scientific reports

Check for updates

OPEN

The spatial and cellular portrait of transposable element expression during gastric cancer

Braulio Valdebenito-Maturana

Gastric Cancer (GC) is a lethal malignancy, with urgent need for the discovery of novel biomarkers for its early detection. I previously showed that Transposable Elements (TEs) become activated in early GC (EGC), suggesting a role in gene expression. Here, I follow-up on that evidence using single-cell data from gastritis to EGC, and show that TEs are expressed and follow the disease progression, with 2,430 of them being cell populations markers. Pseudotemporal trajectory modeling revealed 111 TEs associated with the origination of cancer cells. Analysis of spatial data from GC also confirms TE expression, with 204 TEs being spatially enriched in the tumor regions and the tumor microenvironment, hinting at a role of TEs in tumorigenesis. Finally, a network of TE-mediated gene regulation was modeled, indicating that ~ 2,000 genes could be modulated by TEs, with ~ 500 of them already implicated in cancer. These results suggest that TEs might play a functional role in GC progression, and highlights them as potential biomarker for its early detection.

Gastric Cancer (GC) is one of the leading causes of cancer death, with an estimated 800,000 demises per year[1,2]. Although it has a global incidence, it is now considered an endemic malignancy in South America and some parts of Asia and Europe[1,3]. The 5-year survival rate of advanced stages GC is ~ 5%, while for early GC is > 70%[2,4], highlighting the importance of detecting this malignancy in a timely manner[3]. Persistent inflammation to the stomach is one of the main factors associated with the origin of GC. Particularly, chronic gastritis and Intestinal Metaplasia (IM) are precursor lesions, and the overall cascade progression of the disease can be recapitulated from Chronic Non-Atrophic Gastritis (NAG), Chronic Atrophic Gastritis (CAG), IM, early GC (EGC) and GC[1,4–6]. Taking this into account, several groups have studied this progression using different modalities of RNA-Sequencing (RNA-Seq) to profile the changes in genome-wide gene expression[6–8].

RNA-Seq has remained the gold standard in gene expression studies due to its large-scale throughput, and nucleotide-level profiling of gene expression[9,10]. Traditional RNA-Seq is also known as "bulk", because it captures a homogenized portrait of gene expression[11]. Although this has allowed for many advances in our knowledge, it has been somewhat limited for studying cancer due to intratumor heterogeneity[12]. In turn, recent works using higher-resolution modalities of RNA-Seq, such as single-cell (scRNA-Seq) and spatially-resolved (srRNA-Seq), have been published[6,13,14]. In addition to the identification of tumor subpopulations, scRNA-Seq allows for the reconstruction of cell trajectories through Trajectory Inference (TI) methods. TI corresponds to the modeling of dynamic cellular processes through the pseudotemporal arrangement of cells based on their transcriptional similarity[15]. This methodology has been successfully studied to understand cancer evolution[6,13,16,17] and how different cancer subpopulations respond to treatment[18]. In addition, by using this methodology, gene expression changes involved in cell fate decisions can also be studied, which can help understand what drives the changes in cell subpopulations towards a malignant genotype[18–20]. Despite the many breakthroughs of scRNA-Seq, one drawback is that tissue topology and organization is lost. In cancer studies, this makes it difficult to understand the interactions between tumor and its microenvironment. In this regard, srRNA-Seq preserves the two-dimensional spatial architecture and allows the profiling of in situ expression across a tissue section[21]. This technique has accelerated our understanding of cancer by allowing the study of gene expression in the tumor and surrounding microenvironment, and has been applied to GC[14]and other types of cancer[21].

Transposable Elements (TEs) are genetic agents with the ability to move and increase their copy number[22]. They are present in every eukaryotic genome known to date, and in humans they occupy about 50% of the genome[23,24]. Broadly, they can be classified into retrotransposons, which transpose via an RNA intermediate, and into DNA transposons, which transpose via a DNA intermediate. Retrotransposons are further subdivided into Long Terminal Repeats (LTRs), Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs), while DNA transposons are subdivided into DNA and Rolling-Circle (RC) TEs[22,25]. Although most TEs are now genetically fixed, they can still become transcriptionally active, which can impact

Centro de Genómica Avanzada de Talca, Talca, Chile. email: bvaldebenitom@gmail.com

gene regulation[26]. Current evidence supports their role as regulatory elements in health and disease[27–29]. In particular, there are many examples of their global derepression and subsequent activation in a wide array of cancer types (reviewed in[30]). Despite this, there are scarce works studying them using either single-cell or spatially-resolved methods. For example, only recently a group profiled gallbladder cancer using scRNA-Seq and applied TI to reveal that human endogenous retroviruses (HERVs, a type of LTR TEs) are associated with the transition of epithelial cells to the malignant status, and that they might act as regulators of gene expression in cancer cells[13].

Previously, I hinted at a regulatory role of TEs in EGC[29]. An outstanding question in that study was whether TEs were expressed in previous timepoints, and whether their expression continues during GC. Moreover, it is unclear if TEs are expressed in the GC tumor microenvironment. Here, by leveraging single-cell and spatially-resolved RNA-Seq data, I provide an extended analysis of the role of TEs in GC (Fig. 1). In this work, using the single-cell data, I show that TEs become up-regulated during the progression from gastritis to EGC, and their expression is associated with the acquisition of malignancy. Then, using the spatially-resolved transcriptomes, I provide evidence of TE activation in both the tumor and its microenvironment. Additionally, network analysis suggests that they might be involved in the regulation of genes, highlighting a functional role in the GC cascade. Overall, these findings propose TEs as potential biomarkers for the detection of GC at its early stages.

Results

TE expression during the progression of gastritis to early GC

The single-cell raw sequencing data generated at the NAG (3 samples), CAG (3 samples), IM (3 samples, one wild and 2 severe) and EGC (1 sample) stages was aligned to the human genome. Afterwards, the resulting BAM alignment files were processed using SoloTE[29] to get matrices containing gene and TE expression per each cell. In order to get a preliminary overview on TE expression across the early GC cascade, a Principal Component Analysis (PCA) was carried out using a pseudobulk approach in which total expression was summarized at the sample level. This procedure was done 3 times, by using a matrix subsetted only to genes, another subsetted only to TEs, and one containing gene and TE expression (Fig. 2a). This analysis illustrates that gene expression alone recapitulates the differences between each time point, with the exception of the IM-Wild sample, which seems closer to the CAG samples. This suggests that only subtle changes occur during the IM-Wild timepoint. Using only TE expression, a similar distinction between timepoints can be seen. Although less variance is explained, the samples are still reasonably grouped in terms of their stage. As expected, the analysis using the complete Gene + TE matrix shows a result consistent with the ones done independently.

Next, to understand the influence of TEs at the cellular level, t-distributed stochastic neighbor embedding (tSNE) dimensional reduction analysis was carried out similarly, using the gene expression matrix, the TE expression matrix and the combined gene and TE expression matrix (Fig. 2b). The full Gene + TE matrix was processed first to generate a cell type annotation, resulting in 10 epithelial types, and 7 non-epithelial types. The epithelial population is comprised by Pit Mucous Cells (PMC), Neck-like, Gland Mucous Cells (GMC), Proliferative Cells (PC), Enteroendocrine, Chief, Enterocytes, Goblet, Metaplastic Stem-like Cells (MSCs) and Cancer cells. On the other hand, the non-epithelial population is comprised by T cells, B cells, Endothelial, Macrophages, Mast cells, Smooth Muscle Cells (SMC) and Fibroblasts. This cell type annotation was later transferred to the single-cell analyses done on genes and on TEs independently. In contrast to result obtained at the Gene level, the tSNE dimensional reduction of TE expression depicts more heterogeneity between cell types, in which PMCs, GMCs and MSCs seem to be more spread-out. It was previously shown that PMCs decrease and MSCs increase along the early GC cascade, and that there are transcriptional similarities between PMCs with both GMCs and MSCs, and between MSCs and Cancer cells[6]. Indeed, comparison of the cell type annotations with the unbiased clustering shows that some TE clusters are comprised by different populations (Supplementary Fig. 1). For example, cluster 0 is comprised by PMCs and GMCs, while cluster 1 and 2 span several types from PMCs, GMCs up to MSCs and Cancer cells. It has been proposed that mixed cells in the single-cell projection could be predictive of intermediate cell states[31], and thus, it can be hypothesized that TE expression could represent such states. Interestingly, the opposite can be seen in some non-epithelial subtypes, such as T cells and B cells. These cells exhibit a well-defined clustering pattern, hinting at a differential activation of a subset of TEs that might be modulating gene expression[32], with potential implications in T cell exhaustion[33].

Overall, these results show that although there is some variability in TE expression at the single-cell level, they are globally expressed throughout the progression from gastritis to EGC and recapitulate changes in each timepoint similar to those observed at the gene level.

The single-cell expression of TEs in early GC

Given the global TE expression revealed by the previous analyses, I investigated what TEs could exhibit consistent or increasing expression from the premalignant gastritis status to EGC, and in what cell populations they might be enriched. To this end, the pseudobulked matrix was used, and a series of differential expression (DE) analyses were carried out with DESeq2[34], using NAG as the baseline condition. Then, TEs significantly up-regulated (having $\log_2$(Fold Change) > 0 and adjusted P-value ≤ 0.05) in CAG, IMW, IMS and EGC timepoints were selected. This resulted in a total of 2,581 TEs (Supplementary data 1, Supplementary Fig. 2). Although the EGC DE analysis might be statistically underpowered due to only having one sample, it still recapitulates changes reported in previous works (Supplementary Fig. 3). To also consider TEs that might already be activated in NAG, and whose expression remains constant throughout the EGC cascade (and thus, not appearing in the DE analysis), the TEs in the top 5% of highest expression across all stages were also selected (Supplementary data 1). This added 14,566 TEs to the set, resulting in a total of 17,147 TEs potentially associated with EGC progression (Fig. 3a). Furthermore, there is a consistent percentage of TE expression across the single-cell transcriptomes: NAG—37.822%, CAG—41.247%, IMW—50.282%, IMS—41.989%, EGC—40.931% (Fig. 3b).

