7. APPENDICES

7.1. Appendix I. Supplementary tables.

Supplementary Table 1*. Risk factors associated with oesophageal squamous cell carcinoma and adenocarcinoma.* GERD: Gastroesophageal Reflux Disease; +: associated risk; -: no risk associated. Retrieved from Domper Arnal *et al.*, 2015.

Risk factor	Oesophageal squamous cell carcinoma	Oesophageal adenocarcinoma
Geography	South-eastern Africa, Asia, Iran, South America.	Western Europe, North America, Australia.
Race	Black > White	White > Black
Gender	Male > Female	Male > Female
Alcohol	++++	-
Tobacco	++++	++
Obesity	-	+++
GERD	-	++++
Diet low in fruits and vegetables	++	+
Socioeconomic conditions	++	-
Genetic aspects	++	+

Supplementary Table 2. *Comparison of the advantages and limitations of conventional tissue biopsy and liquid biopsy.* CTCs: Circulating Tumour Cells. Adapted from Poulet *et al.*, 2019.

Conventional tissue biopsy	Liquid biopsy
Gold standard.	Clinical interest under investigation.
Accessible to histological analysis and staging.	The possibility of a histological analysis is limited to obtention of CTCs.
Sometimes unavailable.	Easy to obtain & faster turn-around time. Low level of tumour-derived products in body fluids, increasing risk of false negative results.
Invasive procedure, discomfort for patients.	Minimally invasive.
Potential high yield of DNA but risk of DNA degradation/cross-link. DNA quantity highly variable with sampling method.	Quantity and quality of DNA strongly dependent on pre-analytical and analytical processes.
Localised analysis, no characterization of intra- or inter-tumour heterogeneity (metastasis), especially in advanced stages.	Allows, if enough DNA is available, to highlight both intra- and inter-tumour heterogeneity.
Not applicable to serial monitoring. No possibility of dynamic follow-up of cancer molecular modifications.	Applicable to serial monitoring. Dynamic follow-up of tumour evolution.

***Supplementary Table 3**. Main reasons why candidate genes of the initial common gene selection were discarded.* This initial selection comprises overexpressed genes in tumour tissue of all 4 cancer types assessed (LADC, LSqCC, HNSqCC and OC) with a $\log_2$FC > 1.5. In the case of low expression in cancer types of interest, < 25 transcripts per million (tpm) was set as threshold. For high expression in normal tissues a > 50 tpm cutoff was set and in the case of expression in white blood cells (WBCs), > 15 tpm was used as threshold. Primer availability was not checked (-) for genes failing to meet two or more of the other selection criteria. The criteria used to establish primer availability are listed in section 3.5. IS: immune system; WBCs: white blood cells.

Discarded gene	Low expression in cancer types of interest	Not specific for cancer types of interest	High expression in normal tissues	Expression in WBCs	IS gene	Primers availability
ALG1L	X	X				-
ANLN	X	X		X		-
AURKA	X	X		X		-
CA9	X		X			-
CDKN2A		X				X
CEP55	X	X		X		-
COL1A1		X	X			-
CXCL10	X	X				-
DSG2		X		X		-
DTL	X	X		X		-
FOXM1		X		X		-
IGF2BP3	X			X		-
IGFBP3		X	X			-
IGHG1					X	-
IGHG2		X			X	-
IGHG3		X			X	-
IGHG4		X			X	-
IGHV1-69-2					X	-
IGHV4-34		X			X	-
ITGB4		X	X			-
MARCKSL1		X	X	X		-
MCM2	X	X		X		-
MMP7	X		X			-
MMP9		X	X			-
MYBL2		X		X		-
PLK1	X	X		X		-
SPP1		X	X			-
SULF1		X	X			-
TOP2A		X		X		-
TPX2	X	X		X		-
TRIP13	X	X		X		-
UBE2T		X		X		-

***Supplementary Table 4**. Candidate genes' official symbol, name and forward and reverse primers for their amplification.* Primer design details are explained in section 3.5. Official gene full names retrieved from HUGO (https://www.genenames.org).

