# Fusion transcript detection using spatial transcriptomics

**Stefanie Friedrich** ( stefanie.friedrich@scilifelab.se )
Stockholms Universitet https://orcid.org/0000-0002-3889-5589
**Erik LL Sonnhammer**
Stockholms Universitet

**Technical advance**

**Keywords:** Fusion transcript detection, Spatial Transcriptomics, gene fusion, cis-SAGE, oncogene

**Posted Date:** June 15th, 2020

**DOI:** https://doi.org/10.21203/rs.2.19314/v4



**Version of Record:** A version of this preprint was published on August 4th, 2020. See the published version at https://doi.org/10.1186/s12920-020-00738-5.

# Abstract

*Background* Fusion transcripts are involved in tumourigenesis and play a crucial role in tumour heterogeneity, tumour evolution and cancer treatment resistance. However, fusion transcripts have not been studied at high spatial resolution in tissue sections due to the lack of full-length transcripts with spatial information. New high-throughput technologies like spatial transcriptomics measure the transcriptome of tissue sections on almost single-cell level. While this technique does not allow for direct detection of fusion transcripts, we show that they can be inferred using the relative poly(A) tail abundance of the involved parental genes.

*Method* We present a new method STfusion, which uses spatial transcriptomics to infer the presence and absence of poly(A) tails. A fusion transcript lacks a poly(A) tail for the 5´ gene and has an elevated number of poly(A) tails for the 3´ gene. Its expression level is defined by the upstream promoter of the 5´ gene. STfusion measures the difference between the observed and expected number of poly(A) tails with a novel C-score.

*Results* We verified the STfusion ability to predict fusion transcripts on HeLa cells with known fusions. STfusion and C-score applied to clinical prostate cancer data revealed the spatial distribution of the cis-SAGe *SLC45A3-ELK4* in 12 tissue sections with almost single-cell resolution. The cis-SAGe occurred in disease areas, e.g. inflamed, prostatic intraepithelial neoplastic, or cancerous areas, and occasionally in normal glands.

*Conclusions* STfusion detects fusion transcripts in cancer cell line and clinical tissue data, and distinguishes chimeric transcripts from chimeras caused by trans-splicing events. With STfusion and the use of C-scores, fusion transcripts can be spatially localised in clinical tissue sections on almost single cell level.


# Background

A fusion transcript is a merging of different fragment transcripts. This molecule can be translated into a chimeric protein that possesses either a new or adapted function. Many fusion transcripts are classified as oncogenes that have the potential to cause cancer but also contribute to tumourigenesis as driver mutations. Mitelman et al. [1] estimated that 20% of cancer morbidity is caused by such fusions. Currently, 21,477 fusion transcripts have been identified; almost all can be found in neoplastic cells [2]. The majority of chimeric transcripts were detected in recent years by deep-sequencing technologies and the application of bioinformatics tools [2]. The occurrence of fusion transcripts is used as a cancer biomarker [3] as well as a cancer treatment target [4].

Fusion transcripts, also termed chimeric transcripts or chimeric RNA, are usually detected at the RNA level. If its underlying cause is known, it is further specified as either a genetic (i.e., gene fusion) or transcription-induced chimera. Gene fusions are caused by a genetic mutation – deletion, inversion or translocation event – of the DNA sequence from both parental genes. The mutated parental gene DNA

sequences are translated into a hybrid messenger RNA (mRNA). Whereas transcription-induced chimeras are caused by an abnormal mechanism either cis-splicing or trans-splicing [5]. The parental gene sequences remain intact.

Cis-splicing fusion transcripts are termed cis-splicing of adjacent genes (cis-SAGe) or read-through transcripts. cis-SAGe result from neglected gene boundaries; instead the DNA sequences of two adjacent genes are read and transcribed into a hybrid mRNA transcript. cis-SAGe were not exclusively identified in cancer cells. Despite intense research on clinical tissues and cancer cell lines that harbour a cis-SAGe, there is no convincing genetic mutation behind their production and thus a molecular mechanism is suspected. Chimeric transcription of cis-SAGe requires two genes on the same strand within 30 kilobase pairs (kbp) [6]. Chwalenia et al. [7] summarised features of cis-splicing chimeras: (i) active transcription of the 5´ gene, (ii) multi exonic neighbouring parental genes, (iii) absence of interstitial DNA deletion, (iv) presence of transcripts between the two neighbouring genes, (v) presence of CTCF binding sites between parental genes and (vi) induction by CTCF knockdown. They further suggest that cis-SAGe are an additional element of biological processes that increases the diversity of gene products. This supposition is consistent with the fact that cis-SAGe are also found in healthy tissue samples. Different cis-SAGe fusion variants, which are characterised by the different fusion points of the parental genes, are observed. Dominant among the so far detected cis-SAGe is the fusion point at the second exon of the 3´ gene [7].

Transcription induced fusion transcripts that result from trans-splicing are hybrids of two mRNA transcripts of separately transcribed genes. These molecules are rare, but if they occur, they can contribute to neoplastic transformation due to their pro-proliferative effects [8].

The poly(A) tail is a 100 – 250 bp sequence of adenine nucleotides. The poly(A) tail is not part of the DNA sequence; it is attached to the 3´ gene post-transcriptionally. A poly(A) signal at the 3´ untranslated region (UTR) of the DNA sequence defines the point of the poly(A) tail synthesis. Poly(A) polymerases, which are RNA polymerases, encode and attach the poly(A) tail to the pre-mRNA. Cleavage and polyadenylation specificity factor (CPSF) is one representative protein that recognises the poly(A) signal in the pre-mRNA sequence (in eukaryotes, often AAUAAA). Although the function of the poly(A) tail is not completely known, it increases mRNA stability during export from the nucleus to the ribosome, protects the mRNA from degradation and regulates its half-life. The poly(A) tail shortens with mRNA age. Further, the poly(A) tail, together with the counterpart at the 5´ end (the 5´ cap), initiates protein translation [9]. Recent studies by Park et al. [10] investigated its regulatory role in the somatic cell cycle. Cell cycle dysregulation is a hallmark of cancer.

Chimeric RNA can be detected experimentally (e.g. fluorescence in situ hybridisation [FISH] or Southern blot) or computationally. The latter is based on RNA-sequencing (RNA-seq) data and the application of software tools that identify the chimeric RNA. The tools search for encompassing reads, i.e., read pairs with each read aligned to one of the parental genes, and spanning reads, i.e., partially aligned reads that span the fusion point. Sensitivity and specificity depend on the tool as well as read length, read quality score and the number of reads that support each fusion transcript [11,12]. The results are often compared

to known fusion transcripts and then categorised into potential gene fusions, cis-SAGe or trans-splicing fusion transcripts. This search strategy, however, provides no information about the transcription direction and is limited in terms of spatial information.

In this study, we present a novel method to detect fused mRNA molecules using the poly(A) tail presence at the 3´ gene and its absence at the 5´ gene. Applying this method, and the C-score, to clinical tissue sections analysed with spatial transcriptomics allowed the detection of the cis-SAGe *SLC45A3-ELK4* on almost single cell level. The cis-SAGe clearly overlapped with disease areas within the tissue samples. Further, we emphasise that increased cis-SAGe correlates with elevated levels of transcriptional stress.

# Methods

### Fusion transcript detection using STfusion and poly(A) tail presence

During the correct transcription of two adjacent genes, the poly(A) tail is attached to each of the genes. The poly(A) tail serves as a proxy for the transcription level of each gene.