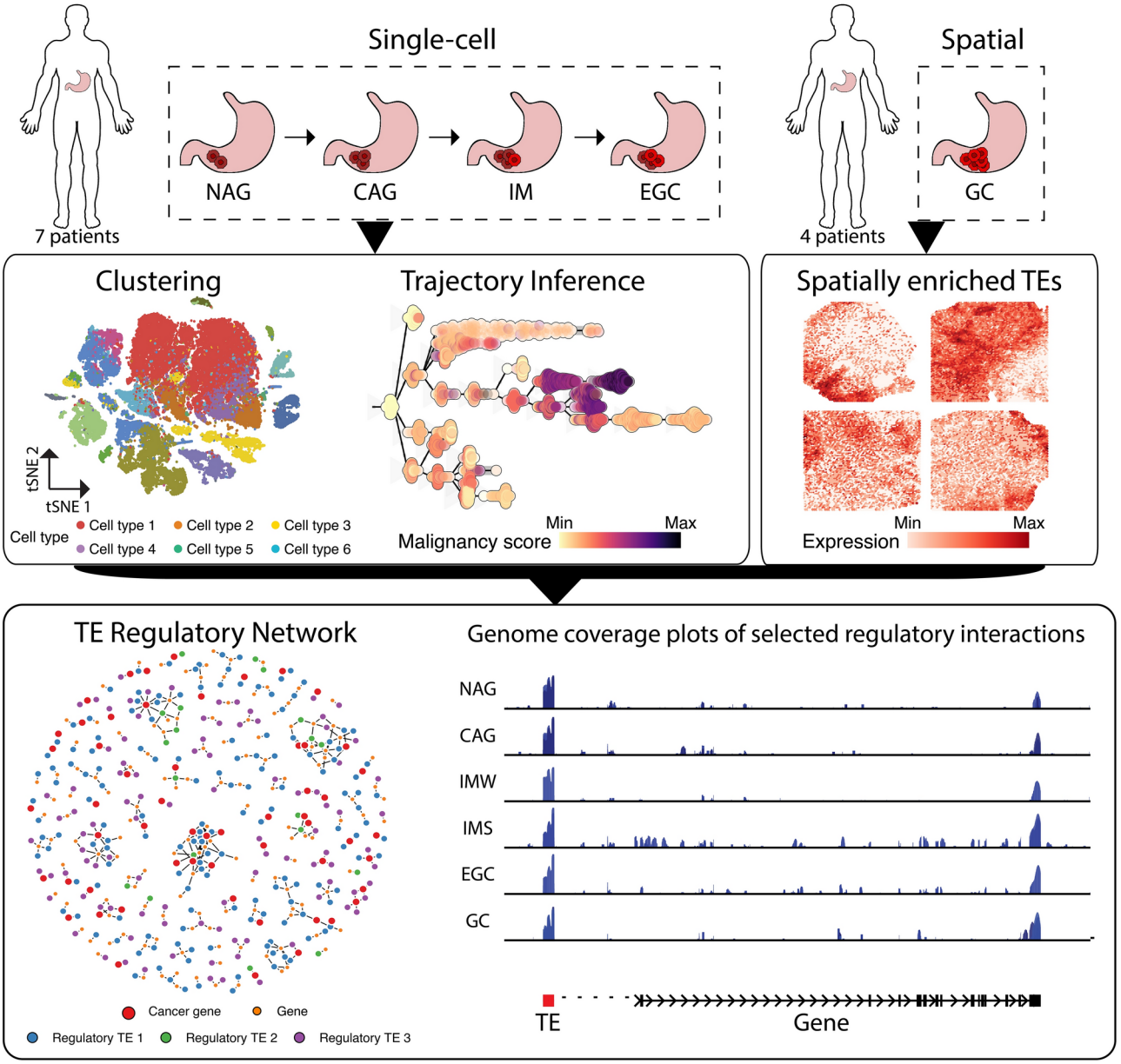

Fig. 1. Overview of the analysis protocol used in this study. In the first row, the studied datasets are depicted: the single-cell RNA-Seq data collected from 7 patients, spanning the progression from gastritis to cancer (Non-atrophic gastritis—NAG, Chronic atrophic gastritis—CAG, Intestinal metaplasia—IM and Early Gastric Cancer—EGC), and the spatial RNA-Seq data collected from 4 patients with Gastric Cancer. The second row shows the main analysis steps: for single-cell, "Clustering" was carried out to annotate cell types and then, by using this result, TEs enriched in cell populations were identified. "Trajectory Inference" was later performed to predict the pseudotemporal progression of the cells along the EGC cascade. Malignancy scores were calculated to validate the inferred trajectory. Afterwards, the impact of TE expression on the acquisition of the malignant status was assessed. For spatial data, by taking advantage of the pathologist annotations, TEs with increased expression in tumor regions and the tumor microenvironment were identified and labeled as "Spatially enriched TEs". Finally, the TEs identified from both analyses were used to build the TE regulatory network, where TEs were characterized according to their potential regulatory impact. Genes associated with TEs were labeled as "Cancer genes" in the network if they have been previously implicated in cancer. Schematics were drawn with Inkscape 1.3.2 (https://inkscape.org), and the plots produced with ggplot2.

Marker analysis revealed that out of the TEs selected in the previous step, 2,430 have increased expression in the different cell populations (Table 1). Some of the TE markers appear as enriched in more than one cell population, as evidenced in the numbers after de-duplication (Table 1, "Unique marker TEs"). Interestingly, a high proportion of these markers have locus resolution, which is essential to analyze their potential influence in gene regulation. Additionally, marker TE distribution per timepoint indicates that for most part they are

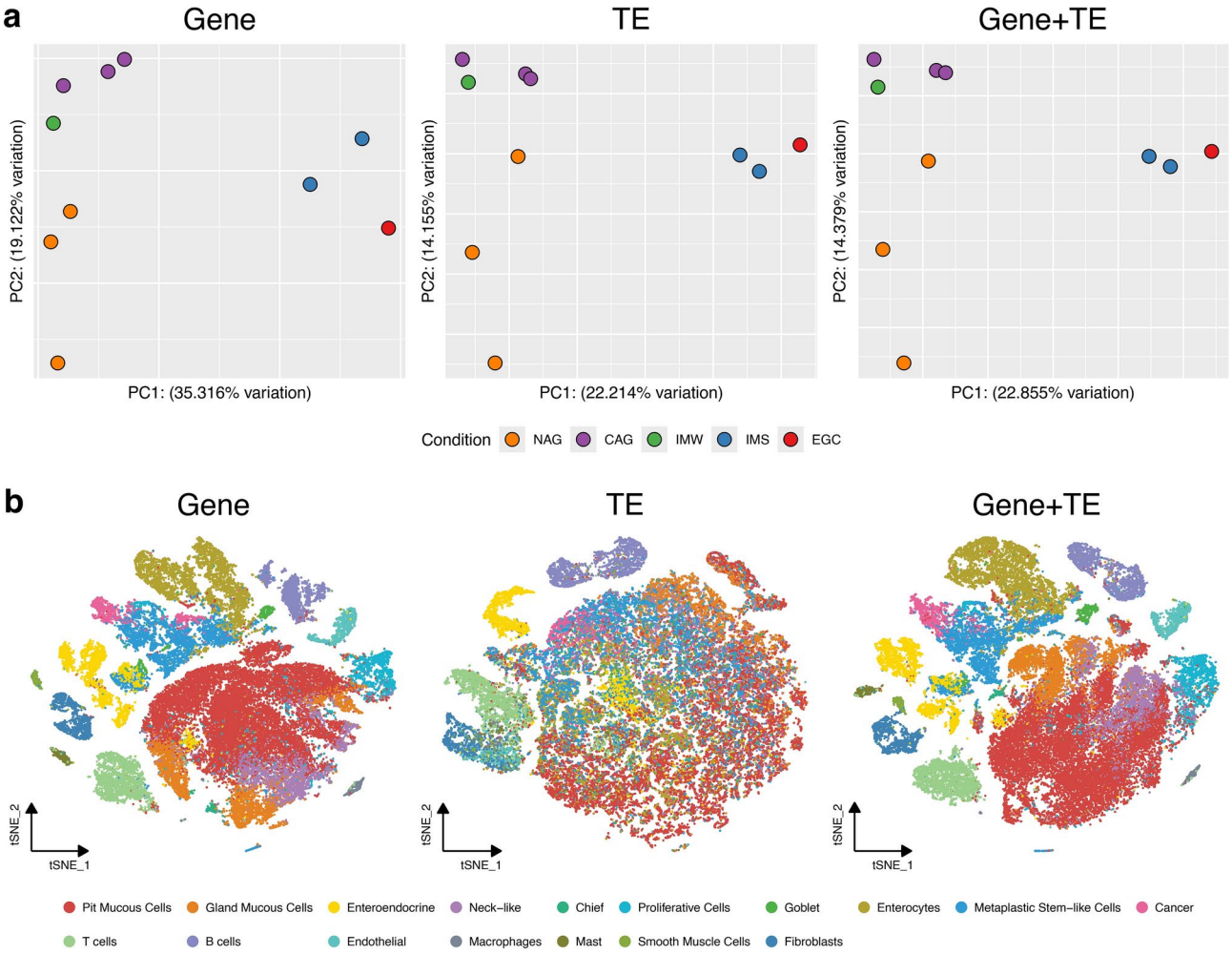

Fig. 2. Dimensional reduction analyses of the progression from gastritis to early gastric cancer. (**a**) Principal Component Analysis of the pseudobulked scRNA-Seq data, using gene expression only (first panel), TE expression only (second panel), and gene and TE expression (third panel). Each sample is color-coded according to the condition to which they belong. (**b**) Dimensional reduction plots using the t-distributed stochastic neighbor embedding (tSNE). Similarly, the first panel shows the tSNE plot using gene expression only, the second using TE expression only, and the third using gene and TE expression. The plots are color-coded according to the cell types identified.

equally distributed (Fig. 3c), suggesting that once their expression is increased in a cell type, it remains stable throughout the early GC progression. In Chief cells, a predominance at the CAG stage is observed, while Proliferative cells, Enterocytes, and Metaplastic Stem-like Cells seem to be predominantly expressed at the IMS stage, consistent with the emergence of these cell types during IM[6]. Analysis of the major TE types revealed that they are all present to varying degrees in the set of markers (Fig. 3d), with a clear predominance of SINEs. Although previous works have pointed out a role of LINE L1 and LTR HERV TEs in cancer[30], there is evidence showing that SINE and DNA TEs are also involved by driving oncogene activation[35]. This is also consistent with a recent work in colorectal cancer in which TEs were found to modulate gene expression and that SINEs were predominantly expressed[36].

A caveat here is that the 10X single-cell data has a 3' bias, and thus usually the terminal region of transcripts is captured and sequenced. Amongst TEs, the Alu family, part of the SINE group, has been reported to lead to truncated isoforms by acting as alternative transcription end sites[37,38]. In turn, out of the 927 marker SINE TEs, 875 (94.4%) were part of the Alu family, suggesting that in EGC these TEs could be effectively acting as premature transcription end sites. It has been proposed that the impact of these events could be associated with disease progression[37,38], and indeed there is an example of such events in liver cancer, where Alu TEs were identified as the major TE becoming a terminal exon[39]. Alternatively, the predominance of Alu TEs could be explained by their genetic structure: these elements harbor both a linker and a terminal A-stretch[40], which in turn could be causing internal poly-A priming during 10X 3' single-cell sequencing.

Marker TEs, which are defined as TEs with high expression, and expressed in a high proportion of the cells of a specific type (labeled as "pct.1"), could be classified in 2 groups based on their percentage expressed in the remaining cell types (labeled as "pct.2"): group 1—high pct.2; group 2—low pct.2. For example, some of the top

