1 **Gene expression profiling of The Cancer Genome**

2 **Atlas supports an inverse association between body**

3 **mass index (BMI) and major oesophageal tumour**

4 **subtypes**

5 Inge Seim^{1,2,3,4*}, Penny L. Jeffery^{1,2,3}, Lisa K. Chopin^{1,2,3}

6 ¹ Ghrelin Research Group, Translational Research Institute – Institute of Health and
7 Biomedical Innovation, School of Biomedical Sciences, Queensland University of
8 Technology, Brisbane, Queensland, Australia.

9 ² Comparative and Endocrine Biology Laboratory, Translational Research Institute –
10 Institute of Health and Biomedical Innovation, School of Biomedical Sciences,
11 Queensland University of Technology, Brisbane, Queensland, Australia.

12 ³ Australian Prostate Cancer Research Centre - Queensland, Institute of Health and
13 Biomedical Innovation, Queensland University of Technology, Princess Alexandra
14 Hospital, Translational Research Institute, Brisbane, Queensland, Australia.

15 ⁴ Integrative Biology Laboratory, College of Life Sciences, Nanjing Normal
16 University, Nanjing, Jiangsu, China.

17

18 *Corresponding author

19 E-mail: inge@seimlab.org (I.S.).

20

21

22 **Abstract**

23 In the last decade the Cancer Genome Atlas (TCGA) program has revealed
 24 significant insights into molecular events of dozens of cancers. These data sets are
 25 continuously updated, providing an unprecedented resource to the research
 26 community. There is now an emerging link between obesity and the development and
 27 progression of cancer. In this study we wished to identify genes related to body mass
 28 index (BMI) in TCGA datasets. Supporting epidemiological data, our gene expression
 29 profiling analyses suggest that oesophageal adenocarcinoma (EAC) can be considered
 30 a true obesity-associated cancer subtype, presenting avenues for prevention and
 31 treatment.

32

33

34

35 **Introduction**

36 Obesity is now considered to be an epidemic. Epidemiological data links
 37 overweight and obesity, dietary fats, and a sedentary lifestyle with the risk of
 38 developing many types of cancer [1], however, mechanistic data has been relatively
 39 sparse. In recent years, the development of powerful experimental methods (such as
 40 CRISPR [2]) and new technologies (such as RNA-sequencing [3]) have fuelled
 41 research aiming to better understand how obesity affects cells and tissues at the
 42 whole-body level. Genetic, epidemiological, and experimental data show that a high-
 43 fat diet and obesity can mediate tumour progression in diverse cancers [4, 5]. To
 44 contribute to this emerging research effort, we here interrogated The Cancer Genome
 45 Atlas (TCGA) gene expression data (RNA-sequencing), with the aim of identifying
 46 genes whose expression correlates with body mass index (BMI).

48 **Material and methods**

49 **Datasets**

50 Using the Genomics Data Commons (GDC) data portal [6, 7] we
 51 downloaded clinical, exposure (includes BMI), and gene expression (RNA-
 52 sequencing) data using the companion GDC Data Transfer Tool from the following
 53 12 The Cancer Genome Atlas (TCGA) datasets with corresponding BMI records:
 54 TCGA-BLCA (bladder urothelial carcinoma; $N=358$), TCGA-CESC (cervical
 55 squamous cell carcinoma and endocervical adenocarcinoma; $N=259$), TCGA-CHOL
 56 (cholangiocarcinoma; $N=35$), TCGA-COAD (colon adenocarcinoma; $N=231$),

