NAR Cancer, 2024, **6**, zcae041
https://doi.org/10.1093/narcan/zcae041
Advance access publication date: 16 October 2024
Cancer Gene Regulation, Chromatin, and Epigenetics

OXFORD

Massively parallel reporter assays identify enhancer elements in oesophageal Adenocarcinoma

Shen-Hsi Yang†, Ibrahim Ahmed†, Yaoyong Li†, Christopher W. Bleaney and Andrew D. Sharrocks *

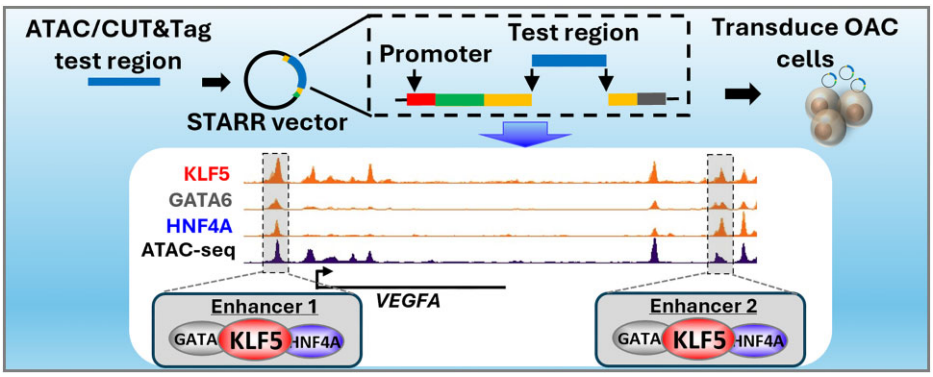
School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, UK

*To whom correspondence should be addressed. Tel: +44 161 275 5979; Fax: +44 161 275 5082; Email: andrew.d.sharrocks@manchester.ac.uk
†The first three authors are considered to be Joint First Authors.

Abstract

Cancer is a disease underpinned by aberrant gene expression. Enhancers are regulatory elements that play a major role in transcriptional control and changes in active enhancer function are likely critical in the pathogenesis of oesophageal adenocarcinoma (OAC). Here, we utilise STARR-seq to profile the genome-wide enhancer landscape in OAC and identify hundreds of high-confidence enhancer elements. These regions are enriched in enhancer-associated chromatin marks, are actively transcribed and exhibit high levels of associated gene activity in OAC cells. These characteristics are maintained in human patient samples, demonstrating their disease relevance. This relevance is further underlined by their responsiveness to oncogenic ERBB2 inhibition and increased activity compared to the pre-cancerous Barrett's state. Mechanistically, these enhancers are linked to the core OAC transcriptional network and in particular KLF5 binding is associated with high level activity, providing further support for a role of this transcription factor in defining the OAC transcriptome. Our results therefore uncover a set of enhancer elements with physiological significance, that widen our understanding of the molecular alterations in OAC and point to mechanisms through which response to targeted therapy may occur.

Graphical abstract



Introduction

Oesophageal adenocarcinoma (OAC) is a gastrointestinal cancer that is one of the leading global causes of cancer-associated deaths (1). A number a genome-wide DNA-sequencing studies have identified potential disease driver events, including amplifications of genes encoding receptor tyrosine kinases (RTKs) such as ERBB2 (2–4). Despite this, there is still a paucity of information regarding the specific mechanisms by which disease occurs. We have previously demonstrated that changes to the accessible chromatin landscape are critical in both the onset of disease, and the response to targeted therapeutic intervention (5–8). These chromatin changes often occur in regions of the genome that have the potential to function as regulatory elements, such as enhancers, and may represent an underappreciated mechanism by which disease and therapy-resistance is facilitated. This makes the study of enhancer elements in OAC of potential clinical importance.

Enhancers are distal regulatory elements that affect the expression of their target genes through long-range interactions with their promoters (9). The identification of enhancers has largely focused on the use of correlative approaches, linking the presence of histone marks such as H3K27ac and

Received: June 12, 2024. **Revised:** August 9, 2024. **Editorial Decision:** September 24, 2024. **Accepted:** October 9, 2024

H3K4me1, as well as chromatin accessibility, to enhancer potential (10,11). Whilst these approaches have revolutionised the study of enhancer elements, they remain hampered by the inability to provide a robust readout of genuine enhancer activity. Indeed, there is a body of evidence to suggest that correlative enhancer marks have no bearing on enhancer activity itself (12–14). Low-throughput enhancer reporter assays, which link enhancer function to a reporter gene such as green fluorescent protein (GFP) or luciferase, remain the most commonly used method by which enhancer activity is assessed (15). However, these approaches generally interrogate a single enhancer at time, limiting their use in genome-wide enhancer identification. To this end, genome-wide screening such as STARR-seq is an important advance in the understanding of enhancer biology (16)]. STARR-seq is a form of massively parallel reporter assay (MPRA), where regions of DNA are inserted downstream of a promoter, and into the 3′-UTR of a reporter gene. Provided that the region of DNA functions as an enhancer, it will drive expression of the reporter transcript, and itself, which may be detected using RNA-seq (16). By generating genome-wide libraries of potential regulatory DNA regions, and inserting these into the STARR reporter, it is possible to interrogate global enhancer activity.

Our previous studies identified regions of the genome that change accessibility during the transition to OAC from the precursor condition Barrett's oesophagus (5–7), but these provide only correlative evidence for enhancer activity based on one modality (chromatin accessibility). Here, we employed STARR-seq to expand on these studies and provide direct evidence of enhancer activity in OAC. By using genome-wide DNA libraries generated using a combination of ATAC-seq and CUT&Tag, we identified chromatin regions in OAC that function as active enhancer elements. Integration with other genome-wide datasets from both OAC cells and patient samples validated their designation as *bona fide* regulatory elements and their association with regulatory events in OAC patients. These findings therefore expand our current understanding of the gene regulatory mechanisms underlying OAC.

Materials and methods

Cell culture and treatments

OE19 cells were cultured in RPMI 1640 (ThermoFisher Scientific, 52400), HEK293T cells were cultured in DMEM (ThermoFisher Scientific, 11965084). All media was supplemented with 10% foetal bovine serum (ThermoFisher Scientific, 10270). Cell lines were cultured at 37°C, 5% CO_2 in a humidified incubator. Lapatinib (Selleckchem, S1028) treatments were performed for 24 h and at a 500 nM final concentration.

Luciferase reporter assays

Regions containing *VEGFAe*, *ELF3e*, *GATA6e* or *PAX8e* were amplified from OE19 genomic DNA using primers containing 20 bp overlap regions with the multiple cloning site of the pGL3 Promoter vector (Promega, E1761) for luciferase assays (Supplementary Table S4). Final vectors were assembled using HiFi assembly (NEB, E5520S) according to the manufacturer's instructions to create pGL3 plasmids containing *VEGFAe*, *ELF3e*, *GATA6e* or *PAX8e* enhancer regions (pAS5014-pAS5017). Enhancer vectors were transfected into OE19 cells using the Amaxa™ Nucleofector™ II (Lonza) with Cell Line NucleofectorTM Kit V (Lonza, VCA-1003) using program T-020, according to manufacturer's instructions. To conduct luciferase assays, 250 ng of enhancer vector was co-transfected alongside 50 ng of pCH110 (Amersham). Enhancer activity was assessed using the Dual-Light™ Luciferase & β-Galactosidase Reporter System (ThermoFisher Scientific, T1003) according to the manufacturer's instructions.

STARR-seq vector production

An integrating STARR-seq vector was designed based on the pLenti-FKBP-delCasp9-Puro vector (17) and the hSTARR-seq_ORI vector (Addgene, 99296) (18). Briefly, the lentiviral machinery was amplified from the pLenti-FKBP-delCasp9-Puro vector to create two PCR products (primer pairs ADS6967/ADS6968 and ADS6969/6970). The STARR reporter machinery was amplified from the hSTARR-seq_ORI vector to create two PCR products using primer pairs ADS6963/ADS6964 and ADS6965/6976. The four PCR products were then assembled in a 1:1:1:1 ratio using HiFi assembly (NEB, E5520S) according to the manufacturer's instructions to create the pLenti-STARR vector (pAS5018). PCR primers are listed in Supplementary Table S4.

STARR-seq plasmid library generation

To maintain library complexity, 8 Omni-ATAC or CUT&Tag tagmentation reactions of 50000 OE19 cells each were conducted per library, as previously described (19,20) except Nextera sample barcodes were only introduced using the forward primer (Supplementary Table S4) and with an altered nuclear extraction step for CUT&Tag libraries. For the CUT&Tag nuclear extraction, OE19 cells were initially lysed in Nuclei EZ lysis buffer (Sigma-Aldrich, NUC-101) at 4°C for 10 min followed by centrifugation at 500 g for 5 mins. The subsequent clean-up was performed in a buffer composed of 10 mM Tris-HCl pH 8.0, 10 mM NaCl and 0.2% NP40 followed by centrifugation at 1300 g for 5 min. Nuclei were then lightly cross-linked in 0.1% formaldehyde for 2 mins followed by quenching with 75 mM glycine followed by centrifugation at 500 g for 5 mins. Cross-linked nuclei were resuspended in 20 mM HEPES pH 7.5, 150 mM NaCl and 0.5 M spermidine at a concentration of 4–8 × 10^3/μl (2–4 × 10^4 total). For 2–4 × 10^4 nuclei, 0.5 μg of primary and secondary antibody were used with 1 μl of pA-Tn5 (Epicypher, 15–1017). Antibodies used for CUT&Tag: anti-BRD4 (Abcam, ab128874), anti-H3K27ac (Abcam, ab4729) and anti-MED1 (Antibody-Online, A98044/10UG). Subsequent CUT&Tag stages were as previously described (20).