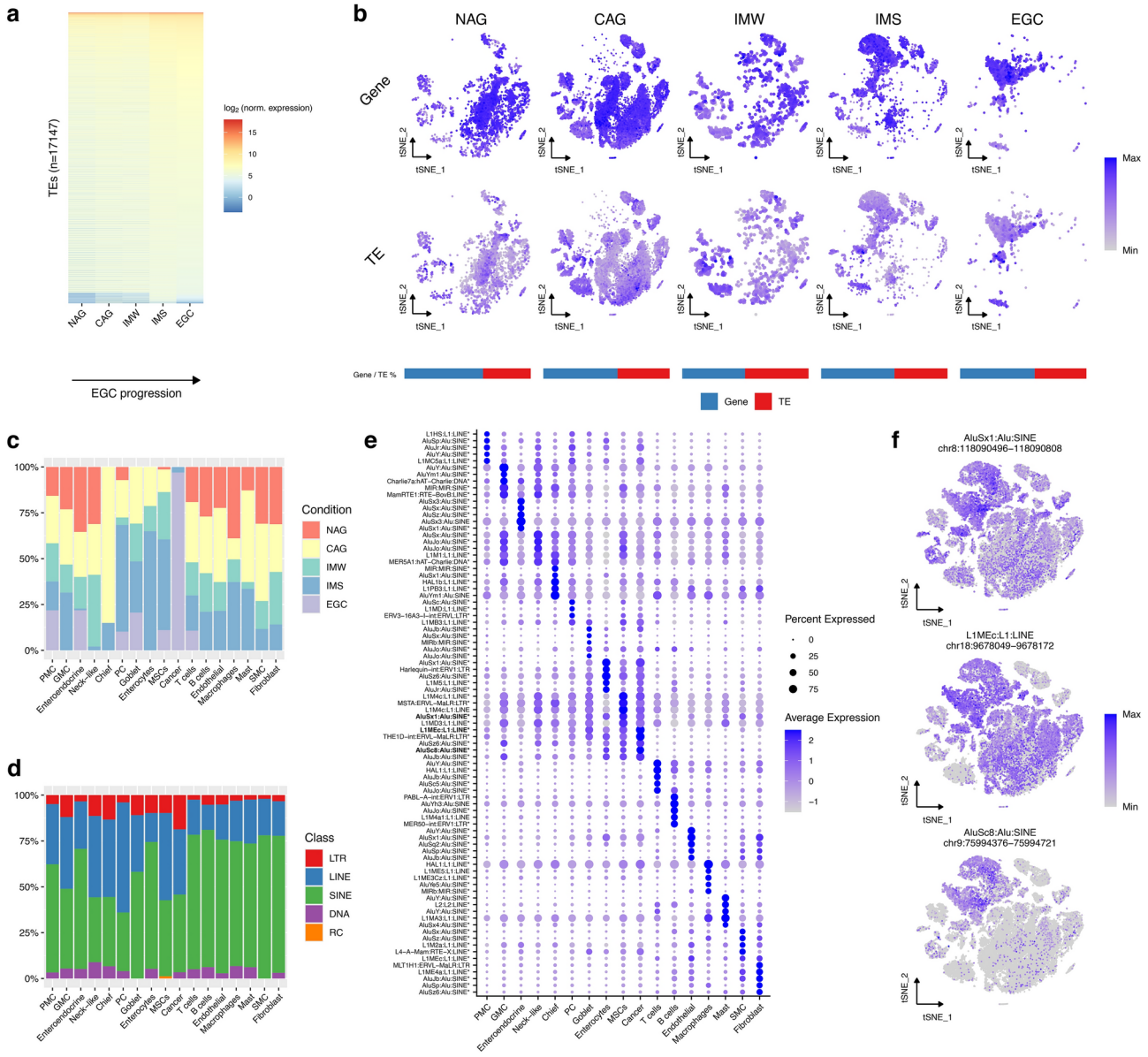

Fig. 3. The single-cell profile of TEs across EGC progression. (**a**) Pseudobulked $\log_2$-normalized expression of TEs that were either differentially up-regulated at any time point with respect to NAG, or that were in the upper 5% of normalized expression at all time points. (**b**) tSNE plots showing the per-cell expression percentage of genes (first row) or TEs (second row) across all timepoints. The horizontal bars show the percentage of the transcriptome at each timepoint corresponding to genes (blue) or TEs (red). (**c**) Distribution of marker TEs across stages. (**d**) Class distribution of marker TEs of each cell type. (**e**) Dot plot depicting the top 5 marker TEs of each cell type. In bold, TEs selected to show as example in the following tSNE plots. (**f**) tSNE plots of selected TEs.

PMC, MSCs, Cancer, Chief and Neck-like markers belong to group 1, while the top T cells and B cells markers belong to group 2 (Fig. 3e, Supplementary Fig. 4, Supplementary data 2). In turn, group 1 would define markers that have a gradient of expression, and in some cases, might be statistically enriched in more than one cell type (Table 1, Supplementary Data 2). As mentioned earlier, in the original work, it was observed that MSCs have a high transcriptional similarity to Cancer cells[6]. Here, I also noted that in terms of enriched TEs, there is also transcriptional similarity between these cell types, as indicated by the number of shared markers between them (29 common marker TEs, Supplementary Data 2). Similarly, it has been previously indicated that neck cells differentiate into chief cells[6], and marker expression seem to coincide with this: the top neck cell markers are broadly expressed, while the top chief cells markers seem a bit more specific. To further illustrate these results, tSNE dimensional reduction expression is depicted for 3 examples (Fig. 3f): AluSx1, located in chr8:118,090,496-118,090,808; L1MEc, located in chr18:9,678,049-9,678,172; and AluSc8 located in chr9:75,994,376-75,994,721. AluSx1 can be classified to group 1 considering that it is expressed in a high number of cells, but enriched in

Cluster	Total marker TEs	Locus-specific marker TEs	Unique marker TEs	Unique Locus-specific marker TEs
Pit mucous cells	61	55 (90.164%)	45	44 (97.778%)
Gland mucous cells	92	73 (79.348%)	38	31 (81.579%)
Enteroendocrine	174	115 (66.092%)	105	85 (80.952%)
Neck-like	79	61 (77.215%)	29	22 (75.862%)
Chief	45	39 (86.667%)	23	19 (82.609%)
Proliferative cells	25	22 (88.000%)	4	4 (100.000%)
Goblet	55	53 (96.364%)	21	21 (100.000%)
Enterocytes	133	126 (94.737%)	114	110 (96.491%)
Metaplastic stem-like Cells	82	69 (84.146%)	33	28 (84.848%)
Cancer	59	53 (89.831%)	26	25 (96.154%)
T cells	356	312 (87.640%)	270	249 (92.222%)
B cells	227	209 (92.070%)	146	131 (89.726%)
Endothelial	214	158 (73.832%)	105	94 (89.524%)
Macrophages	163	135 (82.822%)	76	66 (86.842%)
Mast	163	158 (96.933%)	121	119 (98.347%)
Smooth muscle cells	146	129 (88.356%)	55	55 (100.000%)
Fibroblasts	356	278 (78.090%)	198	156 (78.788%)
Total	2430	2045 (84.156%)	1409	1259 (89.354%)

Table 1. Marker TEs per cell cluster. For each cluster, the number of total or unique marker TEs is shown, along with the number, and the respective proportion of locus-specific marker TEs.

MSCs. The same argument applies to L1MEc with high expression in Cancer cells, but expressed at lower levels in other cells. Finally, AluSc8, also appearing as a Cancer marker, shows expression restricted to the region comprised by Cancer, MSCs and Enterocytes. It is worth noting that the expression of these TEs was measured with locus resolution (i.e., having only uniquely mapped reads), making unlikely that the observed expression heterogeneity could be attributed to ambiguity in read assignment. As suggested earlier, such gradient of TE expression spanning from PMC, GMC to MSC and then Cancer could suggest intermediate status between these cell types that could be mediated by TEs. It has been reported that as EGC progresses, PMCs decrease and MSCs increase, and that they have some transcriptional similarity[6], though it is unclear whether some PMC cells might be undergoing a malignant transformation to MSCs. If this is the case, these results would suggest that TEs are playing a role in that event.

Altogether, the evidence presented shows that TEs are expressed throughout the progression from gastritis to GC, and that their expression characterizes the different cell types, potentially highlighting intermediate status. In addition, the high proportion of TEs identified with locus resolution suggest that TEs becoming transcriptionally active in EGC have accumulated discriminative mutations, allowing the unambiguous assignment of sequencing reads. In turn, this would support the idea that TEs are playing a role in EGC progression via epigenetic polymorphisms, where changes in the transcriptional activity of fixed TE copies characterize cellular differences[30].

TE expression is associated with the origin of cancer cells

After getting a global overview of TE expression, and assessing their cellular profile, I then asked if their expression is associated with the origin of cancer cells. To this end, I applied Trajectory Inference (TI) to model the pseudotemporal progression of cells from the normal to the malignant status using the *dynverse* R package[15]. Briefly, this package evaluates more than 50 TI methods and identifies the one most suitable to the dataset, which in this case was PAGA-Tree[41]. Then, *dynverse* represents the trajectory topology in a network of milestones where the cells are placed. The resulting trajectory was rooted at the milestone containing the highest number of NAG cells, and further analyzed (Fig. 4a).

The trajectory broadly recapitulates the progression from NAG to EGC: the majority of cancer cells appear in a lineage that originates from IMS cells, which in turn, originates from CAG cells. Notably, there are 2 milestones with a high number of IMS cells, with one of them continuing directly to the Cancer milestone. To add support to the trajectory in the context of cancer evolution, I also applied inferCNV[42] to generate per-cell scores of copy number variations (CNVs), which has been used as proxy of malignancy development[13]. The inferCNV scores projected in the trajectory are also in concordance with the progression of CAG to IM, and subsequently to EGC (Fig. 4b). Cell type and marker gene analysis of the trajectory shows that the two branches with the highest inferCNV scores depict the transitions between different cell types (Supplementary Figs. 5 and 6). The first branch reveals a transition from Enterocytes, to MSCs and Cancer, while the second branch reveals a transition from a milestone comprised by PMCs and some Enterocytes to one comprised by MSCs and Goblet cells. Although these cell types share a link on the basis of their transcriptional similarity[6], it is unclear whether they represent the actual cell evolution in EGC. In cancer, plasticity allows tumor cells to change between cell status[43]. Thus, these observations might be indicative of cellular plasticity occurring during EGC, which would explain the cell type diversity in some branches with a malignant profile as revealed by their high CNV scores.

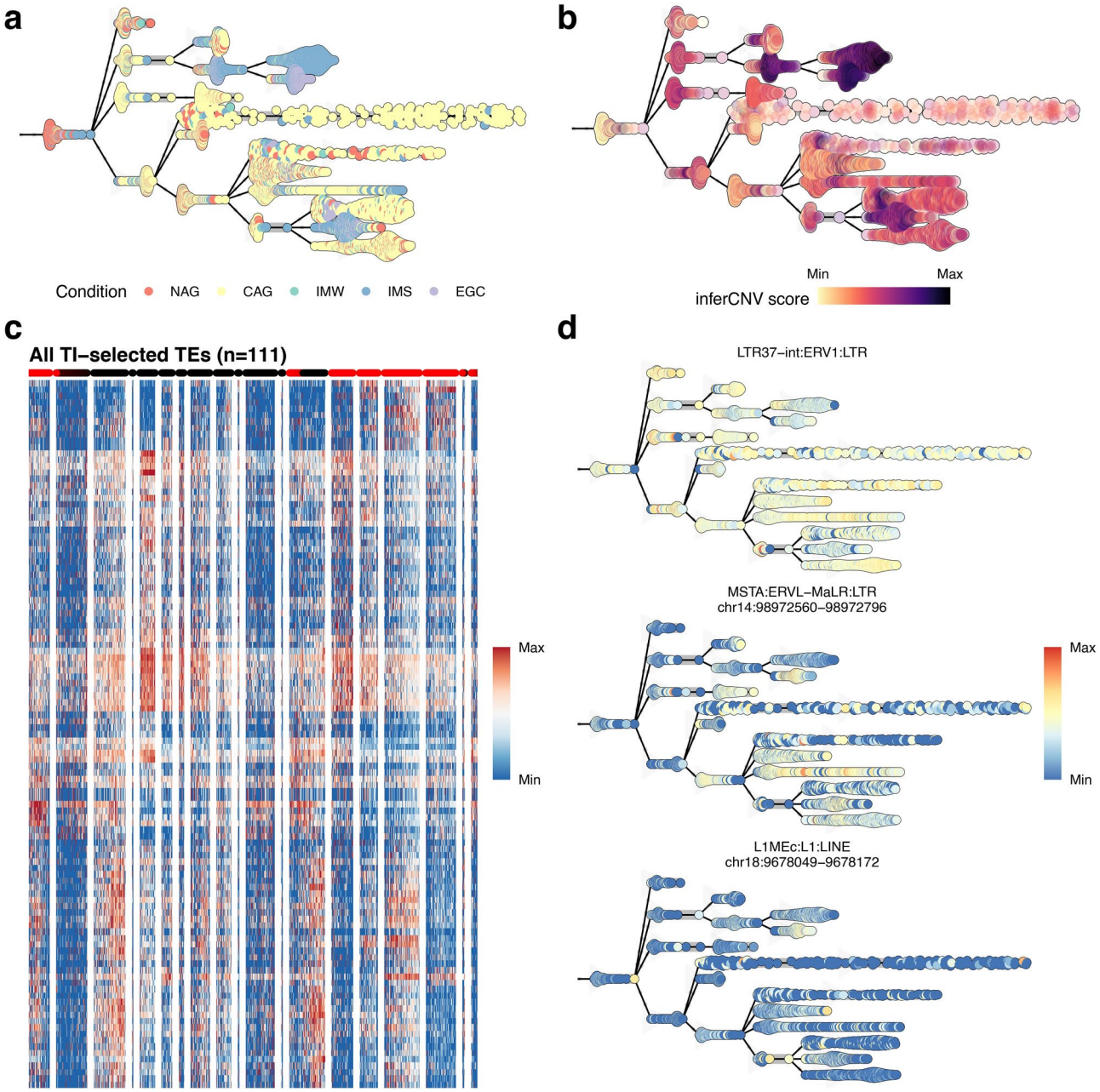

Fig. 4. The single-cell trajectory of EGC. (**a**) Inferred pseudotemporal trajectory of the EGC dataset colored by condition: NAG in orange, CAG in yellow, IMW in cyan, IMS in light blue and EGC in purple. (**b**) Pseudotemporal trajectory colored by inferCNV score: minimum values in light yellow, and maximum values in dark purple. (**c**) Heatmap showing the expression of TEs associated with the increase in cell malignancy as measured by inferCNV scores. Colored dots above the heatmap correspond to a one-dimensional representation of the trajectory. Highlighted in red are the cells that go from the beginning of the trajectory to the Cancer milestone (**d**) Example of TEs that are expressed in the malignant (i.e., high inferCNV score) branches of the trajectory.

