G-quadruplex formation in *GCK* and *TM6SF2* are targets for differential DNA methylation in metabolic dysfunction-associated fatty liver disease and type II diabetes mellitus patients

Angelika Lahnsteiner[1, 2*], Victoria Ellmer[1], Anna Oberlercher[1], Zita Lutkeviciute[1], Esther Schönauer[3], Bernhard Paulweber[4], Elmar Aigner[4,5], Angela Risch[1,2]

[1]Division of Cancer (Epi-)Genetics, Department of Biosciences and Medical Biology, Center for Tumor Biology and Immunology (CTBI), Paris Lodron University Salzburg, Austria

[2]Cancer Cluster Salzburg, Austria

[3]Division of Structural Biology, Department of Biosciences and Medical Biology, Paris Lodron University Salzburg, Austria

[4]First Department of Medicine, University Clinic Salzburg, Austria

[5]Paracelsus Medical University Salzburg, Austria

*Corresponding author

angelika.lahnsteiner@plus.ac.at

University of Salzburg

Hellbrunnerstraße 34,

5020 Salzburg

Austria

Abstract

Background. The alarming increase in global rates of metabolic diseases (MetDs) and their association with an increased cancer risk renders them a considerable burden on our society. The interplay of environmental and genetic factors in causing MetDs may be reflected in DNA methylation patterns, particularly in non-canonical (non-B) DNA structures, such as G-quadruplexes or R-loops. To gain insight into the mechanisms of MetD progression, we focused on DNA methylation and functional analyses on intergenic regions of two MetD risk genes, glucokinase (*GCK)* exon 7 and transmembrane 6 superfamily 2 (*TM6SF2)* intron 2-exon 3 boundary, which are overlapped by long non-coding RNAs.

Results. Pyrosequencing of 148 blood samples from a nested cohort study revealed significant differential methylation in *GCK* and *TM6SF2* in MetD patients versus healthy controls. DNA methylation and gene expression data from The Cancer Genome Atlas (TCGA) for liver tumor versus normal tissue confirmed these observations. Detailed analysis including histone marks, chromatin conformation capture data, and luciferase reporter assays, highlighted the cell-type specific regulatory function of the target regions. The newly established method permanganate/S1 nuclease footprinting with direct adapter ligation (PDAL-Seq), as well as standard methods such as native DNA gel electrophoresis and circular dichroism (CD) spectrophotometry confirmed the formation of G4 structures in these regions. Based on our analysis, we propose that the elevated blood glucose and fatty acid levels as observed in MetD patients could lead to reformation or topological changes in the non-B DNA landscape causing differential methylation and altered transcription factor binding.

Conclusion. Our analyses provide a completely new view on the mechanisms underlying MetDs and their link to cancer, unveiling non-B DNA structures as novel targets for therapeutic interventions.

Key words: Metabolic disease, DNA methylation, permanganate/S1 nuclease footprinting, PDAL-Seq, non-B DNA structures, cancer,

50 **1. Introduction**

51 Overweight and obesity represent a major health issue of the 21st century. Their worldwide rate
52 tripled from 1975 to 2016 (1), affecting approximately 39% of the global population (2). Alarmingly,
53 childhood obesity also increased from 0.7-0.9% in 1975 to 5.6-7.8% in 2016 (3). Caloric imbalance
54 causing obesity is related to dietary habits, physical inactivity, stress, sleep patterns, social and
55 educational status, and endocrine disorders (4). Excessive fat accumulation increases the risk of
56 developing metabolic diseases (MetDs) such as the metabolic syndrome (MetS), type 2 diabetes
57 mellitus (T2DM), metabolic dysfunction-associated fatty liver disease (MAFLD) (5, 6), and
58 cardiovascular diseases (7). Furthermore, MetD patients have an elevated risk for gastric (8),
59 pancreatic (9), hepatocellular (10), colorectal (11), and lung cancer (12, 13).