TCGA-DLBC (lymphoid neoplasm diffuse large B-cell lymphoma; $N=48$), TCGA-ESCA (oesophageal carcinoma; $N=152$), TCGA-KIRP (kidney renal papillary cell carcinoma; $N=201$), TCGA-LIHC (liver hepatocellular carcinoma; $N=334$), TCGA-READ (rectum adenocarcinoma; $N=72$), TCGA-UCEC (uterine corpus endometrial carcinoma; $N=511$), TCGA-UCS (uterine carcinosarcoma; $N=51$), and TCGA-UVM (uveal melanoma; $N=53$). As we interrogated the entire gene set, we employed normalised gene expression values (FPKM-UQ; Fragments Per Kilobase of transcript per Million mapped reads upper quartile). FPKM-UQ is a modified FPKM calculation in which the total protein-coding read count is replaced by the 75th percentile read count value for the sample to scale the reads (see https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/). GDC data is annotated with Ensembl gene IDs. To obtain classical gene symbols and to classify genes into ‘gene types’ (e.g. protein-coding, lincRNA, antisense, and pseudogene), a gene annotation file (GTF; v22; GRCh38.p2) was downloaded from the GENCODE [8] web site (<https://www.encodegenes.org/releases/22.html>) and merged with our output data using a custom R script.

Identification of BMI-related genes

Gene expression data was parsed using custom bash and R scripts, to generate data frames that could be merged with the associated clinical and exposure GDC data files (includes BMI). BMI was calculated as weight (kg) divided by height (m^2). In each data set, only primary tumour samples were retained, as recommended [7]. Clinical information was manually interrogated to remove obvious outliers (e.g. extreme BMI values in the 132-415 range which could stem from inputting incorrect

81 patient height or the patient has an unspecified subnormal height condition) in the
82 case of TCGA-UZ-A9PU (TCGA-KIRP), TCGA-NH-A8F7 (TCGA-COAD), TCGA-
83 WN-AB4C (TCGA-KIRP), TCGA-SJ-A6ZI (TCGA-UCEC), and TCGA-DD-A4NR
84 (TCGA-LIHC). Additional TCGA-ESCA (oesophageal cancer) clinical information
85 (including histology) was obtained from Supplementary Table 1 of the associated
86 dataset manuscript [9].

87 To identify body mass index (BMI)-related genes Spearman's correlation
88 was used for examining the associations between BMI and gene expression values on
89 an ordinal scale [10, 11]. For each dataset, genes with no expression in $\geq 80\%$ of
90 tumours were excluded from analysis [11] and BMI was transformed on a log-scale.
91 A general rule of thumb is that one requires about 30 samples to assume a normal
92 distribution. A sample size of 30 would result in a Spearman correlation (ρ) cutoff
93 of 0.36, with ρ decreasing as sample size increases (see [12]). In our analysis we
94 employed a blanket ρ cutoff of 0.30 and retained genes with a false-discovery-
95 corrected P -value less than 0.01 (Bonferroni-corrected to account for a total of 60,483
96 genes tested). We also compared groups of cancer patients with a normal body mass
97 index value (lean; BMI <25) to overweight (BMI of 25-30) and obese (BMI >30)
98 patients using a non-parametric Mann-Whitney U -test (also known as Wilcoxon rank-
99 sum test; corrected for multiple-testing as above). Groups (e.g. cancer subtype) by
100 BMI category (e.g. lean versus high-BMI) were compared using a Pearson Chi-square
101 (χ^2) test. Scripts are available at GitHub (https://github.com/sciseim/BMI_MS).

102

103

104 **Survival outcome analysis**

105 Survival data was obtained from a recent comprehensive clinical endpoint
 106 outcome (OS, overall survival; PFI, progression-free interval; DFI, disease-free
 107 interval; DSS, disease-specific survival) curation and analysis effort of 33 TCGA-
 108 datasets [7]. Kaplan-Meier survival analysis [13] was performed with the R package
 109 ‘survival’ [14], fitting survival curves (*survfit*) and computing log-rank P -values using
 110 the *survdif* function, with $\rho=0$ (equivalent to the method employed by UCSC Xena;
 111 see <https://goo.gl/4knf62>). Survival curves were plotted when survival was
 112 significantly different between two groups (log-rank $P \leq 0.05$). Datasets with fewer
 113 than 10 events per group can be considered unreliable [7].