To better understand the impact of TE expression in the EGC cascade, the enrichment of TEs in the malignant milestones was assessed. TE expression seems to occur throughout the entire trajectory, and 111 were TI-selected TEs (Fig. 4c, Supplementary Fig. 7), with 50 of them corresponding to markers of different cell populations and the remaining 61 without significant cell type specific expression (Supplementary data 3). Broadly speaking, three types of expression patterns throughout the trajectory were revealed by this analysis: 1. TE showing high and consistent expression (Fig. 4d, "LTR37-int"), 2. TE showing moderate levels of expression (Fig. 4d, "MSTA" located in chr14:98,972,560-98,972,796) and 3. TE expression mostly restricted to the malignant milestone (Fig. 4d, "L1MEc", located in chr18:9,678,049-9,678,172). For example, "L1MEc" also appeared in the marker analysis as a Cancer cells-enriched TE (depicted in Fig. 3d), indicating some agreement

between the 2 analyses. Notably, the 50 TEs that are also cell population markers all have locus resolution, while none of the other 61 have. The lack of locus resolution is usually indicative of expression of TE copies with a negligible number of discriminative mutations (i.e., evolutionary younger copies), hence the difficulty on accurately assigning reads to specific instances in the genome[26]. In this line, a possible scenario is that many of these copies become transcriptionally activated concurrently, pointing out to a pattern of global activation. For example, the widespread expression of "LTR37-in", a TE belonging to the group of retroviruses, would be in agreement with evidence indicating that HERVs undergo global activation in cancer[44].

Collectively, this analysis showed that by leveraging TI methods, TEs are potentially contributing to the evolution of cancer cells and to the transition and interplay between cell status during EGC progression. In turn, the detection of TEs in the premalignant milestones might be informative to their role as biomarkers for the early detection of GC.

The spatial portrait of TE expression in GC

The single-cell findings indicate that TEs are expressed in early pre-malignant GC stages, and that their expression occurs in the different cell types, hinting at a role of TEs in the tumor microenvironment. However, two outstanding questions from these results are (1) is TE expression also occurring in the malignant GC? and (2) are TEs expressed in the tumor microenvironment? To address them, I studied 4 spatially-resolved transcriptomes from GC sections[14]. In the original study, the tumor and surrounding regions in these sections were annotated by pathologists. By leveraging the annotations, I assessed if TE expression is associated with the tumor. To calculate TE expression, I processed the dataset with SoloTE, and the resulting gene + TE expression matrices were further analyzed with STutility[45]. With STutility, I characterized the in situ expression of TEs and contrasted it with the pathologist-annotated tumor and normal epithelium regions. First, I studied global TE expression across all tissue sections to assess the extent of TE activation in GC.

All 4 samples exhibit TE expression that is higher in the tumor regions and lower in the normal epithelium, indicating that enhanced TE expression is a hallmark of GC (Fig. 5a, "All TEs"). Furthermore, activation of TEs in the tumor region suggests that these elements could be involved in tumorigenesis. When observing the expression at the major TE type level, subtle patterns can be seen. For example, a clear enrichment of LTRs in tumors is revealed (Fig. 5, "LTR"), with LINEs and SINEs having a moderate increase in the tumor area, and higher expression in regions not annotated as normal nor as tumor (Fig. 5, "LINE" and "SINE", respectively). Interestingly, DNA TEs are also enriched in tumors, despite their lower presence in the human genome compared to retroelements. Nonetheless, at the statistical level their expression is still relatively increased when comparing tumor versus normal regions (Fig. 5b). Collectively, these results bridge the findings obtained in the single-cell section of this work, by showing that, in addition to being activated in early GC stages, TEs are also expressed in the malignant tumor regions.

The spatial transcriptomics analysis also revealed that TE expression extends beyond the tumor regions. Activation of TEs in the tumor microenvironment could also play a role in GC initiation and progression. As mentioned earlier, in the single-cell results it was also observed that TEs were enriched in non-tumor cell populations, hinting at this hypothesis. Furthermore, as observed in the hematoxylin and eosin-stained (H&E) tissue sections, the profiled sections display variability in their morphology (Fig. 5). In the original study, the pathologists selected the most heterogeneous gastric cancer sections in order to characterize intratumor heterogeneity[14]. Therefore, in addition to Normal epithelium (NE) and Tumor Tissue (TT), other regions identified in the tissue histological images were: Tumor tissue and glands (TTG), Intestinal metaplasia (IM), Serrated glandular structure (SGS), Lymphoid follicle (LF), Muscularis mucosa (MM), Peritumoral muscularis (PM), Muscle tissue (MT), Heterotopic cystic malformation (HCM), Lamina propria (LP), Blood – containing tissue (BCT), Connective tissue (CNT) (Fig. 6). Some of these regions were identified in one patient only: for example, Intestinal metaplasia in patient JJ, Serrated glandular structure in patient JJ62, and Heterotopic cystic malformation in patient ZL716, further highlighting intertumor diversity.

By leveraging the GC tissue annotations, I then investigated the spatial enrichment of specific TEs, in order to address whether TEs are also expressed in the tumor microenvironment. To this end, I applied the FindMarkers function to compare expression in the different tissue regions using the normal epithelium as control. This analysis showed substantial variability in the number of spatially enriched TEs detected on each sample: JJ62 – 148, JJ – 75, ZL716 – 57, ZL69 – 38 (Fig. 6, Supplementary Fig. 8, Supplementary data 4), which makes sense given the heterogeneity observed between the GC tissue sections. In agreement with the global TE analysis, specific TEs were enriched in the tumor regions of all samples. Sample JJ exhibits an overall increase in TE activity in the tumor and regions surrounding it, such as Lamina propria, Muscularis mucosa, Peritumoral muscularis, Lymphoid follicle, Connective tissue and Blood-containing tissue (Fig. 6a). Interestingly, this sample also has a region of Intestinal metaplasia, where enriched TEs were also identified. This finding provides further evidence of TE activation during intestinal metaplasia, as observed in the single-cell analyses. Additionally, Sample JJ62 shows a Serrated glandular structure next to the tumor region, and it has been reported that the region represents an intermediate step in the alteration of the normal epithelium during dysplasia, resulting in GC. TE activation between Serrated glandular structure and the tumor tissue was observed, implicating that TEs might be contributing to GC development (Fig. 6b). Interestingly, the Lymphoid follicles and muscular tissue surrounding the tumor also seems to show activation of TEs. The enhanced TE expression in the muscle tissue would match the single-cell result for smooth muscle cells. Sample ZL79 corresponds mostly to tumor tissue and Muscularis mucosa, with both regions having a signature of TE expression (Fig. 6c). Sample ZL716 display a large Heterotopic cystic malformation region with several Lymphoid follicles. Heterotopic cystic malformation might be associated with early GC[46], and for most part, TEs detected in this sample seem to be expressed across it and the tumor regions (Fig. 6d). In terms of TE types, both intra- and inter-tumor, there seems to be as similar distribution, in which LINEs and SINEs are the most represented, followed by LTRs, and then DNA TEs.

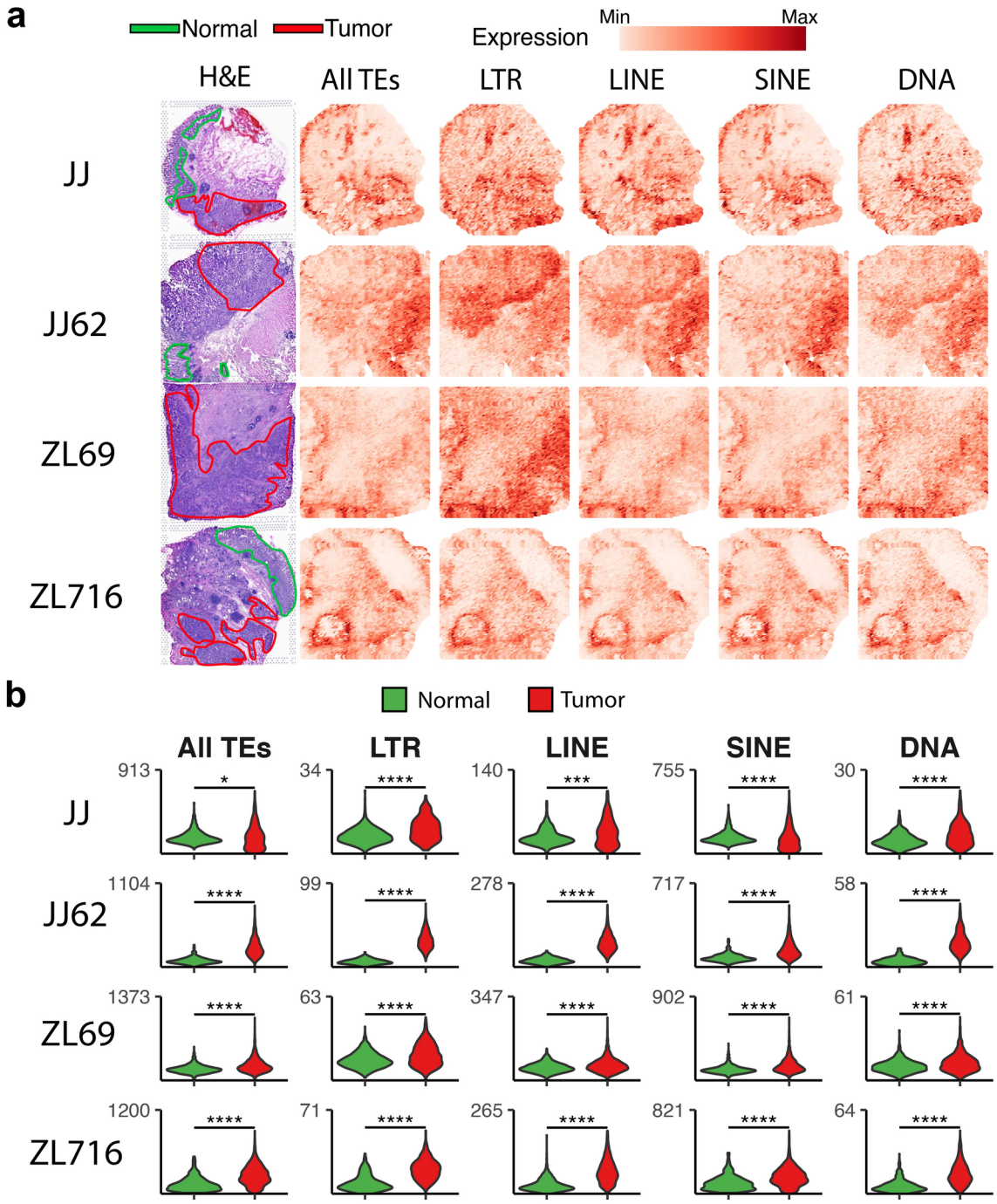

Fig. 5. TEs are enriched in the tumor regions of GC tissue sections. (**a**) Spatially-resolved expression of TEs when analyzed collectively ("All TEs), or at the class level (LTR, LINE, SINE, DNA). "H&E" column shows the pathologist-annotated normal and tumor regions in green and red, respectively. (**b**) Violin plots showing TE expression in the normal and tumor regions. Asterisks denote statistical significance at the following levels: $^{****}p \leq 0.0001$, $^{***}p > 0.0001$ and $p \leq 0.001$, $^{**}p > 0.001$ and $p \leq 0.01$, $^{*}p > 0.01$ and $p \leq 0.05$. Annotation of H&E images was done with Inkscape 1.3.2 (https://inkscape.org).