Omni-ATAC or CUT&Tag libraries assembled into the pLenti-STARR vector using HiFi assembly (NEB, E5520S), as previously described (16) and according to manufacturer's instructions. 25 assembly reactions per library were conducted to maintain library complexity. Assembly reactions were pooled and cleaned using Ampure XP beads (Beckman Coulter Agencourt, A63881) at a 1.8× ratio of beads to input. Assembled libraries were eluted into 10 μl nuclease-free H_2O. MegaX DH10B T1R electrocompetent bacteria (20 μl; ThermoFisher Scientific, C640003) were transformed with 150 ng of library, with a total of 8 transformations performed per library to maintain complexity. Bacteria were transformed using the Ec1 setting on the BioRad MicroPulser (BioRad, 165-2100) in 1 cm cuvettes, ensuring time constants between 4.5–5.0 ms. Bacterial cells were recovered in 750 μl of pre-warmed

SOC medium and incubated for 1 h in a 250 RPM shaker at 37°C. Post-recovery, bacterial cells were pooled and incubated overnight in 4 l of Luria broth containing 100 μg/ml carbenicillin (Sigma, C1613). Plasmid libraries were collected using the Plasmid Giga Kit (Qiagen, 12191). An aliquot of the plasmid library was sequenced on an Illumina HiSeq 4000 System to assess complexity (University of Manchester Genomic Technologies Core Facility).

STARR-seq library transduction and screening

Lentiviral particles were generated as previously described (21). Briefly, 2×10^6 HEK293 cells were transfected with 15 μg pLenti-STARR library, 10 μg psPAX2 (Addgene, 12260) and 5 μg pMD2.G (Addgene, 12259) using PolyFect (Qiagen, 301107). Media was collected at 48 and 72 h post-transfection and viral particles were precipitated using PEG-it™ Solution (System Biosciences, LV810A-1). To transduce, 1×10^7 OE19 cells were treated with virus (MOI 0.7) and 5 μg/ml Polybrene (EMD Millipore, TR-1003). Polyclonal cells were selected for 2 weeks in 500 ng/μl puromycin (Sigma, P7255). After selection and growth to 5×10^7 cells, 50% of cells were processed for gDNA isolation using the DNEasy Blood and Tissue Kit (Qiagen, 69504) and 50% processed for RNA isolation using the RNEasy Midi Kit with optional on-column DNAse digestion (Qiagen, 75144), as per manufacturer's instructions. Sequencing-ready gDNA and RNA libraries were generated as previously described (22). Briefly, polyA + mRNA was isolated using the Oligotex mRNA Midi kit (Qiagen, 70042), as per manufacturer's instructions. RNA was reverse-transcribed using Superscript III Reverse Transcriptase (ThermoFisher Scientific, 18080085) with 4 reactions per library (ensuring <4 μg per reaction) using a specific primer and according to manufacturer's instructions (Supplementary Table S4). cDNA was pooled and cleaned up using Ampure XP beads (Beckman Coulter Agencourt, A63881) at a 1.8X ratio of beads to input, eluting in 80 μl nuclease-free H_2O. cDNA and gDNA libraries were amplified by PCR using the NEBNext 2X Master Mix with the remaining sample barcode introduced using the reverse primer (Supplementary Table S4). 4 reactions were performed per library. Libraries were pooled and sequenced on an Illumina HiSeq 4000 System (University of Manchester Genomic Technologies Core Facility).

STARR-seq data analysis

Initial STARR-seq data processing was performed similarly to ATAC-seq data processing previously described (5). Briefly, reads were mapped to GRCh38 (hg38) using Bowtie2 v2.3.0 (23) with the options: -X 2000 -dovetail. Mapped reads (>q30) were retained using SAMtools (24). Reads mapping to blacklisted regions were removed using BEDtools (25). Peaks were called using MACS2 v2.1.1 (26) with the following parameters: -q 0.01, -nomodel -shift -75 -extsize 150 -B –SPMR. A union peakset was formed from all plasmid library samples, using HOMER v4.9 mergePeaks.pl -d 250 (27) as described previously (6). STARR signal was ranked using featureCounts (28) by taking the sum of RNA libraries at regions above the DESeq2-defined count threshold and calculating signal over plasmid (ATAC-STARR-seq) or gDNA (CUT&Tag-STARR-seq). The changepoint package in R v3.6.0 was used to determine RNA and plasmid ranks, as well as the top-ranked regions for subsequent analysis. Differentially active STARR regions were subsequently determined from top-ranked regions using DESeq2 (29). A $\log_2$-fold change of ± 0.3 and $P\text{-value}_{adj} < 0.1$ defined differential expression.

HOMER v4.9 was used for de novo transcription factor motif enrichment analysis on the identified STARR + regions. The default settings of HOMER were used, where the background regions were randomly selected that match the GC-content distribution of the STARR + regions. STARR + regions were annotated to genes by the nearest gene model and genomic distribution profiled using HOMER v4.9 annotatePeaks.pl. A custom peakset of high-confidence STARR + intersect regions was generated as described previously (5). Differentially accessible/H3K27ac-marked high-confidence STARR + intersect regions upon lapatinib treatment were determined using DESeq2 (29). A $\log_2$-fold change of ±0.2, and $P\text{-value}_{adj} < 0.075$ and 0.05 defined differential accessible and H3K27ac-marked regions, respectively.

CUT&tag processing and data analysis

OE19 cells were treated with 500 nM lapatinib or DMSO. After 24 h, CUT&Tag library generation was performed as described above using an anti-H3K27ac antibody (Abcam, ab4729). CUT&Tag libraries were pooled and sequenced on an Illumina HiSeq 4000 System (University of Manchester Genomic Technologies Core Facility). CUT&Tag data processing was performed as for ATAC-seq analysis described above. A union peakset was generated using HOMER v4.9 mergePeaks.pl -d 250 (27) as described previously (5) and biological replicates were assessed for concordance ($r > 0.80$).

KAS-seq processing and data analysis

24 h-lapatinib or DMSO treated OE19 cells/cells from OAC tissue were prepared for KAS-seq. OE19 KAS-seq library generation was performed as described previously for bulk low input KAS-seq (28) except for altered nuclear extraction and labelling reactions, and using home-made Tn5 transposase as described previously (30). Nuclei were extracted and washed as described for CUT&Tag. Nuclei were then resuspended in nuclease-free H_2O at a concentration of 1×10^4/μl (2×10^5 total) Labelling reactions were carried out in DNA LoBind® tubes (Eppendorf, 0030108051) using 5 mM N3-kethoxal (a gift from Chuan He) in PBS to a final volume of 50 μl for 15 min at 37°C with 1000 RPM mixing in a thermomixer. Labelled gDNA was isolated using the PureLink™ Genomic DNA Mini kit (ThermoFisher Scientific, K182001) and eluted twice with 21.5 μl 25 mM K_3BO_3 pH 7.0. Purified DNA underwent tagmentation reaction with 2 μM of Tn5 in 1× TD buffer (33 mM Tris–OAc pH 7.8, 66 mM KOAc, 10 mM $Mg(OAc)_2$ and 16% dimethylformamide) in a 25 μl reaction volume at 37°C for 30 min. Subsequent enrichment was performed with 5 μl of Dynabeads Streptavidin C1 (ThermoFisher Scientific, 65001) and resuspended in 19.5 μl H_2O.

Library amplification was performed by PCR with 20 μl beads, 0.5 μM i5 and i7 Illumina index primers (Illumina, 20027213) and NEBNext Ultra II Q5 Master Mix (NEB, M0544S) in a 50 μl final reaction volume. The PCR reaction was carried out at 72°C, 5 min; 95°C 10 min, 15 cycles at 98°C 10 s, 63°C 30 s, 72°C 60 s and a final 72°C step for 2 min. The final libraries were cleaned up using Zymo DNA Clean & Concentrator kit 5 (Zymo Research, D4014). KAS-seq libraries were pooled and sequenced on an Illumina HiSeq 4000 System (University of Manchester Genomic Technolo-

gies Core Facility). Three biological replicates were sequenced and checked for concordance ($r > 0.80$). KAS-seq data processing was performed as described previously (28), but with the MACS2 v2.1.1 –broad peak calling option.

For OAC and Barrett's tissue KAS-seq, fresh frozen OAC 3 mm biopsies were collected at Cambridge University Hospitals NHS Trust (Addenbrooke's Hospital) from patients undergoing endoscopy. The study was approved by the Institutional Ethics Committees and all patients gave individual informed consent. Tissue nuclei were extracted as previously described (5) and treated as described for OE19 cells.

KAS-seq data processing was performed as described previously (28). Briefly, reads were mapped to the human genome GRCh38 (hg38) using Bowtie2 v2.3.0 (23). Mapped reads (>q30) were retained using SAMtools (24). Reads mapping to blacklisted regions were removed using BEDtools (25). Peaks were called using MACS2 v2.1.1 (26) with the following parameters: -q 0.01, -nomodel -shift -75 -extsize 150 -B –SPMR –broad. A union peakset was generated as described previously (5) and biological replicates were assessed for concordance ($r > 0.80$).

Bioinformatics

Genome browser data were visualised using the UCSC Genome Browser (31) Heatmaps and tag density plots of epigenomic data were generated the using deepTools (32) computeMatrix, plotProfile, plotCorrelation and plotHeatmap functions. Metascape (33) was used for gene and disease ontology analysis of closest protein-coding genes of the STARR + regions. Metascape utilizes the hypergeometric test and Benjamini-Hochberg p-value correction algorithm to identify all ontology terms that contain a statistically greater number of genes in common with an input list than expected by chance.

The eulerr package in R v3.6.0 was used for generating Venn diagrams. The Fisher exact test was used to calculate the p-values for the genomic region overlaps shown in the Venn diagrams in Figure 3B and Supplementary Figure S3C, and the Monte Carlo method was used for obtaining a *P*-value in Supplementary Figure S4A. In order to use the Fisher exact test, we created a genomic background (namely a set of genomic regions) as a possible sample space for the genomic regions involved in the overlapping. For Figure 3B, the genomic background was the whole genome. In detail, two peaks from the two sets were regarded as overlapping if the distance between their mid-points is <250 bp. From such overlapping definition we obtained the number of peaks in one set overlapping with peaks in another set, and the number of peaks in one set which does not overlap with any peak in another set, giving us three numbers of peaks. To use the Fisher exact test to calculate *P* value, a fourth number was calculated by dividing the whole human genome into contiguous and non-overlapping 1 kb regions. After excluding those 1 kb regions which overlap with the hg38 blacklisted regions downloaded from UCSC genome browser web site, we counted the number of 1 kb regions which do not overlap with any peak in the two sets, which was used as the fourth number for calculating the p-value in Figure 3B. For the overlapping shown in Supplementary Figure S3C, the genomic background was the non-redundant ATAC-seq peaks in OE19 cells. In detail, two peaks were regarded as overlapping with each other if their mid points were in the same ATAC-seq peak, and the number of ATAC-seq peaks which did not contain any peaks in the two overlapping sets was regarded as the fourth number for Fisher exact test. The calculation of the p-values shown in Supplementary Figure S4A also used the non-redundant ATAC-seq peaks in OE19 cells. In detail, we randomly selected 1549 peaks from the set of ATAC-seq peaks, and counted the number of these 1549 ATAC-seq peaks overlapping with the peaks in one ChIP-seq peak set (either GATA6, HNF4A or KLF5). We repeated the random selection 10 000 times, and ordered the number of overlaps obtained in the 10 000 experiments, and finally compared the ordered numbers with the overlap numbers shown in Venn diagrams to obtain an estimation of *P*-value for the overlap.