114

Results and discussion

As of May 2018, The Cancer Genome Atlas (TCGA) program had amassed clinical and molecular data of more than 30 different cancer types. Of these, 12 cancer types had associated body mass index (BMI) data (Table 1). We employed Spearman's correlation test to assess the correlation between gene expression (RNA-sequencing data) of protein- and non-protein coding genes and BMI (cut-off set at an absolute ρ of 0.30 and a multiple testing adjusted P -value, Q , less than 0.01). Using this threshold, we observed significant correlation between gene expression and BMI in four TCGA datasets (summarised in Table 1): TCGA-CESC (cervical squamous cell carcinoma and endocervical adenocarcinoma; 1 gene with negative correlation), TCGA-LIHC (liver hepatocellular carcinoma; 4 positive and 3 negative), TCGA-UCEC (uterine corpus endometrial carcinoma; 6 positive), and TCGA-ESCA (oesophageal carcinoma; 771 positive and 583 negative) (Table 1). Please see Table 2 and S1-S3 Tables for details.

Table 1. BMI-related genes identified by Spearman's correlation test.

Abbreviations: BMI, body mass index; N_{lean} , the number of patients with a BMI below 25; $N_{\text{overweight}}$, the number of patients with a BMI between 25 and 30; N_{obese} , the number of patients with a BMI above 30; N_{combined} , the number of genes identified by Spearman correlation at a ρ cut-off of 0.30 and a multiple-testing adjusted P -value (Bonferroni) less than 0.01; Up, the number of genes with a positive correlation with BMI; Down, the number of genes with a negative correlation with BMI. BMI distribution is shown as mean (lower quartile-upper quartile). The Cancer Genome Atlas (TCGA) datasets are TCGA-BLCA (bladder urothelial carcinoma), TCGA-CESC (cervical squamous cell carcinoma and endocervical adenocarcinoma), TCGA-CHOL (cholangiocarcinoma), TCGA-COAD (colon adenocarcinoma), TCGA-DLBC (lymphoid neoplasm diffuse large B-cell lymphoma), TCGA-ESCA (oesophageal carcinoma), TCGA-KIRP (kidney renal papillary cell carcinoma), TCGA-LIHC (liver hepatocellular carcinoma), TCGA-READ (rectum adenocarcinoma), TCGA-UCEC (uterine corpus endometrial carcinoma), TCGA-UCS (uterine carcinosarcoma), and TCGA-UVM (uveal melanoma).

Dataset	BMI distribution (kg m^{-2})	N_{lean}	$N_{\text{overweight}}$	N_{obese}	N_{cor}		
					All	Up	Down
TCGA-BLCA	26.99 (23.20-29.63)	149	124	85	0	0	0
TCGA-CESC	27.95 (22.62-32.41)	98	75	86	1	0	1

TCGA- CHOL	27.99 (24.82-30.22)	10	16	9	0	0	0
TCGA- COAD	28.37 (23.85-31.83)	78	80	73	0	0	0
TCGA- DLBC	25.91 (22.01-28.37)	25	15	8	0	0	0
TCGA- ESCA	25.07 (21.30-27.47)	84	46	22	1294	711	583
TCGA- KIRP	28.92 (25.26-32.34)	50	86	74	0	0	0
TCGA- LIHC	25.82 (21.76-28.66)	177	90	67	4	2	2
TCGA- READ	27.12 (24.60-29.15)	20	37	15	0	0	0
TCGA- UCEC	33.43 (26.37-38.83)	95	114	302	3	3	0
TCGA- UCS	29.71 (23.95-32.86)	17	14	20	0	0	0

TCGA- UVM	28.68 (23.32-30.32)	17	22	14	0	0	0
--------------	----------------------------	----	----	----	---	---	---

147

148

149 **Table 2. BMI-related genes identified by Spearman correlation test in TCGA-**
150 **UCEC (uterine corpus endometrial carcinoma).** Cutoff set at a Spearman ρ of
151 0.30 and a multiple-testing corrected P -value less than 0.01 (Bonferroni-adjusted P ,
152 $BH-FDR$).