Interestingly, Muscularis mucosa seem to be characterized by expression of LTRs and SINEs, and in the case of sample ZL716, only LTRs. LTRs have long been associated with cancer[47], and alterations in the muscularis mucosa might play a role in early GC[48]. Thus, this finding could suggest that aberrant activation of LTRs in this region could be associated with GC. In sum, these results indicate that TEs are expressed in both the tumor tissue and in the tumor microenvironment, which suggests that TEs might also contribute to GC progression through changes to the regions surrounding the tumors.

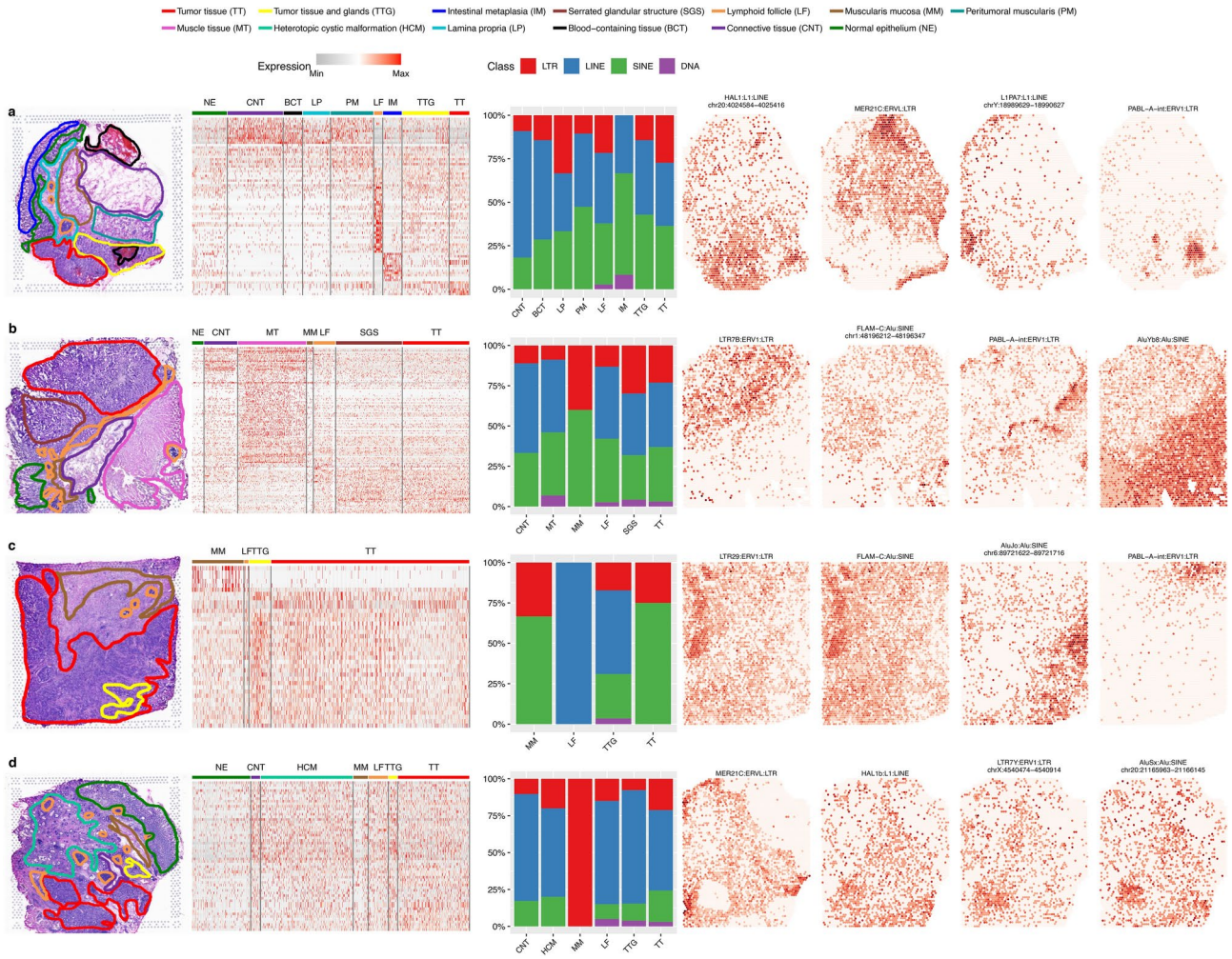

Fig. 6. The spatial portrait of Transposable Element expression in Gastric Cancer. (**a**) Annotated H&E image of GC tissue section of sample JJ, followed by the heatmap depicting the expression of spatially enriched TEs, the class distribution bar plots, and the spatial expression plots of representative TEs. (**b**) Annotated H&E image of GC tissue section of sample JJ62, followed by the heatmap depicting the expression of spatially enriched TEs, the class distribution bar plots, and the spatial expression plots of representative TEs. (**c**) Annotated H&E image of GC tissue section of sample ZL69, followed by the heatmap depicting the expression of spatially enriched TEs, the class distribution bar plots, and the spatial expression plots of representative TEs. (**d**) Annotated H&E image of GC tissue section of sample ZL716, followed by the heatmap depicting the expression of spatially enriched TEs, the class distribution bar plots, and the spatial expression plots of representative TEs. Annotation of H&E images was done with Inkscape 1.3.2 (https://inkscape.org).

Next, I asked if there are TEs conserved across the studied tissue sections. To this end, the number of TEs common between different combinations of samples was visualized in an upset plot (Fig. 7a). By using this result as guide, it was observed that 33 spatially enriched TEs appear in at least 3 out of the 4 samples, and these were selected to build the set of "top" spatial TEs. Visualization of the TE type distribution indicates that despite the difference in number of enriched TEs detected on each sample, LINEs seem to be predominant, followed by SINEs, and then LTR and DNA (Fig. 7b). In the top set, DNA TEs do not appear, indicating that their activation follows the inter-patient GC heterogeneity captured in these samples. Conversely, there are 9 leading TEs in the top set because they were detected in all the samples (Fig. 7c). These 9 TEs correspond to 1 LTR, 4 LINEs and 4 SINEs, with 6 of them having locus resolution. In contrast with the single-cell analysis, these results show more sparsity in terms of the number of TEs whose location is detected unambiguously in the genome. For instance, in the total 33 TEs of the top set, only 15 (45.5%) have locus resolution, with 6 of them being amongst the leading 9 TEs present in all datasets. Nonetheless, when assessing the overlap between the top spatial TEs and the single-cell results, 29 (87.9%) out of the 33 top TEs were also detected in the single-cell analysis, with 22 (66.7%) of these TEs also associated with the malignant milestones detected in the trajectory analysis (Supplementary Fig. 9).

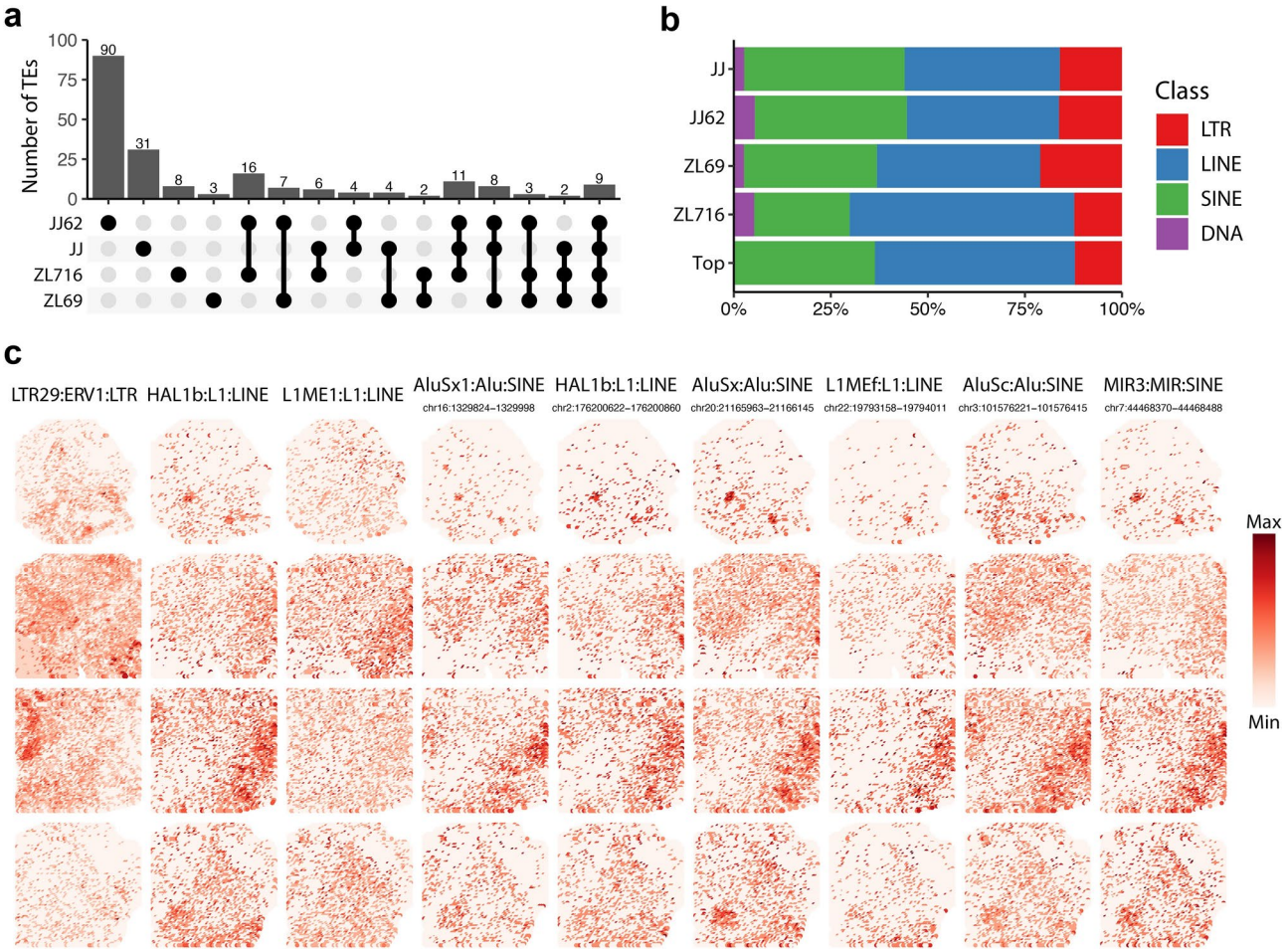

Fig. 7. Spatially enriched TEs common between the GC tissue sections. (**a**) Upset plot of the spatially enriched TEs. The number of TEs at each given overlap, represented by a set of connected dots in the lower half, is indicated as a bar plot. (**b**) Class distribution of spatially enriched TEs on each sample, and for the top TEs ("Top") which corresponds to those identified in at least 3 out of the 4 samples. (**c**) Feature plots of spatially enriched TEs identified in all samples, depicting their expression across the tissue sections. For TEs identified with locus resolution, their genomic location is indicated under their name.