60 MAFLD is the most common form of chronic liver disease in industrialized societies (14) with a
61 worldwide prevalence of 25% (15), and is strongly associated with T2DM (16). One hallmark of MAFLD
62 is the development of large cytoplasmic lipid droplets in hepatocytes with the accumulation of proteins
63 involved in lipid homeostasis on their surface. The key disease mechanism in MAFLD is very likely
64 related to peripheral insulin resistance (16, 17). In addition, fatty acid β-oxidation is impaired in MAFLD
65 owing to mitochondrial respiratory chain dysfunction, which, in combination with the accumulation of
66 incompletely metabolized fatty acids, contributes to oxidative and lipotoxic stress in hepatocytes. This
67 facilitates the progression from simple steatosis to advanced stages of MAFLD (18) which can further
68 progress to liver fibrosis, cirrhosis, and to hepatocellular carcinoma (HCC) (19).

69 Currently, we do not fully understand the molecular mechanisms leading to the development
70 and progression of MAFLD, but they clearly involve a complex interplay between genetic predisposition
71 and environmental factors influencing epigenetic processes, such as DNA methylation (5-
72 methylcytosine, 5mC). Studies investigating 5mC have identified differentially methylated positions
73 (DMPs) associated with hepatic fat or key signaling pathways regulating glycolysis and *de novo*
74 lipogenesis (20-24). Nevertheless, a mechanistic explanation of the impact of DMPs on gene
75 expression, is lacking. In particular, if these DMPs are located in intergenic regions or exons without
76 overlapping annotated regulatory elements, interpreting their biological functions remains difficult.

77 New insights arise from the observation that G-rich sequences with the ability to form non-
78 canonical (non-B) DNA structures called G-quadruplexes (G4s) are sites of methylation instability (25),
79 or actively influence DNA methylation (26). G4s are formed via the consensus motif $G_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}$
80 and can adopt intra- or intermolecular planar G4 tetrads held together by Hoogsteen hydrogen bonds
81 with a 1-7 bp long linker of any base (N) and stabilized by a monovalent cation *in vivo,* typically
82 potassium (27-29). Their manifold biological roles and occurrences have been extensively reviewed
83 (30-32). They often co-occur with RNA:DNA hybrids (R-loops) and stabilize each other (33). Recent
84 studies have shown that their simultaneous formation plays an important role in regulatory processes

85 (34-36). Interestingly, G4s have been identified as binding hubs for transcription factors (TFs) (37, 38),
86 and DNA methyltransferases (DNMTs) (26, 39). Therefore, changes in the G4 landscape could lead to
87 alterations in DNA methylation and transcription factor binding, causing aberrant gene expression
88 patterns. This could ultimately lead to MAFLD, which in turn supports the development of more severe
89 diseases such as cancer.

90 To investigate the link between differential DNA methylation and non-B DNA formation in
91 MetDs, we performed locus-specific 5mC analysis in the gene body of two key metabolic genes, exon
92 7 of glucokinase (*GCK*) and the intron 2-exon 3 boundary of transmembrane 6 superfamily 2 (*TM6SF2*)
93 using 148 blood samples from a well-characterized cohort including lean and obese healthy subjects,
94 lean and obese MAFLD, and patients suffering from advanced MetDs including uncontrolled T2DM and
95 severe MAFLD, collected at the University Hospital Salzburg, Austria. *GCK* and *TM6SF2* are not only
96 known risk genes (*GCK* (40-43), *TM6SF2* (44-48)), but are also targets for drug design (49). In this study,
97 we focused on intragenic regions overlapped by long non-coding RNAs, which potentially form R-loops
98 in these genes and enhance G4 formation. *GCK* encodes for glucokinase, which is expressed only in the
99 pancreas and in hepatocytes. It serves as the first rate-limiting enzyme of the glycolysis pathway by
100 converting glucose to glucose-6-phosphate in hepatocytes and as a sensor of blood glucose levels in
101 the pancreas (reviewed in (50)). *GCK* also interacts with the fatty acid metabolism, since glucose-6-
102 phosphate can in addition serve as a precursor for glycerol-3-phosphate, an essential component in
103 the synthesis of triglycerides. This connection means that glucokinase activity can indirectly influence
104 fatty acid metabolism by contributing to the production of glycerol-3-phosphate, thus affecting fat
105 storage and energy balance in the body (51). TM6SF2 plays a major role in lipid metabolism via
106 processing and secretion of triglycerides and very low-density lipoproteins (VLDL) in hepatocytes
107 (reviewed in (52)). Impaired TM6SF2 function leads to disruptions of the lipid export process, causing
108 lipid accumulation within hepatocytes and the development and progression of MAFLD (53).