In the case of a fusion transcript caused by a cis-SAGe mechanism or chromosomal rearrangement, however, the poly(A) tail is absent from the 5´ gene. In this case, the 5´ gene expression level defines the fusion transcript expression by the promoter of the 5´ UTR of the 5´ gene. The poly(A) tail is attached to the 3´ end of the parental 3´ gene. In the proposed method, the number of sequenced poly(A) tails that can be mapped to the 3´ gene, and the absence of poly(A) tails at the 5´ gene, is used to indicate a fusion transcript. The number of poly(A) tails aligned to the 3´ gene mirrors the expression level of the fusion transcript (Figure 1, Table S12).

### STfusion verification in HeLa cells

In order to verify STfusion, sequenced mRNA from HeLa cancer cells were analysed. The occurrence of poly(A) tails aligned to the parental genes of experimentally confirmed fusion transcripts was tracked (Tables 1, 2, S1-S11).

The sequenced mRNA data produced by TAIL-seq [13], and the results published in the same paper, were used. This tool sequenced the end of the mRNA and thus includes the poly(A) tail sequence. Chang et al. [13] applied TAIL-seq to HeLa and found 4,000 genes have a poly(A) tail.

Additionally, for parental genes of the known HeLa fusion transcripts with no poly(A) tails, according to the results by Chang et al. [13], the raw RNA-seq data from the same publication were aligned and analysed in-depth. We were only interested in reads that contain a poly(A) tail sequence, which is the file that contains the second read (read 2, read length 230 bp). First, the potential poly(A) tail sequence nucleotides were removed. Second, the reads were further trimmed by 64 nucleotides (nt) to remove additional adapter sequences. The trimming was performed with Cutadapt [14]. The resulting reads were aligned with Tophat2 [15] and STAR [16] to the human assembly GRCh38 Ensembl (release 84) [17]. Uniquely mapped reads (Tophat2 mapping quality $\geq 10$, STAR mapping quality = 255) were kept. Finally,

reads that were aligned to a reference sequence with multiple adenine or thymine nucleotides in the direct vicinity were removed, because the eliminated sequence assumed to be a poly(A) tail is part of the genome. All reads that mapped to the 3´ UTR of the parental genes were considered (Table S13).

## STfusion applied to clinical tissue samples

### Spatial transcriptomics data, transcriptomic factors and activity maps

Spatial transcriptomics [18] (The Spatial Transcriptomics method, Suppl.), a novel technology, allows one to obtain expression levels throughout tissue samples while maintaining spatial information. Spatial transcriptomics opens new possibilities for the investigation of altered expression levels, especially under modified conditions (e.g., cancerous cells within tissue samples). In a recent publication, Berglund et al. [19] showed that the cells in the centre, periphery and vicinity of prostate cancer areas develop a unique expression pattern that is clearly differentiated from areas with healthy cells of a similar type. Thus, this technique can provide insights into cancer progression.

Spatial transcriptomics produces very rich expression levels data throughout a tissue sample. In order to identify hidden patterns of gene expressions that characterise cell types, spatial transcriptomics decomposition (STD) was developed by Maaskola et al. [20]. This method calculates expected gene expression (read counts) as the matrix product of observed gene expression and spatial activity matrices. The revealed unique expression profiles, i.e. transcriptomic factors, across tissue sections, represent different cell types, microenvironments or tissue components. For each identified expression profile, the method provides a spatial activity map that represents where the transcriptomic factor is active in the tissue sample. For example, the transcriptomic factor that represents “cancerous epithelial” cells exhibits a unique expression pattern that reveals genes strongly or differentially expressed compared to another transcriptomic factor. The activity map for the transcriptomic factor “cancerous epithelial” shows where the expression pattern is active within a tissue sample.

Berglund et al. [19] applied spatial transcriptomics to 12 tissue sections obtained from a patient diagnosed with prostate cancer. The spatial transcriptomics data comprised the expression levels of 5,053 protein-coding genes in 1,007 spots in each of the tissue sections. Further, the 12 tissue sections were analysed with STD in different joint approaches: (i) samples 1.2, 2.4 and 3.3 joined and (ii) the 12 tissue sections joined. The spatial transcriptomics data, STD transcriptomic factors and activity maps were used in this paper to localise the cis-SAGe *SLC45A3-ELK4*, link it to disease areas, calculate differentially expressed genes and perform pathway annotation (Figures 2-4, S2, S5).

### Fusion transcript localisation using STfusion and C-scores

In spatial transcriptomics, the poly(A) tail of a transcript is captured and measured as a proxy for the expression level of a gene in a tissue sample on an almost single-cell level. However, for a gene that is abnormally transcribed, as it is the case for a fusion transcript, the amount of poly(A) tails provides shifted results. This deviation is measured.

For each parental gene, the ratio (R) of the gene expression in each spot divided by the sample mean expression was calculated. The C-score of a spot is the maximum value of both ratios and presents the presence or absence of the fusion transcript. In the case of absence, the C-score level mirrors the 5´ gene expression level. In the case of a fusion transcript, the C-score level mirrors the fusion transcript expression level. (see Equations 1 and 2 in the Supplementary Files)

The proposed poly(A) tail detection method, STfusion using C-scores, was applied to the 12 clinical tissue samples analysed with spatial transcriptomics. The level of the C-score mirrors an abnormally high amount of poly(A) tails on the 3´ gene *ELK4* and predicts the cis-SAGe *SLC45A3-ELK4* (Figures 2 and S2).

To avoid divisions of 0, a pseudo-count of 1 can be added to both dividend and divisor of the ratios R5´ and R3´. The spatial distribution of the C-scores then changes slightly which can be circumvented using a threshold (C-score with pseudo count, Suppl., Tables S20, Figures S4, S6 and S7).

### Differentially expressed genes and pathway annotation

Spots with fusion transcript presence and absence were compared to investigate differentially expressed and co-expressed genes and activated pathways (Figures 4 and S5). Spots were only chosen according to their C-score, thus the likelihood and expression level of the cis-SAGe, and regardless of an annotation as stroma or epithelial.

To assign a spot to the group ‘occurrence’ or ‘absence’, C-score thresholds were applied:

(i) Absence $C\text{ - }score < 0$

(ii) Occurrence $0 < C\text{ - }score$

The spatial transcriptomics data with read counts for the 5,053 protein-coding genes across the spots, were checked for quality. Spots with a log-library size smaller than three median absolute deviations below the median log-library size were removed. Low-abundance genes with a read count of zero or close to zero among the spots were removed. The resulting data set was normalised per tissue sample using the R package “scran” [21]. The optimal pool size was calculated with the R package “scater” [22]. Genes with a very low standard deviation (sd) for the normalised expression levels among the chosen spots (sd < 10% of the expression mean) were removed.

The fold change per gene was calculated as gene expression mean of spots with C-scores > *0* (occurrence) divided by gene expression mean of spots with C-scores < 0 (absence). Differentially expressed genes were calculated with a two-sample t-test (confidence level 0.95) [23]. P-values were corrected for multiple testing with the Benjamini-Hochberg procedure[24]. Significantly differentially expressed genes (false discovery rate [FDR], q-value < 0.1) were submitted to PathwAX [25] on the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database [26].

