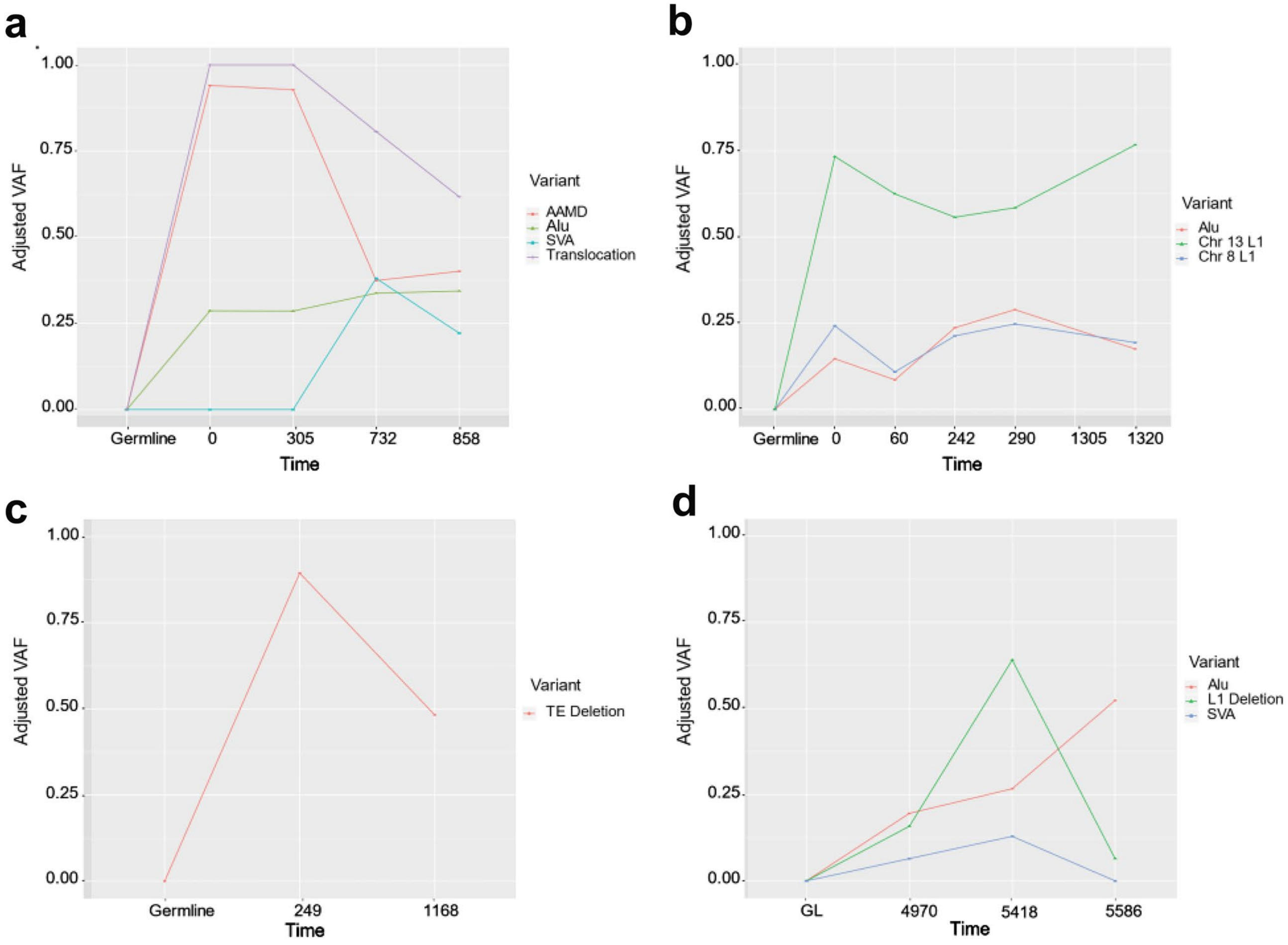

Figure 2. Adjusted VAF for each of the 4 patients. Parts (**A**–**D**) correspond to patients 1–4, respectively. The VAF for each patient has been adjusted to better reflect the percentage of cells in the sample that appear to be from the tumor. VAF is shown on the Y axis with each sampling timepoint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