109 In this study, luciferase reporter assays, analysis of histone marks, and chromosome
110 conformation capture data (HiC) revealed the important regulatory roles of the intergenic *GCK* and
111 *TM6SF2* regions. Furthermore, using an updated version of permanganate/S1 nuclease footprinting
112 with direct adapter ligation (PDAL-Seq) (54), native polyacrylamide (PAA) gel electrophoresis and
113 circular dichroism spectrophotometry (CD), we derived information about G-quadruplex formation in
114 these two regions. In summary, our observations elucidate not only the dynamics of 5mC in the gene
115 bodies of these two MetD risk genes in our patient cohort, but also result in a novel perspective of
116 non-B DNA involvement in the development of MAFLD. Furthermore, these new insights could help
117 explain the elevated cancer risk observed in MetD patients, emphasizing the critical need to
118 understand the interplay between DNA methylation and non-B DNA structures.
119

120 ## 2. Results

121 **Cohort demographics and description**

122 To investigate the relationship between DNA methylation and G4 formation in the selected
123 genes, we performed locus-specific 5mC analysis on the Qiagen Pyromark Q24 Platform for 148 blood
124 donors which were classified as lean-healthy (LH), obese-healthy (OH), lean-MAFLD (LM), obese-
125 MAFLD (OM), and advanced MetD (aMetD) according to their metabolic parameters, as described in
126 the *Methods* section. The mean values of the clinical parameters are provided in Table 1 and
127 Supplement Fig. 1.

128 The age of our study cohort ranged from 42-77 years, albeit the OM and aMetD groups showed
129 significantly elevated age compared to the LH group (LH vs. OM p=.0152, LH vs aMetD p= 1.53e-05,
130 Student´s t-test, and Supplement Table 1). This trend was not unexpected since MAFLD and
131 uncontrolled T2DM, which are represented in the aMetD group, occur at a higher frequency in the
132 elderly (aging-related diseases) (55). Fasting blood glucose, insulin, and aspartate transferase (AST)
133 levels were significantly increased in OH, OM and aMetD, but not in LM compared to LH (Students t-
134 test, Supplement Table 1). No significant change was detected in total cholesterol levels in any group
135 compared to healthy controls. High-density lipoprotein (HDL) was significantly decreased in OM and
136 aMetD, and low-density lipoprotein (LDL) was only decreased in aMetD, whereas total triglycerides
137 were increased in LM, OM, and aMetD. The detailed p-values are shown in the Supplement Table 1.

138 **Table 1. Study characteristics.** Data are expressed as means ± standard deviation (SD)
139