Ensembl gene ID	gene symbol (NCBI)	description	chr	gene type	ρ	P	BH- FDR
ENSG00 0000785 49.13	<i>ADCYAP1R1</i>	ADCYAP receptor type I	7	protein _coding	0.3 0	8.98×10^{-12}	5.43×10^{-7}
ENSG00 0000821 75.13	<i>PGR</i>	progesteron e receptor	11	protein _coding	0.3 0	9.44×10^{-12}	5.71×10^{-7}
ENSG00 0001014 93.9	<i>ZNF516</i>	zinc finger protein 516	18	protein _coding	0.3 1	3.50×10^{-12}	2.12×10^{-7}
ENSG00 0001962 08.12	<i>GREB1</i>	growth regulating estrogen receptor binding 1	2	protein _coding	0.3 0	1.67×10^{-11}	1.01×10^{-6}
ENSG00 0002041 21.2	<i>ECELIP1</i>	endothelin converting enzyme like 1 pseudogene 1	2	unproce ssed_ps eudoge ne	0.3 1	1.82×10^{-12}	1.10×10^{-7}
ENSG00 0002442	<i>ECELIP2</i>	endothelin converting	2	transcri bed_un	0.3 1	1.51×10^{-12}	9.10×10^{-8}

80.1		enzyme like 1 pseudogene 2		process ed_pse udogen e			
------	--	-------------------------------------	--	----------------------------------	--	--	--

153

154

155 In a previous study which employed an earlier iteration of the TCGA-UCEC
156 endometrial carcinoma dataset, analysis was limited to endometroid tumours and
157 linear regression (with no r^2 cut-off) was performed on protein-coding genes and
158 BMI. Comparing 185 obese (BMI >30) and 105 non-obese patients, 214 genes were
159 associated with body mass index. In our analysis the strength of the relationship
160 between a gene's expression and BMI was quantified using Spearman correlation
161 analysis. Interrogating a total of 40,669 non-coding and 19,814 protein-coding genes
162 in 302 obese and 213 non-obese TCGA-UCEC patients we identified four protein-
163 coding genes and two non-coding genes with significant associations between
164 expression and BMI. Two protein-coding genes identified are critical regulators of
165 hormone-dependent tumour growth and also identified in the previous study [15]:
166 growth regulating oestrogen receptor binding 1 (*GREB1*) and progesterone receptor
167 (*PGR*) were both positively associated (Spearman $r=0.30$, $Q=5.71 \times 10^{-7}$ and $r=0.30$,
168 $Q=1.01 \times 10^{-6}$, respectively) with body mass index (Table 2). A very recent study on
169 an independent clinical data set revealed that PGR expression is indeed correlated
170 with BMI in endometrial cancer [16, 17]. GREB1 is essential for progesterone-driven
171 endometrial decidualisation [18]. Hormone-targeted therapies are a common approach
172 for advanced endometrial cancer and are more effective in patients with progesterone
173 receptor positive tumours [19]. These data suggest that *PGR* and *GREB1* should be
174 further studied to understand the link between endometrial cancer and increasing rates
175 of obesity worldwide [20].

176 Because a large number of TCGA-ESCA (oesophageal carcinoma) genes
177 (1,294 of a total of 60,483 genes; 2.14%) (Table 1) correlated with BMI, we
178 performed additional analyses on this dataset (see S1 Fig.). First, we repeated the
179 analyses with a stricter Spearman ρ (0.50) and adjusted P -value cut-off ($Q \leq 0.001$).