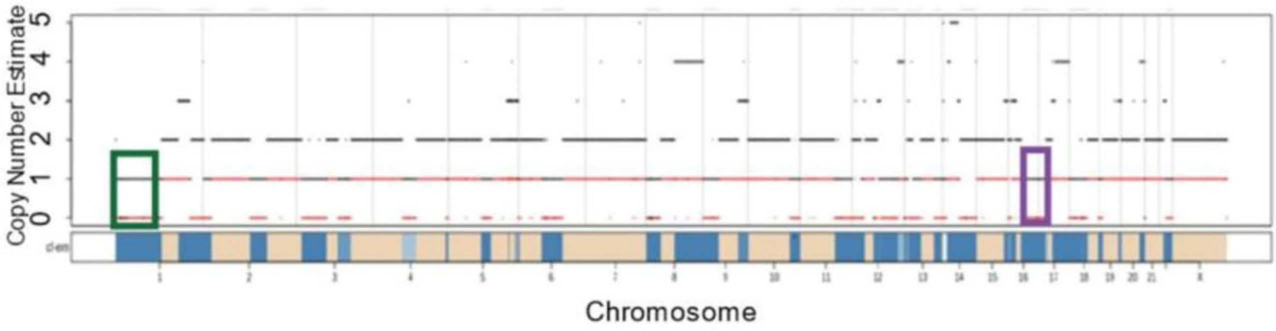

Figure 3. Copy number estimates from FACETS for Patient 1. A decrease in copy number along part of Chromosome 1 is shown in the green box. This green box includes *YTHDF2*, a gene that is partially deleted by a mobile element associated event. The purple box is highlighting a decrease in copy number along Chromosome 17. Included in this purple box is *MAP2K4,* a gene that is interrupted by a translocation event associated with mobile elements. The black lines in the figure represent the total copy number, while the red lines show the minor copy number for each segment.

the median value found by GTEx (median 18.05 TPM; the GTEx Portal on 08/28/2020). Additionally, GEPIA[39], which uses data from both GTEx and TCGA shows an average of 13.3 TPM for *MAP2K4* in cancer and 11.43 TPM for control samples (http://gepia.cancer-pku.cn/detail.php?gene=MAP2K4). Analyzing RNA-seq data from another patient that was not suspected to have any disruption to *MAP2K4*, we found an average of 14.80 TPM.

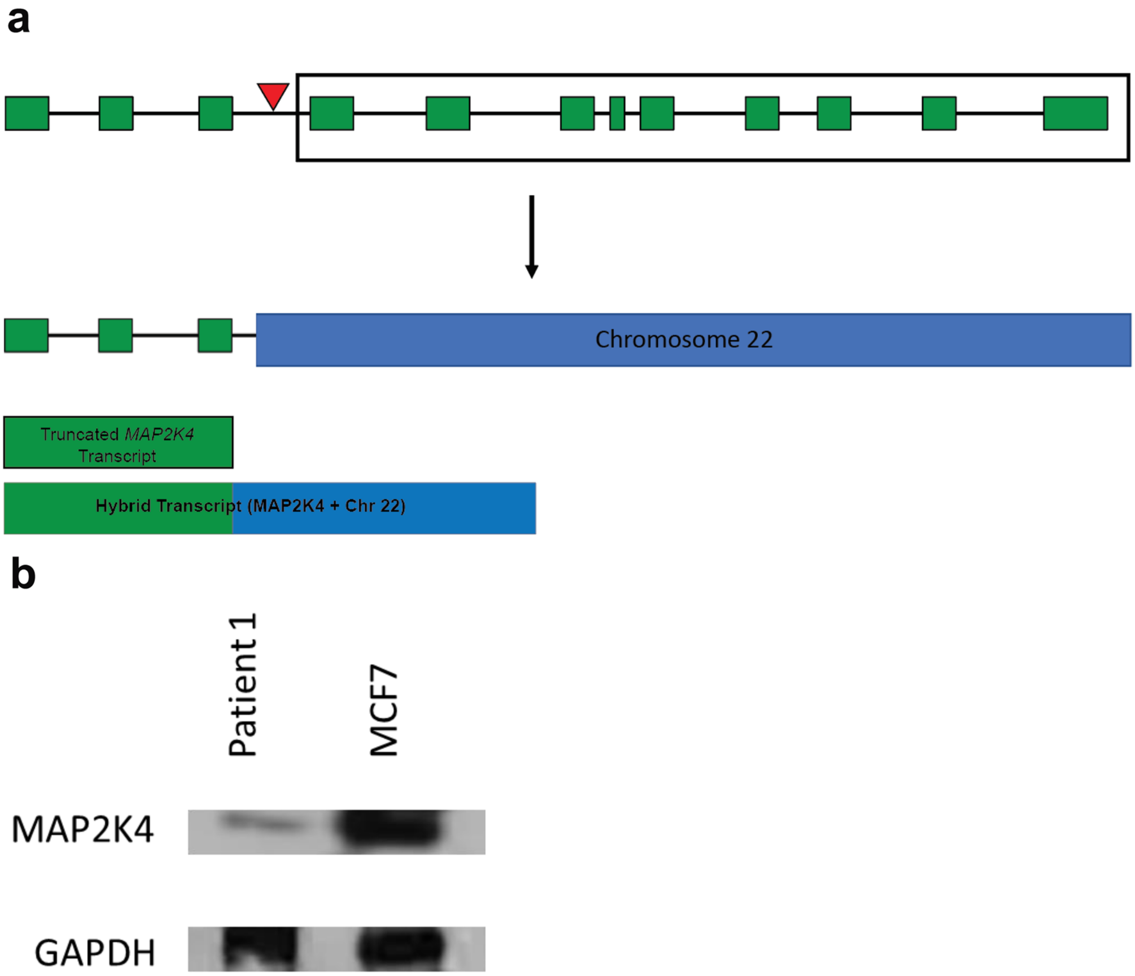
a
Chromosome 22
Truncated *MAP2K4* Transcript
Hybrid Transcript (MAP2K4 + Chr 22)
b
Patient 1
MCF7
MAP2K4
GAPDH