	lean-healthy (LH) n=24		obese-healthy (OH) n=31		lean-MAFLD (LM) n=27		obese-MAFLD (OM) n=32		advanced MetD (aMetD) n=34	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Gender [f/m]	14/10 (58% f)		21/10 (68% f)		12/15 (44% f)		10/22 (31% f)		15/19 (44% f)	
Age [y]	52	8	53	6	54	4	57	3	63	8
BMI [kg/m2]	21.9	2.0	33.1	3.1	23.5	1.3	33.4	3.4	30.1	3.9
HbA_{1c} [%]	5.3	0.2	5.5	0.3	5.4	0.3	5.8	0.5	7.3	1.4
HOMA	1.4	0.5	2.4	1.1	1.6	0.6	4.9	3.2	11.0	15.3
FIB4	1.2	0.4	0.9	0.2	1.1	0.2	1.1	0.3	3.0	1.6
Fast. Glucose [mg/dL]	89.6	7.0	93.5	6.9	91.3	6.3	105.3	16.1	166.0	54.2
Fast. Glucose [mmol/L]	5.0	0.4	5.2	0.4	5.1	0.3	5.8	0.9	9.2	3.0
Fast. Insulin [mU/L]	6.2	2.4	10.5	4.6	7.2	2.8	18.4	9.3	29.3	47.8
GGT [U/L]	15.6	6.5	22.3	12.1	28.3	20.5	55.5	71.6	210.1	219.6
AST [U/L]	20.4	5.5	22.0	5.5	23.7	5.6	29.3	12.7	60.2	41.7
ALT [U/L]	18.8	6.9	24.2	10.1	22.8	9.1	39.9	26.1	76.4	48.0
HDL-C [mg/dL]	69.8	15.0	63.9	19.3	62.7	16.5	52.1	15.8	48.6	13.7
LDL-C [mg/dL]	149.6	32.5	144.0	35.1	152.9	27.8	157.6	44.1	120.7	45.8
TG [mg/dL]	83.6	21.7	92.9	27.7	121.4	62.6	178.3	139.8	198.9	98.7
TChol [mg/dL]	216.8	38.8	211.0	35.6	221.3	28.7	222.3	45.2	195.2	50.5

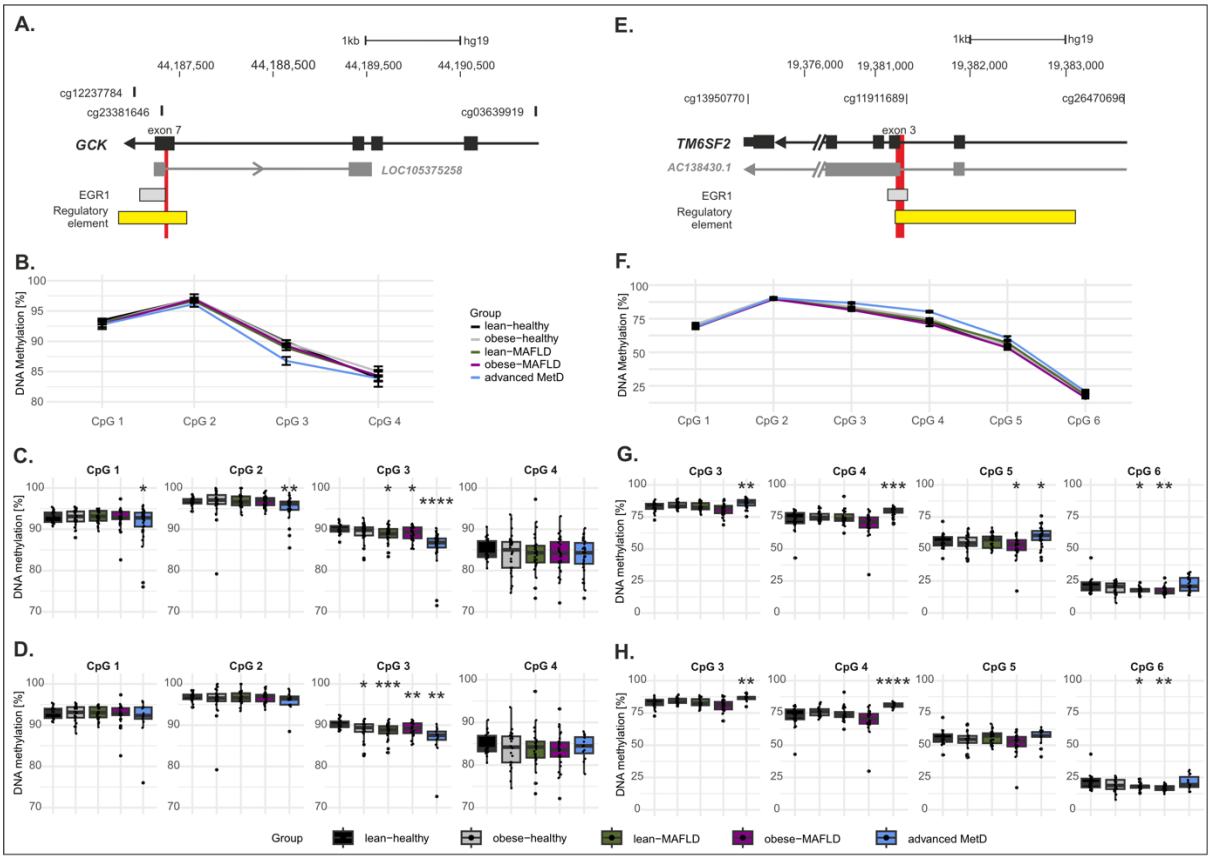
140
141 BMI = Body mass index; HOMA = homeostasis model assessment; HbA_{1c} = Glycated haemoglobin A_{1c}; FIB4 = Fibrosis4 score; Fast. Glucose = Fasting glucose; Fast. Insulin =
142 Fasting insulin; GGT = γ-glutamyl transpeptidase; AST = aspartate aminotransferase; ALT = alanine aminotransferase; HDL-c = High density lipoprotein cholesterol; LDL-c = Low
143 density lipoprotein cholesterol; TG = Triglycerides; TChol = total cholesterol;

