International Journal of
Molecular Sciences

MDPI

Article

Identification of Recurrent Fusion Genes in Esophageal Squamous Cell Carcinoma Patients from Kazakhstan

Aibek Karimov [1,2], Rakhimbek Bektayev [1,3], Nazerke Satvaldina [1], Ruslan Kalendar [1], Anel Ordabayeva [1], Asset Daniyarov [1,4], Aigul Sharip [1], Ainur Ashenova [1], Ulan Kozhamkulov [1], Dinara Begimbetova [1], Mukhtar Tuleutaev [5], Dos Sarbassov [1], Saule Rakhimova [1,*], Ainur Akilzhanova [1] and Ulykbek Kairov [1,*]

[1] Center for Life Sciences, National Laboratory Astana, Nazarbayev University, Astana 010000, Kazakhstan; aibek.karimov@nu.edu.kz (A.K.); rakhimbek.bektayev@nu.edu.kz (R.B.); nazerke.satvaldina@nu.edu.kz (N.S.); ruslan.kalendar@nu.edu.kz (R.K.); anel.ordabayeva@nu.edu.kz (A.O.); asset.daniyarov@nu.edu.kz (A.D.); aigul.sharip@nu.edu.kz (A.S.); ainur.seisenova@nu.edu.kz (A.A.); dos.sarbassov@nu.edu.kz (D.S.); akilzhanova@nu.edu.kz (A.A.)
[2] Department of Biological and Biomedical Sciences, School of Medicine, Nazarbayev University, Astana 010000, Kazakhstan
[3] Medical Genetics and Molecular Biology Department, Astana Medical University, Astana 010000, Kazakhstan
[4] Department of General Biology and Genomics, Institute of Natural Sciences, L.N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan
[5] Multidisciplinary Medical Center, Astana 010000, Kazakhstan; oncoastana@mail.ru
* Correspondence: saule.rakhimova@nu.edu.kz (S.R.); ulykbek.kairov@nu.edu.kz (U.K.)

Abstract

Esophageal squamous cell carcinoma (ESCC) in Kazakhstan has high mortality and is often diagnosed late, yet its structural transcriptome is uncharacterized. We performed deep RNA sequencing of tumor tissue from 23 treatment-naive ESCC patients from Kazakhstan. Fusion genes were called with a consensus pipeline requiring detection by both Arriba and STAR-Fusion and were validated by RT-PCR and Sanger sequencing. Five chimeric transcripts were confirmed: *MAP4K5–L2HGDH*, *ADAMTS2–ENSG00000253652*, *BAZ2B–WDSUB1*, *HUWE1–SUPT3H* and *PLEKHA5*–intergenic (*KRAS*-adjacent); four were computationally recurrent; and none has previously been validated in ESCC. Reconstruction of the fusion transcripts, combined with AlphaFold 3 structure prediction, showed that none encodes a chimeric protein retaining an intact catalytic or signaling domain. *MAP4K5–L2HGDH* was the only in-frame chimera, joining a catalytically incomplete MAP4K5 kinase N-lobe to an L2HGDH fragment lacking its mitochondrial targeting sequence and FAD-binding motif; three fusions truncated their 5′ partner, and *BAZ2B–WDSUB1* exchanged untranslated regions without altering the coding sequence. In one tumor, the *PLEKHA5* fusion and two further *KRAS*-associated candidates mapped to two discrete blocks of strongly co-overexpressed 12p genes including *KRAS*, consistent with a complex amplicon. Validated fusions in this cohort are therefore more consistent with inactivating or passenger events than with canonical oncogenic drivers.

Keywords: esophageal squamous cell carcinoma; gene fusions; RNA-seq; AlphaFold 3; protein structure prediction; *KRAS* amplification; Kazakhstan

Academic Editor: Hidekazu Suzuki

Received: 6 November 2026
Revised: 24 November 2026
Accepted: 25 November 2026
Published: 26 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Esophageal cancer (EC) remains a major public health challenge globally, with its disease burden exhibiting substantial geographical variation [1,2]. According to the World Health Organization's International Agency for Research on Cancer (IARC) GLOBOCAN 2022 database, the overall incidence age-standardized rate (ASR) for esophageal cancer in Kazakhstan is 4.1 per 100,000 population [3]. When evaluated by sex, the incidence ASR stands at 6.1 per 100,000 among males and 2.8 per 100,000 among females. Notably, the corresponding mortality rates are nearly equivalent to incidence levels, with an overall mortality ASR of 3.6 per 100,000 (5.6 for males and 2.3 for females) [3]. This narrow gap between incidence and mortality underscores the highly aggressive nature of the tumor and its typically late-stage diagnosis in the region, highlighting the critical need for novel molecular markers, such as fusion genes, to enable earlier detection [4,5].

Initial insights into the molecular landscape of esophageal squamous cell carcinoma (ESCC) within the Kazakhstani population have recently been established by our research group through advanced transcriptomic profiling [6]. This pioneering study utilized deep, high-throughput RNA sequencing of tumor tissues from local patients, focusing primarily on differential gene expression analysis. The findings revealed widespread regulatory disruptions across key KEGG signaling pathways, notably characterized by the upregulation of oxidative phosphorylation and the suppression of extracellular matrix degradation mechanisms. Our group has since extended this work with an integrative analysis of lncRNA–miRNA–mRNA expression profiles in ESCC tissues [7]. Despite these insights, the landscape and prevalence of oncogenic fusion genes within this specific patient cohort remain entirely uncharacterized [8–10].

In contrast to these preceding studies, which were confined to the quantitative assessment of gene expression, the present study shifts the focus toward a qualitative evaluation of structural transcriptome alterations within an expanded iteration of the same cohort. This approach represents a fundamentally deeper layer of molecular analysis, specifically designed to identify discrete genetic aberrations—namely, recurrent chimeric transcripts that may function as direct oncogenic drivers. By characterizing these population-specific gene fusions, this study aims to bridge the existing knowledge gap and establish a framework for identifying robust biomarkers for early diagnostic intervention.

To maximize the benefits and minimize the toxicity of an oncotherapy, targeted therapy has emerged as one of the major modalities for cancer control, with impressive results obtained for several cancers [11]. Gene fusions drive approximately 20% of the global cancer burden [12]. Comprehensive characterization of these chimeric transcripts has mapped their functional trajectory from initial genomic formation to downstream oncogenic signaling. Consequently, targeting these specific molecular drivers with precision therapies has significantly improved patient outcomes in clinical practice [13].

Cancer fusion (or chimeric) genes typically arise from structural chromosomal aberrations that combine fragments of unrelated normal genes. Most commonly, the 5′ upstream fusion partner retains its native promoter and initial exons, whereas the 3′ downstream partner provides the terminal exonic sequence [14]. The genomic junction between these two segments, usually situated within an intronic region, is defined as the fusion breakpoint [13].

Large-scale pan-cancer genomic analyses have established fusion genes as a major class of somatic driver alterations. A systematic study across 9624 tumors spanning 33 cancer types identified oncogenic fusions in 16.5% of cases and estimated that 6% of all cancers harbor fusions representing tractable therapeutic targets, with those retaining an intact kinase domain being among the most clinically actionable [15]. However, ESCC remains a molecularly heterogeneous malignancy characterized by multiple genomic

subtypes and poorly defined prognostic markers, which underscores the need for a systematic characterization of recurrent molecular alterations, including fusion genes [16].

The documentation of fusion genes in ESCC has recently yielded clinically relevant targets. Among the earliest characterized events in ESCC clinical material was the *FGFR3–TACC3* fusion, detected at a frequency of 1.4% and associated with an adverse clinical course in a patient with stage IV disease [17]. Notably, *FGFR3–TACC3* belongs to the class of kinase-domain-retaining fusions highly amenable to targeted intervention, indicating that low-frequency fusions in ESCC can carry therapeutic implications.

Yang et al. [18] identified a chromosomal rearrangement producing the *PLEKHA1–TACC2* fusion gene, detected in 16.3% of ESCC tumor samples and correlated with poor patient prognosis. The fusion protein upregulates the EphA2/AKT/MMP2 signaling pathway and promotes vascular mimicry by reducing EphA2 ubiquitylation. EphA2 inhibitors dasatinib and ALW II-41-27 markedly suppressed progression of *PLEKHA1–TACC2*-expressing tumors in vivo, while regorafenib—an FDA-approved inhibitor—showed suppressive effects in a transgenic ESCC mouse model, supporting the fusion's role as both a prognostic marker and a therapeutic target.

Despite these findings, the number of functionally characterized, recurrent fusions in ESCC remains limited, and the comprehensive landscape of these genomic events—including their precise frequencies, mechanistic consequences, and clinical relevance—has yet to be systematically defined.

This structural heterogeneity of rearrangements in ESCC further complicates the development of uniform targeted therapies [19] and necessitates comprehensive transcriptome profiling in previously uncharacterized cohorts, such as the Kazakhstani population. Given this mosaic mutational background, examining underrepresented demographic groups remains essential to identifying rare or population-specific molecular drivers that may otherwise bypass detection in larger, homogeneous studies.

Most genomic studies on fusion genes in ESCC have focused on East Asian patients, with Chinese cohorts dominating the literature [20–23]**.** Populations from Central Asia, Kazakhstan included, carry a different genetic background and face a distinct set of environmental and dietary risk factors [6]. Whether this translates into a different fusion gene profile has not been tested—no study to date has looked at fusion events in Kazakhstani ESCC tumors. As a result, molecular data from this region remain absent, and the findings from elsewhere may not fully apply here.

This study aimed to identify and characterize recurrent fusion genes in tumor transcriptomes of treatment-naive ESCC patients from Kazakhstan using RNA sequencing, to predict their protein-level consequences, and to place them in the context of the clinicopathological features of the cohort and of the genome-wide landscape of candidate fusion calls.

2. Results

2.1. Patient Cohort and Clinical Characteristics

The study cohort comprised 23 patients diagnosed with esophageal squamous cell carcinoma (ESCC) from Astana, Kazakhstan. This cohort largely overlaps with our group's previous transcriptomic profiling study in the Kazakhstani population [6], with the addition of one patient whose RNA-seq data met quality requirements for fusion transcript detection. Baseline clinical and demographic characteristics are summarized in Table 1, with comprehensive individual patient profiles provided in Supplementary Table S1. The cohort comprised 14 males (60.9%) and 9 females (39.1%), with a mean age of 66.0 ± 8.2 years (range 48–79). Regarding ethnicity, 18 patients (78.3%) were Kazakh, 4 (17.4%) were Russian, and 1 (4.3%) was Tatar. Disease staging revealed predominantly advanced presentations: 1 patient had T1 stage, 2 had T2, 18 had T3, and 2 had T4. Regional lymph

node status was N0 in 6 patients (26.1%), N1 in 2 (8.7%) and N2 in 1 (4.3%), and could not be assessed (NX) in 14 (60.9%); all tumors were classified as M0. Most tumors were moderately differentiated (G2, n = 22; G3, n = 1) and keratinizing (17/23, 73.9%), and most were located in the middle (12/23, 52.2%) or lower (6/23, 26.1%) thoracic esophagus; four further tumors spanned the upper–middle (n = 2) or middle–lower (n = 2) thirds, and one was located in the upper third. All patients presented with dysphagia. Body mass index, available for 13 patients, was 24.9 ± 3.6 kg/m^2 (range 19.1–29.5). All 23 patients were treated surgically by Ivor-Lewis esophagectomy between 2013 and 2017, and none had received neoadjuvant chemotherapy or radiotherapy; all tumor specimens analyzed here were therefore treatment-naive (Section 4.1). Data on postoperative treatment and long-term outcome were not available for this retrospective cohort.

Table 1. Demographic and Clinical Characteristics of the Study Cohort (N = 23).

Characteristic	Category	Value
age (years)	mean ± SD	66.0 ± 8.2
	range	48–79
sex, n (%)	male	14 (60.9%)
	female	9 (39.1%)
ethnicity, n (%)	Kazakh	18 (78.3%)
	Russian	4 (17.4%)
	Tatar	1 (4.3%)
pathological T stage, n (%)	T1	1 (4.3%)
	T2	2 (8.7%)
	T3	18 (78.3%)
	T4	2 (8.7%)
N category, n (%)	N0	6 (26.1%)
	N1	2 (8.7%)
	N2	1 (4.3%)
	NX	14 (60.9%)
M category, n (%)	M0	23 (100%)
histological grade, n (%)	G2	22 (95.7%)
	G3	1 (4.3%)
keratinization, n (%)	present	17 (73.9%)
	absent	6 (26.1%)
tumor location, n (%)	upper third	1 (4.3%)
	upper–middle third	2 (8.7%)
	middle third	12 (52.2%)
	middle–lower third	2 (8.7%)
	lower third	6 (26.1%)
dysphagia at presentation, n (%)	yes	23 (100%)
BMI (kg/m^2), n = 13	mean ± SD (range)	24.9 ± 3.6 (19.1–29.5)
surgical treatment, n (%)	Ivor-Lewis esophagectomy	23 (100%)
neoadjuvant therapy, n (%)	none	23 (100%)

All 23 samples were histopathologically confirmed as ESCC, with tumors displaying malignant squamous cells arranged in infiltrating nests and sheets. A hallmark feature of keratinization, manifesting as pearl-like structures, was observed alongside disorganized growth patterns typical of this carcinoma subtype (Figure 1).