Official gene symbol	Official gene full name	Forward primer	Reverse primer
AGR2	Anterior gradient 2, protein disulphide isomerase family member	AAGGCAGGTGGGTGAGGAAATC	TGGGTCGAGAGTCCTTTGTGT
CDH3	Cadherin 3	GGGAGCCTGTGTGTGTCTAC	GTCTCTCAGGATGCGGTAGC
CEACAM6	Carcinoembryonic antigen cell adhesion molecule 6	ACTCAGCGTCAAAAGGAACG	GACGGTAATTGGCCTTTGAG
COL10A1	Collagen type X alpha 1 chain	AAAGGCCCACTACCCAACAC	GTGGACCAGGAGTACCTTGC
CST1	Cystatin-SN	CCCGGGTGGCATCTATAACG	GGTCTGTTGCCTGGCTCTTA
CTHRC1	Collagen triple helix repeat containing 1	GATCCCCAAGGGGAAGCAAA	GGCCCTTGTAAGCACATTCC
CXCL9	CXC motif chemokine ligand 9	GTGCAAGGAACCCCAGTAGT	GGTGGATAGTCCCTTGGTTGG
CXCL13	CXC motif chemokine ligand 13	CAGCCTCTCTCCAGTCCAAG	ATTCCCACGGGGCAAGATTT
CYP2S1	Cytochrome P450 family 2 subfamily S member 1	GGCTATACCCTCCTGCTCCT	CTCCCGATTCAGCTCCTCAC
EPCAM	Epithelial cell adhesion molecule	TACAAGCTGGCCGTAAACTG	GCCAGCTTTGAGCAAATGAC
HAS3	Hyaluronan synthase 3	ATCCCCAAGTAGGGGGAGTC	CAGCCAAAGTAGGACTGGCA
KRT16	Keratin 16	ACGAGCAGATGGCAGAGAAAAA	GCTGCTCTGTACCAGTTCGC
KRT17	Keratin 17	AATCCTGCTGGATGTGAAGACG	GTACTGAGTCAGGTGGGCATC
LAMB3	Laminin subunit beta 3	CTTCTACAACAACCGGCCCT	CAAACACAGCGGGGTCAAAG
LAMC2	Laminin subunit gamma 2	GGAGCTGGAGTTTGACACGA	CAGCGTTCTTGGCTCTGGTA
MALAT1	Metastasis associated lung adenocarcinoma transcript 1	CTGGGGCTCAGTTGCGTAAT	CTCACAAAACCCCCGGAACT
MMP1	Matrix metallopeptidase 1	AGAGCAGATGTGGACCATGC	TTGTCCCGATGATCTCCCCT
MMP10	Matrix metallopeptidase 10	AGTTTGGCTCATGCCTACCC	CAGGGAGTGGCCAAGTTCAT
MMP11	Matrix metallopeptidase 11	AAGAGGTTCGTGCTTTCTGG	ATCGCTCCATACCTTTAGGG
MMP12	Matrix metallopeptidase 12	TTTGGTGGTTTTTGCCCGTG	TCGAAATGTGCATCCCCTCC
NTS	Neurotensin	GCAGGGCTTTTCAACACTGG	TCATACAGCTGCCGTTTCAGA
PHC3	Polyhomeotic homolog 3	GCTGCTGTTCAGGCAAGTTT	GAAGCCTGGGAACGGCTTAT

Official gene symbol	Official gene full name	Forward primer	Reverse primer
PLEC	Plectin	ACCAAGTGGGTCAACAAGCA	CCAGCAGGGAGATGAGGTTG
PTHLH	Parathyroid hormone like hormone	GGAGACTGGTTCAGCAGTGG	CCCTTGTCATGGAGGAGCTG
SLC2A1	Solute carrier family 2 member 1	TGGCATCAACGCTGTCTTCT	AGCCAATGGTGGCATACACA
SMIM22	Small integral membrane protein 22	CCCCAGGAAGGAAAGACCCA	CAGACGGGGACTGGAAGACA
SOX2	SRY-box transcription factor 2	AGGATAAGTACACGCTGCCC	TAACTGTCCATGCGCTGGTT
TP63	Tumour protein p63	CTGCCCTGACCCTTACATCC	TGGGACATGGTGGATCGGTA
UBD	Ubiquitin D	AGATGGCTCCCAATGCTTC	TCACGCTGTCATATGGGTTG
UBE2C	Ubiquitin conjugating enzyme E2C	TTCCTGTCTCTCTGCCAACG	CTCCTGCTGTAGCCTTTTGC
WFDC2	WAP four-disulfide core domain 2	CCCTAGTCTCAGGCACAGGA	CTGTCCGAGACGCACTCTTG