144 ***GCK* exon 7 and the *TM6SF2* intron 2-exon 3 boundary are differentially methylated in MetD**
145 **patients**

146 *GCK* is encoded on the minus strand of chromosome 7 and consists of ten exons (Fig. 1A). 14
147 different transcripts are annotated for *GCK* in the ENSEMBL genome browser (56). Our region of
148 interest is located in exon 7 (chr7:44,187,339-44,187,380; hg19) and is overlapped by an antisense
149 lncRNA LOC105375258 (XR_927223.3) on the + strand. Furthermore, genome segmentation by
150 automating chromatin state discovery and characterization (ChromHMM) in the UCSC genome
151 browser (57) predicts a "weak enhancer or open chromatin cis regulatory element" in hepatoblastoma
152 cell line HepG2 and a "promoter region including TSS" in the lymphoblast cell line K562 (Fig. 1A, yellow
153 rectangle). In addition, an EGR1 TF binding site can be found in this region (Fig. 1A, grey rectangle) (57).
154 Significant hypomethylation was detected in CpG 1 and 2 in aMetD patients, and in CpG 3 in LM, OM
155 and aMetD compared to LH subjects. No changes were detected in CpG 4 (Fig. 1B). Given that
156 alterations in DNA methylation are associated with aging (58) and metabolic diseases occur at higher
157 frequency with increasing age (reviewed in (55)), we analyzed the magnitude of age-relation compared
158 to published aging markers located in *EDARADD* (cg09809672), *IPO8* (cg19722847), *NHLRC1*
159 (cg22736354), *P2RXL1* (cg05442902) and *SCGN* (cg06493994). These five markers were selected based
160 on their consistent identification in three distinct epigenetic clock studies utilizing Illumina Methylation
161 arrays (58-60). Serving as reference genes, these markers were employed to assess the presence of
162 age correlation with 5mC in our study cohort. Consistent with published values, we detected a
163 significant positive (*NHLRC1* and *SCGN*) and negative (*IPO8*, *EDARADD* and *P2RXL1*) association
164 between DNA methylation a chronological age (Supplement Fig. 2). Since we wanted to discern
165 whether the observed effects in the two target genes arise primarily from aging processes, disease-
166 related mechanisms, or a synergistic combination of both factors, we limited our analysis to 42–60-
167 year-old subjects. In this subset, we still observed a significant decrease in DNA methylation in all
168 groups compared to lean-healthy subjects at CpG 3, which is even already detectable in obese-healthy
169 and lean-MAFLD subjects (Fig. 1C). Apart from CpG 3, all other CpG positions did not show a significant
170 change in DNA methylation, indicating that aging also plays a role in this region.