Datasets

All data were obtained from ArrayExpress, unless stated otherwise and are listed in Supplementary Table S5. OAC tissue total RNA-seq data were obtained from: OCCAMS consortium (European Genome-Phenome Archive, EGAD00001007496) (34). Lapatinib-treated OE19 ATAC-seq, lapatinib-treated OE19 RNA-seq and OE19 H3K27ac ChIP-seq was obtained from: E-MTAB-10334, E-MTAB-10304 and E-MTAB-10302, respectively (8). OE19 KAS-seq and CUT&Tag data were obtained from E-MTAB-11357 and E-MTAB-11356, respectively 35. The 4600 test eRNA set was obtained from published data (35).

Results

Identification of enhancer-like regulatory regions using STARR-seq in OAC cells

To identify genomic regions with enhancer activity in OAC cells, we applied STARR-seq in the OAC cell line, OE19 which we have previously demonstrated to faithfully recapitulate OAC at the chromatin level (5–7). This technique leverages the ability of putative enhancer elements to drive transcription (15,16). Putative enhancer elements are placed into the 3′-UTR of a truncated GFP reporter under the control of a minimal promoter (Figure 1A). If these elements function as enhancers, they may upregulate the production of the reporter upon introduction into cells. Genomic sequencing of these cells can determine the extent of the plasmid introduction into the cells, whilst RNA sequencing can determine the extent of reporter RNA production. As the enhancers are within the 3′-UTR, they serve as their own barcode in the generated transcripts. By calculating the ratio of RNA to plasmid insertion, it is possible to calculate relative enhancer strength (15,16).

We adopted a multi-pronged approach to identify putative enhancers by creating STARR-seq libraries from DNA fragments generated from OE19 cells, with accessible chromatin (via ATAC) or linked to the enhancer-associated chromatin factors and marks BRD4, H3K27ac and MED1 (via CUT&Tag). Libraries were created in lentiviral vectors to enable integration into the host genome. This created four independent libraries; ATAC-STARR, BRD4-STARR, H3K27ac-STARR and MED1-STARR libraries respectively. We then transduced OE19 cells with these libraries and used RNA-seq to identify enhancer derived transcripts alongside sequencing of the integrated gDNA as well as input plasmid library for determining signal enrichment relative to overall library abundance (Figure 1A).

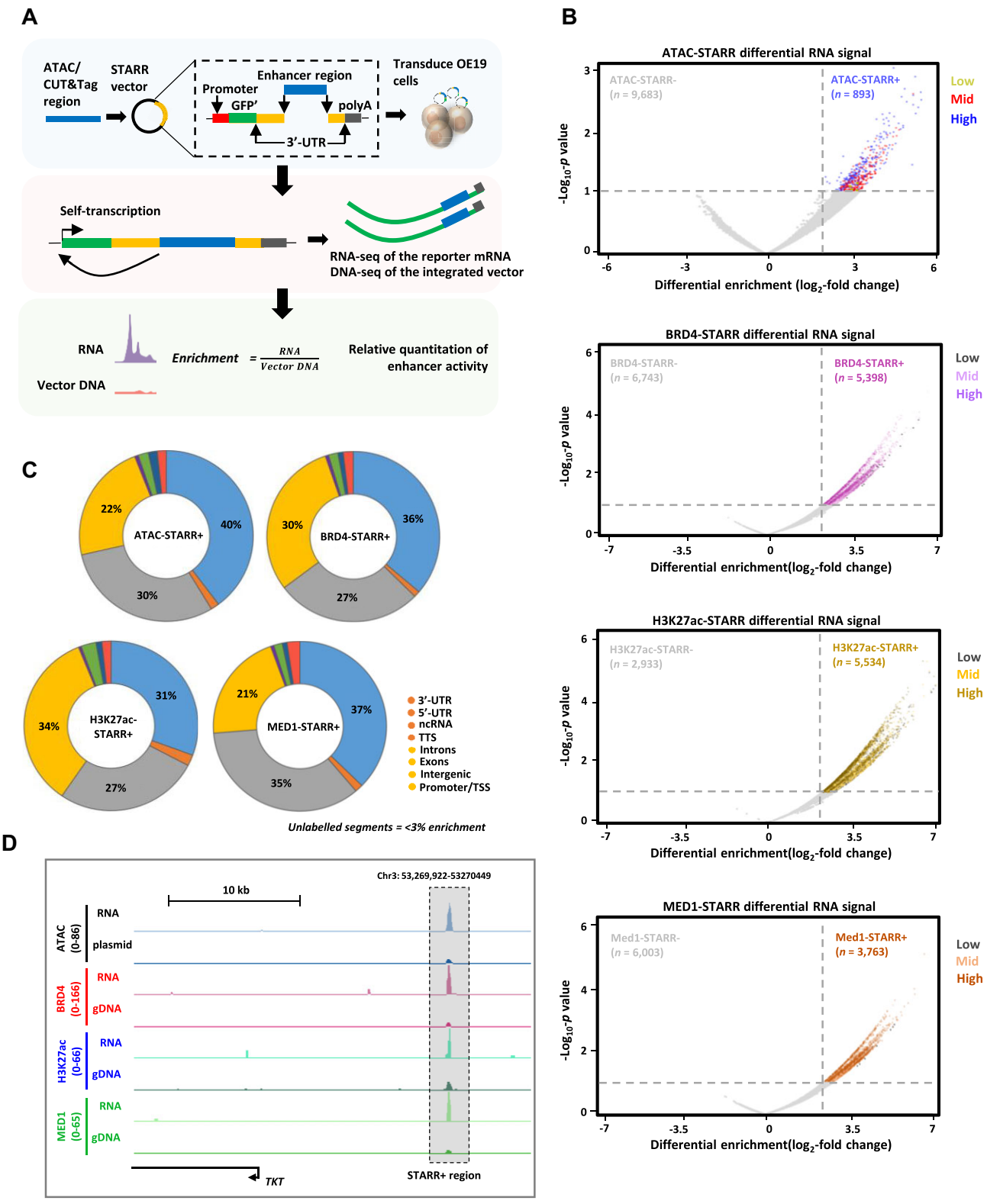


Figure 1. STARR-seq identifies potential enhancer regions. (**A**) STARR-seq strategy. Putative enhancer regions from ATAC-seq or CUT&Tag assays are inserted downstream from a truncated GFP reporter (GFP') in a lentiviral vector. Regulatory regions which represent active enhancers are revealed by RNA-seq where they are 'self transcribed' along with the GFP' segment. Enrichment is calculated relative to input DNA from the library or integrated into the host genome. (**B**) Volcano plots displaying the differential (± Log_2FC 2.0, < –Log_{10}-*P*-value = 1.0) RNA signal over plasmid (ATAC) or gDNA (BRD4, H3K27ac, MED1) for ATAC, BRD4, H3K27ac and Med1-STARR-seq assays. Regions are colour coded by RNA signal strength. (**C**) Genomic distribution of the STARR + regions for ATAC, BRD4, H3K27ac and MED1-STARR-seq assays. (**D**) Genome browser view of ATAC, BRD4, H3K27ac and MED1-STARR-seq assays, with plasmid (ATAC) or gDNA (BRD4, H3K27ac, MED1) and RNA tracks shown. Displayed is the *TKT* locus with an upstream STARR + region highlighted.

Differential STARR-derived RNA enrichment relative to plasmid library representation was calculated for each STARR-seq experiment and predominantly demonstrated higher levels of reporter-derived RNA signal enrichment, as expected (Figure 1B). Genomic regions showing significant enrichment ($\log_2 \geq 2$ fold; $-\log_{10}$-*P*-value ≤ 1) were taken forward as positive hits for demonstrating enhancer activity and henceforth be referred to as STARR + regions (conversely those not exhibiting high signal are referred to as STARR- regions). This analysis revealed, 893, 5398, 5534 and 3763 STARR + regions representing potential enhancers, from the ATAC-STARR (1%), BRD4-STARR (28%), H3K27ac-STARR (20%) and MED1-STARR (40%) plasmid libraries respectively. These regions were predominantly from mid-high RNA producing regions rather than low activity, potentially high abundance regions as indicated by colour coding each region (Figure 1B) based on whether in a low, mid or high expressing category (Supplementary Figure S1A). Furthermore, we determined the impact of plasmid count on RNA production, by categorising STARR regions into low, mid and high plasmid DNA copy number regions (Supplementary Figure S1B). This analysis demonstrated no disproportional enrichment of any plasmid DNA copy number categories, precluding plasmid drop-out or integration bias as drivers of RNA signal strength.

The genomic distribution of the STARR + regions showed consistency across all STARR libraries (Figure 1C) which is similar to that observed in all accessible ATAC-seq derived regions (Supplementary Figure S1C). However, by comparison to input libraries (7–16%) an enrichment was observed for promoter elements (21–34%) (Figure 1C). This phenomenon has previously been reported for other high throughput enhancer screens (36) and is consistent with the notion that a subset of promoters can function as enhancer elements in MPRAs (37). These observations confirm that the STARR + elements are not unrepresentative elements predisposed to transduction and are genuinely randomly sampled from the libraries from which they are generated.