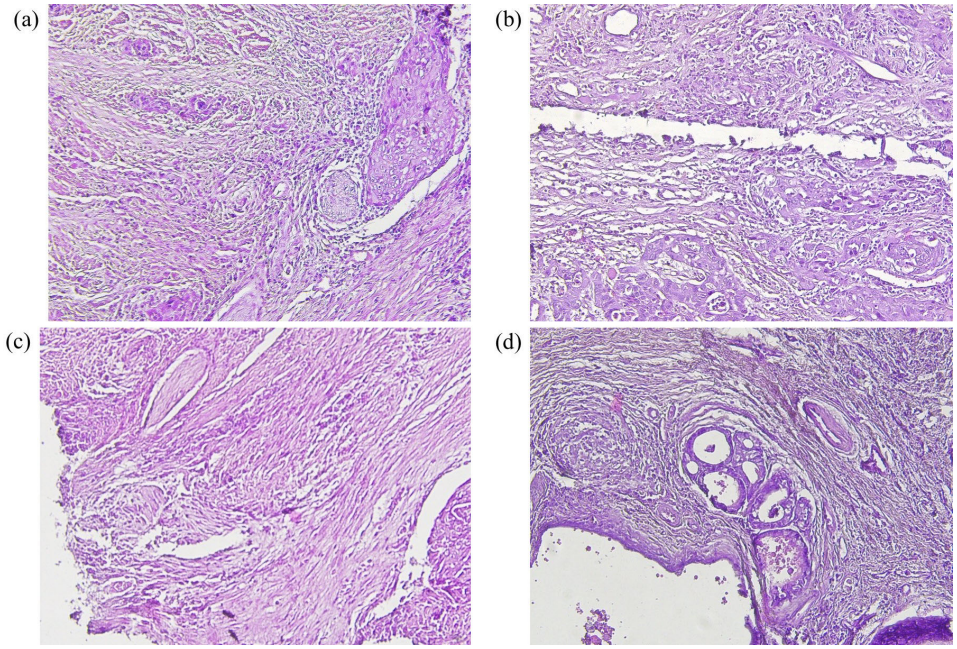

Figure 1. Representative histological images of tumor centers from four ESCC patients, showing infiltrating nests and sheets of malignant squamous cells with keratin pearl formation and disorganized growth patterns characteristic of squamous cell carcinoma in samples 14T (**a**), 11T (**b**), 13T (**c**) and 8T (**d**).

2.2. Fusion Gene Detection by Dual-Caller Consensus Approach

Fusion gene detection was performed on RNA-seq data from all 23 tumor samples using Arriba and STAR-Fusion. Arriba identified a total of 517 candidate fusion events across the cohort, with a range of 0 to 74 calls per patient. STAR-Fusion identified a total of 178 candidate fusion events, with a range of 0 to 28 calls per patient, consistent with the more stringent built-in filtering applied by this algorithm relative to Arriba. To generate a high-confidence set of fusion candidates, results from both tools were cross-referenced at the per-sample level, retaining only fusion events independently detected by both algorithms within the same patient. This consensus filtering approach yielded 14 fusion events across 9 of the 23 patients (39.1%), corresponding to 8 unique fusion gene pairs (Supplementary Tables S2 and S3). Of these, four fusion gene pairs were recurrent, detected in two or more patients—*MAP4K5–L2HGDH* (3 patients), *ADAMTS2–ENSG00000253652* (3 patients), *BAZ2B–WDSUB1* (2 patients), and *HUWE1–SUPT3H* (2 patients)—while the remaining four were detected in a single patient each: *RAB8A–CIB3* in patient 1T, and *KRAS–ETFRF1*, *KRAS–PIK3C2G* and a *PLEKHA5* fusion reported by both callers as *PLEKHA5–KRAS* in patient 8T; because the Sanger-confirmed junction of the latter lies in intergenic sequence 7.8 kb upstream of *KRAS* (Section 2.6.4), it is referred to hereafter as *PLEKHA5*–intergenic (*KRAS*-adjacent). The overall dual-caller consensus pipeline and downstream validation workflow are summarized in Figure 2.

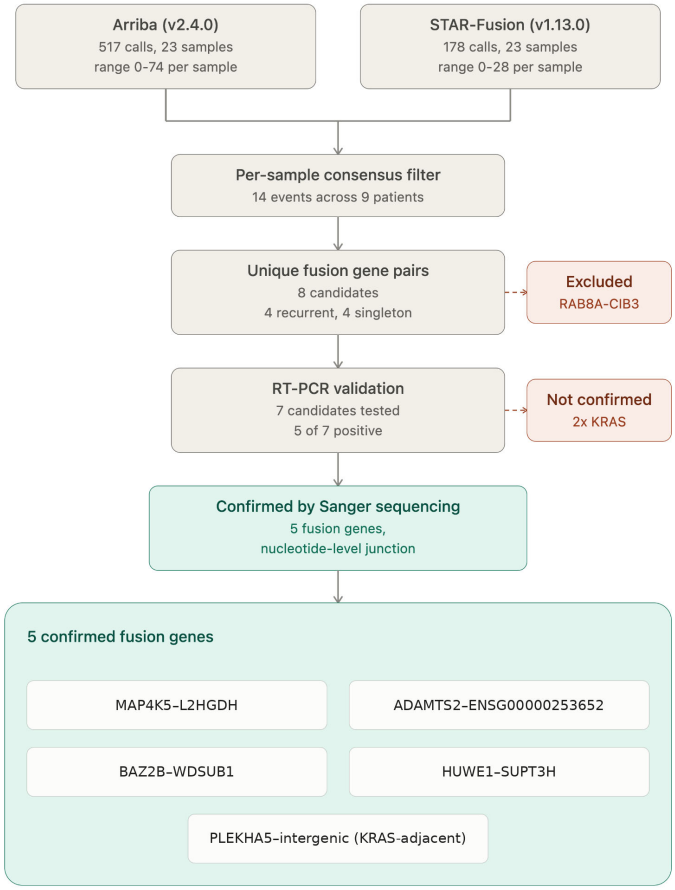

Figure 2. Schematic representation of the fusion gene discovery, consensus filtering, and experimental validation pipeline.

2.3. Selection of Candidates for Experimental Validation

From the eight unique fusion gene pairs identified by consensus filtering, candidates were prioritized for experimental validation based on recurrence across patients. Four fusion gene pairs detected in two or more patients—*MAP4K5–L2HGDH*, *ADAMTS2–ENSG00000253652*, *BAZ2B–WDSUB1*, and *HUWE1–SUPT3H*—were selected on this basis. The remaining four candidates were each detected in a single patient. Of these, *RAB8A–CIB3* (patient 1T) was not taken forward, as singleton events were not prioritized for validation in the absence of additional supporting evidence. However, patient 8T was notable for harboring three independent *KRAS*-involving fusion candidates—*KRAS–ETFRF1*, *KRAS–PIK3C2G*, and *PLEKHA5*–intergenic (*KRAS*-adjacent)—all passing the consensus filter in the same sample. Given the recognized oncogenic significance of *KRAS* and the unusual clustering of multiple *KRAS* fusion events within a single tumor, all three were selected for experimental validation despite their non-recurrent status.

In total, seven fusion gene candidates were taken forward for RT-PCR validation. Fusion-spanning primers were designed for each candidate (Supplementary Table S4).

2.4. RT-PCR Validation

Of the seven fusion candidates selected for experimental validation, five produced amplicons of the expected size in at least one tested sample. ADAMTS2–ENSG00000253652 produced the expected product in all three patients in which it was detected computationally (11T, 19T, 33T); for sample 33T, a second primer pair was

designed to generate an extended amplicon spanning the junction, which also produced the expected product. MAP4K5–L2HGDH produced amplicons in all three patients with a computational call (12T, 13T, 14T), and HUWE1–SUPT3H in both (18T, 19T). BAZ2B–WDSUB1 produced an amplicon in sample 14T but not in sample 12T, despite being identified by both fusion-calling algorithms in this patient. Among the three KRAS-involving candidates from patient 8T, only the PLEKHA5 fusion produced a confirmatory amplicon; KRAS–ETFRF1 and KRAS–PIK3C2G did not yield detectable products and were excluded from further analysis.

The calls in 12T (MAP4K5–L2HGDH and BAZ2B–WDSUB1) and 18T (HUWE1–SUPT3H) were weakly supported, and none of them was confirmed by Sanger sequencing (Section 2.5). Moreover, 12T shared most of its Arriba calls with 14T, and 18T with 19T, with identical breakpoints and markedly weaker read support than in the partner sample (Supplementary Table S2). This pattern suggests low-level transfer of reads or material between samples rather than independent events, and these calls were therefore not counted as independent occurrences. In total, five fusion gene pairs produced RT-PCR products in at least one patient sample and were confirmed by Sanger sequencing: MAP4K5–L2HGDH, ADAMTS2–ENSG00000253652, BAZ2B–WDSUB1, HUWE1–SUPT3H, and PLEKHA5–intergenic (KRAS-adjacent). Two of them were recurrent: ADAMTS2–ENSG00000253652 (11T, 19T, 33T) and MAP4K5–L2HGDH (13T, 14T).

2.5. Sanger Sequencing Confirmation

To confirm the predicted fusion junctions at the nucleotide level, RT-PCR-positive amplicons were subjected to Sanger sequencing. For each of the five validated fusion gene pairs, the resulting sequence aligned to the human reference genome (GRCh38) across the predicted breakpoint, confirming the presence of the chimeric junction in at least one patient sample. *MAP4K5–L2HGDH* was confirmed by Sanger sequencing in two of the three PCR-positive samples. *ADAMTS2–ENSG00000253652* was confirmed in all three PCR-positive samples. *BAZ2B–WDSUB1* was confirmed in its single PCR-positive sample. *HUWE1–SUPT3H* was confirmed in one of the two PCR-positive samples. *PLEKHA5*–intergenic (*KRAS*-adjacent) was confirmed in sample 8T. The individual patient distribution, computational recurrence, and experimental validation status across all consensus-filtered fusion events are summarized in Figure 3.

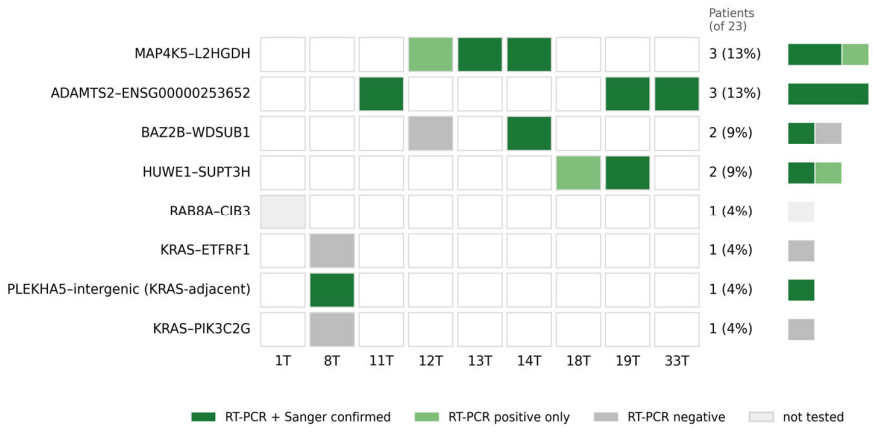

Figure 3. Patient-level landscape and experimental validation status of consensus-filtered fusion gene candidates in ESCC. Grid showing the distribution of the 8 unique consensus fusion gene pairs across the 9 fusion-positive patients (39.1% of the cohort). Cell color indicates the validation status in each patient. The number of patients carrying each candidate (percentage of all 23 patients) and a bar summarizing validation status are shown on the right.

In total, nucleotide-level confirmation of the predicted fusion junction was obtained for all five fusion gene pairs identified by consensus computational detection and RT-PCR. Representative Sanger sequencing chromatograms aligned to the predicted junction sequence are shown in Figure 4.

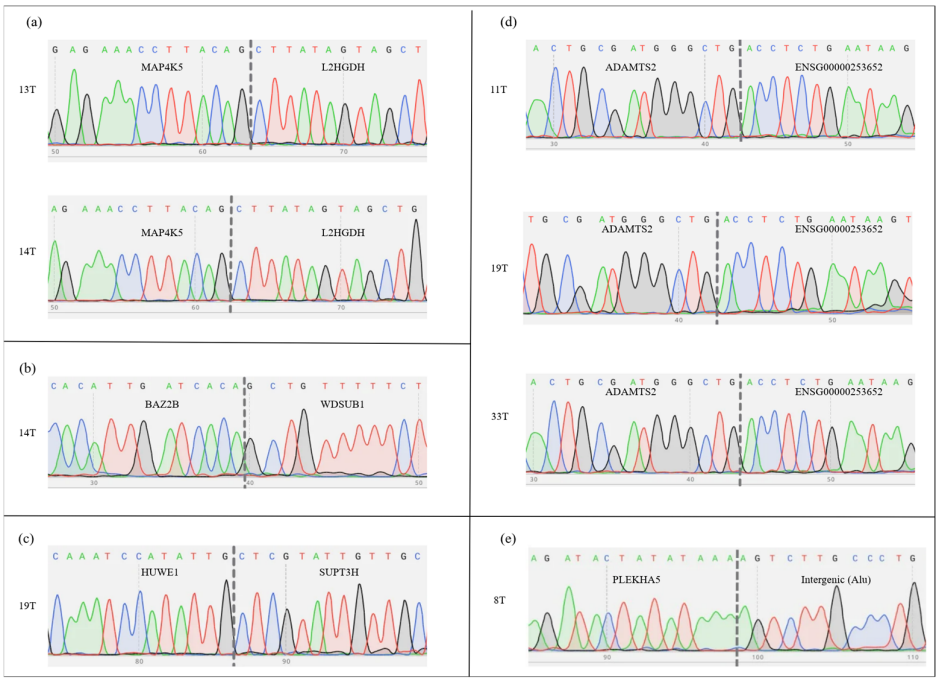

Figure 4. Sanger sequencing chromatograms confirming nucleotide-level junction sequences for validated fusion gene pairs. Representative chromatograms showing the exact junction sequences for (**a**) *MAP4K5–L2HGDH* (samples 13T, 14T), (**b**) *BAZ2B–WDSUB1* (sample 14T), (**c**) *HUWE1–SUPT3H* (sample 19T), (**d**) *ADAMTS2–ENSG00000253652* (samples 11T, 19T, 33T), and (**e**) *PLEKHA5*–intergenic (*KRAS*-adjacent), in which the 3′ sequence is an intergenic Alu element (sample 8T). Vertical dashed lines indicate predicted chimeric breakpoints.