Figure 4. Translocation within *MAP2K4* gene decreases protein production 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,* leaving only the first three exons remaining on Chromosome 17. The portion of *MAP2K4* shown in the box is translocated to Chromosome 22. The translocation is depicted below the arrow. Truncated and hybrid transcripts are shown below the translocation. (**B**) Western blot of MAP2K4 in control MCF7 cells and ascites-derived cells from Patient 1. The control cells show clear expression of *MAP2K4*, while the cells taken from Patient 1 show decreased production of the protein. GAPDH blotting was performed as a loading control. Uncropped images of the Western blot are shown in Supplemental Figure 7.

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 (Supplemental 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. We saw no subfamily-level change in expression in these three patients for L1, SVA, or *Alu* elements. 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 in Supplemental Table 4. Approximately half of the differentially expressed mobile elements overlapped with genes that were also significantly differentially expressed. The location of these mobile elements that overlap with genes is shown in Supplemental Table 5.

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 (Fisher's exact, $p > 0.012$). These 87 regulatory regions include a number of CCCTC-binding factor (CTCF) binding sites, open chromatin regions, promoter flanking regions, enhancers, promoters, and transcription factor binding sites. These overlaps are shown in Supplemental Table 6.

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 (Supplemental Figure 1). We find that the visual validation step (IGV or other similar tools) reduces the risk of false positive calls and significantly improves the overall accuracy of variant calls (similar to[7]) (Fig. 1). This is particularly important for cancer genomes because most mobile element detection software is not designed for their complexity. This increased complexity makes the visual inspection step necessary, as thousands of potential mobile element insertions suggested by these programs 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 fraction and very low variant allele fraction (Fig. 2), showing that the insertions or variants did not have to be present at a particularly high frequency to be detected. The SV events identified with this pipeline are largely the result of recombination occurring between mobile elements that are already present in the reference genome[26,40,41] and not a product of somatic retrotransposition[42,43].

Large changes in VAF shown in the current study generally correlate with bottleneck events or large shifts in subclone frequency found previously[38]. The tumor subclones also change frequency in response to treatment[38], which can lead to less stable VAF, a finding shared here when examining mobile element-associated variants over time. 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) (Fig. 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. This decrease in frequency may be due to changes in the number of unique subclones present at later timepoints. The insertions and variants that are present at very high frequency (near 100%) may have occurred early in tumor progression, but these events could also be the result of a bottleneck event or a selective clonal advantage. The identification of these early events may also be influenced by decreased tumor purity in the samples during later timepoints (Supplemental Table 2). 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[44–47]. Using these markers in conjunction with SNPs may increase the level of resolution for tumor subclones.