### Detection of fusion transcript candidates

To identify a fusion transcript candidate caused by a cis-SAGe mechanism or a structural mutation (gene fusion) among random gene pair combinations, the diversity index D can be used; a higher value can indicate a fusion transcript. For each gene pair *i*, the diversity index is calculated as (eq 3), where $N_i$ is the number of C-scores ≠ 0, and $U_i$ is the number of unique C-scores ≠ 0, in both cases rounded to one decimal point: (see Equation 3 in the Supplementary Files)

To increase the amount of data, the sample-wise calculated C-scores were concatenated for the 12 tissue samples to calculate $D_i$ (Figure S8).

### Fusion transcript detection in bulk sequenced RNA from prostate cancer tissue samples

Bulk-sequenced mRNA from each of the 12 tissue sections were used to confirm the cis-SAGe *SLC45A3-ELK4*. The sequenced reads were aligned using two aligners to increase the possibility of identifying the cis-SAGe. The alignments of fastq reads were performed using Tophat2 (b2-sensitive and otherwise default parameters) [15] and STAR [16], both alignments against the human assembly GRCh38 Ensembl (release 84) [17]. Conversion from sam to bam format and indexing was done using Samtools [27].

Fusioncatcher [28] using Blat [29], Star and Bowtie2 [30] was applied to the aligned RNA-seq data to confirm the cis-SAGe. Additionally, the alignments were searched for encompassing reads, i.e., read pairs with each of the reads mapped to one of the parental genes, and for spanning reads that covered the fusion points identified with Fusioncatcher. The search was performed using Samtools (Tables S16 and S17).

## Results

### STfusion verification in HeLa cells

To verify STfusion accuracy, we applied it to HeLa cancer cells. P. Wu et al. [31] experimentally identified nine chimeric RNAs in HeLa cells. Further detected in Hela cancer cells were the cis-SAGe *SLC45A3-ELK4* by Zhang et al. [32] and the trans-splicing fusion event *VMP1-RPS6KB1* by L. Wu et al. [33]. Of these 11 fusion transcripts, the number of poly(A) tails per parental gene were considered. If the concept is correct and a chimeric transcript caused by a cis-SAGe mechanism or a chromosomal rearrangement is transcribed, the 5´ genes should not have a poly(A) tail, but the 3´ genes will have one.

STfusion verification was performed using the sequenced mRNA and the number of poly(A) tails produced by TAIL-seq, and the published results of counted poly(A) tails per gene [13] (Tables 1 and 2).

### STfusion verification for gene fusions and cis-SAGe

*LHX6-NDUFA8*, *SLC2A11-MIF*, and *SLC45A3-ELK4* are confirmed cis-SAGe events in HeLa [31,32]. The parental genes *SLC45A3* and *ELK4* were not listed as having poly(A) tails in Chang et al. [13]. With an in-

depth search in the sequenced mRNA published in the same paper, five poly(A) tails attached to the 3´ genes were identified (Tables 1 and S13). However, *TXNDC9-LYG1* did not seem to follow the proposed hypothesis. An inversion on Chr2:87-111 megabase pairs (Mbp) was identified by Breakdancer [34] (Table S14) and experimentally confirmed by Landry et al. [35]. Both parental genes are located within this region.

The results shown in Table 1 confirm our assumption that a fusion transcript caused by a cis-SAGe mechanism or a chromosomal rearrangement lacks a poly(A) tail at the 5´ gene and instead has an elevated number of poly(A) tails attached to the 3´ gene.

**Table 1**. Poly(A) tail occurrences of the parental genes in HeLa cells. For two fusion transcripts, *GFOD2-ENKD1* and *MFSD7-ATP5I*, no poly(A) tails were detected. Consistently, no tails were detected by Chang et al. [13] nor with an in-depth search in the TAIL-seq data.

| is-SAGe/ gene fusion | Poly(A) tails | |
|---|---|---|
| (5´ gene - 3´ gene) | 5´ gene | 3´ gene |
| *FOXRED2-TXN2* | 0 | 340 |
| *LHX6-NDUFA8* | 0 | 218 |
| *SLC2A11-MIF* | 0 | 842 |
| *SLC45A3-ELK4* | 0 | 5 |
| *TXNDC9-LYG1* | 103 | 0 |
| *UBE2Q2-FBXO22* | 0 | 97 |

## Distinction among fusion transcripts caused by trans-splicing

In the case of a fusion transcript caused by trans-splicing, both parental genes were transcribed and poly-adenylated. Poly(A) tails for both parental genes were observed (Table 2).

The transcription-induced chimera *VMP1-RPS6KB1* is assumed to occur via trans-splicing [33] in HeLa cells. Indeed, this event was confirmed with STfusion. The fusions *TINF2-NEDD8* and *DHRS13-FLOT2* were experimentally confirmed [31], but both parental genes were polyadenylated. This data suggests that these fusions are caused by a trans-splicing event. The latter fusion transcript, *DHRS13-FLOT2,* is suggested to be transcription induced [36], because no genetic cause could be identified.

**Table 2**. Poly(A) tails occurrences of the parental genes assumed to occur via trans-splicing in HeLa cells.

| 'rans-splicing fusion transcript | Poly(A) tails | |
|---|---|---|
| (5´ gene - 3´ gene) | 5´ gene | 3´ gene |
| *DHRS13-FLOT2* | 64 | 71 |
| *TINF2-NEDD8* | 35 | 69 |
| *VMP1-RPS6KB1* | 51 | 48 |

## STfusion and C-scores applied to clinical tissue samples

Spatial transcriptomics data published by Berglund et al. [19] was used to localise the cis-SAGe *SLC45A3-ELK3*. In this study, 12 tissue sections taken from a patient with prostate cancer were analysed; each section harboured epithelial areas annotated as healthy, inflamed, prostatic intraepithelial neoplasia (PIN, a precancerous lesion), cancerous with a Gleason Score (Gs) 3 + 3, or cancerous with Gs 3 + 4. The Gs is a grading system used to classify the aggressiveness of prostate cancer, scales range from 1 (appears healthy) to 5 (appears abnormal). The total Gs is a combination of two grades, one each for the dominant and minor area [37]. The tissues harbour the cis-SAGe *SLC45A3-ELK4* (Tables S15-S17) which contributes to cell proliferation in prostate cancer [32], Two fusion variants were identified in the bulk RNA-sequenced tissue sections: *SLC45A3-ELK4* exon 4-exon 2 and *SLC45A3-ELK4* exon 5-exon 2 (Figure S1, Table S18).

### Fusion transcript localisation using STfusion and C-scores

The C-score measures the fold change in the numbers of poly(A) tails on the parental gene compared to the parental gene sample mean expression. A higher C-score indicates that the occurrence of the cis-SAGe *SLC45A3-ELK4* is likely. This difference was caused by the chimeric mRNA and elevated 3´ gene *ELK4* expression, which is defined by the promoter of the 5´ UTR of the 5´ gene *SLC45A3*. A low C-score, however, represents a large number of poly(A) tails for the 5´ gene *SLC45A3* compared to the sample mean *SLC45A3* expression. This data indicates the occurrence of the cis-SAGe is very unlikely. The C-score mirrored the likelihood of cis-SAGe absence or occurrence in a spot as well as the expression levels of *SLC45A3* or cis-SAGe *SLC45A3-ELK3,* respectively (Additional file 1).

The C-scores' spatial distribution per sample was compared to the activity maps of the transcriptomic factors identified in the clinical tissue samples analysed with spatial transcriptomics (Figures 2 and S2). The predicted occurrence of the cis-SAGe *SLC45A3-ELK4* in the 12 tissue sections was dominant in the centre or periphery of disease areas; the predicted absence of the cis-SAGe was dominant in normal glands.