2.6. Predicted Protein Consequences and Structural Modeling of the Validated Fusions

To assess the likely functional consequences of the five validated fusions, we reconstructed each fusion transcript from the Sanger-confirmed junction sequence and the Ensembl reference transcripts used for annotation. We then translated each transcript in silico and mapped the retained and lost protein domains using UniProt annotations (Figure 5; Table 2). For the four events with a junction in coding sequence, the reconstructed sequence reproduced the junction peptide reported by Arriba. Only *MAP4K5–L2HGDH* joined the coding sequences of both partners in frame and was therefore expected to encode a chimeric protein; this chimera was analyzed further by structure prediction. Three fusions were predicted to truncate their 5′ partner (ADAMTS2, HUWE1 and PLEKHA5). The fifth, *BAZ2B–WDSUB1*, joined the untranslated regions of the two genes without altering any coding sequence. Three of the fusions involve genes that are neighbors on the same chromosome and could therefore also arise by read-through in normal tissue: *MAP4K5–L2HGDH*, *ADAMTS2–ENSG00000253652* and *BAZ2B–WDSUB1*. None of these three was detected by Arriba in RNA-seq data from 11 healthy esophageal squamous epithelium samples from a public atlas of normal human tissues (GEO GSE120795) [24], the same reference set used in our previous study [6].

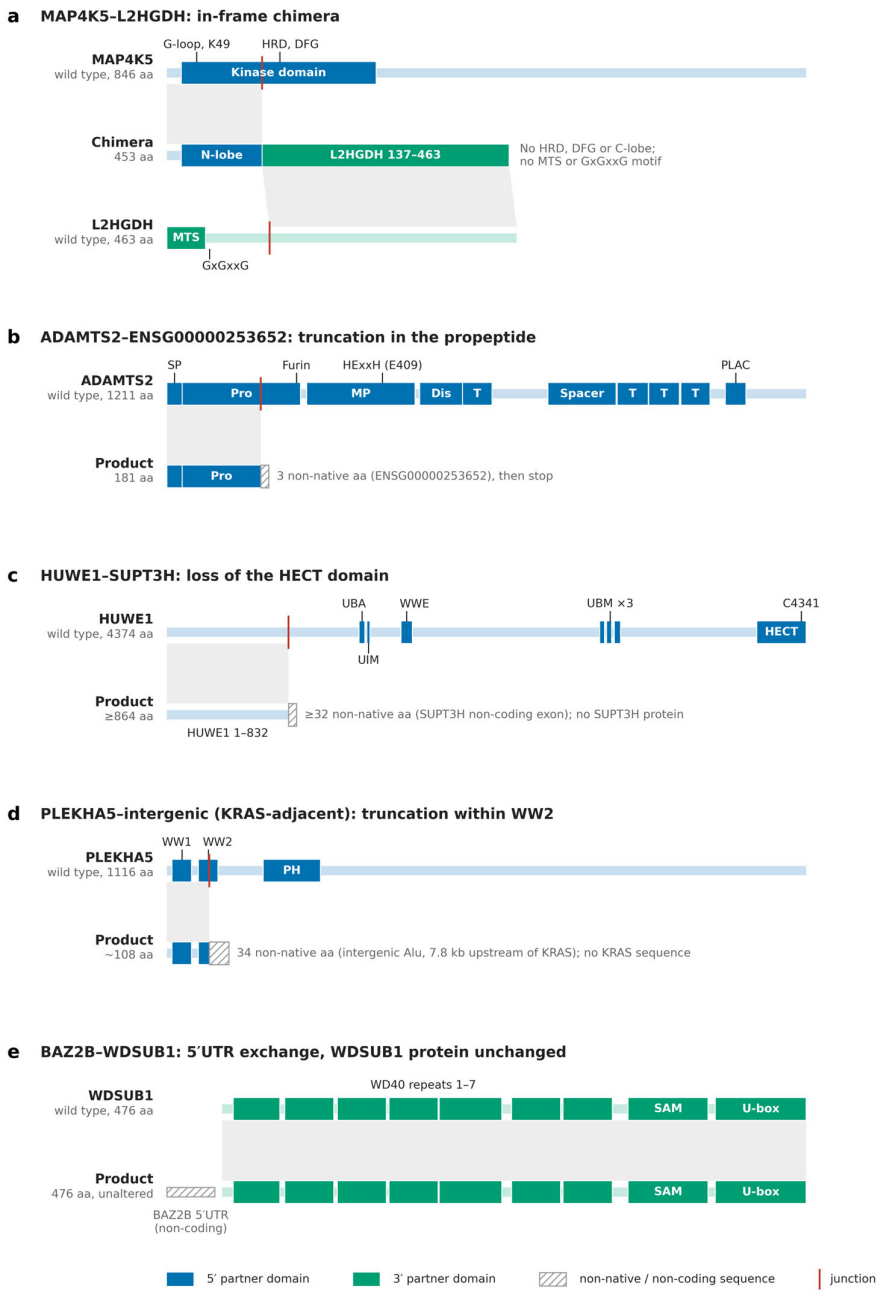

Figure 5. Predicted protein products of the five validated fusions. For each fusion, the wild-type partner protein(s) (top) and the predicted fusion product (bottom) are drawn to scale; grey bands map retained segments onto the product and red ticks mark the junction. (**a**) *MAP4K5–L2HGDH*: the chimera (middle) joins MAP4K5 residues 1–126 to L2HGDH residues 137–463. (**b**) *ADAMTS2–ENSG00000253652*. (**c**) *HUWE1–SUPT3H*. (**d**) *PLEKHA5*–intergenic (*KRAS*-adjacent). (**e**) *BAZ2B–WDSUB1*: the *BAZ2B* contribution is non-coding, and WDSUB1 is unaltered. Blue, 5′-partner domains; green, 3′-partner domains; hatched, non-native or non-coding sequence. Domain positions follow UniProt canonical isoforms. MTS, mitochondrial targeting sequence; SP, signal peptide; Pro, propeptide; MP, metalloproteinase domain; Dis, disintegrin-like domain; T, thrombospondin type-1 repeat; PH, pleckstrin-homology domain.

Table 2. Predicted protein products of the five validated fusions.

Fusion (Sanger-Confirmed Samples)	Junction	Reading Frame	Predicted Product	Retained	Lost
MAP4K5–L2HGDH (13T, 14T)	CDS–CDS; 14q22.1	In frame	453-residue chimera: MAP4K5 1–126 + L2HGDH 137–463	MAP4K5 kinase N-lobe (G-loop, K49); L2HGDH C-terminal region	MAP4K5 HRD, DFG and C-lobe; L2HGDH mitochondrial targeting sequence and GxGxxG FAD motif
ADAMTS2–ENSG00000253652 (11T, 19T, 33T)	CDS–non-coding exon; 5q35.3	Continues into non-coding sequence	181 residues: ADAMTS2 1–178 + 3 non-native	Signal peptide; part of propeptide	Furin site; metalloproteinase domain (HExxH, E409); all C-terminal domains
HUWE1–SUPT3H (19T)	CDS–non-coding exon; t(X;6)	Out of frame	HUWE1 1–832 + ≥32 non-native	N-terminal DUF908/DUF913 region	UBA, UIM, WWE, UBMs, BH3 region, HECT domain (C4341); no *SUPT3H* coding sequence
PLEKHA5–intergenic (*KRAS*-adjacent) (8T)	CDS–intergenic Alu; 12p12.3–p12.1	Continues into non-coding sequence	~108 residues: PLEKHA5 1–74 + 34 non-native	WW1; part of WW2	Rest of WW2; PH domain; residues 75–1116; no *KRAS* sequence
BAZ2B–WDSUB1 (14T)	5′UTR–5′UTR; 2q24.2	Not applicable	Wild-type WDSUB1 (476 residues)	All WDSUB1 domains	No *BAZ2B* coding sequence on the fused transcript

Positions refer to UniProt canonical isoforms: MAP4K5 Q9Y4K4, L2HGDH Q9H9P8, ADAMTS2 O95450, HUWE1 Q7Z6Z7, PLEKHA5 Q9HAU0, WDSUB1 Q8N9V3.

2.6.1. MAP4K5–L2HGDH Encodes an In-Frame Chimera Lacking the Catalytic Machinery of Both Partners

MAP4K5–L2HGDH was detected computationally and by RT-PCR in three tumors (12T, 13T and 14T) and confirmed by Sanger sequencing in 13T and 14T. Read support was strongest in 14T (106 split reads and 26 discordant read pairs; Arriba confidence "high"); 13T had 5 split reads and 1 discordant pair (confidence "low"). *MAP4K5* and *L2HGDH* are neighboring genes on chromosome 14q22.1, and the junction joins a *MAP4K5* coding exon to an *L2HGDH* coding exon, consistent with either an interstitial deletion or transcriptional read-through.

The junction preserves the reading frame and is predicted to encode a 453-residue chimera consisting of *MAP4K5* residues 1–126 fused to *L2HGDH* residues 137–463 (Figure 5a). The *MAP4K5* portion retains the glycine-rich ATP-binding loop (residues 26–34) and the catalytic Lys49. It lacks the HRD catalytic motif (residues 138–140), the DFG motif (158–160) and the entire kinase C-lobe, so the chimera cannot function as a kinase. The *L2HGDH* portion lacks the N-terminal mitochondrial targeting sequence (residues 1–51) and the GxGxxG dinucleotide-binding motif (G55–G60) of the FAD-binding domain. It is therefore expected to remain outside the mitochondria and to lack the structural elements required for FAD binding.

We predicted the structure of the chimera with AlphaFold 3 (Figure 6a). The overall prediction had moderate confidence (pTM 0.69–0.70; ranking score 0.70–0.71; mean pLDDT 86.5). 92% of residues had pLDDT > 70; low-confidence residues (pLDDT < 50; 2%) were limited to the extreme N-terminus (residues 1–7) and a short linker just upstream of the junction (residues 122–124). Each half of the chimera was individually well defined. The *MAP4K5* segment (residues 21–126; mean pLDDT 80.7) was modeled on

MAP4K1/MAP4K3 kinase-domain templates. The *L2HGDH* segment (mean pLDDT 89.6) was predicted without a structural template. The mean predicted aligned error (PAE) was 6.1 Å within the *MAP4K5* segment and 4.5 Å within the *L2HGDH* segment, and the five models agreed to within 1.1 Å and 0.4 Å Cα RMSD, respectively. In contrast, the PAE between the two halves was high (22.6–25.0 Å; Figure 6b), whole-chain RMSD between models reached 16.5 Å, and inter-segment contacts were confined to residues adjacent to the junction. The model therefore does not support a stable interface between the two halves; instead, it suggests two independent units joined by a flexible hinge.

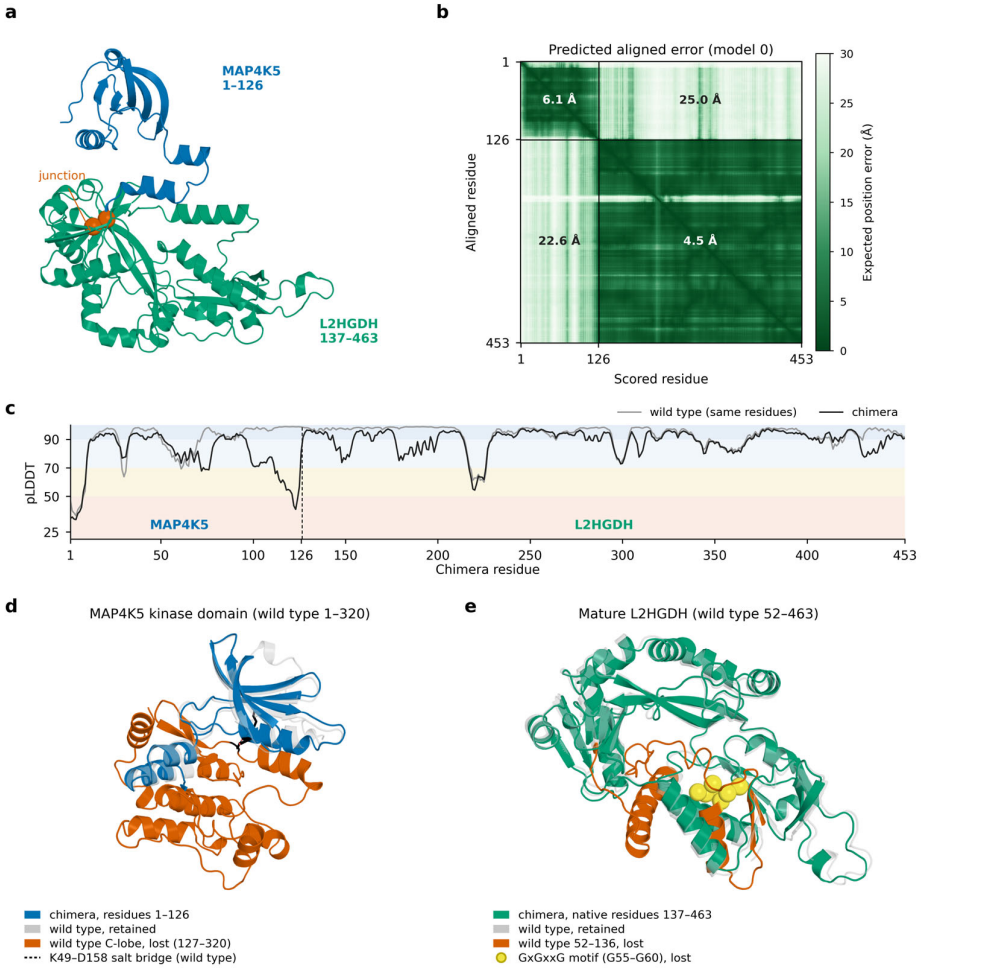

Figure 6. AlphaFold 3 models of the MAP4K5–L2HGDH chimera and its wild-type partners. (**a**) Model of the 453-residue chimera (model 0); the MAP4K5 segment (residues 1–126) is blue, the L2HGDH segment (native residues 137–463) green, and the junction residues are shown as orange spheres. Residues 1–7 (pLDDT < 50) are hidden. (**b**) Predicted aligned error for the chimera; values give the mean expected error within and between the MAP4K5 (residues 21–126) and L2HGDH segments. (**c**) Per-residue confidence (pLDDT) of the chimera (black) and of the corresponding residues in the wild-type models (grey); background bands mark the standard pLDDT confidence categories. (**d**) MAP4K5 segment of the chimera superimposed on the wild-type MAP4K5 kinase domain; the C-lobe absent from the chimera is orange, and the Lys49–Asp158 pair is shown as sticks (C-terminal residues 296–320, predicted with low confidence, are hidden). (**e**) L2HGDH segment of the chimera superimposed on mature wild-type L2HGDH; the deleted segment (residues 52–136) is orange and the Cα atoms of the GxGxxG motif (G55–G60) are yellow spheres. Superpositions used Cα atoms of shared residues.