From the summary of the insertions and SVs involving mobile elements identified in four patients (Table 1), we show activity for not only L1s, which we expect based on previous studies[16–20,22], but also activity from *Alu* and SVA elements. Most previous studies have largely focused on L1s and the structural variants that they mediate[17,20], though others have identified a small number of *Alu* insertions[22]. SVA has been found in a previous study[17] and recent work also supports post-zygotic insertions in normal tissue for this mobile element family[7]. The two SVA insertions identified here are present at low frequency and may have been missed by different filtering methods. We also see variation among patients, with most individuals showing multiple insertions or variants but one showing only a single event that appears to involve an interaction between existing *Alu* elements. The number of identified insertions is relatively low compared to some cancers that show high mobile element activity, but our results are similar to those seen for L1s in breast cancer[20].

The majority of the variants we identified appear to occur early in our sampling timeframe (Table 1, Fig. 2) and likely occur early in tumor progression or development. This suggests that mobile elements may be 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 insertions that occurred in the primary tumor were reflected in metastases[23]. However, ascertainment bias may allow for the increased possibility of identifying high frequency mobile elements and SVs. Further, given the lower sequencing depth of the germline in Patient 2 and Patient 4, it is possible that some of the identified events in these patients could be mosaic. 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 something unique about SVA insertions causes 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 SVA insertions to determine if the trend of insertions at later timepoints remains.

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, we did not observe any instances of subfamily expression changes for *Alu,* SVA, or L1, and only one patient showed a family-level change in expression. This was linked to an increase in HERV activity, which previous studies have reported for HERV-K in various cancers[48,49]. We were also able to examine the expression of specific mobile element loci in which there were informative reads. Here, using RNA-seq data for Patient 2, which spanned multiple years, we showed that many loci did change expression levels over time. Approximately half of the mobile element loci that showed statistically significant differential expression overlapped genes that were also significantly differentially expressed. Many of the remaining 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[50,51]. Some of these expression changes are likely tied to methylation changes due to the high CpG content of mobile elements[52]. However, differentiating between transposable element transcripts and other transcripts can be challenging, and the patterns shown here may be more indicative of general gene expression change than changes in mobile element expression. We were only able to demonstrate expression changes in Patient 2, who 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.

The early SV events identified in Patient 1 both appear to be associated with *Alu* elements and to affect the coding sequence. The affected genes (*MAP2K4* and *YTHDF2*) have both been implicated in cancer development[53–57]. 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)[58], 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[53,59,60]. *MAP2K4* is expressed in mammary tissue (median 18.05 TPM; the GTEx Portal on 08/28/2020), and missense and nonsense mutations have been identified at low frequency as part of The Cancer Genome Atlas (TCGA Research Network: https://www.cancer.gov/tcga). The translocation breakpoint on Chromosome 17 occurs after the third exon of *MAP2K4* (Fig. 4A), with the split reads in this region showing evidence of an *Alu* element. The breakpoint on Chromosome 22 disrupts a reference *Alu* element, and it is likely that this translocation was associated with an *Alu* interaction. There have also been previous examples of translocations and recombination events associated with mobile elements in cancer[61,62]. *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* (Fig. 3). Our findings support a two-hit model, where a copy of *MAP2K4* was lost, and the remaining allele was disrupted by the *Alu*-associated translocation described here. With only a single copy of this gene remaining, the translocation renders *MAP2K4* inactive, leading to the decrease in production of the protein shown in the Western blot (Fig. 4B). Further analysis of the RNA-seq data for this patient showed that there were no complete transcripts for *MAP2K4*. *MAP2K4* is found in the *JNK* pathway, which is responsible for numerous cellular functions[63]. Previous work in breast cancer tissue has identified multiple mutations along this pathway (including *MAP2K4* and *MAP2K7*)[54] and suggests that this mutation could act as a driver by altering function of the *JNK* pathway. *MAP3K1* is another commonly mutated gene in this pathway and is responsible for the signaling step prior to *MAP2K4*[54]. Mutations in these crucial genes can lead to changes in cell proliferation and the ability to escape from apoptosis, both of which are commonly seen in cancer. 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 understand the cause of the mutation.