The TE-mediated gene regulatory network during GC

To ask about the potential role of TEs as regulators of gene expression, a methodology similar to previous works[13,44] was applied: (1) using all the locus-specific TEs detected in both the single-cell and the spatial analysis, a gene-TE dictionary was built, considering all genes within 500 kbp of each TE; (2) Afterwards, to characterize the regulatory potential of TEs, the overlap with regulatory elements in GeneHancer and the ENCODE SCREEN candidate Cis-Regulatory Elements (cCREs) was assessed; (3) the gene-TE dictionary was filtered by considering either pre-defined interactions in GeneHancer or SCREEN, or if the gene was within 50 kbp of the TE; (4) finally, the Spearman correlation values were calculated for each gene-TE pair, and all interactions with correlation ≥ 0.3 were kept.

To get an overview of the genes potentially regulated by TEs, two additional analyses were carried out. First, an automated literature search using the NCBI E-Utilities[49] was carried out, and genes associated with publications in cancer were labeled as "Cancer gene". Also, gene set enrichment analysis using the fgsea R package[50] was performed, using the Kyoto Encyclopedia of Genes and Genomes (KEGG) terms, and the Gene Ontology (GO) terms, and filtering results to those having adjusted p-value ≤ 0.05.

Following the aforementioned approach, a total of 3151 interactions were predicted. Based on the genomic overlap with known regulatory elements from the GeneHancer and ENCODE SCREEN database, the TEs potentially interacting with genes are distributed as follows: 1142 "TE GeneHancer", 572 "TE ENCODE cCREs", and 1437 as "TE coexpression" (i.e., they only met the correlation threshold) (Supplementary data 5). A total of 1992 unique genes are regulated by TEs, with 573 of them labeled as "Cancer gene". 210 unique genes (59 being Cancer genes) have 3 or more interactions with TEs, amounting to a total of 997 interactions distributed in 87 modules (Fig. 8a). TEs in these modules are distributed into 445 "TE GeneHancer", 96 "TE ENCODE cCREs" and 454 "TE coexpression" Out of the 210 genes, 153 interact with either a TE GeneHancer or TE ENCODE cCREs, providing further support to the hypothesis of TEs playing a regulatory role in GC. The remaining

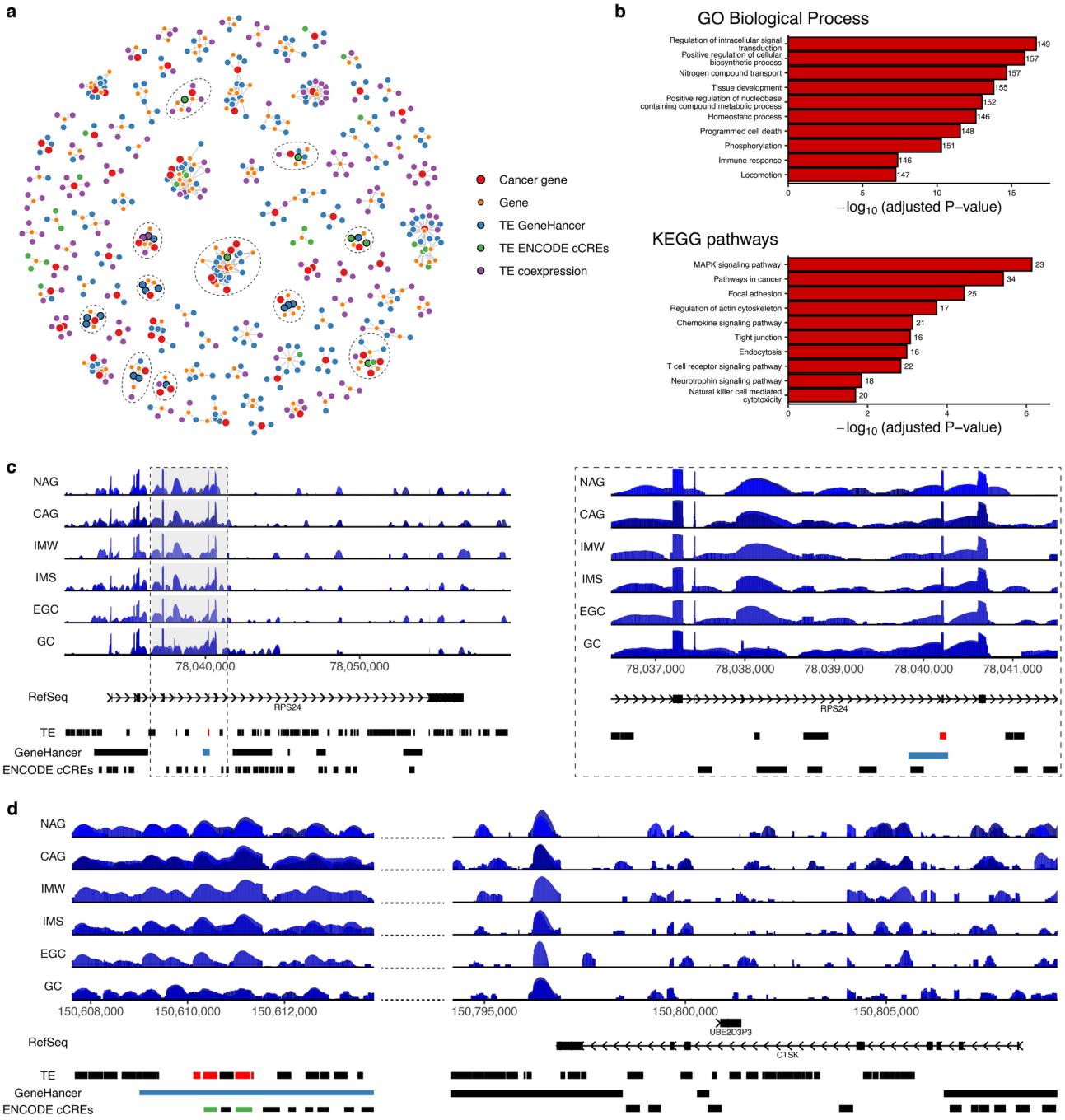

Fig. 8. Network analysis of Gastric Cancer TEs. (**a**) Regulatory network of genes associated with TEs. Genes previously associated with cancer are shown in red, and the remaining genes are shown in orange. TEs are colored according to their predicted regulatory link: TE GeneHancer in blue, TE ENCODE cCREs in green, and TE coexpression in purple. TEs that are module hubs are highlighted with black border, and the respective module is encircled in dashed lines. (**b**) Top 10 enriched gene set terms for the Gene Ontology (GO) Biological Process category (upper half) and KEGG pathways category (lower half). (**c**) Genome coverage plot of the intronic TE MamRTE1:RTE-BovB:LINE in locus chr10:78,040,182-78,040,254 (highlighted in red), having a predicted regulatory link with the RPS24 gene. On the right, the region enclosed with the dashed-lines rectangle is shown zoomed. (**d**) Genome coverage plot of the region chr1:150,610,108-150,809,260, depicting upstream TEs (highlighted in red) having a predicted enhancer regulatory link with the CTSK gene. GeneHancer and ENCODE regulatory elements overlapping the selected TEs are shown in blue and green, respectively.

interactions met the correlation threshold, and the gene is in close genomic vicinity of the TEs, consistent with previous findings[13,36] suggesting such a regulatory role for TEs. In addition, to gain an overview of the regulatory impact of TEs, the network modules were characterized according to the presence of a "TE hub". For a module to be classified as "TE hub", a TE must account for the 50% of more of its interactions. Following, this approach, 11 "TE hub" modules were identified (Fig. 8a, modules encircled in dashed lines). Out of the 87 modules, 44 (50.6%) have cancer genes, with the 11 "TE hub" modules amongst them (Fig. 8a, modules detailed in Supplementary Fig. 10). Interestingly, some of the genes in these modules have been previously associated with GC, such as *CTSK*[51], *CLDN18*[52] , *HORMAD1*[53,54], *MS4A8* and *MS4A15*[55], and *SOCS3*[56].

At a wider scale, gene set enrichment analysis revealed many terms of relevance to cancer (Fig. 8b). For example, intracellular signal transduction, programmed cell death and phosphorylation are all processes long associated with cancer[57–59]. Pathway-level analysis also depicts changes previously associated with cancer. The MAPK signaling pathway plays roles in cell proliferation, differentiation, migration and apoptosis, and is the top pathway likely regulated by TEs. Because of its role in many cell processes, malfunction on the pathway has been linked to cancer[60]. "Pathways in cancer" is the second term, and it encompasses different pathways with evidence associating them with cancer. Thus, this result is in close agreement with the previous literature search analysis carried out independently. Another interesting example is "Endocytosis", that has many functions in nutrient uptake. Disruptions in this pathway also play a significant role in cancer, and there is evidence pointing to a role in GC[61,62]. Furthermore, when performing these analyses on each stage, it is revealed that these terms appear enriched from NAG, suggesting that TEs could potentially contribute to GC development by early alterations to gene expression (Supplementary Figs. 11–16).

Finally, two examples of genes regulated by TEs are depicted: *RPS24* (Fig. 8c) and *CTSK* (Fig. 8d). *RPS24* was chosen because the interacting TE appears in all 3 analyses (single-cell, TI and spatial), has a partial overlap with one of its exons (Fig. 8c, TE highlighted in red), and the regulatory link is based on GeneHancer. In turn, the result would suggest that the TE is acting as an enhancer of the same gene. This would be similar to a documented event occurring in the *EGFR* gene, causing its up-regulation, which in turn, contributes to breast cancer[63]. Also, over-expression of *RPS24* has been reported to be a biomarker of colorectal cancer[64]. In these lines, the TE-mediated up-regulation proposed in this study might also highlight a role of *RPS24* in GC progression. On the other hand, *CTSK* has been previously implicated in GC, and the associated TEs are located about ~ 200,000 kbp away from its locus, and have overlaps with ENCODE cCREs and GeneHancer elements, making this a good example of a long-distance regulatory link (Fig. 8d, Supplementary Fig. 17). *CTSK* has been reported as over-expressed in GC and has been proposed as GC biomarker[51,65]. Additionally, it is unclear how it becomes over-expressed[51]. This result would help bridge that gap, by indicating that up-regulation of *CTSK* is caused by TE-derived enhancers.

In sum, these results show that TEs are potentially regulating genes and pathways associated with cancer in several ways, strongly suggesting that TEs are implicated in the molecular aberrations that occur in GC.

Discussion

GC is commonly detected at advanced stages, point in which the survival rate is low. When detected on initial stages, survival rate is high, prompting the need for better understanding the molecular mechanisms associated with its development and discovery of early biomarkers. In this work, I studied single-cell and spatial RNA data publicly available from different cohorts and provide evidence pointing to TEs as potential biomarkers and regulators of gene expression during GC.

Here, I showed that TE expression is a hallmark of the early GC cascade and during GC. The single-cell analysis revealed that thousands of TEs are cell-type enriched, and their expression increases as gastritis progresses to early GC. Although LTR ERVs have received more attention in cancer studies[13,44], there is increasing evidence depicting changes of several types of TEs in these malignances[30]. Concordant with the latter evidence, in addition to LTRs, I also detected expression of LINE, SINE, and DNA TEs. Marker analysis revealed that TEs are enriched in the different cell populations to varying degrees. For example, the top Cancer markers also exhibit some level of expression in MSCs, in line with the proposed idea that MSCs might provide an environment for the origin of Cancer cells[6]. Afterwards, the application of TI methods revealed a branching trajectory, broadly recapitulating the cascade from NAG, to CAG, to IM and to EGC. Particularly, some cell lineages show a clear association with the acquisition of the malignant status, based on their emergence in EGC and their high CNV score. Analysis of TE expression in the malignant lineages revealed 111 TEs that could potentially be early biomarkers for GC detection due to having expression not only in the later trajectory milestones, but in those leading up to them. Despite *dynverse* allowing a convenient assessment of many TI methods, the trajectory generated for this work still has some limitations, considering that in some instances it seemed to indicate cellular plasticity (phenomenon known to occur in cancer[43]) rather than cell evolution. Nonetheless, the inferred trajectory is still informative and underlines an additional layer in which the study of TEs can provide insights to understanding GC progression.