STARR-seq therefore identifies a catalogue of self-transcribing regions that may represent enhancer elements in OAC cells. This is exemplified by the STARR + region identified by all STARR-seq modalities, found at the *TKT* locus (Figure 1D), which encodes a transketolase which has been associated with the tumourigenic properties of many different cancers (38)

High-confidence STARR + regions are marked with features of active chromatin

To create a high confidence set of STARR + regions, we overlapped the regions found in each dataset (Figure 2A) and found 1549 STARR + intersect regions common to at least two of the ATAC, BRD4, H3K27ac and MED1 STARR + datasets (Figure 2A; Supplementary Table S1). These intersect STARR + regions include promoter proximal elements in addition to both intergenic and intragenic regions (Supplementary Figure S2A). Intragenic enhancers, in particular, can be challenging to identify through other approaches such as eRNA profiling in patient samples (35,39). This relatively low level of overlap can be partially explained by the lack of saturation in our screen, but also by the fact that although there was high correlation between plasmid input and gDNA libraries (Supplementary Figure S2B), each gDNA library represented 32–42% coverage of all possible fragments from the input libraries. Moreover, when comparing across modalities, we found good correlations between gDNA libraries but numerically, while the majority of MED1-defined regions are also present in in the two other datasets (76 and 96%), there was a lower between H3K27ac- and BRD4-derived gDNA libraries (42 and 63%). Collectively, this means that the theoretical possible numbers of overlaps between modalities is substantially less than 100%, and the gDNA libraries are subsampling all possible genomic fragments. This is not unexpected, given that we are using different antibodies and biologically, all enhancers would not necessarily have the same levels of each modification, and different combinations of modifications and binding proteins likely produce different activity levels.

We next sought to identify the features associated with these STARR + intersect regions, and the implications on regulatory potential. We previously generated CUT&Tag, ChIP-seq and ATAC-seq data for a range of chromatin-associated factors in OE19 cells (6,7,35) and we used this data to examine the overall levels of a broad range of chromatin marks and binding proteins across all of the STARR + regions. Strikingly, the high confidence 1549 STARR + intersect peaks showed widespread enrichment for chromatin features associated with active regulatory potential, including BRD4, H3K27ac, MED1 and PolII as expected, but also more overt enhancer-associated marks such as H3K4me1/2 (Figure 2B). This is exemplified by an intergenic region located upstream from *TKT* (Figure 2C). Chromatin architectural marks were also detected (CTCF and SMC1) but there was little evidence of the repressive chromatin mark H3K27me3 (Figure 2B). While the STARR + regions identified in all of the individual assays exhibited enrichment for active chromatin features, the high confidence intersecting regions showed higher levels than any of the regions identified uniquely in a particular dataset (Figure 2D; Supplementary Figure S2C). This mirrors the data from the STARR-seq assays where there is high signal resulting from the assay in which the data is generated but high signal is only uniformly observed across all marks in the intersecting regions (Supplementary Figure S2D and E). Overall, the high confidence intersect set of 1549 STARR + regions exhibit enrichment of features indicative of active enhancers in OE19 cells.

To further explore the characteristics of regions exhibiting enhancer activity, we compared the ratios of H3K4me1:H3K4me3 and H3K4me2:H3K4me3 found in OE19 cells between STARR + intersect regions annotated as promoter-proximal (–1 kb to +0.1 kb) and the remaining intergenic or intragenic regions (representing distal enhancers). Both ratios were significantly higher for distal enhancer elements, than for promoter-proximal elements (Figure 2E; Supplementary Figure S2F). These analyses imply that the subset of intersect STARR + regions located in intergenic or intragenic regions, may represent bona fide enhancers. This is corroborated by metaplots which show a clear demarcation of higher H3K4me1 signal at putative enhancers and H3K4me3 signal at promoter proximal regions (Supplementary Figure S2G).

To explore the potential functional consequences of enhancer activity at the STARR + intersect regions, we linked STARR + regions to the nearest coding gene TSS using HOMER (27) (Supplementary Table S1) and compared the expression of the 1372 genes linked to the STARR +

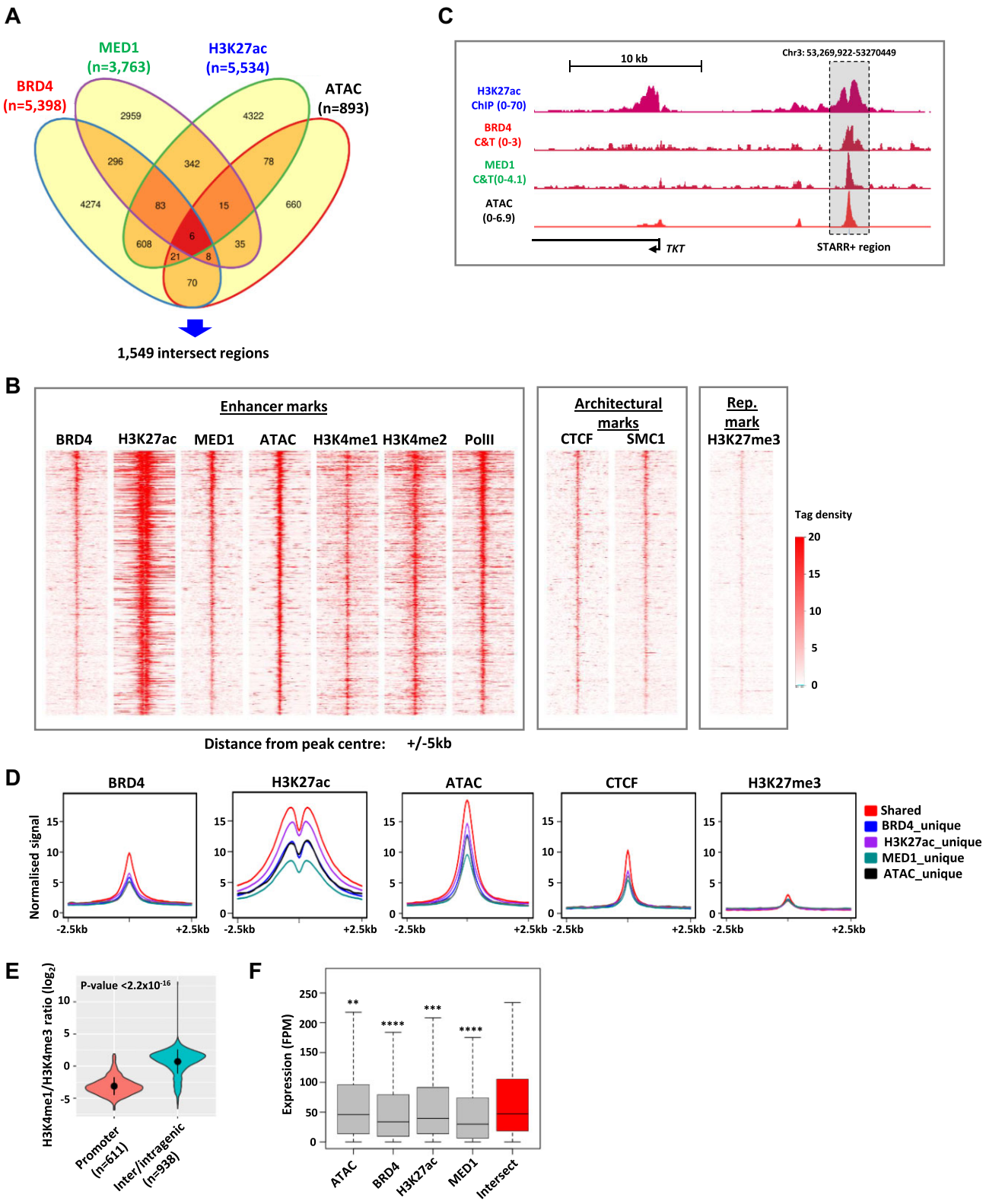


Figure 2. High confidence STARR + regions are associated with active chromatin and transcriptional features. (**A**) Venn-diagram displaying the intersect between ATAC-, BRD4-, H3K27ac- and MED1-STARR + regions. (**B**) Heatmaps showing CUT&Tag (BRD4, MED1, H3K4Me1, H3K4Me2, PolII, CTCF, SMC1 and H3K27me3), ChIP-seq (H3K27ac) and ATAC-seq signal in OE19 cells at the 1549 STARR + intersect regions. (**C**) Genome browser view of H3K27ac ChIP-seq, BRD4 CUT&Tag, MED1 CUT&Tag and ATAC-seq signal at an intergenic STARR + intersect region highlighted upstream from the *TKT* locus. (**D**) Metaplots of CUT&Tag, ChIP-seq and ATAC-seq signal in OE19 cells at the 1549 STARR + intersect regions compared to STARR + unique regions in ATAC, BRD4, H3K27ac and MED1-STARR-seq assays. (**E**) Violin plots displaying the ratio of H3K4me1:H3K4me3 ratios at distal regulatory and promoter proximal regions (–1 kb to +0.1 kb) within the 1549 STARR + intersect regions (*P*-value is shown; Welch's *t*-test). (**F**) Box plots comparing the expression of genes in OAC patient tissue total RNA-seq samples from the OCCAMS dataset (n = 210) annotated to unique BRD4, H3K27ac, MED1-STARR + regions against genes annotated to the 1549 STARR + intersect regions (*P*-value is shown; **≤0.01, ***≤0.001, ****≤0.0001 Student's *t*-test).