To define what each partner loses, we modeled wild-type *MAP4K5* (residues 1–320, spanning the kinase domain) and mature *L2HGDH* (residues 52–463) with the same settings. Both were predicted with high confidence (*MAP4K5*: pTM 0.85–0.86, mean pLDDT 85.7; *L2HGDH*: pTM 0.94–0.95, mean pLDDT 94.3). In wild-type *MAP4K5*, the N- and C-lobes form a compact kinase domain (inter-lobe PAE 5.3 Å), and Lys49 forms a salt bridge with Asp158 of the DFG motif (2.7 Å) (Figure 6d). In the chimera, the *MAP4K5* N-lobe retained its wild-type fold (Cα RMSD 1.6 Å over residues 21–126), but it was predicted with lower confidence (mean pLDDT 80.7 vs. 91.7 in the wild type; Figure 6c). Lys49 lost its DFG partner, and approximately 2700 Å^2 of the N-lobe surface that is normally buried against the C-lobe became exposed; in the wild type, 52 of the 126 retained residues contact the C-lobe.

The deleted *L2HGDH* segment (residues 52–136) is not a disordered region. In the wild-type model, it was the most confidently predicted part of the protein (mean pLDDT 97.5) and packed against 93 of the 327 retained residues, 45 of them hydrophobic (Figure 6e). Its removal exposes approximately 3600 Å^2 (23%) of additional surface on the retained fragment, including the residues that surround the GxGxxG motif in the wild type (residues 195–202, 257–260 and 442–450). AlphaFold nevertheless placed the retained *L2HGDH* fragment in its wild-type conformation (Cα RMSD 1.06 Å over residues 137–463). Deviations above 2 Å were limited to residues 230–237, 306 and 439, and confidence dropped only modestly (mean ΔpLDDT −4.0). The largest decrease (−24.1) was at residue 158, which contacts the deleted segment in the wild type.

2.6.2. ADAMTS2–ENSG00000253652 Truncates ADAMTS2 Within Its Propeptide

The *ADAMTS2–ENSG00000253652* junction was identical at the nucleotide level in all three fusion-positive tumors (chr5:179,343,767/chr5:179,503,740). It was supported by 16, 82 and 39 split reads in 11T, 19T and 33T, respectively (Arriba confidence "high" in each). The junction joins *ADAMTS2* codon 178, located within the propeptide, to the second and terminal exon of *ENSG00000253652*, a non-coding transcript located ~160 kb upstream of *ADAMTS2* on the same strand. The *ADAMTS2* reading frame continues into non-coding sequence for three codons before reaching a stop codon (…SNCDGL-TSE*). The predicted product is therefore a 181-residue fragment that retains the signal peptide (residues 1–29) and part of the propeptide (residues 30–178) (Figure 5b). It lacks the furin cleavage site, the metalloproteinase domain (residues 266–470) with its zinc-binding HExxH motif and catalytic Glu409, and all C-terminal ancillary domains (disintegrin-like domain, thrombospondin type-1 repeats, spacer and PLAC). The event therefore cannot give rise to an active procollagen N-proteinase. Because the stop codon lies in the terminal exon of the non-coding partner, the fusion transcript is not expected to be targeted by nonsense-mediated decay.

2.6.3. HUWE1–SUPT3H Removes the HECT Domain of HUWE1

In tumor 19T, *HUWE1* (Xp11.22) was joined to *SUPT3H* (6p21.1), consistent with a t(X;6) rearrangement (9 split reads; Arriba confidence "high"). *HUWE1* is the 5′ partner, so the fusion transcript is expressed from the *HUWE1* promoter. The junction follows *HUWE1* codon 832 and continues out of frame into a non-coding *SUPT3H* exon. The predicted product therefore contains the N-terminal 19% of HUWE1 (residues 1–832, including the DUF908 and DUF913 regions), followed by a non-native C-terminal extension of at least 32 residues (Figure 5c). It lacks the UBA (residues 1316–1355), UIM (1370–1389) and WWE (1603–1680) domains, the three ubiquitin-binding motifs, the BH3 region, and the entire C-terminal HECT domain (4038–4374), including the catalytic Cys4341. No *SUPT3H* coding sequence is retained. The fusion is therefore predicted to abolish HUWE1 E3 ubiquitin-ligase activity on the rearranged allele without generating any SUPT3H

protein. The non-native extension is encoded by sequence that appears to be derived from an Alu element. The reverse primer used for validation was placed in the unique *SUPT3H* sequence immediately downstream of the junction, upstream of the Alu-derived segment.

2.6.4. The PLEKHA5 Fusion in Tumor 8T Lies Within a KRAS-Associated 12p Expression Block

In tumor 8T, the Sanger-confirmed junction joined *PLEKHA5* codon 74, within its second WW domain, to an intergenic Alu element 7.8 kb upstream of *KRAS* (chr12:25,258,749; 24 split reads; Arriba confidence "high"). We refer to this event as *PLEKHA5*–intergenic (*KRAS*-adjacent). The validated fusion transcript contains no *KRAS* sequence. Translation continues for 34 codons into the Alu sequence before a stop codon, yielding a predicted ~108-residue product (Figure 5d). This product retains the first WW domain and part of the second, but lacks the pleckstrin-homology domain (residues 169–268) and the remaining 93% of PLEKHA5. A lower-support alternative junction joining the same *PLEKHA5* exon to the 5′UTR of *KRAS* exon 2 was also called (8 split reads; confidence "medium") but was not tested experimentally. In that transcript, *KRAS* would lie out of frame with *PLEKHA5*, and the *PLEKHA5*-frame stop codon overlaps the *KRAS* start codon. Neither transcript is therefore expected to encode a PLEKHA5–KRAS chimeric protein.

Because 8T carried two further *KRAS*-associated fusion calls (*KRAS–ETFRF1* and *KRAS–PIK3C2G*) and showed marked overexpression of *KRAS* and *PLEKHA5*, we examined the expression of all protein-coding genes across chr12:15–30 Mb in 8T relative to the other 22 tumors (Figure 7). Two discrete blocks of co-overexpressed genes were observed. The first spanned chr12:18.2–19.7 Mb, from *PIK3C2G* to *AEBP2*, and included *PLEKHA5* (*PLEKHA5* and *AEBP2* were +3.6 and +3.4 VST units above the mean of the other tumors, respectively). The second spanned chr12:25.1–25.65 Mb, from *DNAI7* to *LMNTD1*, and included *KRAS* and *ETFRF1* (+4.4 and +7.0 VST units, respectively). The 5.4-Mb interval between the two blocks was not overexpressed. All three 8T fusion calls mapped within or at the boundaries of these blocks, and the validated junction directly connects them. This pattern is consistent with a complex amplicon joining two non-contiguous segments of 12p, with the fusion transcripts arising at the junctions between segments; confirmation of copy-number gain would require DNA-level analysis.

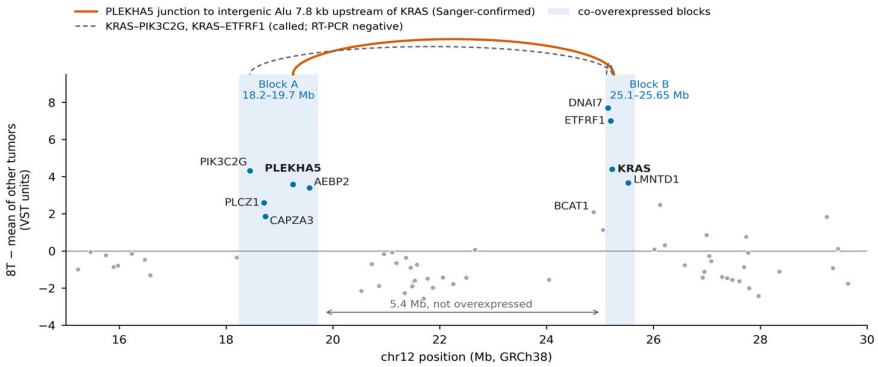

Figure 7. Two co-overexpressed chr12p12 segments in tumor 8T. For each protein-coding gene in chr12:15–30 Mb, the difference between VST-normalized expression in 8T and the mean of the other 22 tumors is plotted against genomic position (GRCh38). Shaded areas mark the two co-overexpressed blocks (18.2–19.7 Mb and 25.1–25.65 Mb). Arcs connect the partners of the three 8T fusion calls (approximate gene positions): the solid arc is the Sanger-confirmed *PLEKHA5* junction to an intergenic Alu element 7.8 kb upstream of *KRAS*; dashed arcs are calls not confirmed by RT-PCR (the *KRAS–ETFRF1* arc is very short because the genes are adjacent).

2.6.5. BAZ2B–WDSUB1 Exchanges Untranslated Regions Without Altering the WDSUB1 Protein

In tumor 14T, the *BAZ2B–WDSUB1* junction (23 split reads and 19 discordant pairs; Arriba confidence "high") joins the 5′-untranslated exon(s) of *BAZ2B* to the 5′UTR of *WDSUB1*, 24 nt upstream of the *WDSUB1* start codon. The two genes are adjacent on chromosome 2q24.2, and the event is consistent with a deletion or transcriptional read-through. The fusion transcript contains no *BAZ2B* coding sequence; the only upstream ATG in the *BAZ2B*-derived sequence opens a four-codon open reading frame that terminates before the junction. The transcript is therefore predicted to encode full-length, unaltered WDSUB1 (476 residues, including all seven WD40 repeats, the SAM domain and the U-box) under the control of the *BAZ2B* promoter (Figure 5e). No chimeric protein was modeled for this event.

2.7. Genome-Wide Characteristics of Candidate Fusion Calls

To place the validated events in the context of all candidate calls, and to obtain independent evidence on their origin, we characterized the complete Arriba and STAR-Fusion call sets (Supplementary Tables S2, S3 and S5; Section 4.11). After alternative gene annotations of the same breakpoint pair were collapsed, the 517 Arriba calls corresponded to 495 distinct events (median 17 per tumor; range 0–73), and STAR-Fusion reported 178 calls (median 5 per tumor; range 0–28) (Figure 8a). The per-tumor numbers of calls from the two algorithms were moderately correlated (Spearman $\rho = 0.53$, $p = 0.009$; Figure 8b), and the three tumors with the most Arriba events (8T, 2T and 14T) were also among the four tumors with the most STAR-Fusion calls. Of the 495 Arriba events, 202 (40.8%) were classified as high confidence and 126 (25.5%) were interchromosomal, and 231 of the 369 intrachromosomal events (62.6%) joined breakpoints less than 1 Mb apart. A further 127 events (25.7%) joined the two genes head-to-head (5′–5′) or tail-to-tail (3′–3′), an orientation in which one partner is transcribed in antisense and no conventional sense–sense chimeric transcript can be formed (Figure 8c).

The FusionAnnotator annotations produced by STAR-Fusion showed that 40 of the 178 STAR-Fusion calls (22.5%) involved at least one gene annotated as an oncogene or included in commercial cancer gene panels (for example, *KRAS*, *STAT3*, *SETD2*, *EPHA2* and *YES1*), and that 21 calls (11.8%) had been reported previously in cancer fusion resources (TCGA, CCLE, DepMap, ChimerSeq, Klijn cell lines or TumorFusions). The latter included *MAP4K5–L2HGDH* (TCGA and ChimerSeq) and *BAZ2B–WDSUB1* (TCGA), consistent with the database entries discussed in Section 3, whereas *ADAMTS2–ENSG00000253652*, *HUWE1–SUPT3H* and the *PLEKHA5* fusion carried no such annotation (Supplementary Table S5). Only one Arriba call matched a gene pair in the Mitelman Database-derived list of known cancer fusions: *MTAP–CDKN2B-AS1* in 33T, a low-confidence call at the 9p21.3 *CDKN2A/B* locus supported by a single split read. No STAR-Fusion call matched this list.

Breakpoints were not uniformly distributed across the genome. In tumor 8T, 70 of 146 breakpoint ends (47.9%) mapped to chromosome 12, all within chr12:18.5–26.3 Mb. Of these, 58 lay inside the two co-overexpressed blocks defined from expression data (Figure 7) and the remaining 12 within 1 Mb of them, whereas none of the 843 breakpoint ends from the other 22 tumors fell within chr12:15–30 Mb (Figure 8d). This breakpoint-level signal is independent of the expression analysis and supports the interpretation that 8T carries a complex 12p rearrangement joining the *PIK3C2G–PLEKHA5* and *KRAS* segments. Across tumors, the cytobands with breakpoints in the largest number of tumors were 9q21.33 (8 tumors) and 20p13 (7 tumors) (Figure 8e). This recurrence was driven largely by a single medium-confidence interchromosomal call that joined *SMOX* (20p13) to an intergenic, pseudogene-containing region of 9q21.33 at identical coordinates in six

tumors (4T, 7T, 16T, 21T, 23T and 29T) and was not reported by STAR-Fusion. A somatic translocation is very unlikely to recur at base-pair resolution in six unrelated patients, so this call most probably reflects a mapping artifact. The next most frequently affected region was 11q13.3, with 18 breakpoint ends in six tumors (1T, 4T, 11T, 12T, 14T and 16T) involving *CCND1*, *CTTN*, *PPFIA1* and the region adjacent to *FGF19*; 13q14.2 (*RB1*, *DLEU2*, *RCBTB1*/*RCBTB2*) was also affected in six tumors, mostly by calls with low read support (Supplementary Table S5).