The three spatial transcriptomics samples with clearly identifiable cancer areas and transcriptomic factors related to cancer, resulting from the joint STD analyses are shown in Figures 2, together with the activity of selected factors and C-scores. For the transcriptomic factor “Cancer”, the predicted cis-SAGe occurs intensely at its activity centre in these samples, as well as at its periphery. In the cancer areas of samples 2.4 and 3.3 there are no spots with much higher expression of *SLC45A3* than *ELK4* (dark blue), hence no strong absence of the cis-SAGe is predicted in these areas. The cis -SAGe occurs only occasionally in the periphery of the PIN area in sample 3.3. Normal glands are dominated by absence of cis-SAGe. We note a few spots with strong cis-SAGe intensity scattered in various areas, often in the direct vicinity of spots with strong cis-SAGe absence.

To provide a statistical test for the coherence of disease areas and cis-SAGe occurrence, Spearman and Pearson correlations  were calculated (Tables 3 and S19, Figure S3). The strongest correlation of cis-SAGe occurrence is to the cancerous areas in sample 2.4 ($\rho_{Pearson}$ = 0.25,  p= 1.07E-07), sample 3.3 ($\rho_{Pearson}$ = 0.14,  p = 1.55E-03), and sample 3.2 ($\rho_{Pearson}$ = 0.12,  p = 4.32E-03), and to the PIN areas in sample 2.4 ($\rho_{Spearman}$ = 0.14, p = 3.15E-03), sample 1.2 ($\rho_{Pearson}$ = 0.23, p = 4.22E-06), sample 4.1 ($\rho_{Pearson}$ = 0.15, p = 1.17E-04), and sample 1.3 ($\rho_{Pearson}$ = 0.13, p = 1.51E-03).

The areas with active transcriptomic factors (“Normal glands”, “PIN glands”, “Inflammation”, and “Cancer”) were further analysed concerning the share of spots with predicted present or absent fusion transcripts (Figure 3). In normal glands, the cis-SAGe is dominantly absent. In the cancerous areas of the sample 3.3, the share of spots with mild occurrence (0 < C-score < 1) is increased compared to the other factors, whereas in the PIN areas of the same samples, the share of spots with strong occurrence (C-score > 1) is increased.

**Table 3.** Correlation of C-score and factor activities per sample shown in Figure 2.

| Sample | Factor | Sample-wide | | | | |
|---|---|---|---|---|---|---|
| | | # spots | Correlation $\rho_{Spearman}$ | Correlation $\rho_{Pearson}$ | p-value for $\rho_{Spearman}$ | p-value for $\rho_{Pearson}$ |
| Sample 2.4 | Normal glands | 451 | -0.10 | -0.13 | 4.19E-02 | 6.20E-03 |
| Sample 2.4 | PIN glands | 451 | 0.14 | 0.02 | 3.15E-03 | 6.21E-01 |
| Sample 2.4 | Cancer | 452 | 0.26 | 0.25 | 3.74E-08 | 1.07E-07 |
| Sample 3.3 | Normal glands | 500 | -0.02 | -0.04 | 6.51E-01 | 3.53E-01 |
| Sample 3.3 | PIN glands | 500 | -0.03 | 0.00 | 4.91E-01 | 9.59E-01 |
| Sample 3.3 | Cancer | 500 | 0.20 | 0.14 | 4.16E-06 | 1.55E-03 |
| Sample 3.2 | Inflammation | 560 | 0.17 | 0.11 | 1.32E-04 | 1.50E-02 |
| Sample 3.2 | Normal & PIN | 560 | -0.17 | -0.04 | 8.28E-05 | 3.86E-01 |
| Sample 3.2 | Cancer | 560 | 0.16 | 0.12 | 2.61E-04 | 4.32E-03 |

### Differentially expression and pathway annotation for cis-SAGe occurrence

The combination of spatial transcriptomics data and STfusion using C-scores offers new possibilities to explore fusion transcript occurrence, differences in cis-SAGe transcription levels and their spatial relation in clinical tissue samples.

Areas with absent and present *SLC45A3-ELK4* fusion transcripts were compared with regards to differentially and co-expressed genes and enriched pathways (Figure 4). In areas without fusion transcripts, there were pathways activated which are related to higher transcriptional stress (protein processing in the endoplasmic reticulum and lysosome). The pathways focal adhesion and regulation of actin cytoskeleton are highly active in the areas with cis-SAGe occurrence and are known to play a crucial role in cancer cell motility and invasion [38,39]. Phosphatidylinositol-3-kinase (PI3K)-AKT signalling is linked to treatment resistance [40,41].

### Detectability of fusion transcript candidates

We found that the C-scores of fusion transcripts (cis-SAGe and gene fusions) identified in the twelve tissue sections are more dispersed than those calculated for random gene pairs. Based on a diversity index, true fusion transcripts are top ranked (Figure S8).

To summarise the results, the proposed method, STfusion, identified fusion transcripts caused by a cis-SAGe mechanism or chromosomal rearrangement. It also distinguished these fusion transcripts from those caused by a trans-splicing event. Applying STfusion and the C-score to clinical tissue samples analysed with spatial transcriptomics demonstrated the spatial distribution of the fusion transcripts within the tissue section. Further, the fusion transcript was linked to the disease areas (inflammation, PIN, and cancer).

# Discussion

We propose a novel computational method to detect fusion transcripts. It is based on poly(A) tail presence or absence. We proved that a fusion transcript that lacks the poly(A) tail at the 5´ parental gene contains one at the 3´ gene. The number of poly(A) tails attached to the 3´ gene indicates the expression level of the fusion transcript defined by the 5´ gene promoter region. The novel method was verified on the chimeric transcripts caused by an incorrect cis-splicing of adjacent genes (cis-SAGe) mechanism or a chromosomal rearrangement (gene fusion) in HeLa cells. Fusion transcripts caused by trans-splicing with both genes that are poly-adenylated cannot be detected with the proposed method. However, our method helps to identify trans-splicing fusions among fusion transcripts identified with alternative methods (Tables 1 and 2).

The proposed method, STfusion, and the use of C-scores were applied to clinical tissue sections analysed with spatial transcriptomics. The tissue samples harbour areas annotated as inflammation, prostatic intraepithelial neoplastic and prostate cancer with different Gleason scores. Spatial transcriptomics, which uses the poly(A) tail as a proxy for expression levels, offers the opportunity to measure the unexpected amount of chimeric transcript parental gene expression levels at an almost single cell level. Fusion transcripts caused by cis-SAGe of the parental genes *SLC45A3* and *ELK4* were confirmed in the bulk sequenced RNA of the tissue sections. The identification of this fusion transcript in clinical tissues that harbour cancerous cells, and a spatial correlation to the disease areas was lacking. With the proposed method, we localised this fusion transcript in 12 tissue sections on almost single-cell level and detected a high variance of cis-SAGe expression in healthy and diseased areas which was reported earlier [42]. We showed the spatial expansion of the fusion transcript in the clinical samples and correlated the fusion transcript occurrence to areas annotated as diseased (Figure 2 and S2, Table 3). Very high cis-SAGe expression levels were observed in the periphery or centre of the disease areas; in one sample, spots with very low fusion transcript expression levels were found in the cancerous area. Occasionally, the cis-SAGe occurred in normal gland areas. Very high *SLC45A3* expression increases the likelihood of the cis-SAGe *SLC45A3-ELK4* occurrence. This observation indicates that *SLC45A3-ELK4* occurrence is an early and local event in the course of prostate cancer development. It appears to commence with higher *SLC45A3* expression, continue with high cis-SAGe expression, and finally end in very low cis-SAGe expression.