Supplementary Table 5*. Common selection genes' average expression data in LADC, LSqCC, HNSqCC and OC cell lines.* The number of cell lines for which candidate genes' expression data was available is also indicated. Data retrieved from the Cancer Cell Line Encyclopedia, EMBL-EBI (https://www.ebi.ac.uk/gxa). LADC: Lung Adenocarcinoma; LSqCC: Lung Squamous Cell Carcinoma; HNSqCC: Head and Neck Squamous Cell Carcinoma; OC: Oesophageal Carcinoma; tpm: transcripts per million.

Gene	LADC expression (tpm)	LADC cell lines	LSqCC expression (tpm)	LSqCC cell lines	HNSqCC expression (tpm)	HNSqCC cell lines	OC expression (tpm)	OC cell lines
COL10A1	0.51	51	0.30	20	0.17	10	0.27	21
CST1	27.40	43	39.60	13	0.26	8	0.39	12
CTHRC1	44.83	57	25.50	22	9.16	13	10.18	24
CXCL9	0.33	4	0.10	5	0.23	4	0.10	2
CXCL13	0.18	19	0.13	4	0.15	4	0.31	10
EPCAM	389.28	57	332.88	22	239.00	13	296.68	25
KRT17	172.05	57	340.58	22	2543.38	13	1241.20	25
LAMB3	220.80	57	193.28	22	405.92	13	286.04	25
MMP1	107.84	57	39.10	22	316.54	13	24.52	25
MMP11	2.23	57	1.85	22	0.96	13	2.86	25
MMP12	0.24	19	0.21	8	2.23	12	1.50	13
UBE2C	278.23	57	302.18	22	180.54	13	243.64	25

***Supplementary Table 6**. Selected cell lines' genetic alterations and cancer type of origin*. Data retrieved from ATCC (https://www.atcc.org) and DepMap (https://depmap.org). LADC: Lung Adenocarcinoma; LSqCC: Lung Squamous Cell Carcinoma; OADC: Oesophageal Adenocarcinoma; OSqCC: Oesophageal Squamous Cell Carcinoma; SqCC: Squamous Cell Carcinoma.

Cell line	Cancer type	Genetic alterations
A549	LADC	Mutated *CDKN2A* and *KRAS*.
NCI-H1395	LADC	Mutated *BRAF*.
NCI-H1975	LADC	Mutated *CDKN2A, EGFR, PIK3CA* and *TP53*.
NCI-H2228	LADC	*EML4-ALK* fusion.
NCI-H2170	LSqCC	Mutated *CDKN2A* and *TP53*.
HCC-95	LSqCC	*PIK3CA* amplification.
SW 900	LSqCC	Mutated *CDKN2A, KRAS* and *TP53*.
FaDu	Hypopharyngeal SqCC	Mutated *CDKN2A, SMAD4* and *TP53*.
HSC-2	Oral cavity SqCC	Mutated *CASP8, CDKN2A, PIK3CA, TP53* and *TP63*.
HSC-3	Tongue SqCC	Mutated *CASP8, CDKN2A, NOTCH1, TP53* and *SMAD4*.
OE19	OADC	Mutated *SMAD2* and *TP53*.
TE-1	OSqCC	Mutated *ERBB2, KRAS, SMAD4* and *TP53*.
KYSE-30	OSqCC	Mutated *CDKN2A* and *TP53*.

***Supplementary Table 7**. LADC candidate genes' expression data in selected LADC cell lines.* An average expression value was obtained using expression data of the genes of interest in several LADC cell lines. The number of cell lines for which candidate genes' expression data was available is also indicated. Data retrieved from the Cancer Cell Line Encyclopedia, EMBL-EBI (https://www.ebi.ac.uk/gxa). LADC: Lung Adenocarcinoma; tpm: transcripts per million.