Finally, we explored whether the number of candidate calls per tumor was associated with clinicopathological variables (T category, N category, sex, ethnicity, keratinization, tumor location and body-mass index; Supplementary Table S5). Non-keratinizing tumors had more Arriba events than keratinizing tumors (median 28 vs. 13; nominal $p = 0.017$), and tumors from Kazakh patients had more high-confidence Arriba events than those from patients of other ethnicities (median 7.5 vs. 3; nominal $p = 0.030$). Neither difference remained significant after correction for multiple testing (Benjamini–Hochberg q = 0.31 for both), and none of the other 19 comparisons reached nominal significance.

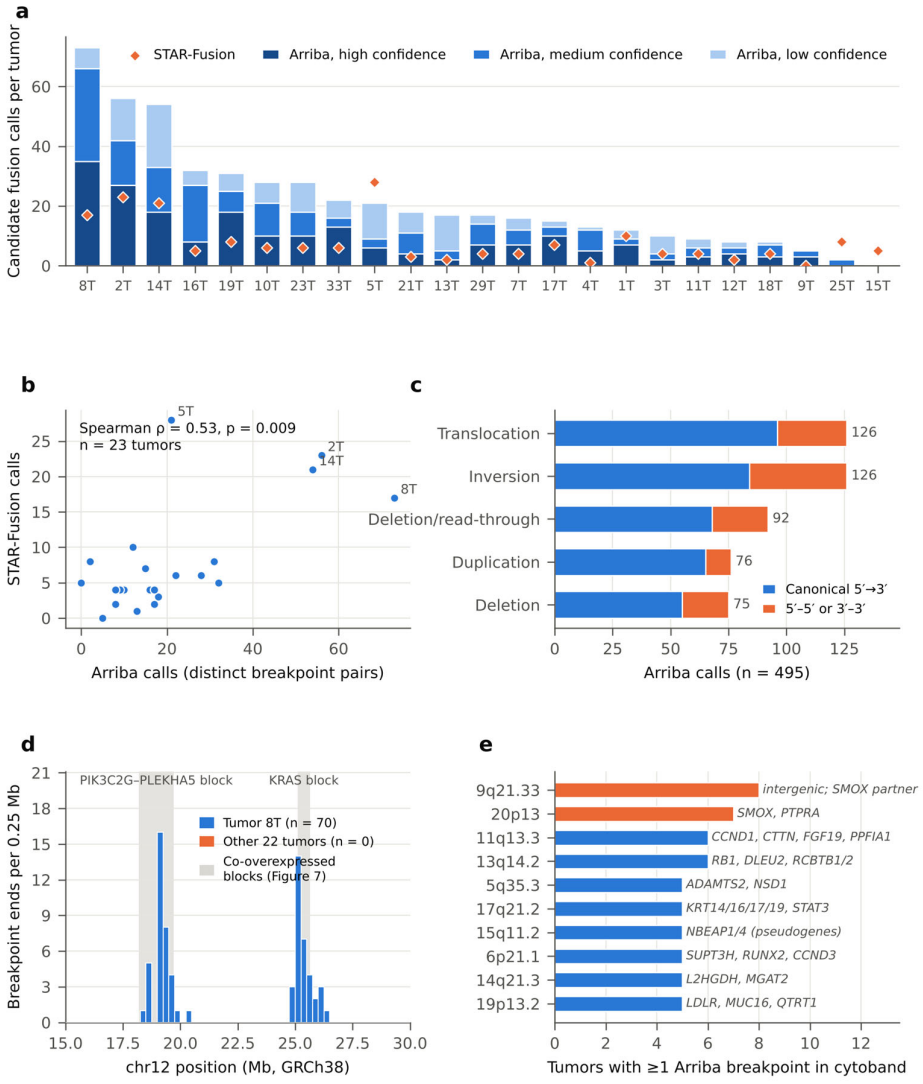

Figure 8. Genome-wide characteristics of candidate fusion calls. (**a**) Number of candidate calls per tumor: Arriba events (distinct breakpoint pairs) stacked by Arriba confidence level, with the number of STAR-Fusion calls shown as diamonds; tumors are ordered by the number of Arriba events. (**b**)

Per-tumor numbers of Arriba and STAR-Fusion calls. (**c**) Arriba event classes; orange segments denote events joining the two genes in head-to-head (5′–5′) or tail-to-tail (3′–3′) orientation. (**d**) Breakpoint ends in chr12:15–30 Mb in tumor 8T (blue) and in the other 22 tumors (none observed), in 0.25-Mb bins; grey bands mark the two co-overexpressed blocks shown in Figure 7. (**e**) The ten cytobands with Arriba breakpoints in the largest number of tumors, with representative genes at the breakpoints; orange bars mark cytobands whose recurrence is driven largely by an identical-coordinate *SMOX*–9q21.33 call, a suspected artifact.

3. Discussion

This study identified and experimentally confirmed five chimeric transcripts in esophageal squamous cell carcinoma (ESCC) from Kazakhstan: *MAP4K5–L2HGDH*, *ADAMTS2–ENSG00000253652*, *BAZ2B–WDSUB1*, *HUWE1–SUPT3H* and *PLEKHA5*–intergenic (*KRAS*-adjacent), in which *PLEKHA5* is joined to intergenic sequence 7.8 kb upstream of *KRAS* (hereafter, the *PLEKHA5* fusion). Each was detected independently by two fusion-calling algorithms and confirmed across the breakpoint by RT-PCR and Sanger sequencing. Four were computationally recurrent, and none has previously been validated in ESCC. The central finding of the protein-level analysis is that none of the five fusions is predicted to produce a chimera that retains an intact catalytic or signaling domain. *MAP4K5–L2HGDH* is the only in-frame chimera, and it lacks the catalytic machinery of both partners; ADAMTS2, HUWE1 and PLEKHA5 are truncated shortly after the junction; and *BAZ2B–WDSUB1* joins untranslated regions and leaves the *WDSUB1* coding sequence unchanged. This contrasts with the kinase-retaining fusions, such as *FGFR3–TACC3*, that have attracted most clinical attention in ESCC, and suggests that the validated fusions in this cohort act, if at all, through loss of function or altered regulation rather than through new oncogenic activity.

The gap between raw computational output and the final validated set shows why fusion detection from RNA sequencing requires more than a single algorithm. Arriba returned 517 candidate calls across the cohort (up to 74 in one sample), and STAR-Fusion, which applies stricter filtering against reference annotations, returned 178 (up to 28 per sample). Both figures reflect the known tendency of short-read fusion callers to generate false positives from alignment ambiguity, pseudogene homology and template switching during reverse transcription [25]. Restricting the analysis to calls detected by both algorithms reduced this set to 14 events across 9 patients (39.1% of the cohort), consistent with published multi-caller workflows [26,27]. Wet-lab testing then acted as a second filter. Of the seven candidates carried into RT-PCR, two *KRAS*-involving calls from the same singleton patient (8T) failed to amplify (*KRAS–ETFRF1* and *KRAS–PIK3C2G*). Weakly supported calls also failed validation in individual samples (*BAZ2B–WDSUB1* in 12T and *HUWE1–SUPT3H* in 18T, each supported by only two to four reads), and two validated junctions involve an Alu-derived sequence (*HUWE1–SUPT3H* and the *PLEKHA5* fusion). Junction-level sequencing, together with inspection of read support and sequence context, therefore remains necessary before any individual prediction is treated as confirmed.

MAP4K5–L2HGDH is the only validated event that joins the coding sequences of both partners in frame, and it was the focus of the structural analysis. Its MAP4K5 portion keeps only the kinase N-lobe, with the glycine-rich loop and catalytic Lys49, and lacks the HRD and DFG motifs and the entire C-lobe; in the wild-type model, Lys49 pairs with the DFG aspartate, a partner that is lost in the chimera. MAP4K5 belongs to the MAP4K kinases that activate JNK signaling and, together with MST1/2, phosphorylate LATS1/2 in the Hippo pathway [28]; the truncated kinase cannot contribute to either pathway. The L2HGDH portion lacks the mitochondrial targeting sequence and the GxGxxG dinucleotide-binding motif, and is therefore expected to remain outside the mitochondria and to be unable to bind FAD. L2HGDH clears the oncometabolite L-2-hydroxyglutarate (L-

2HG), and its loss raises L-2HG, which inhibits α-ketoglutarate-dependent TET and JmjC demethylases [29–31]. AlphaFold 3 predicted the two halves of the chimera as independent units joined by a flexible hinge, with no stable interface between them. It placed the L2HGDH fragment in its wild-type fold, even though the deleted segment is a well-ordered core element that contacts more than a quarter of the retained residues. This reflects how AlphaFold works: it predicts a fold from sequence and homologues and does not assess whether a truncated protein is stable. The large hydrophobic surface exposed by the deletion suggests that the fragment is more likely to misfold or be degraded in cells. Taken together, the more plausible consequence of this event is loss of one functional copy of each gene rather than a gain of function, which is particularly relevant for *L2HGDH*, whose deletion is a recognized route to L-2HG accumulation [31]. *MAP4K5* and *L2HGDH* are neighbors on 14q22.1, and read-through chimeras between adjacent genes also occur in non-malignant tissue [32]. The fusion was, however, absent from 11 public normal esophageal RNA-seq samples. Whether it arises from an interstitial deletion or from transcriptional read-through remains to be established by genomic breakpoint analysis.

ADAMTS2–ENSG00000253652 was the most strongly supported recurrent event, with the same junction found in three patients (11T, 19T and 33T) and high read support in each. It truncates ADAMTS2 within its propeptide: the reading frame runs three codons into a non-coding sequence before a stop codon, leaving a 181-residue product with the signal peptide and part of the propeptide. The furin cleavage site, the metalloproteinase domain and all C-terminal ancillary domains are lost [33,34]. Because the stop codon lies in the terminal exon of the non-coding partner, the fusion transcript is not expected to undergo nonsense-mediated decay, consistent with the elevated expression of both loci in the fusion-positive tumors (Supplementary Figure S1). The rearrangement therefore behaves as a stable, non-catalytic truncation and, given the role of ADAMTS2 in procollagen processing and matrix organization [35,36], could disrupt extracellular matrix remodeling. The origin of this event needs caution. The junction corresponds to a tandem duplication of the ~160-kb interval between the two loci, and rare duplications spanning both breakpoints are reported in the general population (gnomAD-SV v4; combined allele frequency <0.05%) [37,38]. Without matched non-tumor material, a germline origin cannot be excluded. Finding a variant this rare in 3 of 23 unrelated patients would be unexpected, unless such duplications are enriched in the Kazakh population, which is under-represented in current reference databases. The fusion was absent from 11 public normal esophageal RNA-seq samples, but these derive from other individuals and cannot address a germline origin in our patients. Distinguishing a recurrent somatic event from a population-enriched germline variant is therefore a priority for follow-up.

HUWE1–SUPT3H, confirmed in 19T, joins the X-linked HECT E3 ubiquitin ligase HUWE1 to SUPT3H, a subunit of the SAGA/STAGA coactivator complex [39,40]. *HUWE1* is the 5′ partner, so the fusion transcript is driven by the *HUWE1* promoter. The junction follows codon 832 and continues out of frame, so the product retains only the N-terminal 19% of HUWE1, followed by a short non-native tail, and loses the ubiquitin-binding domains, the BH3 region and the catalytic HECT domain. HUWE1 regulates the turnover of substrates including MCL1 and p53 [39], and the fusion is expected to abolish HUWE1 ligase activity on the rearranged allele. No *SUPT3H* coding sequence is retained, and the non-rearranged *SUPT3H* allele is presumably intact; the event is therefore best described as a truncating *HUWE1* alteration rather than a dual loss of function.

BAZ2B–WDSUB1, confirmed in 14T, differs from the other events in that both breakpoints lie in 5′-untranslated regions. The fusion transcript contains no *BAZ2B* coding sequence and is predicted to encode full-length, unaltered WDSUB1 under the control of the *BAZ2B* promoter. Promoter-exchange events can change the level or regulation of the downstream gene without altering its protein, but *WDSUB1* expression in the fusion-

positive sample did not differ clearly from fusion-negative tumors (Supplementary Figure S1). Together with the adjacency of the two genes on 2q24.2, this suggests that the event may be a transcriptional read-through with limited functional consequence, although it was not detected in 11 public normal esophageal samples.

The *PLEKHA5* fusion in 8T is best interpreted as a structural by-product of a complex 12p rearrangement rather than as an oncogenic fusion. The confirmed junction joins *PLEKHA5* exon sequence encoding the N-terminal WW domains to an intergenic Alu element 7.8 kb upstream of *KRAS*; it retains neither the PLEKHA5 pleckstrin-homology domain nor any *KRAS* coding sequence, and yields a predicted ~108-residue truncated product. Tumor 8T showed two discrete blocks of strongly co-overexpressed genes on 12p, one containing *KRAS* and *ETFRF1* and the other containing *PIK3C2G*, *PLEKHA5* and *AEBP2*, and all three 8T fusion candidates mapped within or at the boundaries of these blocks (Figure 7). This pattern is typical of focal amplicons that combine distant chromosomal segments, in which fusion transcripts arise at segment boundaries as passengers. The marked *KRAS* overexpression in this tumor is therefore more plausibly explained by copy-number gain than by the fusion. *KRAS* amplification has been reported in esophageal cancers [41], and DNA-level copy-number analysis would be required to confirm amplification in 8T.

We cross-referenced the five validated fusion pairs against global repositories (TCGA-ESCA [41], ChimerDB 4.0 [42], FusionGDB 2.0 [43] and COSMIC [44]). *ADAMTS2–ENSG00000253652*, *HUWE1–SUPT3H* and the *PLEKHA5* fusion had no prior entries. Unvalidated bioinformatic predictions exist in FusionGDB [43] for *BAZ2B–WDSUB1* (from a single adrenocortical carcinoma sample) and for *MAP4K5–L2HGDH* (from intrahepatic cholangiocarcinoma in CCLE [45] and a single TCGA esophageal carcinoma sample, TCGA-2H-A9GQ-01A [41]). None of these five events had previously been confirmed experimentally in esophageal cancer. Recurrent fusions with established clinical relevance, such as *FGFR3–TACC3* and *PLEKHA1–TACC2*, show that structural rearrangements can carry therapeutic and prognostic weight in ESCC. The events described here are of a different kind: truncating, regulatory or passenger alterations. They extend the documented spectrum of structural variation in ESCC and show that experimental validation of a fusion transcript should be followed by assessment of its predicted protein product before biological significance is inferred.