Under these conditions, 145 genes (0.24%) (S3 Table) were significant. Of these, 86 (59.3%) showed a positive correlation with BMI and 59 (46.7%) a negative correlation. These genes were also significant when BMI was dichotomised into lean and high-BMI (BMI above 25; overweight and obese) (Mann-Whitney *U*-test with $Q \leq 0.05$). Because GENCODE [8] annotations were employed, BMI-related genes included protein-coding genes (69 with a positive correlation with BMI; 45 with a negative correlation) as well as non-coding RNA genes (e.g. antisense lncRNAs). In an attempt to identify clinico-biologically meaningful groups, we performed unsupervised hierarchical clustering (Euclidian, average) of the expression of the 145 BMI-related genes in the TCGA-ESCA dataset. The resulting heat map revealed a striking difference between the two major subtypes of oesophageal carcinoma, ESCC (squamous cell carcinoma) and EAC (adenocarcinoma) (Fig. 1). EAC tumours largely clustered with high-BMI (overweight and obese). In agreement, stratifying EAC and ESCC into lean (20 EAC and 58 ESCC) and high-BMI (46 EAC and 16 ESCC) groups showed that the subtypes are over-represented in distinct BMI strata (Pearson χ^2 -test, $P=1.09 \times 10^{-8}$). Fig. 2 illustrates genes with distinct expression in the dataset: Elevated expression of p63 (*TP63*) [9, 21] and the long non-coding RNA *LINC01415* in ESCC; the converse for mucin 13 (*MUC13*) [22] and the long non-coding RNA *LINC00675* [23] in EAC.

201 **Fig. 1. Heat map of gene expression of 145 BMI-related genes in TCGA-ESCA**
 202 **(oesophageal carcinoma).** Normalised to depict relative values within rows
 203 (samples) with high (yellow) and low expression (blue). Various clinical
 204 stratifications are shown as colour bars on the left-hand side; negative (blue) and
 205 positive (red) correlation with BMI is shown as a bottom colour bar. Samples with no
 206 expression of one or more of the genes were excluded from the heatmap.

207
 208 **Fig. 2. Box plots showing the expression of selected BMI-associated genes in the**
 209 **TCGA-ESCA (oesophageal cancer) data set.** The *x*-axis represents samples grouped
 210 into lean (BMI <25; in green), overweight (BMI of 25-30; in orange), and obese (BMI
 211 >30; in red); the *y*-axis $\log_2(\text{FPKM-UQ})$ gene expression of the gene. The expression
 212 of patients with the oesophageal cancer subtypes and EAC (adenocarcinoma) and
 213 ESCC (squamous cell carcinoma) are shown as distinctly-coloured dots.

214

215

216 A natural next question to ask is what are the implications from this result?

217 We considered the current literature on the relationship between the oesophagus and

218 BMI, and oesophageal carcinoma and obesity. Recent comparative analysis

219 (employing Spearman correlation) on normal tissues revealed that oesophageal tissue

220 is one of six tissues (the other being adipose, nerve, pancreas, pituitary, and skin)

221 where gene expression is strongly associated with BMI [11]. Oesophageal carcinoma

222 is one of the most aggressive human cancers [24]. The TCGA-ESCA companion

223 manuscript argued that EAC (oesophageal adenocarcinoma) and ESCC (squamous

224 cell carcinoma) should be considered inherently distinct subtypes, developed by

225 lineage-specific alterations [9]. EAC and the gastric tumour subtype chromosomal

226 instability (CIN) are remarkably homogenous, both at the histopathological and

227 molecular level, while ESCC is more similar to squamous cell carcinoma derived

228 from other tissues [9]. Of the two subtypes, ESCC is more prevalent worldwide and

229 high BMI is associated with reduced risk of developing this disease [25, 26]; while

230 EAC is highly prevalent in populations where obesity is common [27]. There has been

231 a shift in the prevalence of EAC in the USA correlating with an increase in obesity

232 and sedentary lifestyle, with EAC increasing from 10% of cases in 1975 to 80% in

233 2014 [24]. Furthermore, multiple, large population-based (epidemiological studies)

234 studies show a strong correlation between increased BMI and the risk of developing

235 EAC [25, 27, 28]. EAC is prevalent in men, and potentially associated with masculine

236 fat distribution (primarily abdominal) and sex hormones [9, 27]. These

237 epidemiological observations fit well with our data. We also compared the prognosis

238 of the BMI groups using various endpoints, however, there were too few patients with

239 clinical endpoint events when patients were divided on the basis of histological

240 subtypes (S4 Table). Nevertheless, correlating TCGA-ESCA data with BMI identifies

a gene set which supports distinct diet-associated origins of oesophageal cancer subtypes.