Studying the spatial dataset generated from GC patients also confirmed TE expression. The importance of this is two-fold: first, it allowed bridge the gap between EGC and GC in terms of TEs, and second, it serves as additional and independent confirmation that changes in TE expression during GC development might indeed be representative of the malignancy. A caveat here is that differences were observed between patients, and overall, with the single-cell data. The diversity of TEs detected on each sample can be attributed to inter-patient and intra-tumor heterogeneity, and indeed in the original work it is reported that the regions with the most significant heterogeneity were selected for spatial sequencing[14]. On the other hand, the lower TE detection observed in spatial data versus the single-cell data could be attributed to differences in the techniques: Visium spatial RNA-Seq captures about 5–10 cells per spot, and by sequencing a combination of cells could reduce the detection of several transcripts, including TE-derived ones. Nonetheless, almost 90% of the top spatial TEs were

found in the single-cell data, and close to 70% are found in the list of TI-selected TEs, strongly suggesting that TE expression is a hallmark of GC. Furthermore, TE expression was also observed in the tumor microenvironment suggesting that they might be playing a significant role in GC tumorigenesis.

Finally, I adopted a gene-TE correlation approach to predict the impact of TEs in gene expression, which was coupled with the assessment of TEs as regulators via the overlap with GeneHancer and ENCODE cCREs. Close to 2,000 genes are likely regulated by TEs, and ~ 500 of them have been previously linked with cancer. Gene enrichment analysis also depict the global impact of these potential regulatory events, and several of the top terms and pathways have been also associated with cancer. This gene-TE correlation approach is similar to the one applied in gallbladder cancer, where they validated the regulatory potential of selected TEs[13]. There is a growing body of evidence supporting the role of TEs as regulators of gene expression either in their genomic vicinity or in long-distance locations by acting as enhancers[24,30,44], thus the findings reported here are in strong agreement with the idea of TEs involved in GC-related gene aberrations.

In conclusion, I show that TEs become activated during the progression of gastritis toward EGC, and in GC itself by leveraging datasets publicly provided by independent studies. I present evidence of their activation both in the tumor tissue and in the tumor microenvironment, pointing to a role in tumorigenesis. Furthermore, in addition to becoming activated, TEs might influence gene regulation. In turn, this could contribute to the progression of GC. These findings highlight the biological and functional importance of studying TEs in this malignancy. The portrait of TE expression during GC development shown here advances our understanding of the disease, and pinpoints these elements as potential biomarkers for its early detection.

Methods

Raw sequencing data

The single-cell and spatial RNA-Seq data used in this study were obtained from publicly available databases. Single-cell FASTQ files were obtained from[6], made publicly available at the Sequence Read Archive (SRA) under accession SRP215370. On the other hand, Spatial data was obtained from[14], publicly available at the National Genomics Data Center Genome Sequence Archive (GSA) under accession HRA003070.

TE expression analysis

To calculate TE expression on each dataset, the raw sequencing data was aligned to the human genome using STAR[66]. First, the hg38 genome FASTA and ncbiRefSeq GTF annotation were downloaded from UCSC Genome Browser database[67] and used to generate the genome index. Then, the sequencing data of both the single-cell and spatial experiments was aligned with the following options to generate BAM files compliant with SoloTE (described later): *--outSAMattributes NH HI nM AS CR UR CY UY CB UB GX GN sS sQ sM* to include cell barcode and UMI information in the output files, *--outFilterMultimapNmax 100 --winAnchorMultimapNmax 100* to increase sensitivity of alignment to TEs, *--outSAMmultNmax 1 --outMultimapperOrder Random* to keep only one random alignment for multimapped reads, *--runThreadN 21* to set the number of process threads to 21 and *--runRNGseed 777* to set a random number generator seed to a fixed value for reproducibility. Each alignment file was then processed with SoloTE v1.09[29], using the human genome hg38 version TE annotation in BED format obtained with the helper script *SoloTE_RepeatMasker_to_BED.py*. This process resulted in the raw count matrices that include TE expression.

Single-cell analysis

Analysis of single-cell count matrices generated above was carried out using the Seurat v4.1.0 package[68] of the R statistical computing environment[69] version 4.1.1, as described next. First, single-cell matrices were filtered to keep only the quality-control filtered cells reported in the original work. Afterwards, the matrices were merged in a single object and processed using the default Seurat workflow. Briefly, the object was used as input to NormalizeData, FindVariableFeatures and ScaleData, in order to prepare it for principal component analysis using the RunPCA function. Then, 30 dimensions were used for FindNeighbors, and RunTSNE, and the clustering was obtained with FindClusters. Per-sample pseudobulk count matrices were generated with the AggregateExpression function, using the sample identifier as "group.id". Then, the pseudobulked matrices were processed with DESeq2[34], and the log-normalized counts were obtained with the rlog function, which were used to produce a per-sample PCA. For Fig. 2, these steps were repeated 2 more times using a single-cell matrix with only genes and another one with only TEs.

To identify TEs whose expression increases in the progression towards early gastric cancer, DESeq2 was used again to test for differences between CAG, IM and EGC with respect to NAG, using adjusted p-value ≤ 0.05 as threshold for significance. In addition, to also include TEs highly expressed throughout all time points, those within the top 5% of expression were also selected. This list of EGC progression-associated TEs was used for marker analysis in the FindAllMarkers function to identify the specific cell populations in which they were enriched. Results of this step were also filtered using a threshold of adjusted p-value ≤ 0.05.

Trajectory Inference (TI) analysis was carried out using *dyno,* and related plots were generated using *dynplot*, both from the *dynverse* collection of R packages[15]. The single-cell expression matrix was subsetted to the epithelial subtypes (PMC, GMC, Enteroendocrine, Neck-like, Chief, PC, Goblet, Enterocytes, MSCs, and Cancer) and then processed with the Seurat integration protocol. The "paga_tree" TI method was used for TI as it was identified as the most suitable for the dataset using the "guidelines_shiny" function. *inferCNV v1.8.1*[42] was run to generate per-cell copy number variation scores, which were used as an additional validation of the inferred trajectory. The "branch_feature_importance" function was used to assess TEs enriched in the cell lineages associated with cancer progression.

Spatial analysis

Spatial data was processed with STutility v1.1.1[45], which is built in top of the Seurat package. First, for each of the 4 samples an object was created, and processed with the SCTransform function to obtain the normalized expression. Then, total TE and per-class (LTR, LINE, SINE and DNA) TE expression was calculated by aggregating all the normalized counts respectively. Statistical differences in TE expression between tumor and normal regions across the tissue were tested using the Wilcoxon test implemented in the base R function *wilcox.test*. The results of this step were depicted in the violin plots of Fig. 5b, highlighting the Wilcoxon test p-values obtained.

To find TEs with higher expression in tumor and non-normal regions of the tissues, differential expression analysis was carried out with the FindMarkers function. This was done by taking advantage of the pathologist annotations, using the normal epithelium regions as control (or the unannotated region in the case of sample ZL69) and each of the remaining regions as test groups. All TEs with adjusted p-value ≤ 0.05 were then selected as spatially enriched. The overlap between the spatially enriched TEs detected in each sample was assessed via an upset plot generated with the ggupset v0.3.0 package[70].

Network analysis

To build the TE-gene networks, a list of TEs was built from those enriched in the single-cell or spatial data. Selected single-cell TEs were those associated with the progression from NAG to EGC and that also appear as markers of cell populations, whereas selected spatial TEs were those found with spatial enrichment in at least 3 out of the 4 samples.

The selected TEs were then processed to identify potential regulatory links on the basis of their genomic location, using a methodology similar as previously published works[13,44]. First, interactions between regulatory elements and genes were obtained from GeneHancer and the SCREEN database. Particularly, GeneHancer v4.7 interactions were downloaded from https://genecards.weizmann.ac.il/geneloc/index.shtml, and the ENCODE SCREEN Registry of candidate Cis-Regulatory Elements (cCREs) V3[71] from https://screen.encodeproject.org/. Afterwards, a TE-gene dictionary was built containing all genes within 500 kbp from the selected TEs. Using BEDTools v2.30.0[72], the overlap between selected TEs and regulatory elements was assessed to filter and classify the TE-gene dictionary into potential interactions: "TE GeneHancer", if the TE overlapped with GeneHancer elements and the gene was a target of the regulatory element, "TE ENCODE cCREs" if the TE overlapped didn't overlap with GeneHancer elements, but overlapped with ENCODE cCREs and the gene was a target of the regulatory element, and finally into "TE coexpression" if they didn't have any overlap with regulatory elements, but the TE and gene were within 50 kbp of each other. Finally, the Spearman correlation between each TE-gene pair in the dictionary was calculated and those pairs with correlation ≥ 0.3 were selected to build the TE-gene network depicted in Fig. 7.

Gene set enrichment analysis

Gene set enrichment analysis was carried out as described before[44], using the "geseca" function of the fgsea v1.29.1 R package[50]. The genes used as input were the 1992 being associated with TEs during the previous step (i.e., those selected for the network analysis). For the reference gene groups, the msigdbr package[73] was used, which provides seamless integration of MsigDB[74] gene sets with fgsea. Particularly, the reference gene groups utilized were the curated gene sets of Kyoto Encyclopedia of Genes and Genomes (KEGG) terms (contained in the "C2" category), and the gene sets of Gene Ontology terms (contained in the "C5" category). The following parameters were specified: *minSize = 1, maxSize = 500, nPermSimple = 10,000* and *center = FALSE, scale = FALSE*. In turn, the analysis was carried out twice, first using to the KEGG gene sets as reference, and then using the GO terms as references. Enriched terms having adjusted p-value ≤ 0.05 were selected.

Plots

All plots were generated in the R statistical computing environment, using *ggplot2* v3.4.2[75], and extended with packages *ggupset* v0.3.0[76] (Fig. 7a), and *ggcoverage* v1.2.0[77] (Fig. 8c and Fig. 8d). Particularly, *ggcoverage* was used to plot the RNA-Seq coverage at regions chr10:78,032,363-78,058,306 and chr1:150,610,108-150,809,260. The region chr10:78,032,363-78,058,306 depicts the *RPS24* gene and an intronic TE, whereas the region chr1:150,610,108-150,809,260 depicts a potential long-range interaction between an upstream TE and the *CTSK* gene.

Statistics

Statistical analyses were performed using the R statistical computing environment version 4.1.1. Pseudobulk statistical analysis was performed with DESeq2, using as threshold for significance adjusted P-value ≤ 0.05. Similarly, single-cell marker and spatial enrichment was carried out using the Seurat functions FindAllMarkers and FindMarkers, respectively. Both functions use the Wilcoxon rank sum test, and significance was set at an adjusted P-value ≤ 0.05. To test for overall differences in TE expression between the tumor and normal epithelium regions for the analysis shown in Fig. 5, the Wilcoxon test was used, and differences having P-value ≤ 0.05 were labelled as significant.

Data availability

The single-cell data is publicly available at Sequence Read Archive (SRA) under accession SRP215370. The spatial data is publicly available at the National Genomics Data Center Genome Sequence Archive (GSA) under accession HRA003070.