The genome-wide analysis of all candidate calls adds two further observations. First, calls were concentrated where structural instability would be expected. In 8T, nearly half of all breakpoint ends fell within the two co-overexpressed 12p blocks or their immediate flanks, and no other tumor had breakpoints in this region; this supports, from an independent type of evidence, the interpretation of this tumor as carrying a complex *KRAS*-containing amplicon. A breakpoint hotspot was also found at 11q13.3 in six tumors, involving *CCND1*, *CTTN*, *PPFIA1* and the region adjacent to *FGF19*. *CCND1* is frequently amplified in ESCC [41], and fusion transcripts arise preferentially at copy-number transitions within amplified regions and from tandem duplications [46,47]. These calls are therefore best regarded as markers of underlying copy-number alterations rather than as driver fusions, and DNA-level copy-number analysis would be needed to confirm 11q13.3 amplification in these tumors. Second, the most recurrent calls in the raw output were not the most credible: a *SMOX*–9q21.33 junction recurred at identical coordinates in six tumors but was not reported by STAR-Fusion and most probably represents a mapping artifact. Together with the absence of any association between the number of calls and clinicopathological variables after correction for multiple testing, these observations indicate that call counts and raw recurrence are poor proxies for biological relevance in small cohorts, and that copy-number and DNA-level analyses will be needed to interpret RNA-level structural events in ESCC.

The cohort in which these fusions were identified differs from the populations underlying most existing ESCC genomic resources. Of the 23 patients, 78.3% self-identified as Kazakh, and 87% presented at stage T3–T4 without prior systemic or radiotherapeutic treatment. Advanced-stage, treatment-naive disease typically reflects a longer period of tumor evolution, during which structural instability has more opportunity to accumulate; whether this contributes to the recurrence of events such as *MAP4K5–L2HGDH* and *ADAMTS2–ENSG00000253652* cannot be established from a single-stage cohort. Ancestral background is a further variable, and the *ADAMTS2* locus illustrates its practical importance. Structural variants that are rare in current reference databases may be more frequent in under-represented populations, so recurrence observed in tumor RNA alone cannot be equated with recurrent somatic rearrangement. Separating population-specific, stage-driven and germline contributions will require matched normal material and comparison with earlier-stage and ancestry-matched cohorts.

Several aspects of the study constrain how these results should be interpreted. First, validation attrition was observed: *MAP4K5–L2HGDH* amplified in 12T, 13T and 14T but was sequence-confirmed only in 13T and 14T; *HUWE1–SUPT3H* amplified in 18T and 19T but was confirmed only in 19T; and *BAZ2B–WDSUB1* amplified in only one of its two candidate tumors. In each case, the unconfirmed samples had the weakest read support, so actual recurrence rates may be lower than the computational counts suggest. Second, primer placement was constrained by repeat sequence. For *HUWE1–SUPT3H*, the short unique sequence between the junction and an Alu element required placing the reverse primer adjacent to the breakpoint, so the *SUPT3H* side of the Sanger read is largely primer-derived and evidence for the *SUPT3H* sequence rests mainly on junction-spanning RNA-seq reads. For the *PLEKHA5* fusion, the reverse primer lay within an Alu element, and the assignment of that element to the region upstream of *KRAS* rests on RNA-seq alignment. Third, without DNA sequencing or matched normal tissue, we cannot distinguish genomic rearrangements from transcriptional read-through (*MAP4K5–L2HGDH*, *BAZ2B–WDSUB1*) or exclude a germline origin (*ADAMTS2–ENSG00000253652*), and the 12p amplicon in 8T and the 11q13.3 breakpoint hotspot (Section 2.7) are inferred from RNA-level data alone. Fourth, predicted protein products and AlphaFold models indicate which domains are retained and how they may fold, but they do not establish whether a protein is expressed, stable or active. Finally, the cohort size (n = 23) and the number of fusion-positive samples (n = 1–3 per pair) limit the precision of prevalence estimates and preclude meaningful statistical testing of associations between individual fusions and clinicopathological features, treatment or outcome; the study should therefore be regarded as exploratory.

Three lines of work follow directly from these findings. Genomic breakpoint PCR and analysis of matched non-tumor DNA are needed to establish whether each event is somatic, germline or transcriptional, most urgently for the recurrent *ADAMTS2–ENSG00000253652* junction. Protein-level studies in ESCC models, including Western blotting for the predicted truncated products and measurement of L-2HG in *MAP4K5–L2HGDH*-positive tumors, would test the loss-of-function predictions made here. Copy-number analysis of 12p in ESCC would establish how often *KRAS*-containing amplicons and their boundary fusions occur in this population. Until these questions are resolved, the five fusions described here should be regarded as characterized structural events rather than candidate therapeutic targets, and their value as biomarkers will depend on confirmation in larger, ancestry-matched cohorts.

4. Materials and Methods

4.1. Patient Cohort and Ethics

A cohort of 23 patients diagnosed with esophageal squamous cell carcinoma (ESCC) was recruited from the Multidisciplinary Medical Center in Astana, Kazakhstan. This cohort largely overlaps with that described in Sharip et al. [6], and full details of patient recruitment, surgical procedure, tissue collection, and histological validation have been reported therein and in Supplementary Table S1. Between 2013 and 2017, 184 patients with esophageal cancer were hospitalized at the Multidisciplinary Medical Center, of whom 54 underwent surgical resection; patients were eligible if they had histologically confirmed ESCC, had undergone Ivor-Lewis esophagectomy and had provided written informed consent [6]. Diagnoses were histologically confirmed, and tumor samples were classified according to the TNM classification system, 7th edition [48]. Ethics approval was granted by the institutional ethics review board of the National Laboratory Astana (protocol #02-2022, 31 March 2022), and written informed consent was obtained from all participants. The clinicopathological variables recorded for each patient were sex, self-reported ethnicity, height and weight (used to calculate body-mass index), tumor location, presence of dysphagia, TNM category, histological grade and keratinization. Systematic data on postoperative treatment and long-term outcome were not available for this retrospective surgical cohort (2013–2017). Clinical and demographic characteristics are summarized in Table 1, and individual patient data are provided in Supplementary Table S1.

4.2. Histopathology

Histopathological evaluation of each specimen was conducted via hematoxylin and eosin (H&E) staining as previously described in Sharip et al. [6] to confirm the ESCC diagnosis, determine tumor grade, assess metastatic involvement, and analyze tissue cellular composition. Representative images were captured at 4× and 10× magnifications using a CELENA X High Content Imaging System (Logos Biosystems, Anyang, Korea).

4.3. RNA Sequencing Data

Total RNA extraction, library preparation, and sequencing procedures were performed as described in Sharip et al. [6]. Briefly, paired-end RNA sequencing was carried out on an Illumina HiSeq2000 platform. Raw RNA-seq data in FASTQ format are publicly available in the NCBI Sequence Read Archive under accession number PRJNA608223. The same tumor sample FASTQ files were used as input for fusion gene detection in the present study.

4.4. Fusion Gene Detection and Consensus Filtering

Fusion gene detection was performed independently on tumor samples from each patient using two complementary algorithms: Arriba (v2.4.0) [49] and STAR-Fusion (v1.13.0) [26]. For Arriba, reads were first aligned to the human reference genome (GRCh38 primary assembly, Ensembl annotation v107 [50]) using STAR (v2.7.10a) [51] with parameters recommended by the Arriba documentation, including chimeric alignment settings optimized for fusion detection (--chimSegmentMin 10, --chimOutType WithinBAM HardClip, --chimMultimapNmax 50). Arriba was run with its bundled blacklist, known fusions, and protein domain databases (v2.4.0) to filter artifactual fusions and annotate candidates. For STAR-Fusion, reads were processed using the plug-and-play CTAT genome library built on GRCh38/Gencode v44 (October 2023 release) [52]. STAR-Fusion was run with default parameters, including its built-in filters for low read support and artifactual fusions.

To generate a high-confidence set of fusion candidates, results from both tools were cross-referenced at the per-sample level. Only fusion events independently detected by both Arriba and STAR-Fusion within the same patient sample were retained for further analysis. This consensus filtering strategy was chosen to maximize specificity and reduce false positives inherent to any single fusion detection algorithm, accepting a trade-off of potentially reduced sensitivity. Gene-pair orientation (donor–acceptor order) was verified manually to ensure consistent nomenclature across tools prior to matching. Fusions passing this per-sample consensus filter were then examined across all samples to identify recurrent events, defined as fusions detected in two or more patients.

4.5. Design of Fusion-Specific Primers

For each fusion candidate selected for validation, a chimeric reference sequence was assembled from the breakpoint coordinates reported by Arriba and STAR-Fusion by joining the 5′ partner sequence upstream of the breakpoint to the 3′ partner sequence downstream of it (GRCh38, Ensembl release 107). Fusion-specific primer pairs were designed using the PCR primer design module of the freely available web-based platform for advanced PCR design and sequence analysis (https://primerdigital.com/tools/, accessed on 16 April 2025) [53]. The forward primer was placed within the 5′ partner gene and the reverse primer within the 3′ partner gene, so that a product could be generated only from the chimeric transcript and not from either wild-type partner. Primer candidates were evaluated in the same platform for melting temperature, GC content, linguistic complexity, and the absence of stable hairpins, self-dimers, and cross-dimers [53]. The specificity of each primer pair was subsequently verified by in silico PCR against the human reference genome and transcriptome [54], which confirmed that none of the pairs was predicted to amplify a product from the native, non-rearranged partner genes or from off-target loci. For sample 33T, a second primer pair generating an extended amplicon across the *ADAMTS2–ENSG00000253652* junction was designed using the same procedure. All primers used for RT-PCR amplification and cycle sequencing are listed in Supplementary Table S4.

4.6. RT-PCR and Sanger Sequencing

Total RNA was quantified with a NanoDrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA), and 800 ng was reverse-transcribed using TaqMan Reverse Transcription Reagents (Applied Biosystems, Foster City, CA, USA) with random hexamers in a 20 µL reaction, at 25 °C for 10 min, 37 °C for 30 min and 95 °C for 5 min. Fusion junctions were amplified with the fusion-specific primers described in Section 4.5 in 30 µL reactions containing 3 µL cDNA, 0.2 µM of each primer, 0.16 mM dNTPs and 1 U Phusion Hot Start II DNA Polymerase in the supplied reaction buffer (Thermo Fisher Scientific). Cycling conditions were 98 °C for 60 s; 36 cycles of 98 °C for 5 s, 60 °C for 20 s and 72 °C for 10 s; and a final extension at 72 °C for 1 min. PCR products were analyzed by electrophoresis on 1% agarose gels. Amplicons were purified with ExoSAP-IT Express PCR Product Cleanup Reagent (Thermo Fisher Scientific) by incubating 5 µL of PCR product with 2 µL of reagent at 37 °C for 4 min, followed by 80 °C for 1 min. Cycle sequencing was performed in both directions using the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) in 10 µL reactions containing 2 µL BigDye Ready Reaction Mix, 1 µL 5× sequencing buffer, 0.32 µM primer and 2 µL purified product, and products were resolved on an Applied Biosystems 3500 Series Genetic Analyzer. Chromatograms were inspected manually, and sequences were aligned to the human reference genome (GRCh38) to confirm the exact junction breakpoints.

4.7. Fusion Transcript Reconstruction and Protein-Level Annotation

For each validated fusion, the fusion transcript was reconstructed from the Sanger-confirmed junction sequence and the Ensembl v107 reference transcripts reported by Arriba (*MAP4K5* ENST00000013125, *L2HGDH* ENST00000261699, *ADAMTS2* ENST00000251582, *ENSG00000253652* ENST00000520758, *BAZ2B* ENST00000392782, *WDSUB1* ENST00000359774, *HUWE1* ENST00000262854, *SUPT3H* ENST00000459689 and *PLEKHA5* ENST00000429027). Transcripts were translated in silico from the native start codon of the 5′ partner to the first in-frame stop codon, and the resulting sequences were checked against the junction peptides reported by Arriba. Retained and lost regions were defined by comparison with UniProt canonical sequences and annotations [55] (MAP4K5 Q9Y4K4, L2HGDH Q9H9P8, ADAMTS2 O95450, HUWE1 Q7Z6Z7, PLEKHA5 Q9HAU0 and WDSUB1 Q8N9V3). For the junction located in the 5′-untranslated regions (*BAZ2B–WDSUB1*), the *BAZ2B*-derived sequence was scanned for upstream ATG codons. For the *ADAMTS2–ENSG00000253652* junction, the exon structure of ENST00000520758 was inspected to determine whether the premature stop codon lay in the terminal exon, which determines susceptibility to nonsense-mediated decay.

4.8. Protein Structure Prediction

Structures of the MAP4K5–L2HGDH chimera (453 residues), wild-type MAP4K5 (residues 1–320, spanning the kinase domain) and mature wild-type L2HGDH (residues 52–463, lacking the mitochondrial targeting sequence) were predicted with AlphaFold 3 [56] via AlphaFold Server (accessed September 2026), using a single protein chain without ligands, seed 1599762399 and five models per job. Model confidence was assessed from the predicted template modeling score (pTM), per-residue pLDDT (taken from the Cα atom) and the predicted aligned error (PAE); values are reported for the top-ranked model (model 0) unless stated otherwise. Agreement between the five models was assessed by Cα root-mean-square deviation (RMSD). Each segment of the chimera was superimposed on the corresponding wild-type model using the Cα atoms of shared residues. Residue contacts were defined as heavy atoms within 4.5 Å, and solvent-accessible surface area was calculated with the Shrake–Rupley algorithm [57] (Biopython [58]) with and without the deleted segment. Structure figures were rendered with PyMOL (open-source, v3.2) [59]. AlphaFold Server outputs were used in accordance with the AlphaFold Server Output Terms of Use.