171 To validate these data, we investigated 5mC of *GCK* in TCGA from HCC compared to normal
172 tissue (61). The nearest annotated CpG covered by the Illumina Methylation array is cg23381646
173 located 45 bp upstream of CpG 3. Furthermore, we looked at two flanking CpGs, cg12237784 located
174 294 bp upstream and cg03639919 is found 4,000 bp downstream of cg23381646 (Fig. 1A). Although
175 DNA methylation at cg23381646 and cg12237784 was not significantly different (Wilcoxon rank sum
176 test adjusted using Benjamini-Hochberg (BH) method, cg23381646: p=.58, cg12237784: p=.075),
177 cg03639919 located 4 kb downstream was significantly hypomethylated in tumor versus normal tissue

178 (Supplement Fig. 3A, BH adjusted *p*=2.2E-16). Especially cg23381646 and cg03639919 appeared as
179 highly variable in tumor tissue with DNA methylation beta-values ranging from approximately 0.25-
180 0.8. The aggregated 5mC of these three positions displayed significant hypomethylation in tumor
181 versus normal tissue (Supplement Fig. 3A, BH adjusted *p*=2.4E-6), reflecting the same trend as
182 observed in our study.



183

184 **Figure 1. DNA methylation in *GCK* exon 7 and within the *TM6SF2* intron 2-exon 3 boundary.** Detailed information about the
185 **A.** *GCK* and **D.** *TM6SF2* regions including annotated CpG positions on the Illumina Methylation array. The grey rectangle
186 represents the EGR1 binding motif. The yellow rectangle indicates the genome segmentation by ChromHMM in UCSC genome
187 browser predicting a *"weak enhancer or open chromatin cis regulatory element"*. The pyrosequencing region is marked in
188 red. Line plot of median DNA methylation at **B.** *GCK* CpG 1-4 and **E.** *TM6SF2* CpG 3-6. DNA methylation at **C.** *GCK* CpG 1-3 and
189 **G.** *TM6SF2* CpG 3-5 for all ages and **D.** *GCK* CpG 1-3 and **H.** *TM6SF2* CpG 3-5 for patients aged 42-60 years. Pairwise comparison
190 against LH (Students t.test), **p*<.05, ** *p*<.01, *** *p*<.001, **** *p*<.0001; LH: lean healthy, OH: obese healthy, LM: lean
191 MAFLD, OM: obese MAFLD, aMetD: advanced metabolic disease

192

193 Since aberrant DNA methylation in regulatory regions is often linked to changes in gene
194 expression, we checked the *GCK* expression levels in a published dataset of MAFLD patients because
195 transcriptome analysis data are currently not available for our cohort. *GCK* expression was increased
196 by 2.48-fold (false discovery rate (FDR) adjusted *p*=1.08E-05) in MAFLD compared to obese-healthy
197 subjects and by 1.83-fold (FDR adjusted *p*=1.21E-05) in advanced MetD patients (metabolic
198 dysfunction-associated steatohepatitis, MASH) compared to lean-healthy subjects in blood (62). While

199 Ilieva and colleagues found *GCK* overexpressed in MAFLD and MASH patients (62), it was significantly
200 decreased in HCC patients from TCGA (Supplement Fig. 3B, BH adjusted *p*=1.2E-07) (61). Furthermore,
201 the lncRNA LOC105375258 is expressed in liver tissue (63), but currently there are no further data
202 available concerning this lncRNA.