Conclusions

In conclusion, by integrating gene expression and BMI data from the TCGA program, we have gained insights into the association between overweight and obesity and cancer. Integration of genomic and clinical datasets should be carefully performed to identify subtypes within each cancer, as recommended by TCGA [7]. Indeed, such efforts are likely to reveal further BMI-associated cancer subtypes. Nevertheless, when distinct subtypes associated with overweight and obesity are evident, we and others [15] have shown that correlating gene expression and BMI alone across an entire data set is informative. We propose that an obesogenic environment is essential for the development of oesophageal adenocarcinoma (EAC), but not oesophageal squamous cell carcinoma (ESCC) – adding further weight to distinct prevention and treatment strategies for these histological subtypes [9]. Future studies, integrating additional genomic datasets (such as epigenome data) and the release of updated clinical information by TCGA is likely to further define the link between a high BMI and cancer.

262 Funding sources

263 We acknowledge financial support from the National Health and Medical
264 Research Council Australia (grant no. 1059021; to I.S., P.L.J., L.K.C.), Cancer
265 Council Queensland (grant no. 1098565; to I.S. and L.K.C.), a QUT Vice-
266 Chancellor's Senior Research Fellowship (to I.S.), the Movember Foundation and the
267 Prostate Cancer Foundation of Australia through a Movember Revolutionary Team
268 Award, the Australian Government Department of Health, and the Australian Prostate
269 Cancer Research Center, Queensland (L.K.C.), and the Australian Research Council
270 (grant no DP140100249; to L.K.C.).

271

272 References

- 273 1. Wolin KY, Carson K, Colditz GA. Obesity and cancer. *Oncologist*.
274 2010;15(6):556-65. doi: 10.1634/theoncologist.2009-0285.
- 275 2. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for
276 genome engineering. *Cell*. 2014;157(6):1262-78. doi: 10.1016/j.cell.2014.05.010.
- 277 3. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for
278 transcriptomics. *Nat Rev Genet*. 2009;10(1):57-63. doi: 10.1038/nrg2484.
- 279 4. Chen M, Zhang J, Sampieri K, Clohessy JG, Mendez L, Gonzalez-Billalabeitia E,
280 et al. An aberrant SREBP-dependent lipogenic program promotes metastatic
281 prostate cancer. *Nat Genet*. 2018;50(2):206-18. doi: 10.1038/s41588-017-0027-2.
- 282 5. Maugham ML, Thomas PB, Crisp GJ, Philp LK, Shah ET, Herington AC, et al.
283 Insights from engraftable immunodeficient mouse models of hyperinsulinaemia.
284 *Sci Rep*. 2017;7(1):491. doi: 10.1038/s41598-017-00443-x.