Differentially expressed genes between areas with and without the fusion transcript were calculated and activated pathways inferred (Figure 4 and S5). The observed activated pathways in areas of cis-SAGe

occurrence correlated with disease progression. Further, we observed pathways related to higher transcriptional stress during the switch from high 5´ gene expression to high cis-SAGe expression.

The diversity of the C-scores measured with the diversity index can help to indicate fusion transcripts caused by a cis-SAGe mechanism or chromosomal rearrangement among random gene pairs.

The cause of a cis-SAGe occurrence has not yet been identified. Genetic rearrangement can be excluded, and thus epigenetic changes are the primary focus. There is a DNA motif sequence with $(CCA)_n$ repetitions downstream of the intra-exonic fusion point of the 5´ gene *SLC45A3* that is linked to the i-Motif, a non-canonical DNA structure (Table S21). In more acidic environments, the motif sequence folds reversibly into an intramolecular intercalated cytosine tetraplex and thus can serve as a molecular switch [43–45]. This switch might be involved in passing over the stop codon of the terminal exon and thus the poly(A) tail signal for the 5´ gene.

# Conclusions

Fusion transcripts are detectable by their absent poly (A) tails for the 5’ gene and elevated number of poly (A) tails for the 3’ gene. The presented method, STfusion, uses this concept to detect fusion transcripts in data sets for which the number of poly(A) tails per gene is available. The method further distinguishes chimeric transcripts from chimeras caused by trans-splicing events and can localise fusion transcripts in clinical tissue samples at almost single cell level. It can also be used to identify novel fusion transcripts.

# List Of Abbreviations

| | |
|---|---|
| cis-SAGe | cis-Splicing of Adjacent Genes |
| FDR | False Discovery Rate |
| FOSB | FBJ murine Osteosarcoma viral oncogene homolog B |
| Gs | Gleason Score |
| Kbp | Kilo Base Pairs |
| MDS | Minimal Detectable Signal |
| nt | Nucleotide |
| PI3K | Phosphoinositide 3-kinase |
| R | Ratio |
| sd | Standard Deviation |

STD Spatial Transcriptomics Decomposition

# Declarations

### Ethics approval and consent to participate

Not applicable

### Consent for publication

Not applicable

### Availability of data and materials

The code to run STfusion, to calculate C-scores and to plot the spatial distribution, is available on the Bitbucket repository https://bitbucket.org/sonnhammergroup/stfusion/. A Shiny application 'STfusion' was also built and is freely available at https://stfusion.shinyapps.io/stfusion_shiny/.

The spatial transcriptomics datasets analysed in this study were obtained from doi:10.1038/S41467-018-04724-5, available on the spatial research repository https://www.spatialresearch.org/resources-published-datasets/10-1038-s41467-018-04724-5/.

The sequenced reads using TAIL-seq of HeLa cells analysed during the current study are available in the NCBI Gene Expression Omnibus (GEO) database (accession number GSM1242325). These datasets were derived from the following public domain resources: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1242325.

The human assembly GRCh38 Ensembl (release 84) files were downloaded from ftp://ftp.ensembl.org/pub/release-84/ .

### Competing interests

The authors declare that they have no competing interests.


### Funding

Not applicable


### Author contributions

SF and ES conceptualised and designed the project, performed the data analysis and wrote the manuscript. All authors read and approved the final manuscript.


### Acknowledgements

We wish to thank Niklas Schultz and Joakim Lundeberg for helpful comments on the project and manuscript.

# References

1. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer. 2007;7:233–45. doi:10.1038/nrc2091.
2. Mertens F, Johansson B, Fioretos T, Mitelman F. The emerging complexity of gene fusions in cancer. Nat Rev Cancer. 2015;15:371–81. doi:10.1038/nrc3947.
3. Sanguedolce F, Cormio A, Brunelli M, D'Amuri A, Carrieri G, Bufo P, et al. Urine TMPRSS2: ERG fusion transcript as a biomarker for prostate cancer: literature review. Clin Genitourin Cancer. 2016;14:117–21. doi:10.1016/j.clgc.2015.12.001.
4. Zhou J, Liao J, Zheng X, Shen H. Chimeric RNAs as potential biomarkers for tumor diagnosis. BMB Rep. 2012;45:133–40. doi:10.5483/BMBRep.2012.45.3.133.
5. Li Z, Qin F, Li H. Chimeric RNAs and their implications in cancer. Curr Opin Genet Dev. 2018;48:36–43. doi:10.1016/j.gde.2017.10.002.
6. Jia Y, Xie Z, Li H. Intergenically spliced chimeric RNAs in cancer. Trends Cancer. 2016;2:475–84. doi:10.1016/j.trecan.2016.07.006.
7. Chwalenia K, Facemire L, Li H. Chimeric RNAs in cancer and normal physiology. Wiley Interdiscip Rev RNA. 2017;8. doi:10.1002/wrna.1427.
8. Li H, Wang J, Ma X, Sklar J. Gene fusions and RNA trans-splicing in normal and neoplastic human cells. Cell Cycle. 2009;8:218–22. doi:10.4161/cc.8.2.7358.
9. Guydosh NR, Green R. Translation of poly(A) tails leads to precise mRNA cleavage. RNA. 2017;23:749–61. doi:10.1261/rna.060418.116.
10. Park J-E, Yi H, Kim Y, Chang H, Kim VN. Regulation of Poly(A) Tail and Translation during the Somatic Cell Cycle. Mol Cell. 2016;62:462–71. doi:10.1016/j.molcel.2016.04.007.
11. Carrara M, Beccuti M, Cavallo F, Donatelli S, Lazzarato F, Cordero F, et al. State of art fusion-finder algorithms are suitable to detect transcription-induced chimeras in normal tissues? BMC Bioinformatics. 2013;14 Suppl 7:S2. doi:10.1186/1471-2105-14-S7-S2.
12. Kumar S, Vo AD, Qin F, Li H. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. Sci Rep. 2016;6:21597. doi:10.1038/srep21597.
13. Chang H, Lim J, Ha M, Kim VN. TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications. Mol Cell. 2014;53:1044–52. doi:10.1016/j.molcel.2014.02.007.
14. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. 2011;17:10. doi:10.14806/ej.17.1.200.
15. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36. doi:10.1186/gb-2013-14-4-r36.

16. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21. doi:10.1093/bioinformatics/bts635.
17. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018;46:D754–61. doi:10.1093/nar/gkx1098.
18. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353:78–82. doi:10.1126/science.aaf2403.
19. Berglund E, Maaskola J, Schultz N, Friedrich S, Marklund M, Bergenstråhle J, et al. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. Nat Commun. 2018;9:2419. doi:10.1038/s41467-018-04724-5.
20. Maaskola J, Bergenstråhle L, Jurek A, Fernández Navarro J, Lagergren J, Lundeberg J. Charting tissue expression anatomy by spatial transcriptome decomposition. BioRxiv. 2018;[Preprint]. accessed 1.12.2019 Available from https://doi.org/10.1101/362624. doi:10.1101/362624.
21. 21. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res. 2016;5:2122. doi:10.12688/f1000research.9501.2.
22. 22. McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017;33:1179–86. doi:10.1093/bioinformatics/btw777.
23. Welch BL. The generalization of 'student's' problem when several different population varlances are involved. Biometrika. 1947;34:28–35. doi:10.1093/biomet/34.1-2.28.
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995;57:289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.
25. Ogris C, Helleday T, Sonnhammer ELL. PathwAX: a web server for network crosstalk based pathway annotation. Nucleic Acids Res. 2016;44:W105-9. doi:10.1093/nar/gkw356.
26. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. Nucleic Acids Res. 2019;47:D590–5. doi:10.1093/nar/gky962.
27. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93. doi:10.1093/bioinformatics/btr509.
28. Nicorici D, Satalan M, Edgren H, Kangaspeska S, Murumagi A, Kallioniemi O, et al. FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. BioRxiv. 2014. doi:10.1101/011650.
29. Kent WJ. BLAT–the BLAST-like alignment tool. Genome Res. 2002;12:656–64. doi:10.1101/gr.229202.
30. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25. doi:10.1186/gb-2009-10-3-r25.

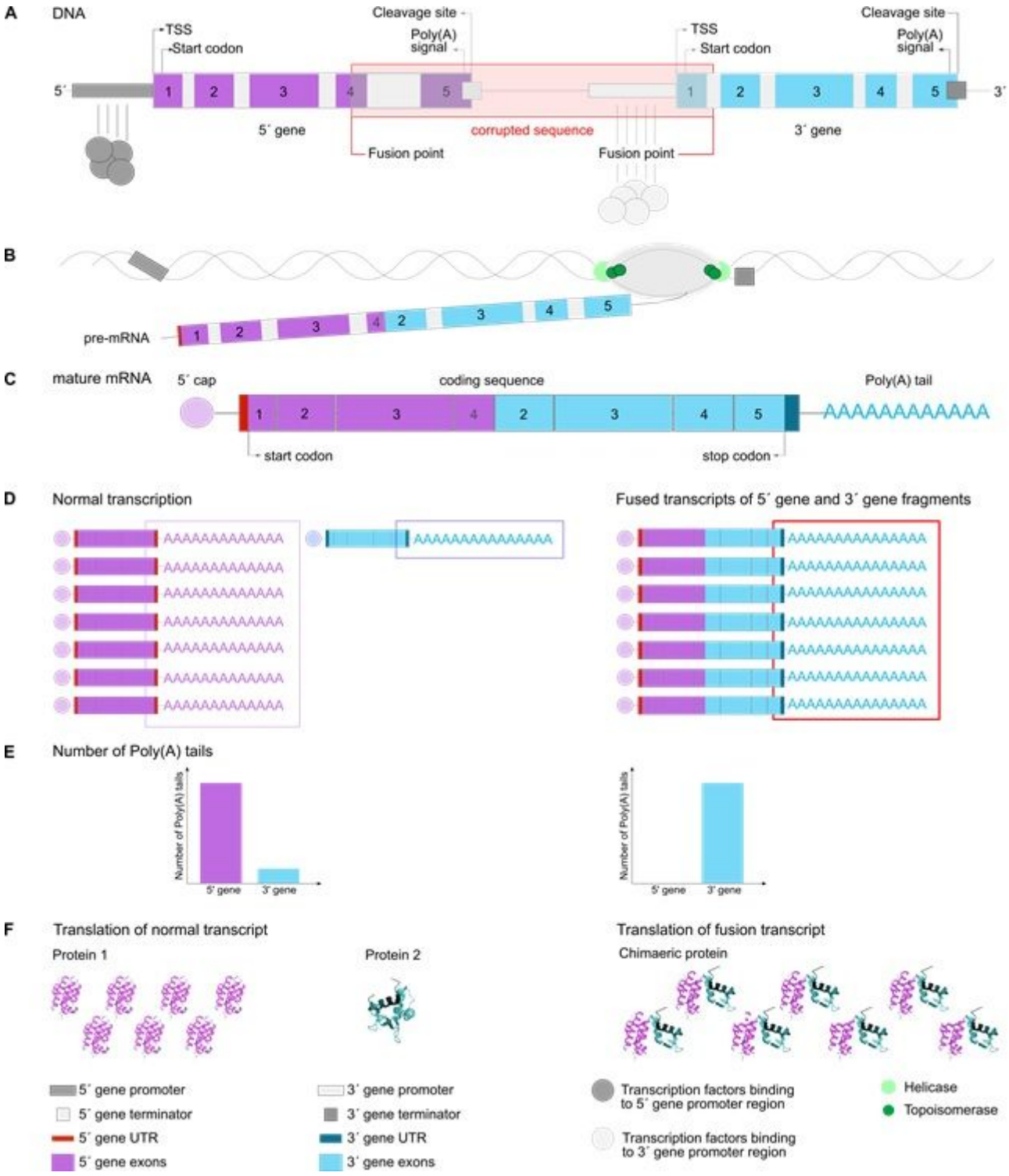
31. Wu P, Yang S, Singh S, Qin F, Kumar S, Wang L, et al. The landscape and implications of chimeric rnas in cervical cancer. EBioMedicine. 2018;37:158–67. doi:10.1016/j.ebiom.2018.10.059.
32. Zhang Y, Gong M, Yuan H, Park HG, Frierson HF, Li H. Chimeric transcript generated by cis-splicing of adjacent genes regulates prostate cancer cell proliferation. Cancer Discov. 2012;2:598–607. doi:10.1158/2159-8290.CD-12-0042..
33. Wu L, Zhang X, Zhao Z, Wang L, Li B, Li G, et al. Full-length single-cell RNA-seq applied to a viral human cancer: applications to HPV expression and splicing analysis in HeLa S3 cells. Gigascience. 2015;4:51. doi:10.1186/s13742-015-0091-4.
34. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6:677–81. doi:10.1038/nmeth.1363.
35. Landry JJM, Pyl PT, Rausch T, Zichner T, Tekkedil MM, Stütz AM, et al. The genomic and transcriptomic landscape of a HeLa cell line. G3 (Bethesda). 2013;3:1213–24. doi:10.1534/g3.113.005777.
36. Huang R, Kumar S, Li H. Absence of Correlation between Chimeric RNA and Aging. Genes (Basel). 2017;8. doi:10.3390/genes8120386.
37. Epstein JI, Zelefsky MJ, Sjoberg DD, Nelson JB, Egevad L, Magi-Galluzzi C, et al. A contemporary prostate cancer grading system: A validated alternative to the gleason score. Eur Urol. 2016;69:428–35. doi:10.1016/j.eururo.2015.06.046.
38. Devreotes P, Horwitz AR. Signaling networks that regulate cell migration. Cold Spring Harb Perspect Biol. 2015;7:a005959. doi:10.1101/cshperspect.a005959.
39. Yamaguchi H, Condeelis J. Regulation of the actin cytoskeleton in cancer cell migration and invasion. Biochim Biophys Acta. 2007;1773:642–52. doi:10.1016/j.bbamcr.2006.07.001.
40. Edlind MP, Hsieh AC. PI3K-AKT-mTOR signaling in prostate cancer progression and androgen deprivation therapy resistance. Asian J Androl. 2014;16:378–86. doi:10.4103/1008-682X.122876.
41. Crumbaker M, Khoja L, Joshua AM. AR signaling and the PI3K pathway in prostate cancer. Cancers (Basel). 2017;9. doi:10.3390/cancers9040034.
42. Ren G, Zhang Y, Mao X, Liu X, Mercer E, Marzec J, et al. Transcription-mediated chimeric RNAs in prostate cancer: time to revisit old hypothesis? OMICS. 2014;18:615–24. doi:10.1089/omi.2014.0042.
43. Kaushik M, Kaushik S, Roy K, Singh A, Mahendru S, Kumar M, et al. A bouquet of DNA structures: Emerging diversity. Biochemistry and Biophysics Reports. 2016;5:388–95. doi:10.1016/j.bbrep.2016.01.013.
44. Li T, Famulok M. I-motif-programmed functionalization of DNA nanocircles. J Am Chem Soc. 2013;135:1593–9. doi:10.1021/ja3118224.
45. Zemánek M, Kypr J, Vorlícková M. Conformational properties of DNA containing (CCA)n and (TGG)n trinucleotide repeats. Int J Biol Macromol. 2005;36:23–32. doi:10.1016/j.ijbiomac.2005.03.005.