Cell line	A549 expression (tpm)	NCI-H1395 expression (tpm)	NCI-H1975 expression (tpm)	NCI-H2228 expression (tpm)	Average expression in tpm (all LADC cell lines)	LADC cell lines
AGR2	156.0	2670.0	2.0	167.0	430.4	57
CEACAM6	11.0	393.0	0.5	97.0	575.1	57
SMIM22	0.1	73.0	2.0	17.0	24.2	53
UBD	0.3	-	-	0.3	8.9	41
WFDC2	0.3	1.0	2.0	83	45.4	55

***Supplementary Table 8**. LSqCC candidate genes' expression data in selected LSqCC cell lines.* An average expression value was obtained using expression data of the genes of interest in several LSqCC cell lines. The number of cell lines for which candidate genes' expression data was available is also indicated. Data retrieved from the Cancer Cell Line Encyclopedia, EMBL-EBI (https://www.ebi.ac.uk/gxa). LSqCC: Lung Squamous Cell Carcinoma; tpm: transcripts per million.

Cell line	NCI-H2170 expression (tpm)	HCC-95 expression (tpm)	SW 900 expression (tpm)	Average expression in tpm (all LSqCC cell lines)	LSqCC cell lines
HAS3	0.5	55.0	4.0	55.5	22
NTS	0.2	7.0	0.3	14.8	18
SOX2	50.0	229.0	25.0	136.5	19
TP63	-	483.0	3.0	61.2	18

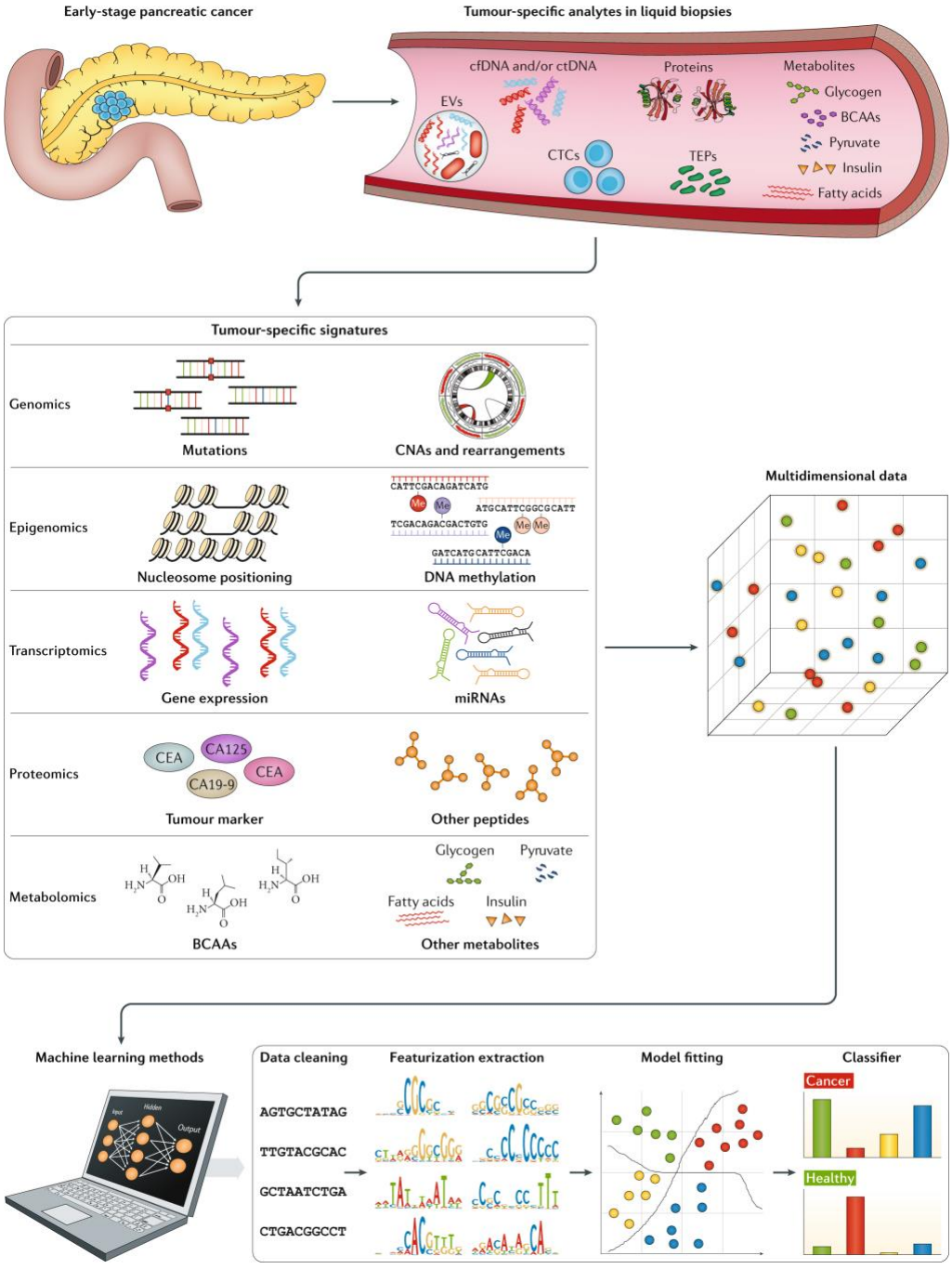
***Supplementary Table 9**. HNSqCC candidate genes' expression data in selected HNSqCC cell lines.* An average expression value was obtained using expression data of the genes of interest in several HNSqCC cell lines. The number of cell lines for which candidate genes' expression data was available is also indicated. Data retrieved from the Cancer Cell Line Encyclopedia, EMBL-EBI (https://www.ebi.ac.uk/gxa). HNSqCC: Head and Neck Squamous Cell Carcinoma; tpm: transcripts per million.