Overall, we find that most mobile element insertions and the structural variation events (between reference mobile elements) 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 SV events involving mobile elements that disrupt the coding sequence of known (*MAP2K4)* and suspected (*YTHDF2)* driver genes in breast cancer. We identified a number of L1, *Alu,* and SVA insertions occurring during tumor progression. As most studies attempt to identify only L1 insertions and structural variants involving L1s, we may be underestimating the impact that mobile elements have on mediating driver mutations in cancer. However, our sample size is small and may not be representative of the activity of these mobile elements in a larger sample size. Future studies should examine other types of cancer for the patterns and impact of mobile element insertions 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 samples, sequencing data, as well as quality control and alignment information, can be found in Brady et al. 2017[38]. 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. Each patient had multiple tumor samples from different timepoints throughout their treatment: Patient 1 had four samples, Patient 2 had six samples, Patient 3 had two samples, and Patient 4 had three samples that were examined. Data from the Brady et al. 2017 publication are available with controlled access on the European Genome-phenome Archive (EGA) under accession EGAS00001002436. Informed consent was obtained from all patients in the original study. Protocols were approved by the University of Utah Institutional Review Board. Data used in this study were publicly available, and experiments were performed in accordance with relevant guidelines and regulations. Coverage for each WGS timepoint was calculated using covstats from the goleft package (https://github.com/brentp/goleft).

Mobile element insertions and variants identification. We used the Mobile Element Locator Tool (MELT) (Version 2.1.5)[64], RUFUS (https://github.com/jandrewrfarrell/RUFUS), where possible, and TranSurVeyor (Version 1.0)[65] 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. TranSurVeyor and MELT are tools used to detect mobile element insertions, but these tools use discordant and split reads as part of their algorithm and returned a number of SVs that were identified during our visual inspection step. RUFUS is an alignment-free, k-mer based algorithm that compares reads between the control samples (germline DNA) and the sample of interest (tumor DNA). Through this approach, RUFUS is capable of detecting many types of variants, including SNVs, SVs, and small indels (Described in[66]). We considered any structural variant (deletions, complex events, and other rearrangements) that included a reference transposable element for further examination. Because these tools use different methods, we required only a single tool to identify a potential insertion. Insertions and variants identified by these programs were filtered to include only those that were called as absent in the germline sequencing data. MELT identified 8825 variants, while TranSurVeyor identified a total of 69,394 variants in these four patients. The total number of variants identified with RUFUS was not calculated because it is designed to look for many types of variants, not only mobile elements. 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)[67]. Through this visual inspection step, > 70,000 IGV images (as there were multiple individuals included in the MELT analysis) were examined from MELT and TranSurveyor. This step helped to prevent the inclusion of reads that had simply mis-mapped to an incorrect region of the genome. Those insertions that showed any evidence of being present in the germline DNA sample, or those loci that did not show any signs of mobile element activity (low complexity repeats or poorly sequenced regions), were discarded. Reads that appeared to have discordant and/or split reads that mapped to a mobile element were included for further validation. To ensure that the potential mobile element insertions or mobile element-associated structural variants were, in fact, mobile element sequences, we used both BLAT[68] and RepeatMasker[69]. 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 involve mobile elements with Lumpy (Version 0.2.13)[70] and we used IGV validation to identify signs of mobile element involvement (breakpoints in or near existing mobile elements, and small regions of microhomology). While the limit of detection of this pipeline is largely determined by the mobile element detection tools, we were able to identify mobile elements and SVs with VAFs of ~ 5%. BEDTools[71] 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/).[72] The UCSC Genome Browser (ENCODE Regulation track) was manually checked to ensure no other regulatory elements overlapped the insertion site.

Variant allele fraction. Variant allele fraction 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 fraction was then adjusted by multiplying the total number of reads at the breakpoint by the tumor purity (or cancer cell fraction) value. Following this, the number of reads containing the variant were divided by the adjusted estimate of the total number of reads that were derived from tumor cells. The tumor purity estimates were calculated using FACETS[73], 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)[74]. 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)[75]; counts of gene expression and mobile element expression were then generated to quantify expression. DESeq2 (Version 1.16.1)[76] 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[77] using BEDTools.

Trinity (Version 2.8.5)[78] 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. Transcripts that aligned to multiple regions of the genome with > 95% accuracy were not included.

Salmon (Version 1.4.0)[79] was run on RNA-seq timepoints from Patient 1 and Patient 2. The data were converted to gene-level annotations, and the transcripts per million (TPM) for *MAP2K4* were counted. The TPM values for each timepoint for Patient 1 were averaged, and the process was repeated for three timepoints in Patient 2.