46. Rickman DS, Pflueger D, Moss B, VanDoren VE, Chen CX, de la Taille A, et al. SLC45A3-ELK4 is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer. Cancer Res. 2009;69:2734–8. doi:10.1158/0008-5472.CAN-08-4926.
47. Makkonen H, Jääskeläinen T, Pitkänen-Arsiola T, Rytinki M, Waltering KK, Mättö M, et al. Identification of ETS-like transcription factor 4 as a novel androgen receptor target in prostate cancer cells. Oncogene. 2008;27:4865–76. doi:10.1038/onc.2008.125.
48. Ostman A, Hellberg C, Böhmer FD. Protein-tyrosine phosphatases and cancer. Nat Rev Cancer. 2006;6:307–20. doi:10.1038/nrc1837.
49. Chmelar R, Buchanan G, Need EF, Tilley W, Greenberg NM. Androgen receptor coregulators and their involvement in the development and progression of prostate cancer. Int J Cancer. 2007;120:719–33. doi:10.1002/ijc.22365.
50. Catalona WJ, Richie JP, Ahmann FR, Hudson MA, Scardino PT, Flanigan RC, et al. Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate cancer: results of a multicenter clinical trial of 6,630 men. J Urol. 2017;197:S200–7. doi:10.1016/j.juro.2016.10.073.
51. Eidelman E, Twum-Ampofo J, Ansari J, Siddiqui MM. The metabolic phenotype of prostate cancer. Front Oncol. 2017;7:131. doi:10.3389/fonc.2017.00131.52. Fennelly C, Amaravadi RK. Lysosomal biology in cancer. Methods Mol Biol. 2017.
52. Aderem A. Phagocytosis and the inflammatory response. J Infect Dis. 2003;187 Suppl 2:S340-5. doi:10.1086/374747.

# Additional File

| File name | Title of data | Description of data |
|---|---|---|
| Additional file 1 | ST_cis-SAGe | ST read counts of the parental genes of the cis-SAGe SLC45A3-ELK4 and C-scores obtained in the clinical tissue samples analysed with Spatial Transcriptomics |

# Figures




## Figure 1

Expression of a fusion transcript. A Gene structure of the parental genes. The corrupted region between both genes leads to the fusion transcript. The fusion point is set to exon 4 of the 5´ gene and the intronic region between exons 1 and 2 for the 3´ gene. B Transcription of the fusion transcript is shown. Only fragments of the 5´ gene and fragments of the 3´ gene are transcribed. During transcription, helicase divides the two strands to allow RNA polymerase to act; both strands are subsequently combined. Topoisomerase avoids rotation and tension of the DNA strands during de-spiralisation. C Polyadenylation

during post-transcriptional modification of the chimeric transcript and splicing to a mature mRNA. D Expression levels are defined by the promoter in the 5´ UTR of each gene. Normal transcription of the adjacent genes involves a different number of poly(A) tails on the parental genes (left). Transcription of the fusion transcript, however, results in an expression level based on the 5´ gene promoter and an elevated number of poly(A) tails attached to the 3´ UTR (right). E The difference in the number of poly(A) tails between normal, not fused (left barplot) and fused (right barplot) transcripts of the parental genes leads to the detection of the fusion transcript. If fused, the number of poly(A) tails on the 5´gene becomes 0 and on the 3´gene equal to the number of poly(A) tails on the normal 5´gene transcript. F mRNA is translated and then folded into a protein. In case of a translated fusion transcript the folded protein is chimeric, i.e., a combination or section of fragments from the parental genes.

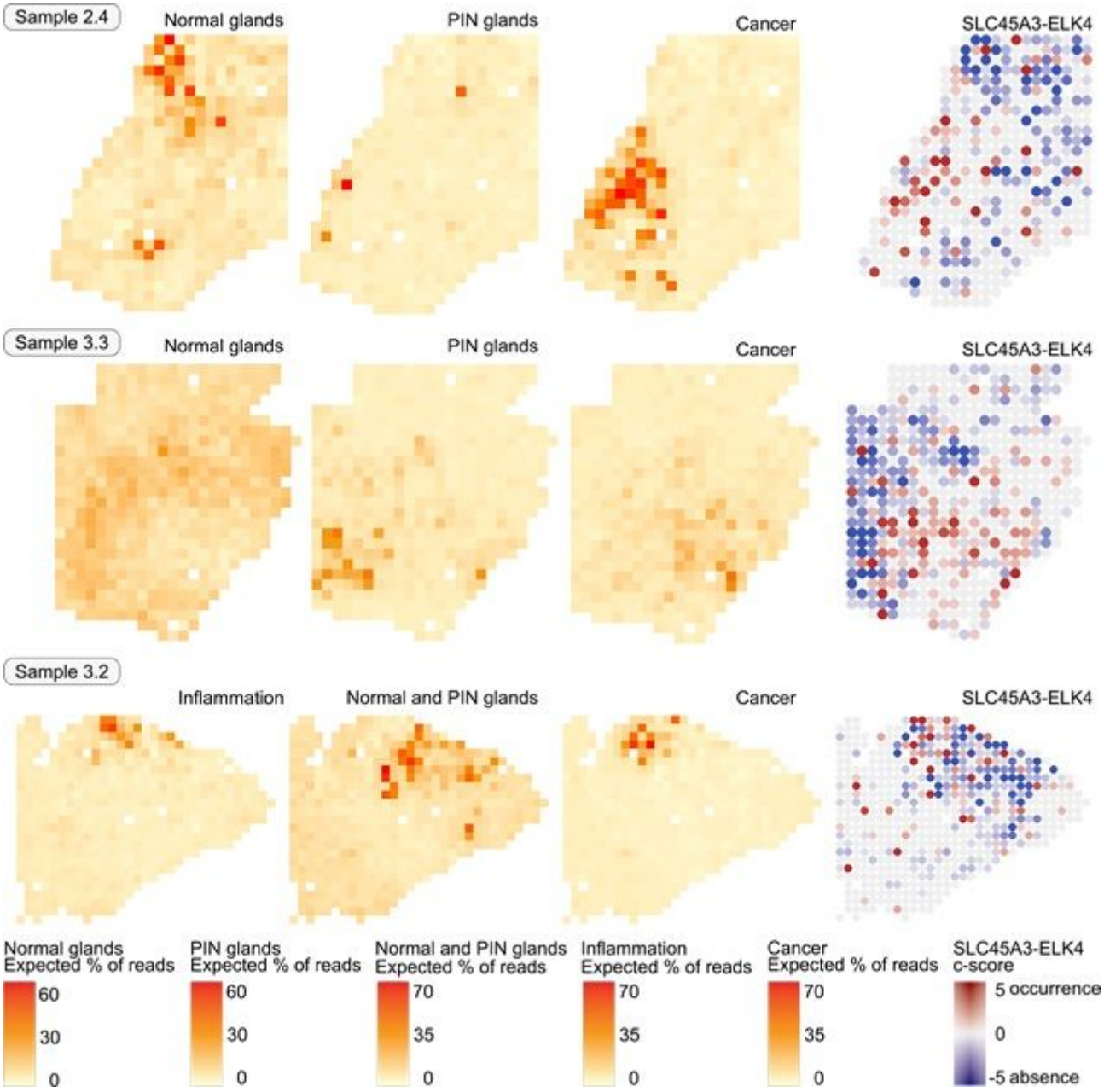
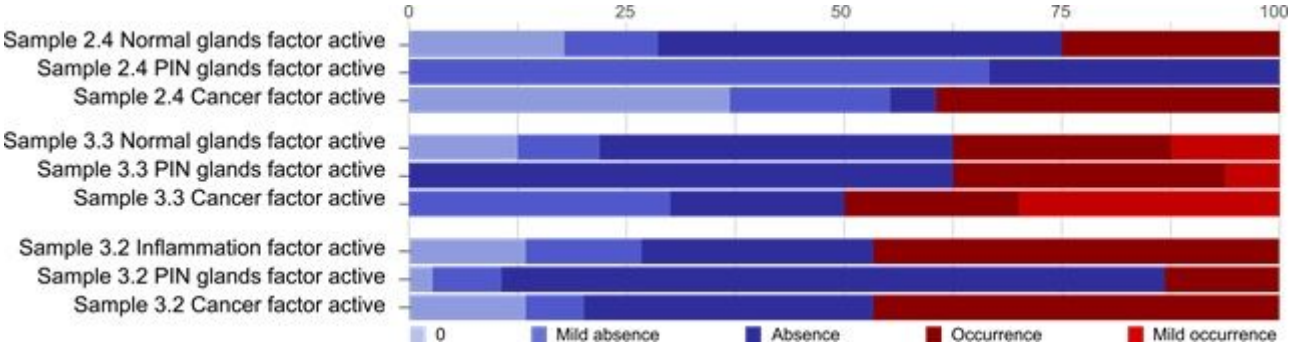