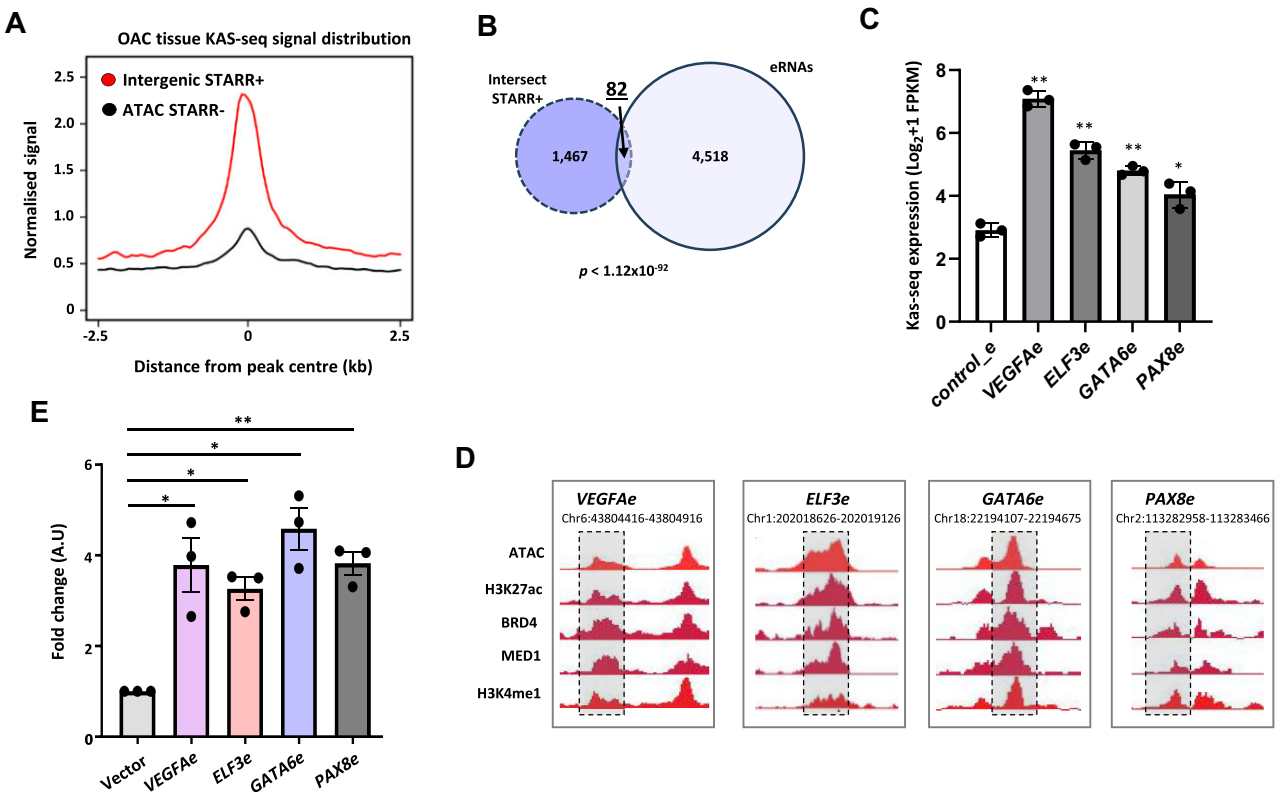


Figure 3. High confidence STARR + regions are associated with enhancer-like activity. (**A**) Metaplot of OAC patient tissue KAS-seq signal at the 566 STARR + intersect regions found in intergenic regions compared to 2303 randomly selected STARR- ATAC peaks from OE19 cells. (**B**) Venn-diagram of overlap between the 1549 STARR + intersect regions and 4600 previously identified BO and OAC patient eRNAs (*P*-value is shown; Fisher's exact test). (**C**) Bar graph displaying the difference in KAS-seq expression between *VEGFAe, ELF3e, GATA6e* and *PAX8e* regions, compared to the average of four negative control regions in OE19 cell KAS-seq data (** $P < 0.01$; * $P < 0.05$; *t*-test relative to controls; $n = 3$). (**D**) Genome browser view of ATAC-seq, BRD4, H3K27ac and MED1 and H3K4me1 CUT&tag signal at the regions associated with the *VEGFAe, ELF3e, GATA6e* and *PAX8e* STARR + regions (boxed). (**E**) Bar graph displaying the difference in luciferase reporter activity between *VEGFAe, ELF3e, GATA6e* and *PAX8e,* compared to vector only negative control (** $P < 0.01$; * $P < 0.05$; one-way ANOVA with Bonferroni's correction).

intersect regions in OE19 cells and OAC patients, to those linked with the unique STARR + regions. The intersect regions exhibited a significantly higher expression of associated genes compared to all of the other 'unique' regions in both OE19 cells (Supplementary Figure S2H) and OAC patients (Figure 2F), consistent with a potential role as active enhancers for these genes in OAC. This may be an underestimation as enhancers are not always linked to the nearest gene.

In summary, we have identified a high confidence dataset of regulatory elements in OAC cells with enhancer activity which are associated with high target gene activity in OAC patients.

High-confidence STARR + regions demonstrate enhancer activity

To further explore the enhancer-like properties of the STARR + regions and more directly link the STARR + regions to clinically relevant active transcription, we analysed kethoxal-assisted single-stranded DNA-sequencing (KAS-seq) data from OE19 OAC cells (35) and performed additional analyses on three OAC patient samples (Supplementary Table S2B). This approach serves as a proxy for active transcription, and potentially eRNA production, thereby indicative of enhancer activity given the close association between eRNA production and enhancer activity (37). The data from OE19 cells shows good correlation with each of the OAC tissue KAS-seq samples (Supplementary Figure S3A). We focussed on intergenic regions to avoid interference from coding gene transcription and found substantially higher levels of KAS-seq signal in OE19 cells at the STARR + intersect regions compared to a random set of accessible chromatin regions not identified in the STARR-seq assays (Supplementary Figure S3B) consistent with enhancer transcription and hence enhancer activity in cancer cells. Next, we turned to patient samples, and again we found higher levels of KAS-seq signal associated with the intersect STARR + regions, indicative of high enhancer activity in OAC patients (Figure 3A). To further explore the relevance of the regulatory regions we have identified in a disease context, we next sought to determine whether any of our previously identified eRNAs in OAC patients (35) were located within the STARR + intersect regions. Comparing the two datasets, we found a significant enrichment of eRNAs amongst the intergenic STARR + intersect regions, consistent with potential enhancer activity in OAC patients (Figure 3B). This overlap of 82 regions is relatively modest which likely arises from a combination of reasons. Firstly, not all patient-specific enhancers are present in OE19 cells. Secondly, we took a highly stringent approach by taking overlap STARR + regions. Indeed, when taking both these two parameters into account, we observed 400 regions that overlapped between the non-redundant total STARR + regions and the eRNA producing regions in patients (Supplementary Figure S3C). Other issues such as the inability of all enhancers to function in isolation in reporter vectors and the filtering out of intragenic enhancers when deriving the

eRNA datasets, further lessens the probability of finding overlaps between the two approaches.

Next, we wished to validate a panel of enhancer regions from the STARR + regions using an independent reporter assay based on a luciferase reporter in OE19 cells. We selected four STARR + intersect regions, annotated to cancer relevant genes: *vascular endothelial growth factor A* (*VEGFA)*, *ELF3*, *GATA6* and *paired box gene 8* (*PAX8)*, which displayed significantly increased KAS-seq signal in OAC tissue relative to a panel of control genomic regions (Figure 3C). These regions all show a range of chromatin features indicative of enhancer activity (Figure 3D). Importantly, all regions demonstrated a significant increase in luminescence in the reporter assay, relative to vector only control (Figure 2E) authentic enhancer activity.

Collectively, this data further supports the conclusion that the STARR + intersect regions we have identified represent *bona fide* regulatory elements with enhancer-like properties and demonstrates their likely clinical relevance due to their association with patient-derived eRNAs and KAS-seq signal in OAC tissue.

Transcription factor and gene activity associated with high-confidence STARR + regions in OAC cells and patients

To further understand the biological activity associated with the high confidence STARR + regions, we performed gene ontology (GO) analysis on their associated genes. This revealed a significant enrichment in processes associated with OAC, including 'VEGFA signalling', 'pathways in cancer' and 'actin cytoskeleton organisation' (Figure 4A). Next, we sought to identify the transcription factors that may bind and regulate the STARR + intersect regions. We identified an enrichment of DNA motifs associated with transcription factors previously shown to play a role in OAC, such as KLF5 as well as AP-1, FOX and ETS family members (Figure 4B) (5-7,40). In the case of KLF5, motif occurrences in the STARR + regions were significantly higher than across all open chromatin regions in OE19 cells, whereas other transcription factors previously implicated in OAC either showed no enrichment (HNF4A) or reduced enrichment (GATA6) relative to the expected frequency (Figure 4C). To provide more direct evidence for regulatory potential, we examined overlaps between ChIP-seq binding datasets for KLF5, GATA6 and HNF4A transcription factors in OE19 cells at the peak (Supplementary Figure S4A) or signal level (Figure 4D). In all cases, significant overlap was observed (Supplementary Figure S4A). Subsequent clustering of binding activity allowed us to identify five broad subgroups among the STARR + regions (Figure 4D), with a relatively small triply bound group (C1:KGH), a group co-bound bound by KLF5 and GATA6 (C2:KG), two groups bound by either KLF5 (C3:K) or GATA6 alone (C4:G) and one group not bound at high levels by any of the transcription factors (C:5). Regions surrounding the *VEGFA* (Figure 4E) and *PIM3* (Supplementary Figure S4B) gene loci exemplify STARR + regions from the C1:KGH and C3:K clusters respectively. All groups showed high levels of open chromatin, H3K27ac and BRD4 binding (Figure 4D, right; Supplementary Figure S4C, top). However, KAS-seq signal (an indicator of active transcription) was highest in the KLF5 only cluster (K) and lowest in the cluster featuring GATA6 binding (G) (Figure 4D, right; Supplementary Figure S4C, bottom left). Importantly, this high signal in the KLF5-specific cluster was maintained when considering only intergenic STARR + regions (Supplementary Figure S4C, bottom right) indicating that this signal is likely generated from eRNA transcription, a feature of active enhancers (9). We therefore analysed KAS-seq from OAC patient samples and found that the KLF5-specific STARR + regions also exhibited the highest transcriptional activity, consistent with active enhancer activity in patients (Figure 4F).

Both motif analysis and ChIP-seq therefore implicate KLF5-associated STARR + regions as active enhancers in OAC cell lines and patients. We therefore asked whether likely target genes also showed enhanced expression in OE19 cells. Compared to a random selection of open chromatin regions, all transcription factor binding-defined clusters showed significantly higher expression of their nearest genes, with the KLF5-specific cluster the highest level and the GATA6-specific cluster the lowest (Supplementary Figure S4D, top). More generally, the genes closest to the KLF5-associated STARR + regions showed higher expression in OE19 cells than those with no associated KLF5 binding (Figure 4G, top). Similarly, in OAC patient samples a similar trend was observed with genes closest to KLF5-associated regions showing the highest expression (Figure 4G, bottom; Supplementary Figure S4D, bottom) but genes associated with GATA6-only regions showed the lowest expression levels in patients (Supplementary Figure S4D, bottom).