DNA extraction and PCR validation. For validation of our detection methods, PCR amplification was run on a potential L1 and a potential *Alu* insertion in Patient 2, the only patient 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 s at 98 °C, with 40 cycles of denaturation for 10 s at 98 °C, the optimal annealing temperature of 60 °C for 30 s, followed by a 2-min extension at 72 °C, and a final extension for 5 min 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 min at 100 V. The gel was imaged using a Fotodyne Analyst Investigator Eclipse imager. The primer sets used for these reactions are listed in Supplemental Table 7, and the corresponding gel images are shown in Supplemental Figures 3 and 4.

Immunoprecipitation and immunoblotting. Patient ascites samples were grown in Human Breast Epithelial Cell Culture Complete Media (Celprogen M36056-01S); MCF7 whole cell lysate was purchased (Novus). Cells were lysed with ThermoFisher Co-IP lysis buffer. MKK4 (1:100 CST) and GAPDH (1:300 ProteinTech) antibody were incubated overnight at 4C with 1000 µg of protein lysate. Immunoprecipitation was performed according to Pierce™ Classic Magnetic IP/Co-IP Kit protocol. 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 (MKK4 1:500, GAPDH 1:000 ThermoFisher) overnight at 4 °C. Blots were washed with PBST and incubated with goat anti-rabbit conjugated to HRP secondary antibody for 60 min at room temperature. Blots were then incubated in chemiluminescent substrate enhancer (BioRad) and visualized using x-ray film. Intensity quantification was performed with Fiji, with the *GAPDH* band for MCF7 quantified as 78 pixels and the *GAPDH* band for the patient as 93 pixels.

Uncropped images of the Western blot are shown in Supplemental Figure 7. The Western blot image shown in the main paper has been cropped to show important bands. *GAPDH* for the patient sample and MCF7 cells were run on the same gel, but not run adjacent to one another. The image has been edited to show them directly adjacent. Membrane edges in the figures are not visible due to the X-ray film used for imaging. The membranes were cut post-antibody incubation to get a clearer image of the bands.

Data availability

Longitudinal data from the Brady et al. 2017 publication are available with controlled access on the European Genome-phenome Archive (EGA) under accession EGAS00001002436.