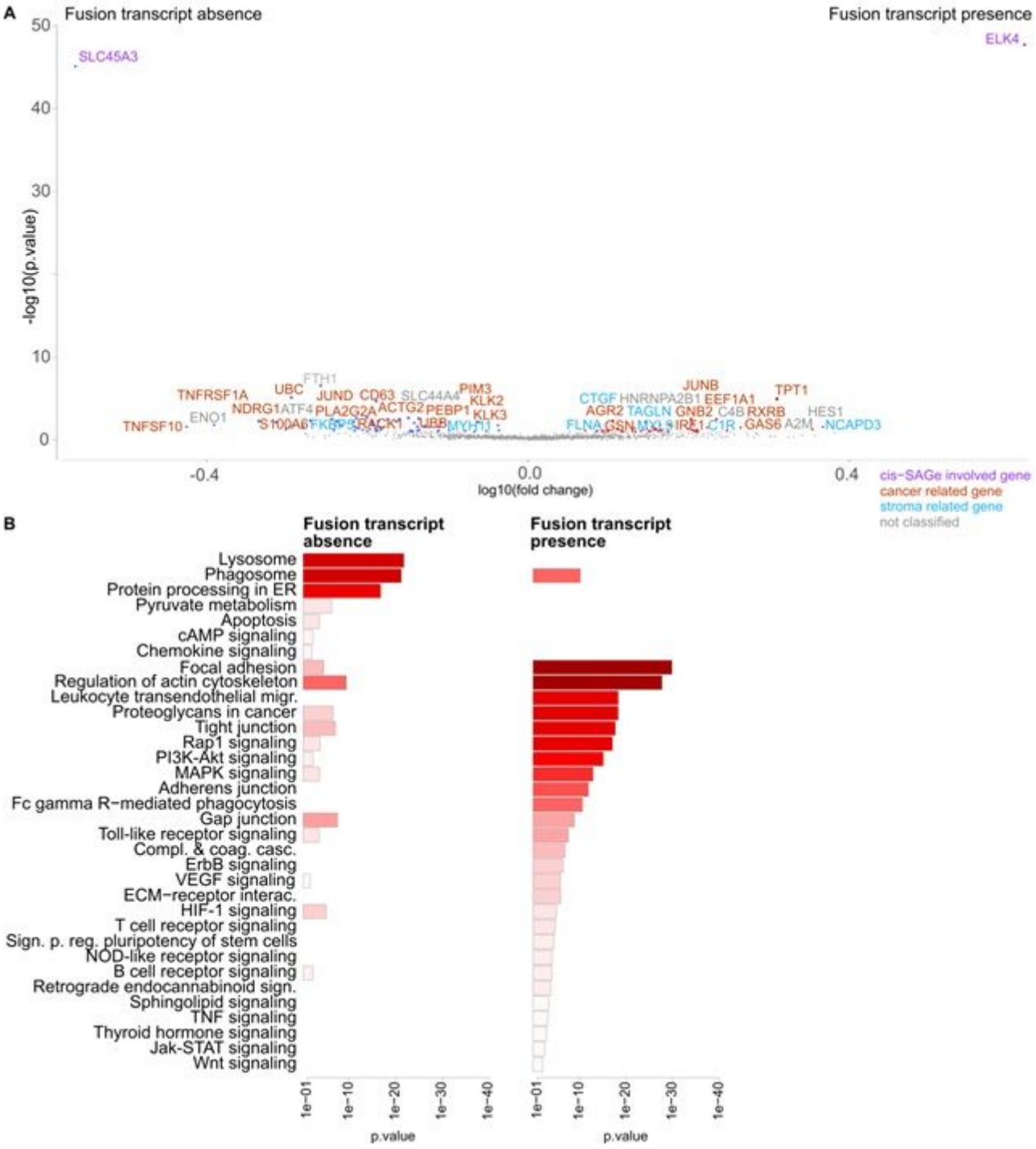

**Figure 2**

Activity maps of transcriptomic factors “Normal glands”, “PIN glands” and Cancer” (from [19]) compared to the predicted occurrence of the fusion transcript SLC45A3-ELK4. The “Normal glands” factor is associated with absence of the fusion, while the “Cancer factor” is associated with occurrence of the fusion.




## Figure 3

Fraction of spots with fusion transcript occurrence (C-score > 0) and absence (C-score ≤ 0) for the factors shown in Figure 2. The factor activity threshold was set to 20%. Mild occurrence and mild absence C-score thresholds were set to 1 and -1, respectively.




## Figure 4

Differential expression and pathway annotation of cis-SAGe occurrence. Areas with absent and present fusion transcripts in sample 3.3 were compared. Besides normal glands, the sample harboured an area annotated as PIN and a large area annotated as cancerous of which some parts were annotated as aggressively cancerous (Gs 3 + 4). A Significantly differentially expressed genes (FDR, $q < 0.1$) are shown. B Significantly differentially expressed genes were submitted to PathwAX on the KEGG database. Enriched pathways are presented.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Fusionsupplvs3reSubmission2.pdf
- Equations.pdf
- STcisSAGe.xlsx