4.9. Screening of Normal Tissue and Population Structural Variant Databases

To assess whether the validated fusions also occur in non-malignant tissue, Arriba was run with the same alignment and filtering settings on total RNA-seq data from healthy esophageal squamous epithelium obtained from the atlas of RNA sequencing profiles for normal human tissues [24] (GEO accession GSE120795; paired-end reads downloaded from the NCBI SRA under BioProject PRJNA494560), as previously described [6]. All 11 esophageal samples used in that study were analyzed. To assess whether a germline structural variant could account for the recurrent *ADAMTS2–ENSG00000253652* junction, the gnomAD structural variant dataset (gnomAD-SV v4, GRCh38) [37,38] was searched for duplications spanning both breakpoints (chr5:179,343,767 and chr5:179,503,740).

4.10. Gene Expression Analyses

Raw gene-level counts from all 23 tumor samples and the 11 healthy esophageal tissue samples described in Section 4.9 (generated with the RNA-seq pipeline described in Sharip et al. [6]) were normalized using the variance-stabilizing transformation (VST) implemented in DESeq2 (v1.44.0) [60], applied across the full dataset to preserve comparability between groups. For each validated fusion, VST-normalized expression of each

partner gene was compared descriptively between normal tissue, fusion-negative tumors and fusion-positive tumors (Supplementary Figure S1). A tumor was classified as fusion-positive for a given pair only if the fusion junction had been confirmed by Sanger sequencing in that tumor (Table 2); all other tumors were classified as fusion-negative for that pair. Given the small number of fusion-positive samples per pair (n = 1–3), no formal statistical testing was performed. To evaluate a possible copy-number gain in tumor 8T, VST values were computed for all protein-coding genes located in chr12:15–30 Mb (Ensembl v107 annotation), and for each gene the difference between the 8T value and the mean of the other 22 tumors was plotted against genomic position (Figure 7). Contiguous groups of genes elevated in 8T were interpreted as candidate co-amplified segments. Analyses were performed in R (v4.4.3) [61] using DESeq2, rtracklayer [62] and ggplot2 [63].

4.11. Characterization of Candidate Fusion Calls and Exploratory Clinical Association Analysis

All Arriba and STAR-Fusion calls (Supplementary Tables S2 and S3) were analyzed in Python (v3.13) with pandas and SciPy (v1.18) [64]. Arriba calls sharing the same pair of genomic breakpoints within a sample (alternative gene annotations of a single event) were collapsed into one event. Event classes and confidence levels were taken from the Arriba "type" and "confidence" fields, and events labeled 5′–5′ or 3′–3′ were classified as having non-canonical orientation. For STAR-Fusion calls, the FusionAnnotator annotations generated with the CTAT genome library described in Section 4.4 [26] were parsed, and calls involving a gene annotated as an oncogene or included in cancer gene panels, and calls previously reported in cancer fusion resources (TCGA, CCLE, DepMap, ChimerSeq, Klijn cell lines or TumorFusions), were counted. Arriba gene pairs were compared, in both orientations, with the Mitelman Database-derived list of known fusions distributed with Arriba v2.4.0 [49]. Breakpoint ends were assigned to GRCh38 cytobands using the cytoband annotation distributed with Arriba v2.4.0, and for each cytoband the number of tumors with at least one breakpoint end was counted. The per-tumor numbers of Arriba events, high-confidence Arriba events and STAR-Fusion calls were compared between T categories (T1–T2 vs. T3–T4), N categories (N0 vs. N1–N2; NX excluded), sexes, ethnic groups (Kazakh vs. other), tumors with and without keratinization, and tumor locations (lower third involved vs. upper or middle third) using two-sided Mann–Whitney U tests, and were correlated with body-mass index using Spearman's rank correlation. *p*-values were adjusted across all 21 tests with the Benjamini–Hochberg procedure [65]. Given the sample size, these analyses were considered exploratory.

5. Conclusions

This first exploratory analysis of fusion genes in esophageal squamous cell carcinoma from Kazakhstan identified five experimentally validated chimeric transcripts, four of them computationally recurrent. Protein-level reconstruction and AlphaFold 3 modeling showed that none of them is predicted to produce a chimeric protein with an intact catalytic or signaling domain: the events are best interpreted as truncating, regulatory or passenger alterations, including fusion junctions arising within a putative *KRAS*-containing 12p amplicon. These findings extend the catalogue of structural variation in ESCC and show that validated fusion transcripts should be assessed at the protein level, and against matched normal and population-specific germline data, before they are considered as drivers or biomarkers. Given the small size of the cohort, these findings are hypothesis-generating and require confirmation in larger, independent and ancestry-matched cohorts.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/ijms272310755/s1.

Author Contributions: Conceptualization, U.K. (Ulykbek Kairov), A.A. (Ainur Akilzhanova) and S.R.; methodology, A.K., R.B., R.K. and U.K. (Ulykbek Kairov); software, A.K. and A.O.; validation, A.K., R.B., S.R., A.S., R.K., D.B., U.K. (Ulan Kozhamkulov) and A.A. (Ainur Ashenova); formal analysis, A.K., A.O. and U.K. (Ulykbek Kairov); investigation, R.B., A.S., N.S. and A.A. (Ainur Ashenova); resources, A.A. (Ainur Akilzhanova), M.T., D.S., S.R. and U.K. (Ulykbek Kairov); data curation, A.K., A.O. and A.D.; writing—original draft preparation, A.K. and U.K. (Ulykbek Kairov); writing—review and editing, U.K. (Ulykbek Kairov), S.R. and A.A. (Ainur Akilzhanova); visualization, A.K.; supervision, U.K. (Ulykbek Kairov) and S.R.; project administration, S.R. and U.K. (Ulykbek Kairov); funding acquisition, U.K. (Ulykbek Kairov), S.R., D.S., M.T. and A.A. (Ainur Akilzhanova). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Program targeted funding #BR24992841, #BR18574184, #BR24993023), Nazarbayev University funding CRP grants 211123CRP1608, 021220CRP2222, 11022021CRP1511, 20122022CRP1602.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Ethics Review Board of National Laboratory Astana (protocol #02-2022, 31 March 2022).

Informed Consent Statement: Written informed consent was obtained from all patients involved in the study.

Data Availability Statement: Tumor RNA-seq data analyzed in this study are available in the NCBI Sequence Read Archive under BioProject PRJNA608223. RNA-seq data of normal esophageal tissue are available in the Gene Expression Omnibus under accession GSE120795 (SRA BioProject PRJNA494560). Fusion calls from Arriba and STAR-Fusion are provided in Supplementary Tables S2 and S3, Sanger sequencing chromatograms of all fusion junctions in Supplementary File S1, and the AlphaFold 3 models in Supplementary File S2. Per-tumor fusion call counts and the results of the landscape and clinical association analyses are provided in Supplementary Table S5. Other data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: We thank all the participants in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Age-standardized rate
BMI	Body-mass index
CDS	Coding sequence
EC	Esophageal cancer
ESCC	Esophageal squamous cell carcinoma
FAD	Flavin adenine dinucleotide
GEO	Gene Expression Omnibus
H&E	Hematoxylin and eosin
IARC	International Agency for Research on Cancer
KEGG	Kyoto Encyclopedia of Genes and Genomes
L-2HG	L-2-hydroxyglutarate
MTS	Mitochondrial targeting sequence
NCBI	National Center for Biotechnology Information
NMD	Nonsense-mediated decay
PAE	Predicted aligned error

PCR	Polymerase chain reaction
pLDDT	Predicted local distance difference test
pTM	Predicted template modeling score
RMSD	Root-mean-square deviation
RNA-seq	RNA sequencing
RT-PCR	Reverse transcription polymerase chain reaction
SD	Standard deviation
SRA	Sequence Read Archive
SV	Structural variant
TCGA	The Cancer Genome Atlas
TNM	Tumor, Node, Metastasis classification system
UTR	Untranslated region
VST	Variance stabilizing transformation

References

1. Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R.L.; Soerjomataram, I.; Jemal, A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2024**, *74*, 229–263. https://doi.org/10.3322/caac.21834.
2. Liu, W.-J.; Zhao, Y.; Chen, X.; Miao, M.-L.; Zhang, R.-Q. Epigenetic modifications in esophageal cancer: An evolving biomarker. *Front. Genet.* **2023**, *13*, 1087479. https://doi.org/10.3389/fgene.2022.1087479.
3. Ferlay, J.; Ervik, M.; Lam, F.; Laversanne, M.; Colombet, M.; Mery, L.; Piñeros, M.; Znaor, A.; Soerjomataram, I.; Bray, F. *Global Cancer Observatory: Cancer Today (Version 1.1)*; International Agency for Research on Cancer: Lyon, France, 2024. Available online: https://gco.iarc.who.int/today (accessed on 5 June 2026).
4. Jemal, A.; Center, M.M.; DeSantis, C.; Ward, E.M. Global patterns of cancer incidence and mortality rates and trends. *Cancer Epidemiol. Biomark. Prev.* **2010**, *19*, 1893–1907. https://doi.org/10.1158/1055-9965.EPI-10-0437.
5. Kim, R.N.; Moon, H.-G.; Han, W.; Noh, D.-Y. Perspective insight into future potential fusion gene transcript biomarker candidates in breast cancer. *Int. J. Mol. Sci.* **2018**, *19*, 502. https://doi.org/10.3390/ijms19020502.
6. Sharip, A.; Rakhimova, S.; Molkenov, A.; Ashenova, A.; Kozhamkulov, U.; Akhmetollayev, I.; Zinovyev, A.; Zhukov, Y.; Omarov, M.; Tuleutaev, M.; et al. Transcriptome profiling and analysis of patients with esophageal squamous cell carcinoma from Kazakhstan. *Front. Genet.* **2024**, *15*, 1249751. https://doi.org/10.3389/fgene.2024.1249751.
7. Ordabayeva, A.; Sarkytbayev, K.; Daniyarov, A.; Molkenov, A.; Sharip, A.; Rakhimova, S.; Satvaldina, N.; Ashenova, A.; Karimov, A.; Bektayev, R.; et al. Integrative analysis of lncRNA–miRNA–mRNA expression profiles in esophageal squamous cell carcinoma tissues. *Sci. Rep.* **2026**, *16*, in press. https://doi.org/10.1038/s41598-026-71282-y.
8. Taniue, K.; Akimitsu, N. Fusion genes and RNAs in cancer development. *Non-Coding RNA* **2021**, *7*, 10. https://doi.org/10.3390/ncrna7010010.
9. Gkountakos, A.; Singhi, A.D.; Westphalen, C.B.; Scarpa, A.; Luchini, C. Fusion genes in pancreatic tumors. *Trends Cancer* **2024**, *10*, 430–443. https://doi.org/10.1016/j.trecan.2024.01.009.
10. Mohammad, T.; Zolotovskaia, M.A.; Suntsova, M.V.; Buzdin, A.A. Cancer fusion transcripts with human non-coding RNAs. *Front. Oncol.* **2024**, *14*, 1415801. https://doi.org/10.3389/fonc.2024.1415801.
11. Jamalinia, M.; Weiskirchen, R. Advances in personalized medicine: Translating genomic insights into targeted therapies for cancer treatment. *Ann. Transl. Med.* **2025**, *13*, 18. https://doi.org/10.21037/atm-25-34.
12. Mitelman, F.; Johansson, B.; Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* **2007**, *7*, 233–245. https://doi.org/10.1038/nrc2091.
13. Dai, X.; Theobard, R.; Cheng, H.; Xing, M.; Zhang, J. Fusion genes: A promising tool combating against cancer. *Biochim. Biophys. Acta Rev. Cancer* **2018**, *1869*, 149–160. https://doi.org/10.1016/j.bbcan.2017.12.003.
14. Sorokin, M.; Rabushko, E.; Rozenberg, J.M.; Mohammad, T.; Seryakov, A.; Sekacheva, M.; Buzdin, A. Clinically relevant fusion oncogenes: Detection and practical implications. *Ther. Adv. Med. Oncol.* **2022**, *14*, 17588359221144108. https://doi.org/10.1177/17588359221144108.
15. Gao, Q.; Liang, W.-W.; Foltz, S.M.; Mutharasu, G.; Jayasinghe, R.G.; Cao, S.; Liao, W.-W.; Reynolds, S.M.; Wyczalkowski, M.A.; Yao, L.; et al. Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep.* **2018**, *23*, 227–238.e3. https://doi.org/10.1016/j.celrep.2018.03.050.