Received: 2 March 2021; Accepted: 7 June 2021
Published online: 22 June 2021

References

1. de Koning, A. P., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **7**, e1002384 (2011).
2. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Boeke, J. D., Garfinkel, D. J., Styles, C. A. & Fink, G. R. Ty elements transpose through an RNA intermediate. *Cell* **40**, 491–500 (1985).
4. Beck, C. R. *et al.* LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159–1170 (2010).
5. Brouha, B. *et al.* Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 5280–5285 (2003).
6. Xing, J., Witherspoon, D. J. & Jorde, L. B. Mobile element biology: New possibilities with high-throughput sequencing. *Trends Genet.* **29**, 280–289 (2013).
7. Feusier, J. *et al.* Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res.* **29**, 1567–1577 (2019).
8. Hancks, D. C. & Kazazian, H. H. Roles for retrotransposon insertions in human disease. *Mob. DNA* **7**, 9 (2016).
9. Kazazian, H. H. Jr. *et al.* Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164–166 (1988).
10. Kazazian, H. H. Jr. & Moran, J. V. Mobile DNA in health and disease. *N. Engl. J. Med.* **377**, 361–370 (2017).
11. Coufal, N. G. *et al.* L1 retrotransposition in human neural progenitor cells. *Nature* **460**, 1127–1131 (2009).
12. Muotri, A. R. *et al.* Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903–910 (2005).
13. Erwin, J. A., Marchetto, M. C. & Gage, F. H. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat. Rev. Neurosci.* **15**, 497–506 (2014).
14. Richardson, S. R., Morell, S. & Faulkner, G. J. L1 retrotransposons and somatic mosaicism in the brain. *Annu. Rev. Genet.* **48**, 1–27 (2014).
15. Doucet, O. *et al.* LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. *Proc. Natl. Acad. Sci.* **112**, E4894 (2015).
16. Doucet-O'Hare, T. T. *et al.* Somatically acquired LINE-1 insertions in normal esophagus undergo clonal expansion in esophageal squamous cell carcinoma. *Hum. Mutat.* **37**, 942–954 (2016).
17. Helman, E. *et al.* Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res.* **24**, 1053–1063 (2014).
18. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
19. Rodriguez-Martin, B. *et al.* Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* **52**, 306–319 (2020).
20. Tubio, J. M. C. *et al.* Mobile DNA in cancer: Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014).
21. Burns, K. H. Transposable elements in cancer. *Nat. Rev. Cancer* **17**, 415–424 (2017).
22. Lee, E. *et al.* Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967–971 (2012).
23. Ewing, A. D. *et al.* Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. *Genome Res.* **25**, 1536–1545 (2015).
24. Rodić, N. *et al.* Long interspersed element-1 protein expression is a hallmark of many human cancers. *Am. J. Pathol.* **184**, 1280–1286 (2014).
25. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703 (2009).
26. Xing, J. *et al.* Mobile elements create structural variation: Analysis of a complete human genome. *Genome Res.* **19**, 1516–1526 (2009).
27. Hsieh, S.-Y., Chen, W.-Y., Yeh, T.-S., Sheen, I. S. & Huang, S.-F. High-frequency Alu-mediated genomic recombination/deletion within the caspase-activated DNase gene in human hepatoma. *Oncogene* **24**, 6584–6589 (2005).
28. Mauillon, J. L. *et al.* Identification of novel germline <em>hMLH1</em> mutations including a 22 kb Alu-mediated deletion in patients with familial colorectal cancer. *Can. Res.* **56**, 5728 (1996).
29. Welcsh, P. L. & King, M.-C. BRCA1 and BRCA2 and the genetics of breast and ovarian cancer. *Hum. Mol. Genet.* **10**, 705–713 (2001).
30. Peixoto, A. *et al.* Genomic characterization of two large Alu-mediated rearrangements of the BRCA1 gene. *J. Hum. Genet.* **58**, 78–83 (2013).
31. Petrij-Bosch, A. *et al.* BRCA1 genomic deletions are major founder mutations in Dutch breast cancer patients. *Nat. Genet.* **17**, 341–345 (1997).
32. Puget, N. *et al.* A 1-kb Alu-mediated germ-line deletion removing <em>BRCA1</em> exon 17. *Can. Res.* **57**, 828 (1997).
33. Rohlfs, E. M. *et al.* 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 (2000).
34. Morse, B., Rotherg, P. G., South, V. J., Spandorfer, J. M. & Astrin, S. M. Insertional mutagenesis of the myc locus by a LINE-1 sequence in a human breast carcinoma. *Nature* **333**, 87–90 (1988).