Cell line	FaDu expression (tpm)	HSC-2 expression (tpm)	HSC-3 expression (tpm)	Average expression in tpm (all HNSqCC cell lines)	HNSqCC cell lines
CDH3	148.0	115.0	154.0	267.6	13
KRT16	37.0	8.0	2.0	270.2	13
LAMC2	52.0	118.0	1245.0	513.5	13
MMP10	17.0	13.0	24.0	74.9	13
PI3	37.0	100.0	2.0	403.3	13
PTHLH	2.0	74.0	241.0	61.2	13
SLC2A1	331.0	145.0	186.0	196.5	13

***Supplementary Table 10**. OC candidate genes' expression data in selected OC cell lines.* An average expression value was obtained using expression data of the genes of interest in several OC cell lines. The number of cell lines for which candidate genes' expression data was available is also indicated. Data retrieved from the Cancer Cell Line Encyclopedia, EMBL-EBI (https://www.ebi.ac.uk/gxa). OC: Oesophageal Carcinoma; tpm: transcripts per million.

Cell line	OE19 expression (tpm)	TE-1 expression (tpm)	KYSE-30 expression (tpm)	Average expression in tpm (all OC cell lines)	OC cell lines
CYP2S1	201.0	13.0	36.0	45.8	25
MALAT1	166.0	113.0	590.0	242.7	25
PHC3	8.0	14.0	11.0	15.2	25
PLEC	113.0	309.0	260.0	132.3	25

7.2. Appendix II. Supplementary figures.



Supplementary Figure 1. *Combination strategies for early detection of cancer from liquid biopsy samples.* Various tumour specific circulating analytes yield different information about the genome (mutations, copy number alterations, etc.), the epigenome, the proteome, the transcriptome or the metabolome. This data is to be combined in innovative ways and used for machine learning purposes. The machine learning workflow comprises the four steps shown in the figure and allows for distinction between tumour and normal states. BCAAs: Branched-Chain Amino Acids; cfDNA: circulating free DNA; CNAs: Copy Number Alterations; CTCs: Circulating Tumour Cells; ctDNA: circulating tumour DNA; EVs: Extracellular Vesicles; TEPs: Tumour-Educated Platelets. Retrieved from Heitzer *et al.*, 2019.