Together these analyses demonstrate that STARR + intersect regions are associated with OAC-specific transcription factors and have target genes that are related to OAC pathology. In particular, the association of the most active enhancers and their likely target genes with KLF5 binding is consistent with our previous finding of an important role for KLF5 in OAC progression from the Barrett's pre-cancer state (6).

High-confidence STARR + regions are associated with clinically relevant regulatory events

The locus encoding the RTK ERRB2 is frequently amplified in OAC, and is therefore thought to be an oncogenic driver event (2,41) and is a target for pharmacological inhibition in the clinic (42). We therefore explored whether the activity of the STARR + regions was altered by treatment of OE19 cells (which contain amplifications of *ERRB2* locus) with the ERBB2 small-molecule inhibitor lapatinib. First, we assessed differential accessibility through ATAC-seq and H3K27ac signal (an activation associated histone mark) by ChIP-seq (8) at STARR + intersect regions upon lapatinib treatment for 24 h. Through using these datasets, we identified 2 and 24 STARR + intersect regions that gained and lost accessibility, respectively, upon lapatinib treatment and 5 and 48 STARR + intersect regions that increased and decreased in H3K27ac signal upon lapatinib treatment (Figure 5A; Supplementary Table S3). Generally consistent changes were seen in chromatin accessibility and histone H3K27 acetylation levels, with decreases in both observed following lapatinib treatment (Figure 5B). An exemplar region is located upstream from the *DUSP5* locus, where chromatin changes indicative of enhancer inactivation (Figure 5C) accompanies reductions in gene expression (Supplementary Figure S5A) following lapatinib treatment. Expression of *DUSP5* has potential clinical and prognostic significance as it is expressed more in OAC compared to Barrett's (Supplementary Figure S5B),

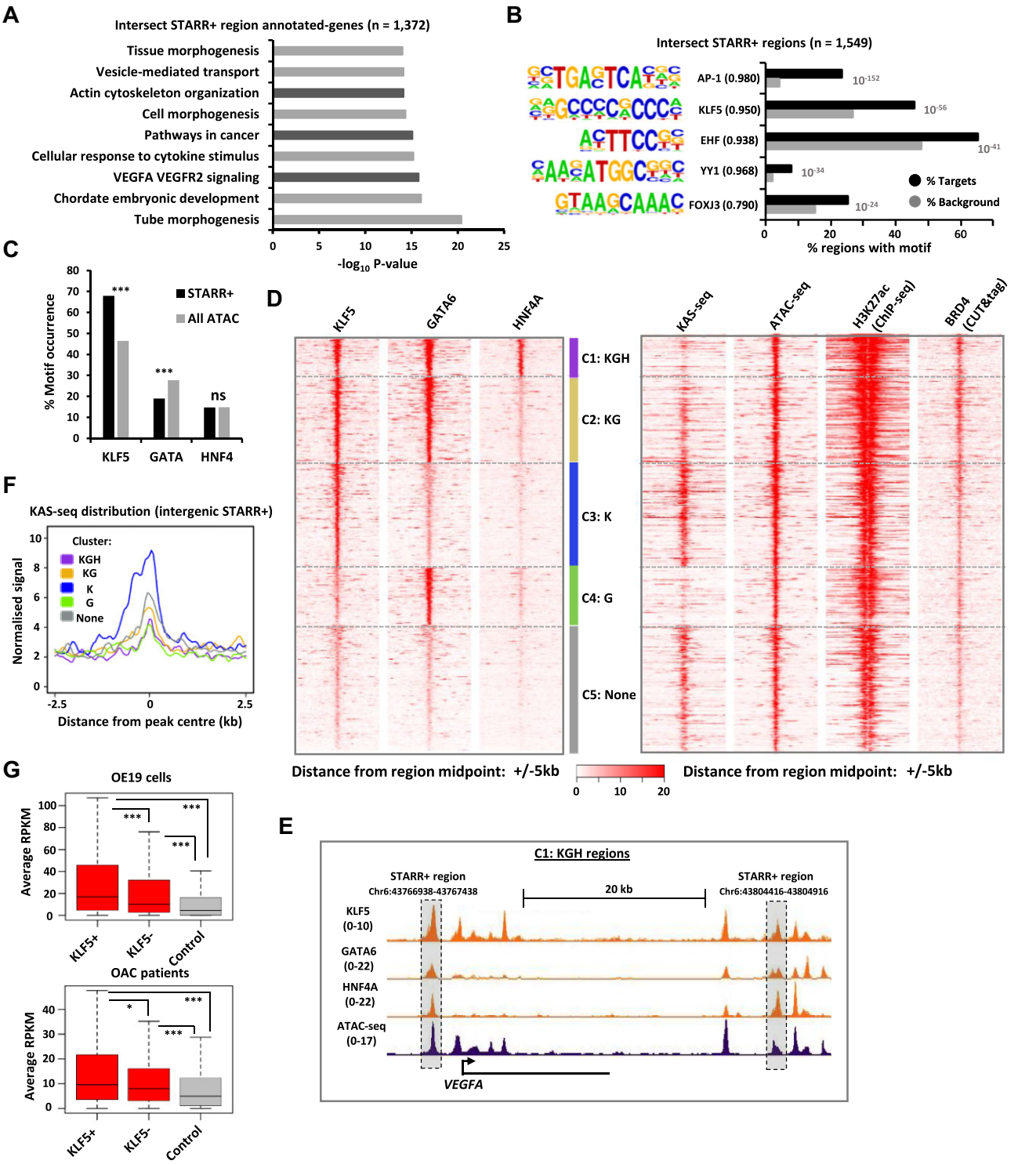


Figure 4. High confidence STARR + regions are enriched for binding to core OAC transcription factors, active chromatin features and are associated with active genes. (**A**) GO-term analysis of genes annotated to the 1549 STARR + intersect regions. (**B**) Transcription factor *de novo* motif enrichment using HOMER for STARR + intersect regions (*P*-values are shown). Matches to the indicated motifs are indicated in brackets. (**C**) Frequency of DNA binding motif occurrence for KLF5 (DGGGYGKGGC), GATA (NBWGATAAGR) and HNF4 (CARRGKBCAAAGTYCA) transcription factors in the 1549 intersect STARR + regions and all 99855 ATAC-seq peaks found in OE19 cells. **** *P*-value < 0.001; ns = non-significant. (**D**) Heatmap of ChIP-seq signal for the transcription factors KLF5 (**K**), GATA6 (**G**) and HNF4A (**H**) after kmeans clustering into five clusters (left) and associated KAS-seq, ATAC-seq, ChIP-seq and CUT&Tag signals for the same regions (right). (**E**) Genome browser view of KLF5, GATA6 and HNF4A ChIP-seq, and ATAC-seq signal at C1:KGH cluster STARR + intersect regions (highlighted) surrounding the *VEGFA* locus. (**F**) Average tag density plots of KAS-seq signal from OAC patients ($n = 3$) for each of the five clusters of the intergenic high confidence STARR + regions. (**G**) Boxplot of average expression values of the genes nearest to the 1549 STARR + intersect regions from clusters C1-3 combined (KLF5+; 860 peaks, 774 genes), clusters C4-5 combined (KLF5-; 689 peaks, 659 genes) or three control sets of randomly selected ATAC-seq peaks, in OE19 cells (top) or 28 OAC patient samples (bottom). Statistical significance is shown; *P*-values < 0.001, **** and < 0.05, *.