- 285 6. Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI Genomic Data
286 Commons as an engine for precision medicine. *Blood*. 2017;130(4):453-9. doi:
287 10.1182/blood-2017-03-735654.
- 288 7. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al.
289 An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality
290 Survival Outcome Analytics. *Cell*. 2018;173(2):400-16 e11. doi:
291 10.1016/j.cell.2018.02.052.
- 292 8. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al.
293 GENCODE: the reference human genome annotation for The ENCODE Project.
294 *Genome Res*. 2012;22(9):1760-74. doi: 10.1101/gr.135350.111.
- 295 9. Cancer Genome Atlas Research N, Analysis Working Group: Asan U, Agency
296 BCC, Brigham, Women's H, Broad I, et al. Integrated genomic characterization of
297 oesophageal carcinoma. *Nature*. 2017;541(7636):169-75. doi:
298 10.1038/nature20805.
- 299 10. Rai MF, Sandell LJ, Cheverud JM, Brophy RH. Relationship of age and body
300 mass index to the expression of obesity and osteoarthritis-related genes in human
301 meniscus. *Int J Obes (Lond)*. 2013;37(9):1238-46. doi: 10.1038/ijo.2012.221.
- 302 11. Hao RH, Yang TL, Rong Y, Yao S, Dong SS, Chen H, et al. Gene expression
303 profiles indicate tissue-specific obesity regulation changes and strong obesity
304 relevant tissues. *Int J Obes (Lond)*. 2018;42(3):363-9. doi: 10.1038/ijo.2017.283.
- 305 12. Ryaboshapkina M, Hammar M. Human hepatic gene expression signature of
306 non-alcoholic fatty liver disease progression, a meta-analysis. *Sci Rep*.
307 2017;7(1):12361. doi: 10.1038/s41598-017-10930-w.

- 308 13. Rich JT, Neely JG, Paniello RC, Voelker CC, Nussenbaum B, Wang EW. A
309 practical guide to understanding Kaplan-Meier curves. *Otolaryngol Head Neck*
310 *Surg.* 2010;143(3):331-6. doi: 10.1016/j.otohns.2010.05.007.
- 311 14. Therneau T. A Package for Survival Analysis in S. version 2.38. 2015.
312 Reference Source. 2016.
- 313 15. Roque DR, Makowski L, Chen TH, Rashid N, Hayes DN, Bae-Jump V.
314 Association between differential gene expression and body mass index among
315 endometrial cancers from The Cancer Genome Atlas Project. *Gynecol Oncol.*
316 2016;142(2):317-22. doi: 10.1016/j.ygyno.2016.06.006.
- 317 16. Peevey JF, Seagle BL, Maniar KP, Kim JJ. Association of body mass index
318 with ER, PR and 14-3-3sigma expression in tumor and stroma of type I and type II
319 endometrial carcinoma. *Oncotarget.* 2017;8(26):42548-59. doi:
320 10.18632/oncotarget.17209.
- 321 17. Mauland KK, Trovik J, Wik E, Raeder MB, Njolstad TS, Stefansson IM, et al.
322 High BMI is significantly associated with positive progesterone receptor status and
323 clinico-pathological markers for non-aggressive disease in endometrial cancer. *Br J*
324 *Cancer.* 2011;104(6):921-6. doi: 10.1038/bjc.2011.46.
- 325 18. Camden AJ, Szwarc MM, Chadchan SB, DeMayo FJ, O'Malley BW, Lydon
326 JP, et al. Growth regulation by estrogen in breast cancer 1 (GREB1) is a novel
327 progesterone-responsive gene required for human endometrial stromal
328 decidualization. *Mol Hum Reprod.* 2017;23(9):646-53. doi:
329 10.1093/molehr/gax045.
- 330 19. Ethier J-L, Desautels DN, Amir E, MacKay H. Is hormonal therapy effective
331 in advanced endometrial cancer? A systematic review and meta-analysis.
332 *Gynecologic oncology.* 2017;147(1):158-66.