16. Gao, J.; Wang, Q.; Fu, F.; Zhao, Y.; Yang, T.; Li, X.; Sun, Y.; Hu, H.; Ma, L.; Miao, L.; et al. Comprehensive genomic and transcriptomic analyses reveal prognostic stratification for esophageal squamous cell carcinoma. *Signal Transduct. Target. Ther.* **2025**, *10*, 223. https://doi.org/10.1038/s41392-025-02306-8.
17. Mizukami, T.; Sakai, K.; Naruki, S.; Taniyama, T.; Horie, Y.; Izawa, N.; Tsuda, T.; Fujino, T.; Boku, N.; Yasuda, H.; et al. Abstract 3818: Identification of a FGFR3-TACC3 fusion in esophageal cancer. *Cancer Res.* **2017**, *77*, 3818. https://doi.org/10.1158/1538-7445.AM2017-3818.
18. Yang, T.; Lin, Z.-R.; Xia, T.-L.; Liu, S.-X.; Yuan, B.-Y.; Luo, Y.-L.; Du, W.-T.; Lei, C.-B.; Nie, Y.-Z.; Zeng, M.-S.; et al. The PLEKHA1-TACC2 fusion gene drives tumorigenesis via vascular mimicry formation in esophageal squamous-cell carcinoma. *Cell Death Differ.* **2025**, *32*, 2323–2339. https://doi.org/10.1038/s41418-025-01536-1.
19. Souza-Santos, P.T.; Soares Lima, S.C.; Nicolau-Neto, P.; Boroni, M.; Meireles Da Costa, N.; Brewer, L.; Menezes, A.N.; Furtado, C.; Moreira, M.A.M.; Seuanez, H.N.; et al. Mutations, differential gene expression, and chimeric transcripts in esophageal squamous cell carcinoma show high heterogeneity. *Transl. Oncol.* **2018**, *11*, 1283–1291. https://doi.org/10.1016/j.tranon.2018.08.002.
20. Chen, X.-X.; Zhong, Q.; Liu, Y.; Yan, S.-M.; Chen, Z.-H.; Jin, S.-Z.; Xia, T.-L.; Li, R.-Y.; Zhou, A.-J.; Su, Z.; et al. Genomic comparison of esophageal squamous cell carcinoma and its precursor lesions by multi-region whole-exome sequencing. *Nat. Commun.* **2017**, *8*, 524. https://doi.org/10.1038/s41467-017-00650-0.
21. An, H.; Cheng, X.; Xue, L.; Wang, G.; Zhu, X.; Li, J.; Xiao, T.; Cheng, S. Genomic instability is widespread in esophageal squamous dysplasia and increases during the progression to cancer. *Front. Biosci. (Landmark Ed.)* **2025**, *30*, 41107. https://doi.org/10.31083/FBL41107.
22. Chang, J.; Zhao, X.; Wang, Y.; Liu, T.; Zhong, C.; Lao, Y.; Zhang, S.; Liao, H.; Bai, F.; Lin, D.; et al. Genomic alterations driving precancerous to cancerous lesions in esophageal cancer development. *Cancer Cell* **2023**, *41*, 2038–2050.e5. https://doi.org/10.1016/j.ccell.2023.11.003.
23. Yu, X.; Xu, L.; Zhang, S.; Pan, P.; Xia, Y. Variations in characteristics and clinical outcomes of esophageal squamous cell carcinoma among Asian American. *Oncology* **2026**, *104*, 51–62. https://doi.org/10.1159/000544841.
24. Suntsova, M.; Gaifullin, N.; Allina, D.; Reshetun, A.; Li, X.; Mendeleeva, L.; Surin, V.; Sergeeva, A.; Spirin, P.; Prassolov, V.; et al. Atlas of RNA sequencing profiles for normal human tissues. *Sci. Data* **2019**, *6*, 36. https://doi.org/10.1038/s41597-019-0043-4.
25. Steijger, T.; Abril, J.F.; Engström, P.G.; Kokocinski, F.; RGASP Consortium; Hubbard, T.J.; Guigó, R.; Harrow, J.; Bertone, P. Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **2013**, *10*, 1177–1184. https://doi.org/10.1038/nmeth.2714.
26. Haas, B.J.; Dobin, A.; Li, B.; Stransky, N.; Pochet, N.; Regev, A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* **2019**, *20*, 213. https://doi.org/10.1186/s13059-019-1842-9.
27. Gaonkar, K.S.; Marini, F.; Rathi, K.S.; Jain, P.; Zhu, Y.; Chimicles, N.A.; Brown, M.A.; Naqvi, A.S.; Zhang, B.; Storm, P.B.; et al. annoFuse: An R package to annotate, prioritize, and interactively explore putative oncogenic RNA fusions. *BMC Bioinform.* **2020**, *21*, 577. https://doi.org/10.1186/s12859-020-03922-7.
28. Meng, Z.; Moroishi, T.; Mottier-Pavie, V.; Plouffe, S.W.; Hansen, C.G.; Hong, A.W.; Park, H.W.; Mo, J.-S.; Lu, W.; Lu, S.; et al. MAP4K family kinases act in parallel to MST1/2 to activate LATS1/2 in the Hippo pathway. *Nat. Commun.* **2015**, *6*, 8357. https://doi.org/10.1038/ncomms9357.
29. Xu, W.; Yang, H.; Liu, Y.; Yang, Y.; Wang, P.; Kim, S.-H.; Ito, S.; Yang, C.; Wang, P.; Xiao, M.-T.; et al. Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases. *Cancer Cell* **2011**, *19*, 17–30. https://doi.org/10.1016/j.ccr.2010.12.014.
30. Intlekofer, A.M.; Wang, B.; Liu, H.; Shah, H.; Carmona-Fontaine, C.; Rustenburg, A.S.; Salah, S.; Gunner, M.R.; Chodera, J.D.; Cross, J.R.; et al. L-2-Hydroxyglutarate production arises from noncanonical enzyme function at acidic pH. *Nat. Chem. Biol.* **2017**, *13*, 494–500. https://doi.org/10.1038/nchembio.2307.
31. Shim, E.-H.; Livi, C.B.; Rakheja, D.; Tan, J.; Benson, D.; Parekh, V.; Kho, E.-Y.; Ghosh, A.P.; Kirkman, R.; Velu, S.; et al. L-2-Hydroxyglutarate: An epigenetic modifier and putative oncometabolite in renal cancer. *Cancer Discov.* **2014**, *4*, 1290–1298. https://doi.org/10.1158/2159-8290.CD-13-0696.
32. Babiceanu, M.; Qin, F.; Xie, Z.; Jia, Y.; Lopez, K.; Janus, N.; Facemire, L.; Kumar, S.; Pang, Y.; Qi, Y.; et al. Recurrent chimeric fusion RNAs in non-cancer tissues and cells. *Nucleic Acids Res.* **2016**, *44*, 2859–2872. https://doi.org/10.1093/nar/gkw032.
33. Colige, A.; Ruggiero, F.; Vandenberghe, I.; Dubail, J.; Kesteloot, F.; Van Beeumen, J.; Beschin, A.; Brys, L.; Lapière, C.M.; Nusgens, B. Domains and maturation processes that regulate the activity of ADAMTS-2, a metalloproteinase cleaving the

aminopropeptide of fibrillar procollagens types I–III and V. *J. Biol. Chem.* **2005**, *280*, 34397–34408. https://doi.org/10.1074/jbc.M506458200.

34. Bekhouche, M.; Leduc, C.; Dupont, L.; Janssen, L.; Delolme, F.; Vadon-Le Goff, S.; Smargiasso, N.; Baiwir, D.; Mazzucchelli, G.; Zanella-Cleon, I.; et al. Determination of the substrate repertoire of ADAMTS2, 3, and 14 significantly broadens their functions and identifies extracellular matrix organization and TGF-β signaling as primary targets. *FASEB J.* **2016**, *30*, 1741–1756. https://doi.org/10.1096/fj.15-279869.
35. Joannes, L.; Dupont, L.; Stock, L.; Arpigny, E.; Hubert, P.; Ancion, M.; Luyckx, M.; Abinet, J.; Peng, W.; Calaldo, D.; et al. The ADAMTS2 metalloproteinase inhibits tumor growth by regulating the innate immune system. *Cancer Cell Int.* **2025**, *25*, 229. https://doi.org/10.1186/s12935-025-03880-1.
36. Tian, C.; Huang, Y.; Clauser, K.R.; Rickelt, S.; Lau, A.N.; Carr, S.A.; Vander Heiden, M.G.; Hynes, R.O. Suppression of pancreatic ductal adenocarcinoma growth and metastasis by fibrillar collagens produced selectively by tumor cells. *Nat. Commun.* **2021**, *12*, 2328. https://doi.org/10.1038/s41467-021-22490-9.
37. Collins, R.L.; Brand, H.; Karczewski, K.J.; Zhao, X.; Alföldi, J.; Francioli, L.C.; Khera, A.V.; Lowther, C.; Gauthier, L.D.; Wang, H.; et al. A structural variation reference for medical and population genetics. *Nature* **2020**, *581*, 444–451. https://doi.org/10.1038/s41586-020-2287-8.
38. Zhao, X.; Collins, R.; Walker, M.; Darnowsky, P.; Chao, K.; Brand, H.; Talkowski, M. Structural Variants in gnomAD v4. gnomAD Blog, 1 November 2023. Available online: https://gnomad.broadinstitute.org/news/2023-11-v4-structural-variants (accessed on 27 September 2026).
39. Giles, A.C.; Grill, B. Roles of the HUWE1 ubiquitin ligase in nervous system development, function and disease. *Neural Dev.* **2020**, *15*, 6. https://doi.org/10.1186/s13064-020-00143-9.
40. Fischer, V.; Hisler, V.; Scheer, E.; Lata, E.; Morlet, B.; Plassard, D.; Helmlinger, D.; Devys, D.; Tora, L.; Vincent, S.D. SUPT3H-less SAGA coactivator can assemble and function without significantly perturbing RNA polymerase II transcription in mammalian cells. *Nucleic Acids Res.* **2022**, *50*, 7972–7990. https://doi.org/10.1093/nar/gkac637.
41. The Cancer Genome Atlas Research Network. Integrated genomic characterization of oesophageal carcinoma. *Nature* **2017**, *541*, 169–175. https://doi.org/10.1038/nature20805.
42. Jang, Y.E.; Jang, I.; Kim, S.; Cho, S.; Kim, D.; Kim, K.; Kim, J.; Hwang, J.; Kim, S.; Kim, J.; et al. ChimerDB 4.0: An updated and expanded database of fusion genes. *Nucleic Acids Res.* **2020**, *48*, D817–D824. https://doi.org/10.1093/nar/gkz1013.
43. Kim, P.; Tan, H.; Liu, J.; Lee, H.; Jung, H.; Kumar, H.; Zhou, X. FusionGDB 2.0: Fusion gene annotation updates aided by deep learning. *Nucleic Acids Res.* **2022**, *50*, D1221–D1230. https://doi.org/10.1093/nar/gkab1056.
44. Sondka, Z.; Dhir, N.B.; Carvalho-Silva, D.; Jupe, S.; Madhumita; McLaren, K.; Starkey, M.; Ward, S.; Wilding, J.; Ahmed, M.; et al. COSMIC: A curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res.* **2024**, *52*, D1210–D1217. https://doi.org/10.1093/nar/gkad986.
45. Ghandi, M.; Huang, F.W.; Jané-Valbuena, J.; Kryukov, G.V.; Lo, C.C.; McDonald, E.R., III; Barretina, J.; Gelfand, E.T.; Bielski, C.M.; Li, H.; et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **2019**, *569*, 503–508. https://doi.org/10.1038/s41586-019-1186-3.
46. Edgren, H.; Murumagi, A.; Kangaspeska, S.; Nicorici, D.; Hongisto, V.; Kleivi, K.; Rye, I.H.; Nyberg, S.; Wolf, M.; Borresen-Dale, A.-L.; et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol.* **2011**, *12*, R6. https://doi.org/10.1186/gb-2011-12-1-r6.
47. Inaki, K.; Hillmer, A.M.; Ukil, L.; Yao, F.; Woo, X.Y.; Vardy, L.A.; Zawack, K.F.B.; Lee, C.W.H.; Ariyaratne, P.N.; Chan, Y.S.; et al. Transcriptional consequences of genomic structural aberrations in breast cancer. *Genome Res.* **2011**, *21*, 676–687. https://doi.org/10.1101/gr.113225.110.
48. Sobin, L.H.; Gospodarowicz, M.K.; Wittekind, C. (Eds.) *TNM Classification of Malignant Tumours*, 7th ed.; Wiley-Blackwell: Chichester, UK, 2009.
49. Uhrig, S.; Ellermann, J.; Walther, T.; Burkhardt, P.; Fröhlich, M.; Hutter, B.; Toprak, U.H.; Neumann, O.; Stenzinger, A.; Scholl, C.; et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* **2021**, *31*, 448–460. https://doi.org/10.1101/gr.257246.119.
50. Cunningham, F.; Allen, J.E.; Allen, J.; Alvarez-Jarreta, J.; Amode, M.R.; Armean, I.M.; Austine-Orimoloye, O.; Azov, A.G.; Barnes, I.; Bennett, R.; et al. Ensembl 2022. *Nucleic Acids Res.* **2022**, *50*, D988–D995. https://doi.org/10.1093/nar/gkab1049.
51. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

52. Frankish, A.; Carbonell-Sala, S.; Diekhans, M.; Jungreis, I.; Loveland, J.E.; Mudge, J.M.; Sisu, C.; Wright, J.C.; Arnan, C.; Barnes, I.; et al. GENCODE: Reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **2023**, *51*, D942–D949. https://doi.org/10.1093/nar/gkac1071.
53. Kalendar, R. Comprehensive web-based platform for advanced PCR design, genotyping, synthetic biology, molecular diagnostics, and sequence analysis. *Mol. Ther. Nucleic Acids* **2025**, *36*, 102716. https://doi.org/10.1016/j.omtn.2025.102716.
54. Kalendar, R.; Shevtsov, A.; Otarbay, Z.; Ismailova, A. In silico PCR analysis: A comprehensive bioinformatics tool for enhancing nucleic acid amplification assays. *Front. Bioinform.* **2024**, *4*, 1464197. https://doi.org/10.3389/fbinf.2024.1464197.
55. The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **2025**, *53*, D609–D617. https://doi.org/10.1093/nar/gkae1010.
56. Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A.J.; Bambrick, J.; et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493–500. https://doi.org/10.1038/s41586-024-07487-w.
57. Shrake, A.; Rupley, J.A. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **1973**, *79*, 351–371. https://doi.org/10.1016/0022-2836(73)90011-9.
58. Cock, P.J.A.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.
59. Schrödinger, LLC. *The PyMOL Molecular Graphics System, Open-Source Version*; Schrödinger, LLC.: New York, NY, USA.
60. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.
61. R Core Team. *R: A Language and Environment for Statistical Computing*, *Version 4.4.3*; R Foundation for Statistical Computing: Vienna, Austria, 2025.
62. Lawrence, M.; Gentleman, R.; Carey, V. rtracklayer: An R package for interfacing with genome browsers. *Bioinformatics* **2009**, *25*, 1841–1842. https://doi.org/10.1093/bioinformatics/btp328.
63. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*, 2nd ed.; Springer: New York, NY, USA, 2016. https://doi.org/10.1007/978-3-319-24277-4.
64. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.
65. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.