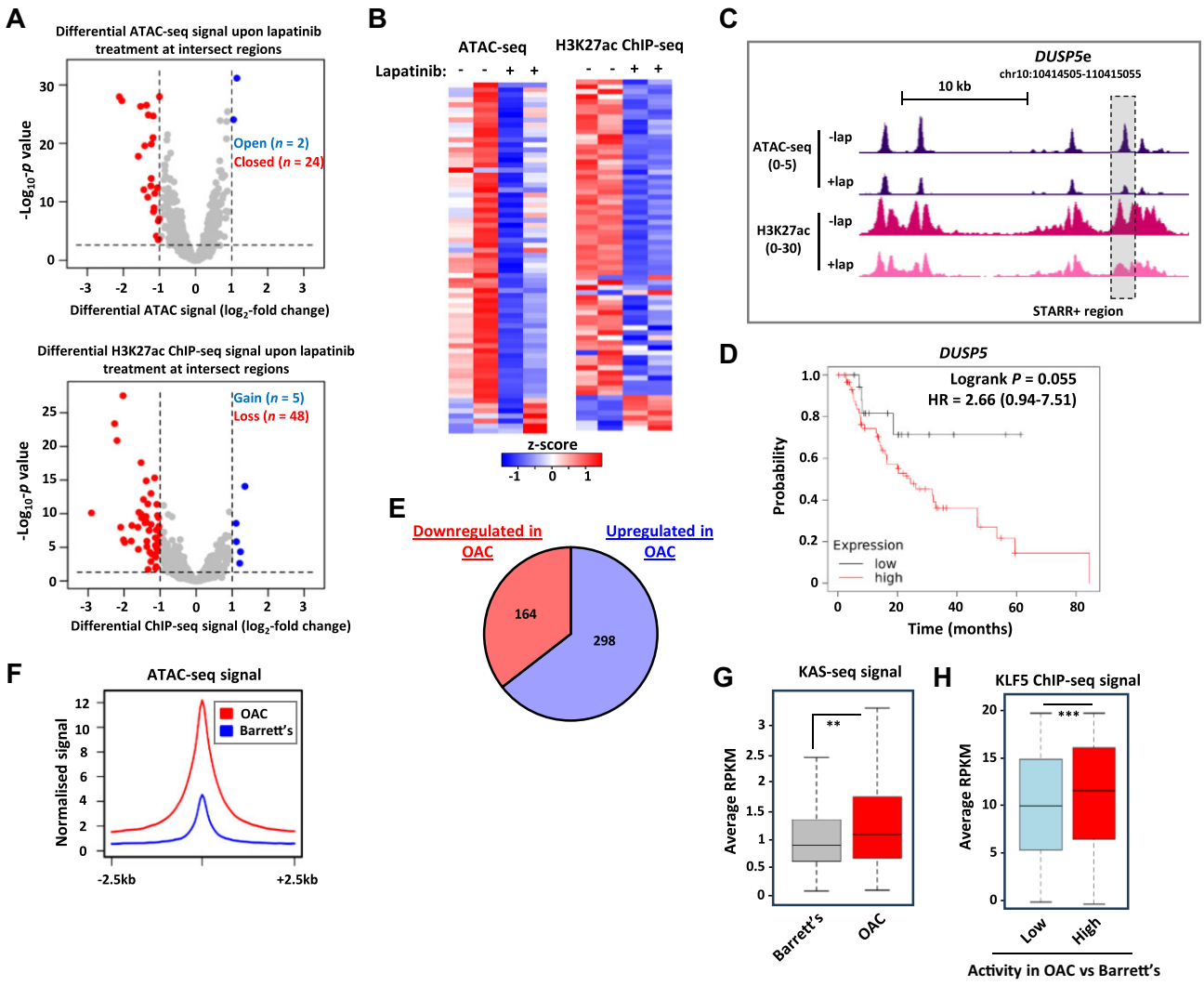


Figure 5. STARR + defined enhancer regions respond to ERBB2 inhibition and are associated with OAC-specific gene regulation. (**A**) Volcano plots displaying the differential OE19 cell ATAC-seq (left; ± $Log_2FC = 1$, < *P*-value = 0.05) or OE19 cell H3K27ac ChIP-seq (right; ± $Log_2FC = 1$, < *P*-value = 0.5) at the 1549 STARR + intersect regions upon 24 h lapatinib treatment. (**B**) Heatmap of the ATAC-seq (left) or H3K27ac ChIP-seq (right) signal following 24hr lapatinib treatment of OE19 cells at the union of significantly closing/opening peaks from part A (i.e. red and blue dots). Two experimental replicates are shown for each condition. Data are row z-normalised, with ATAC-seq and ChIP-seq data normalised separately. (**C**) Genome browser view of ATAC-seq and H3K27ac CUT&Tag signals at the STARR + region upstream from *DUSP5* (*DUSP5*e) in OE19 cells following treatment with DMSO or lapatinib for 24 h. (**D**) Kaplan–Meier plot comparing overall survival between OAC patients ($n = 80$) with low ($n = 19$) or high ($n = 61$) *DUSP5* expression in the TCGA PanCancer Atlas dataset (Log rank *P*-value is shown). (**E**) Pie chart showing the numbers of genes associated with the high confidence 1549 STARR + regions that are up or downregulated in OAC compared to matched Barrett's samples from the same patients ($n = 28$; >1.5-fold change, *P*-value < 0.01). (**F**) Metaplots of ATAC-seq signal within the 1549 STARR + intersect regions in 4 Barrett's or 11 OAC patient samples. (**G**) Boxplot of average KAS-seq signal in the 309 STARR + intersect regions found in intergenic regions, from three matched Barrett's and OAC patient samples. Statistical significance is shown; *P*-values < 0.01, **. (**H**) Boxplot of average KLF5 ChIP-seq signal in the top and bottom terciles of the 1549 STARR + intersect regions (based on fold difference in ATAC-seq signal in OAC relative to Barrett's patients). Statistical significance is shown; *P*-values < 0.001, ***.

and higher expression levels predict poorer patient survival (Figure 5D).

Finally, we asked more globally whether the STARR + regulatory regions we have identified are relevant to OAC by first asking whether accessible chromatin corresponding to these regulatory regions is also present in OAC patient samples. Almost all regions (98%) intersect with accessible chromatin peaks in at least one OAC patient and the majority are found in multiple patients, with 899 (58%) found in eight or more patients (Supplementary Figure S5C). Next, we compared the expression of their closest genes in both Barrett's and OAC patients (Supplementary Table S2A) and there was clear enrichment of genes which are upregulated in OAC patients compared to the pre-cancer state (Figure 5E). Furthermore, consistent with this observation, STARR + regions are more accessible (Figure 5F) and KAS-seq signal is higher at intergenic STARR + regions in OAC patients compared to Barrett's patients (Figure 5G; Supplementary Table S2B), indicating higher enhancer activity in cancer cells. We then partitioned the STARR + regions based on differences in accessibility in OAC versus Barrett's as a proxy for increased activity and examined KLF5 binding levels in the top and bottom terciles. Significantly more KLF5 binding activity was observed within the STARR + regions showing the highest activity in-

crease in OAC patients (Figure 5H), in keeping with our finding that this transcription factor is important for progression to OAC.

Collectively, this data demonstrates that the STARR + regulatory regions we have identified are associated with OAC-specific properties in patients, as many are activated in response to oncogenic ERBB2 signaling and their activity is elevated in OAC patients. Our findings therefore provide us with insights into the active regulatory regions of the genome associated with this disease

Discussion

Cancer can be viewed as a disease characterised by disrupted gene regulatory networks (43). Enhancer elements represent fundamental hubs within these gene regulatory networks. When active, these regions serve as platforms for the integration of signals that dictate a broad range of outcomes influencing cellular phenotype, and when abnormally activated can result in tumourigenesis. Traditionally, histone marks such as H3K27ac and H3K4me1, in addition to chromatin accessibility, have been used to define enhancer elements (10,11). However, while valuable, this approach is correlative; histone marks and accessibility may outline the location of a putative enhancer but these strategies remain limited for the definition of *bona fide* active enhancers. More recently, the finding that when enhancers are active they are themselves transcribed to produce eRNAs, has opened up the possibility of using eRNA profiling to identify enhancers in cancer cells (35,39). However, while promising, Intragenic enhancers can be challenging to identify through this approach due to the interference from ongoing genic transcription. Here, we used STARR-seq to identify representing potentially active enhancers in the OE19 OAC cell line. Using KAS-seq as a validation tool, in addition to our previously published patient eRNA dataset, we confirmed these regions as high confidence enhancers that are also operational in OAC patients. Moreover, we demonstrate the response of a subset of identified enhancers to the ERBB2 inhibitor lapatinib, which highlights the opportunity to approach clinically-relevant questions using our enhancer identification strategy.

A recent study applied STARR-seq to interrogate enhancer activity in gastric cancers (44). Whilst OAC does bear molecular similarities to a specific subtype of chromosomally-instable (CIN) gastric cancers (41), the authors utilise a heterogeneous panel of gastric tumours and normal biopsies, as well as 28 gastric cancer cell lines to generate a STARR reporter library consisting of universal-common H3K27ac-marked elements. This amalgamation of sample types is likely to underestimate any findings given the exquisite cell-type specificity of enhancer activity. Additionally, it is now widely accepted that H3K27ac alone represents a poor indicator of enhancer activity (12,13). In addition to the lack of CIN mutational signature designation for the library source, these caveats preclude the general applicability of these findings to OAC. Here, we focused on using OE19 cells, which faithfully recapitulate OAC at the chromatin level (5–7). Furthermore, our strategy for STARR library generation uses a range of input material that are associated with enhancers, and includes H3K27ac but also BRD4, MED1 and accessible chromatin. Collectively, this approach ensures the relevancy of our results, and broadens the search radius, improving chances of successfully identifying active enhancer regions.

Our approach identified ∼1500 high confidence active enhancers of potential relevance to OAC. These enhancer elements displayed an association with genes linked to processes pertinent to OAC biology, as well as an enrichment in the motifs of transcription factors which have been shown to play a role in the development of OAC (5–7,40). Integration with our previously published genome-wide binding data on a variety of chromatin-associated factors in OE19 cells, as well as our patient-derived eRNA dataset, we verified the enhancer-like nature of our ∼500 regions in addition to highlighting their potential clinical relevance. In support of this, we generated KAS-seq data from OAC patient tumour samples to monitor enhancer activity. KAS-seq measures the transcription bubble (28), serving as a proxy for enhancer transcription. Accordingly, these regions demonstrated high KAS-seq signal in both OE19 cells and patients further confirming their designation as high confidence enhancers. However, we also identified a large number of putative enhancers that are unique to a particular STARR assay. Further data mining alongside newly generated datasets will likely uncover additional *bona fide* enhancers in these cohorts, providing more broader utility of this resource to the community.

An interesting observation from our data was that KLF5 motif presence and chromatin binding is associated with more active enhancer regions and associated target genes whereas GATA6 bound enhancers exhibit the opposite tendency. The former observation is consistent with our previous finding that KLF5 repurposing through redistribution to a novel set of regulatory regions is one of the key molecular events that distinguishes OAC from the Barrett's precursor state (7). Conversely, the implications on GATA6 functionality are less obvious as the gene encoding GATA6 is often amplified during the same transition, so a more prominent role in driving enhancer activity would be expected. It is possible that it acts to moderate the outputs of other transcription factors in the core OAC network to provide the optimal level of gene expression and this enigmatic phenomenon warrants further investigation alongside the outcomes of other combinatorial transcription factor interactions. Our dataset provides candidate regions to test these models.

To explore any clinical associations with our new STARR-seq-defined dataset, we utilised the small-molecule ERBB2 inhibitor, lapatinib. We have previously demonstrated that OE19 cells undergo genome wide chromatin accessibility changes upon lapatinib treatment, and this has important consequences for development of therapeutic resistance (8). These chromatin changes may reflect the activation or inactivation of enhancer elements. By integrating H3K27ac ChIP-seq and ATAC-seq data, we identified a subset of our newly defined enhancer elements that lose accessibility and become less marked by H3K27ac following lapatinib treatment. Furthermore, we show clear associations with increased accessibility and activity in OAC relative to Barrett's patients, further emphasising the relevance of our data to gene regulatory events occurring in OAC patients.

In summary, our STARR-based approach has allowed us to identify a set of enhancers that are active in OAC which has led to a wider understanding of the transcriptional mechanisms and pathways that are operational during the onset of this condition. This provides a compendium of enhancer regions for instigating further mechanistic studies into the gene regulatory networks that are operational in this deadly disease.

Data availability

STARR-seq data (E-MTAB-14083 and E-MTAB-14155), OAC and Barrett's tissue KAS-seq (E-MTAB-14091), CUT&Tag (E-MTAB-14090), RNA-seq (E-MTAB-14093) and KAS-seq data in OE19 cells (E-MTAB-14095 and E-MTAB-14098) are deposited in ArrayExpress and listed in Supplementary Table S5.

Supplementary data

Supplementary Data are available at NAR Cancer Online.