Received: 9 May 2024; Accepted: 20 September 2024
Published online: 30 September 2024

References

1. Hirata, Y., Noorani, A., Song, S., Wang, L. & Ajani, J. A. Early stage gastric adenocarcinoma: Clinical and molecular landscapes. *Nat. Rev. Clin. Oncol.* **20**, 453–469. https://doi.org/10.1038/s41571-023-00767-w (2023).
2. Alsina, M., Arrazubi, V., Diez, M. & Tabernero, J. Current developments in gastric cancer: From molecular profiling to treatment strategy. *Nat. Rev. Gastroenterol. Hepatol.* **20**, 155–170. https://doi.org/10.1038/s41575-022-00703-w (2023).
3. Ajani, J. A. et al. Gastric adenocarcinoma. *Nat. Rev. Dis. Primers* **3**, 1 (2017).
4. Waddingham, W. et al. Recent advances in the detection and management of early gastric cancer and its precursors. *Frontline Gastroenterol.* **12**, 322–331. https://doi.org/10.1136/flgastro-2018-101089 (2021).
5. Rugge, M., Fassan, M. & Graham, D. Y. Epidemiology of gastric cancer. In *Gastric Cancer: Principles and Practice* 23–34 (Springer, 2018). https://doi.org/10.1007/978-3-319-15826-6_2.
6. Zhang, P. et al. Dissecting the single-cell transcriptome network underlying gastric premalignant lesions and early gastric cancer. *Cell Rep.* **27**, 1934-1947.e5 (2019).
7. Kim, J. et al. Single-cell analysis of gastric pre-cancerous and cancer lesions reveals cell lineage diversity and intratumoral heterogeneity. *NPJ Precis. Oncol.* **6**, 9 (2022).
8. Li, A. et al. Identification and validation of key genes associated with pathogenesis and prognosis of gastric cancer. *PeerJ* **11**, e16243 (2023).
9. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
10. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: The teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).
11. Ståhl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **1979**(353), 78–82 (2016).
12. Lee, M. C. W. et al. Single-cell analyses of transcriptional heterogeneity during drug tolerance transition in cancer cells by RNA sequencing. *Proc. Natl. Acad. Sci. U S A* **111**, E4726–E4735 (2014).
13. Wang, J. et al. Single-cell RNA sequencing highlights the functional role of human endogenous retroviruses in gallbladder cancer. *EBioMedicine* **85**, 104319 (2022).
14. Sun, C. et al. Spatially resolved multi-omics highlights cell-specific metabolic remodeling and interactions in gastric cancer. *Nat. Commun.* **14**, 2692 (2023).
15. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
16. Fan, J., Slowikowski, K. & Zhang, F. Single-cell transcriptomics in cancer: Computational challenges and opportunities. *Exp. Mol. Med.* **52**, 1452–1465 (2020).
17. Liu, Z. L. et al. Single cell deciphering of progression trajectories of the tumor ecosystem in head and neck cancer. *Nat. Commun.* **15**, 2595 (2024).
18. Xue, J. Y. et al. Rapid non-uniform adaptation to conformation-specific KRAS(G12C) inhibition. *Nature* **577**, 421–425 (2020).
19. Kim, N. et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.* **11**, 2285 (2020).
20. Sagar, & Grün, D. Deciphering cell fate decision by integrated single-cell sequencing analysis. *Annu. Rev. Biomed. Data Sci.* **3**, 1–22 (2020).
21. Longo, S. K., Guo, M. G., Ji, A. L. & Khavari, P. A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat. Rev. Genet.* **22**, 627–644 (2021).
22. Siomi, M. C., Sato, K., Pezic, D. & Aravin, A. A. PIWI-interacting small RNAs: The vanguard of genome defence. *Nat. Rev. Mol. Cell Biol.* **12**, 246–258 (2011).
23. Senft, A. D. & Macfarlan, T. S. Transposable elements shape the evolution of mammalian development. *Nat. Rev. Genet.* **22**, 691–711 (2021).
24. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: From conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
25. Lemerle, E. & Trompouki, E. Transposable elements in normal and malignant hematopoiesis. *DMM Dis. Models Mech.* **16**, dmm050170 (2023).
26. Lanciano, S. & Cristofari, G. Measuring and interpreting transposable element expression. *Nat. Rev. Genet.* **21**, 721–736 (2020).
27. Valdebenito-Maturana, B., Torres, F., Carrasco, M. & Tapia, J. C. Differential regulation of transposable elements (TEs) during the murine submandibular gland development. *Mob. DNA* **12**, 23 (2021).
28. Valdebenito-Maturana, B., Arancibia, E., Riadi, G., Tapia, J. C. & Carrasco, M. Locus-specific analysis of transposable elements during the progression of ALS in the SOD1G93A mouse model. *PLoS One* **16**, e0258291 (2021).
29. Rodríguez-Quiroz, R. & Valdebenito-Maturana, B. SoloTE for improved analysis of transposable elements in single-cell RNA-Seq data using locus-specific expression. *Commun. Biol.* **5**, 1063 (2022).
30. Liang, Y., Qu, X., Shah, N. M. & Wang, T. Towards targeting transposable elements for cancer therapy. *Nat. Rev. Cancer* **24**, 123–140. https://doi.org/10.1038/s41568-023-00653-8 (2024).
31. MacLean, A. L., Hong, T. & Nie, Q. Exploring intermediate cell states through the lens of single cells. *Curr. Opin. Syst. Biol.* **9**, 32–41 (2018).
32. Ye, M. et al. Specific subfamilies of transposable elements contribute to different domains of T lymphocyte enhancers. *Proc. Natl. Acad. Sci.* **117**, 7905–7916 (2020).
33. Bonté, P.-E. et al. Selective control of transposable element expression during T cell exhaustion and anti–PD-1 treatment. *Sci. Immunol.* **8**, 88 (2023).
34. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
35. Jang, H. S. et al. Transposable elements drive widespread expression of oncogenes in human cancers. *Nat. Genet.* **51**, 611–617 (2019).
36. Lykoskoufis, N. M. R., Planet, E., Ongen, H., Trono, D. & Dermitzakis, E. T. Transposable elements mediate genetic effects altering the expression of nearby genes in colorectal cancer. *Nat. Commun.* **15**, 749 (2024).
37. Chen, C., Ara, T. & Gautheret, D. Using alu elements as polyadenylation sites: A case of retroposon exaptation. *Mol. Biol. Evol.* **26**, 327–334 (2009).
38. Lavi, E. & Carmel, L. Alu exaptation enriches the human transcriptome by introducing new gene ends. *RNA Biol.* **15**, 1–11. https://doi.org/10.1080/15476286.2018.1429880 (2018).
39. Kiyose, H. et al. Comprehensive analysis of full-length transcripts reveals novel splicing abnormalities and oncogenic transcripts in liver cancer. *PLoS Genet.* **18**, e1010342 (2022).
40. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703. https://doi.org/10.1038/nrg2640 (2009).
41. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).

42. Tickle, T., Tirosh, I., Georgescu, C., Brown, M. & Haas, B. *inferCNV of the Trinity CTAT Project*.
43. Pérez-González, A., Bévant, K. & Blanpain, C. Cancer cell plasticity during tumor progression, metastasis and response to therapy. *Nat. Cancer* **4**, 1063–1082. https://doi.org/10.1038/s43018-023-00595-y (2023).
44. Ito, J. et al. Endogenous retroviruses drive KRAB zinc-finger protein family expression for tumor suppression. *Sci. Adv.* **6**, eabc3020 (2020).
45. Bergenstråhle, J., Larsson, L. & Lundeberg, J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC Genomics* **21**, 482 (2020).
46. Niizawa, M., Ishida, H., Morikawa, P., Watanabe, H. & Masamune, O. Diffuse heterotopic submucosal cystic malformation of the stomach: Ultrasonographic diagnosis. *Gastrointest. Radiol.* **17**, 9–12 (1992).
47. Liang, Y., Qu, X., Shah, N. M. & Wang, T. Towards targeting transposable elements for cancer therapy. *Nat. Rev. Cancer* https://doi.org/10.1038/s41568-023-00653-8 (2024).
48. Cho, J. H. & Lee, S. H. Early gastric cancer presenting as a typical submucosal tumor cured by endoscopic submucosal dissection: A case report. *World J. Gastroenterol.* **28**, 2994–3000 (2022).
49. Kans, J. *Entrez Direct: E-utilities on the Unix Command Line*. https://www.ncbi.nlm.nih.gov/books/NBK179288/ (2024).
50. Sergushichev, A. A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv* https://doi.org/10.1101/060012 (2016).
51. Bai, Z., Yan, C. & Chang, D. Prediction and therapeutic targeting of the tumor microenvironment-associated gene CTSK in gastric cancer. *Discov. Oncol.* **14**, 200 (2023).
52. Kim, H.-D. et al. Clinicopathologic features and prognostic value of claudin 18.2 overexpression in patients with resectable gastric cancer. *Sci. Rep.* **13**, 20047 (2023).
53. El-Botty, R. et al. HORMAD1 overexpression predicts response to anthracycline–cyclophosphamide and survival in triple-negative breast cancers. *Mol. Oncol.* **17**, 2017–2028 (2023).
54. Bian, G. et al. The cancer/testis antigen HORMAD1 promotes gastric cancer progression by activating the NF-κB signaling pathway and inducing epithelial-mesenchymal transition. *Am. J. Transl. Res.* **15**, 5808–5825 (2023).
55. Sun, L., Zhang, Y. & Zhang, C. Distinct expression and prognostic value of MS4A in gastric cancer. *Open Med.* **13**, 178–188 (2018).
56. Dai, L. et al. Emerging roles of suppressor of cytokine signaling 3 in human cancers. *Biomed. Pharmacother.* **144**, 112262 (2021).
57. Singh, V. et al. Phosphorylation: Implications in cancer. *Protein J.* **36**, 1–6 (2017).
58. Ouyang, L. et al. Programmed cell death pathways in cancer: A review of apoptosis, autophagy and programmed necrosis. *Cell Prolif.* **45**, 487–498 (2012).
59. Adjei, A. A. & Hidalgo, M. Intracellular signal transduction pathway proteins as targets for cancer therapy. *J. Clin. Oncol.* **23**, 5386–5403 (2005).
60. Dhillon, A. S., Hagan, S., Rath, O. & Kolch, W. MAP kinase signalling pathways in cancer. *Oncogene* **26**, 3279–3290 (2007).
61. Banushi, B., Joseph, S. R., Lum, B., Lee, J. J. & Simpson, F. Endocytosis in cancer and cancer therapy. *Nat. Rev. Cancer* **23**, 450–473 (2023).
62. Yoon, C. et al. Role of Rac1 pathway in epithelial-to-mesenchymal transition and cancer stem-like cell phenotypes in gastric adenocarcinoma. *Mol. Cancer Res.* **15**, 1106–1116 (2017).
63. McInerney, J. M., Wilson, M. A., Strand, K. J. & Chrysogelos, S. A. A strong intronic enhancer element of the EGFR gene is preferentially active in high EGFR expressing breast cancer cells. *J. Cell Biochem.* **80**, 538–549 (2001).
64. Zou, D. et al. Three functional variants were identified to affect RPS24 expression and significantly associated with risk of colorectal cancer. Arch. Toxicol.**94**, 295–303 (2020).
65. Feng, Z. et al. Could CTSK and COL4A2 be specific biomarkers of poor prognosis for patients with gastric cancer in Asia?-a microarray analysis based on regional population. *J. Gastrointest. Oncol.* **11**, 386–401 (2020).
66. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
67. Nassar, L. R. et al. The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res.* **51**, D1188–D1195 (2023).
68. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
69. R Core Team. *R: A Language and Environment for Statistical Computing*. Preprint at https://www.r-project.org/ (2022).
70. Ahlmann-Eltze, C. *ggupset: Combination Matrix Axis for 'ggplot2' to Create 'UpSet' Plots*. Preprint at https://github.com/const-ae/ggupset (2024).
71. ENCODE Project Consortium et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
72. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
73. Dolgalev, I. *msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format*. Preprint at https://CRAN.R-project.org/package=msigdbr (2022).
74. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
75. Wickham, H. *Ggplot2: Elegant graphics for data analysis* (Springer, 2016).
76. Ahlmann-Eltze, C. *Combination Matrix Axis for 'ggplot2' to Create 'UpSet' Plots*. https://github.com/const-ae/ggupset (2020).
77. Song, Y. & Wang, J. ggcoverage: An R package to visualize and annotate genome coverage for various NGS data. *BMC Bioinformatics* **24**, 309 (2023).

Author contributions

B.V.M. conceived the project. B.V.M. initiated the project and wrote the paper. B.V.M performed the bioinformatic analysis. B.V.M. produced all figures. B.V.M. supervised and funded the project.

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/s41598-024-73744-7.

Correspondence and requests for materials should be addressed to B.V.-M.