35. Walsh, T. *et al.* Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. *JAMA* **295**, 1379–1388 (2006).
36. Choi, Y. J. *et al.* Intraindividual genomic heterogeneity of high-grade serous carcinoma of the ovary and clinical utility of ascitic cancer cells for mutation profiling. *J. Pathol.* **241**, 57–66 (2017).
37. Husain, H. *et al.* Cell-free DNA from ascites and pleural effusions: Molecular insights into genomic aberrations and disease biology. *Mol. Cancer Therap.* **16**, 948–955 (2017).
38. Brady, S. W. *et al.* Combating subclonal evolution of resistant cancer phenotypes. *Nat. Commun.* **8**, 1231 (2017).
39. Tang, Z. *et al.* GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* **45**, W98–W102 (2017).
40. Gu, W., Zhang, F. & Lupski, J. R. Mechanisms for human genomic rearrangements. *PathoGenetics* **1**, 4 (2008).
41. Kolomietz, E., Meyn, M. S., Pandita, A. & Squire, J. A. The role of Alu repeat clusters as mediators of recurrent chromosomal aberrations in tumors. *Genes Chromosomes Cancer* **35**, 97–112 (2002).
42. Gilbert, N., Lutz, S., Morrish, T. A. & Moran, J. V. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol. Cell Biol.* **25**, 7780–7795 (2005).
43. Symer, D. E. *et al.* Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* **110**, 327–338 (2002).
44. Steely, C. J. *et al.* Alu insertion polymorphisms as evidence for population structure in baboons. *Genome Biol. Evol.* **9**, 2418–2427 (2017).
45. Watkins, W. S. *et al.* The Simons Genome Diversity Project: A global analysis of mobile element diversity. *Genome Biol. Evol.* **12**, 779–794 (2020).
46. Watkins, W. S. *et al.* Genetic variation among world populations: Inferences from 100 Alu insertion polymorphisms. *Genome Res.* **13**, 1607–1618 (2003).
47. Witherspoon, D. J. *et al.* Mobile element scanning (ME-Scan) identifies thousands of novel Alu insertions in diverse human populations. *Genome Res.* **23**, 1170–1181 (2013).
48. Ma, W. *et al.* Human endogenous retroviruses-K (HML-2) expression is correlated with prognosis and progress of hepatocellular carcinoma. *Biomed. Res. Int.* **2016**, 8201642–8201642 (2016).
49. Wallace, T. A. *et al.* Elevated HERV-K mRNA expression in PBMC is associated with a prostate cancer diagnosis particularly in older men and smokers. *Carcinogenesis* **35**, 2074–2083 (2014).
50. Jang, H. S. *et al.* Transposable elements drive widespread expression of oncogenes in human cancers. *Nat. Genet.* **51**, 611–617 (2019).
51. Xie, M. *et al.* DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat. Genet.* **45**, 836–841 (2013).
52. Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**, 335–340 (1997).
53. Ahn, Y.-H. *et al.* 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 (2011).
54. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
55. Su, G. H., Song, J. J., Repasky, E. A., Schutte, M. & Kern, S. E. Mutation rate of MAP2K4/MKK4 in breast carcinoma. *Hum. Mutat.* **19**, 81 (2002).
56. Xue, Z. *et al.* MAP3K1 and MAP2K4 mutations are associated with sensitivity to MEK inhibitors in multiple cancer models. *Cell Res.* **28**, 719–729 (2018).
57. Zhong, L. *et al.* YTHDF2 suppresses cell proliferation and growth via destabilizing the EGFR mRNA in hepatocellular carcinoma. *Cancer Lett.* **442**, 252–261 (2019).
58. Tate, J. G. *et al.* COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2018).
59. Teng, D. H. *et al.* Human mitogen-activated protein kinase kinase 4 as a candidate tumor suppressor. *Cancer Res.* **57**, 4177–4182 (1997).
60. Pavese, J. M. *et al.* Mitogen-activated protein kinase kinase 4 (MAP2K4) promotes human prostate cancer metastasis. *PLoS ONE* **9**, e102289 (2014).
61. Elliott, B., Richardson, C. & Jasin, M. Chromosomal translocation mechanisms at intronic Alu elements in mammalian cells. *Mol. Cell* **17**, 885–894 (2005).
62. Onno, M., Nakamura, T., Hillova, J. & Hill, M. Rearrangement of the human tre oncogene by homologous recombination between Alu repeats of nucleotide sequences from two different chromosomes. *Oncogene* **7**, 2519–2523 (1992).
63. Johnson, G. L. & Lapadat, R. Mitogen-Activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases. *Science* **2002**, 298 (1911).
64. Gardner, E. J. *et al.* The mobile element locator tool (MELT): Population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).
65. Rajaby, R. & Sung, W. K. TranSurVeyor: An improved database-free algorithm for finding non-reference transpositions in high-throughput sequencing data. *Nucleic Acids Res.* **46**, e122 (2018).
66. Ostrander, B. E. P. *et al.* Whole-genome analysis for effective clinical diagnosis and gene discovery in early infantile epileptic encephalopathy. *NPJ Genom. Med.* **3**, 22 (2018).
67. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
68. Kent, W. J. BLAT—The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
69. RepeatMasker Open-4.0. http://www.repeatmasker.org
70. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
71. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
72. Dimitrieva, S. & Bucher, P. UCNEbase—A database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res.* **41**, D101–D109 (2013).
73. Shen, R. & Seshan, V. E. FACETS: Allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).
74. Yang, W. R., Ardeljan, D., Pacyna, C. N., Payer, L. M. & Burns, K. H. SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Res.* **47**, e27–e27 (2019).
75. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15–21 (2013).
76. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
77. Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Res.* **45**, D635–D642 (2016).
78. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
79. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

Acknowledgements

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.

Author contributions

C.J.S., G.M., and L.B.J. designed the research; Y.Q. assisted with data acquisition and interpretation; C.J.S. and J.E.F. identified and reviewed mobile element calls; K.L.R. performed Western blot validation; C.J.S. performed PCR validation; S.V.T. contributed to experimental design and manuscript editing; C.J.S. wrote the first draft of the manuscript with all authors editing and contributing to the final version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-92444-0.

Correspondence and requests for materials should be addressed to C.J.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.