Acknowledgements

We thank Guanhua Yan for excellent technical assistance, and staff in the Bioinformatics and Genomic Technologies core facilities. We also thank Rebecca Fitzgerald, Cambridge on behalf of the OCCAMS network for providing the patient RNA-seq data and matched Barrett's and OAC samples for KAS-seq analysis. We are grateful to Chuan He for providing us with N_3-kethoxal. .

Author contributions: I.A., S.H.Y., C.W.B., Y.L. performed the experiments and data analysis in this study; A.D.S. contributed to the inception, design and supervision of the project. All authors contributed to manuscript preparation and/or critically appraised manuscript drafts.

Funding

Manchester Cancer Research Centre (MCRC); MRC [MR/V010263/1]; Wellcome Trust [102171/B/13/Z].

Conflict of interest statement

None declared.

References

1. Coleman,H.G., Xie,S.H. and Lagergren,J. (2018) The epidemiology of esophageal adenocarcinoma. *Gastroenterology*, **154**, 390–405
2. Frankell,A.M., Jammula,S., Li,X., Contino,G., Killcoyne,S., Abbas,S., Perner,J., Bower,L., Devonshire,G., Ococks,E., *et al.* (2019) Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium; Fitzgerald RC. The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. *Nat. Genet.*, **51**, 506–516
3. Ross-Innes,C.S., Becq,J., Warren,A., Cheetham,R.K., Northen,H., O'Donovan,M., Malhotra,S., di Pietro,M., Ivakhno,S., He,M., *et al.* (2015) Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. *Nat. Genet.*, **47**, 1038–1046
4. Stachler,M.D., Taylor-Weiner,A., Peng,S., McKenna,A., Agoston,A.T., Odze,R.D., Davison,J.M., Nason,K.S., Loda,M., Leshchiner,I., *et al.* (2015) Paired exome analysis of Barrett's esophagus and adenocarcinoma. *Nat. Genet.*, **47**, 1047–1055.
5. Britton,E., Rogerson,C., Mehta,S., Li,Y., Li,X. and OCCAMS consortiumOCCAMS consortium, Fitzgerald,R.C., Ang,Y.S. and Sharrocks,A.D. (2017) Open chromatin profiling identifies AP1 as a transcriptional regulator in oesophageal adenocarcinoma. *PLoS Genet.*, **13**, e1006879
6. Rogerson,C., Britton,E., Withey,S., Hanley,N., Ang,Y.S. and Sharrocks,A.D. (2019) Identification of a primitive intestinal transcription factor network shared between esophageal adenocarcinoma and its precancerous precursor state. *Genome Res.*, **29**, 723–736.
7. Rogerson,C., Ogden,S., Britton,E. and OCCAMS ConsortiumOCCAMS Consortium, Ang,Y. and Sharrocks,A.D. (2020) Repurposing of KLF5 activates a cell cycle signature during the progression from a precursor state to oesophageal adenocarcinoma. *eLife*, **9**, e57189
8. Ogden,S., Carys,K., Ahmed,I., Bruce,J. and Sharrocks,A.D. (2022) Regulatory chromatin rewiring promotes metabolic switching during adaptation to oncogenic receptor tyrosine kinase inhibition. *Oncogene*, **41**, 4808–4822.
9. Andersson,R., Gebhard,C., Miguel-Escalada,I., Hoof,I., Bornholdt,J., Boyd,M., Chen,Y., Zhao,X., Schmidl,C., Suzuki,T., *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.
10. Creyghton,M.P., Cheng,A.W., Welstead,G.G., Kooistra,T., Carey,B.W., Steine,E.J., Hanna,J., Lodato,M.A., Frampton,G.M., Sharp,P.A., *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *P. Natl. Acad. Sci. U.S.A.*, **107**, 21931–21936.
11. Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W., *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.
12. Sankar,A., Mohammad,F., Sundaramurthy,A.K., Wang,H., Lerdrup,M., Tatar,T. and Helin,K. (2022) Histone editing elucidates the functional roles of H3K27 methylation and acetylation in mammals. *Nat. Genet.*, **54**, 754–760.
13. Pengelly,A.R., Copur,Ö., Jäckle,H., Herzig,A. and Müller,J. (2013) A histone mutant reproduces the phenotype caused by loss of histone-modifying factor Polycomb. *Science*, **339**, 698–699.
14. Dorighi,K.M., Swigut,T., Henriques,T., Bhanu,N.V., Scruggs,B.S., Nady,N., Still,C.D. 2nd, Garcia,B.A., Adelman,K. and Wysocka,J. (2017) Mll3 and Mll4 facilitate enhancer RNA synthesis and transcription from promoters independently of H3K4 monomethylation. *Mol. Cell*, **66**, 568–576.
15. Inoue,F. and Ahituv,N. (2015) Decoding enhancers using massively parallel reporter assays. *Genomics*, **106**, 159–164.
16. Arnold,C.D., Gerlach,D., Stelzer,C., Boryń,Ł.M., Rath,M. and Stark,A. (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, **339**, 1074–1077.
17. Pang,B. and Snyder,M.P. (2020) Systematic identification of silencers in human cells. *Nat. Genet.*, **52**, 254–263.
18. Muerdter,F., Boryń,Ł.M., Woodfin,A.R., Neumayr,C., Rath,M., Zabidi,M.A., Pagani,M., Haberle,V., Kazmar,T., Catarino,R.R., *et al.* (2018) Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods*, **15**, 141–149.
19. Corces,M.R., Granja,J.M., Shams,S., Louie,B.H., Seoane,J.A., Zhou,W., Silva,T.C., Groeneveld,C., Wong,C.K., Cho,S.W., *et al.* (2018) Cancer Genome Atlas Analysis Network; Greenleaf WJ, Chang HY. The chromatin accessibility landscape of primary human cancers. *Science*, **362**, eaav1898
20. Kaya-Okur,H.S., Janssens,D.H., Henikoff,J.G., Ahmad,K. and Henikoff,S. (2020) Efficient low-cost chromatin profiling with CUT&tag. *Nat. Protoc.*, **15**, 3264–3283.
21. Tiscornia,G., Singer,O. and Verma,I.M. (2006) Production and purification of lentiviral vectors. *Nat. Protoc.*, **1**, 241–245.
23. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
24. 1000 Genome Project Data Processing Subgroup, Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, ,**25**, 2078–2079
25. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
26. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137

27. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589
27. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
28. Wu,T., Lyu,R., You,Q. and He,C. (2020) Kethoxal-assisted single-stranded DNA sequencing captures global transcription dynamics and enhancer activity in situ. *Nat. Methods*, **17**, 515–523.
29. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
30. Picelli,S., Björklund,A.K., Reinius,B., Sagasser,S., Winberg,G. and Sandberg,R. (2014) Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.*, **24**, 2033–2040.
31. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006
32. Ramírez,F., Ryan,D.P., Grüning,B., Bhardwaj,V., Kilpert,F., Richter,A.S., Heyne,S., Dündar,F. and Manke,T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160–W165
33. Zhou,Y., Zhou,B., Pache,L., Chang,M., Khodabakhshi,A.H., Tanaseichuk,O., Benner,C. and Chanda,S.K. (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.*, **10**, 1523.
34. Jammula,S., Katz-Summercorn,A.C., Li,X., Linossi,C., Smyth,E., Killcoyne,S., Biasci,D., Subash,V.V., Abbas,S., Blasko,A., *et al.* (2020) Identification of subtypes of Barrett's esophagus and esophageal adenocarcinoma based on DNA methylation profiles and integration of transcriptome and genome data. *Gastroenterology*, **158**, 1682–1697.
35. Ahmed,I., Yang,S.H., Ogden,S., Zhang,W., Li,Y. and OCCAMS ConsortiumOCCAMS Consortium and Sharrocks,A.D. (2023) eRNA profiling uncovers the enhancer landscape of oesophageal adenocarcinoma and reveals new deregulated pathways. *eLife*, **12**, e80840
36. Wang,X., He,L., Goggin,S.M., Saadat,A., Wang,L., Sinnott-Armstrong,N., Claussnitzer,M. and Kellis,M. (2018) High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun.*, **9**, 5380.
37. Andersson,R. and Sandelin,A. (2020) Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.*, **21**, 71–87.
38. Hao,S., Meng,Q., Sun,H., Li,Y., Li,Y., Gu,L., Liu,B., Zhang,Y., Zhou,H., Xu,Z., *et al.* (2022) The role of transketolase in human cancer progression and therapy. *Biomed. Pharmacother.*, **154**, 113607.
39. Chen,H., Li,C., Peng,X., Zhou,Z., Weinstein,J.N. and Cancer Genome Atlas Research NetworkCancer Genome Atlas Research Network and Liang,H. (2018) A pan-cancer analysis of enhancer expression in nearly 9000 patient samples. *Cell*, **173**, 386–399.
40. Chen,L., Huang,M., Plummer,J., Pan,J., Jiang,Y.Y., Yang,Q., Silva,T.C., Gull,N., Chen,S., Ding,L.W., *et al.* (2020) Master transcription factors form interconnected circuitry and orchestrate transcriptional networks in oesophageal adenocarcinoma. *Gut*, **69**:630–640
41. The Cancer Genome Atlas Research Network. (2017) Integrated genomic characterization of oesophageal carcinoma. *Nature*, **541**, 169–175
42. Bang,Y.J., Van Cutsem,E., Feyereislova,A., Chung,H.C., Shen,L., Sawaki,A., Lordick,F., Ohtsu,A., Omuro,Y., Satoh,T., *et al.* (2010) Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. *Lancet*, **376**, 687–697.
43. Zhao,S., Allis,C.D. and Wang,G.G. (2021) The language of chromatin modification in human cancers. *Nat. Rev. Cancer*, **21**, 413–430.
44. Sheng,T., Ho,S.W.T., Ooi,W.F., Xu,C., Xing,M., Padmanabhan,N., Huang,K.K., Ma,L., Ray,M., Guo,Y.A., *et al.* (2021) Integrative epigenomic and high-throughput functional enhancer profiling reveals determinants of enhancer heterogeneity in gastric cancer. *Genome Med*, **13**, 158.

Received: June 12, 2024. **Revised:** August 9, 2024. **Editorial Decision:** September 24, 2024. **Accepted:** October 9, 2024