- 333 20. Lortet-Tieulent J, Ferlay J, Bray F, Jemal A. International Patterns and Trends
334 in Endometrial Cancer Incidence, 1978-2013. *J Natl Cancer Inst.* 2018;110(4):354-
335 61. doi: 10.1093/jnci/djx214.
- 336 21. Hu H, Xia SH, Li AD, Xu X, Cai Y, Han YL, et al. Elevated expression of p63
337 protein in human esophageal squamous cell carcinomas. *Int J Cancer.*
338 2002;102(6):580-3. doi: 10.1002/ijc.10739.
- 339 22. Williams SJ, Wreschner DH, Tran M, Eyre HJ, Sutherland GR, McGuckin
340 MA. Muc13, a novel human cell surface mucin expressed by epithelial and
341 hemopoietic cells. *J Biol Chem.* 2001;276(21):18327-36. doi:
342 10.1074/jbc.M008850200.
- 343 23. Li DD, Fu ZQ, Lin Q, Zhou Y, Zhou QB, Li ZH, et al. Linc00675 is a novel
344 marker of short survival and recurrence in patients with pancreatic ductal
345 adenocarcinoma. *World J Gastroenterol.* 2015;21(31):9348-57. doi:
346 10.3748/wjg.v21.i31.9348.
- 347 24. Napier KJ, Scheerer M, Misra S. Esophageal cancer: A Review of
348 epidemiology, pathogenesis, staging workup and treatment modalities. *World J*
349 *Gastrointest Oncol.* 2014;6(5):112-20. doi: 10.4251/wjgo.v6.i5.112.
- 350 25. Enzinger PC, Mayer RJ. Esophageal cancer. *N Engl J Med.*
351 2003;349(23):2241-52. doi: 10.1056/NEJMra035010.
- 352 26. Smith M, Zhou M, Whitlock G, Yang G, Offer A, Hui G, et al. Esophageal
353 cancer and body mass index: results from a prospective study of 220,000 men in
354 China and a meta-analysis of published studies. *Int J Cancer.* 2008;122(7):1604-10.
355 doi: 10.1002/ijc.23198.
- 356 27. Lagergren J. Influence of obesity on the risk of esophageal disorders. *Nat Rev*
357 *Gastroenterol Hepatol.* 2011;8(6):340-7. doi: 10.1038/nrgastro.2011.73.

358 28. Fang X, Wei J, He X, Lian J, Han D, An P, et al. Quantitative association
359 between body mass index and the risk of cancer: A global Meta-analysis of
360 prospective cohort studies. International journal of cancer. 2018.

361

362 **Supporting information**

363 **S1 Fig. Overview of body mass index (BMI) distribution in the TCGA-ESCA**
364 **(oesophageal cancer) data set.** Lean (BMI <25; in green), overweight (BMI of 25-
365 30; in orange), and obese (BMI >30; in red) patients are shown.

366

367 **S1 Table. BMI-related genes identified by Spearman correlation test in TCGA-**
368 **CESC (cervical squamous cell carcinoma and endocervical adenocarcinoma).**
369 Cutoff set at ρ of 0.30 and a multiple-testing corrected P -value less than 0.01
370 (Bonferroni-adjusted P).

371

372 **S2 Table. BMI-related genes identified by Spearman correlation test in TCGA-**
373 **LIHC (liver hepatocellular carcinoma).** Cutoff set at ρ of 0.30 and a multiple-
374 testing corrected P -value less than 0.01 (Bonferroni-adjusted P).

375

376 **S3 Table. BMI-related genes identified by Spearman correlation test in TCGA-**
377 **ESCA (esophageal carcinoma).** Cutoff set at ρ of 0.50 and a multiple-testing
378 corrected P -value less than 0.001 (Bonferroni-adjusted P). The following genes have

379 a different symbol in GENCODE: *AFDN-DT* (*MLLT4-AS1*), *PHETA1* (*FAM109A*),
 380 and *TMEM238L* (*LINC00675*).

381

382 **S4 Table. Survival outcome analysis of TCGA-ESCA subsets.** BMI denotes body
 383 mass index; EAC, oesophageal adenocarcinoma; ESCC, oesophageal squamous cell
 384 carcinoma; *P*, log-rank *P*-value; OS, overall survival; PFI progression-free interval;
 385 DFI, disease-free interval; DSS, disease-specific survival.

TCGA-ESCA N=142

A

LINC00675 (TMEM238L)

B

MUC13

C

LINC01415

D

TP63