203 In line with *GCK* exon 7, the same analysis was performed for the *TM6SF2* intron 2-exon 3
204 boundary (chr19:19,381,224-19,381,297; hg19, Fig. 1E). *TM6SF2* is encoded by ten exons on the
205 reverse strand of chromosome 19. In total, there are five transcripts annotated, of which two result in
206 a retained intron and one in a skipped exon overlapping our target region (56). Notably, especially exon
207 3-5 are targets of alternative splicing events (Supplement Fig. 4). Furthermore, *TM6SF2* overlaps with
208 the sense lncRNA ENSG00000267629 (AC138430.4), sharing the same transcription start side. This
209 lncRNA represents a readthrough between *TM6SF2* and *HAPLN4* (56). ChromHMM predicts a "weak
210 enhancer or open chromatin cis regulatory element" in HepG2 and a "promoter region including TSS"
211 in K562 in the UCSC genome browser (57) (Fig. 1E, yellow rectangle), as well as an EGR1 binding site.
212 5mC analyses revealed that CpGs 3-5 were significantly hypermethylated in the advanced MetD group.
213 In contrast, CpG 4 and 5 were significantly hypomethylated in OM compared to LH and CpG 5 was also
214 significantly hypomethylated in LM compared to LH (Student´s t-test, Fig. 1F and G). Similar to *GCK*, we
215 observed an effect of age on DNA methylation in *TM6SF2*. The mean DNA methylation calculated from
216 CpG 1-6 in the *TM6SF2* intron 2-exon 3 boundary, as well as for the individual positions CpG 1, 3 and
217 strongest for CpG 5 correlated with the patients´ age (Supplement Fig. 2). Therefore, we restricted the
218 analysis to 42-60-years-old patients and still found a significant hypermethylation at CpG 3 and 4, but
219 not in 5 for the advanced MetD group. CpG 6 showed significant hypomethylation for LM and OM
220 compared to LH (Students t-test, Fig. 1H). Strikingly, we observed two different effects:
221 hypomethylation in lean and obese MAFLD subjects and hypermethylation in patients with advanced
222 MetD compared to lean healthy subjects. This could be caused by the elevated HbA_{1c} >7% as a sign of
223 less controlled T2DM, as well as increased fasting blood glucose, HOMA, ALT and total triglyceride
224 levels (Table 1 and Supplement Table 1 for *p* values, Supplement Fig. 1C, E, F, H and L). TCGA data
225 showed that the nearest annotated CpG cg11911689 (78 bp downstream of CpG 4, Fig. 1H) was
226 significantly hypomethylated in hepatocellular tumor tissue compared to normal tissue (Supplement
227 Fig. 3C, BH adjusted *p*=4.5E-10) (61). Cg26470696, located 2.2 kb downstream was hypermethylated,
228 although this was only barely significant (Supplement Fig. 3C, BH adjusted *p*=.04). The CpG located 5.9
229 kb upstream was also significantly hypomethylated (Supplement Fig. 3C, BH adjusted *p*=6.1E-10), but
230 not informative for our analysis due to the large distance to the investigated region. The nearest
231 annotated CpG and the next CpG located downstream were both highly variable methylated
232 (Supplement Fig. 3C). The aggregated 5mC of all three CpGs was significantly hypomethylated in tumor
233 versus normal tissue (Supplement Fig. 3C, BH adjusted *p*=.0033). Of note, the aggregated 5mC obtained

234 from TCGA was over 8 kb long and therefore, may not be informative. *TM6SF2* expression was
235 significantly reduced in HCC tumor compared to adjacent normal tissue (Supplement Fig. 3D, BH
236 adjusted *p*=6.5E-14). The expression of the overlapping sense lncRNA ENSG00000092054
237 (AC138430.1) has recently been shown to be increased by 3.19-fold (FDR adjusted *p*=.0005898) in
238 MAFLD compared obese-healthy subjects (62). Furthermore, the expression level of this lncRNA was
239 prognostic for the survival of intrahepatic cholangiocarcinoma (iCCA) with a hazard ratio (HR) of 1.582
240 (1.143-2.190, 95% CI, *p*=.006). Patients with a higher survival probability showed decreased expression
241 of this lncRNA transcript (64). Moreover, it was also prognostic for patients suffering from acute
242 myeloid leukemia (AML) (65).

243 **Differentially methylated regions in *GCK* and *TM6SF2* harbor regulatory functions and contact**
244 **neighboring genes**

245 To assess the regulatory potential of the identified DMPs in *GCK* and *TM6SF2* target regions
246 highlighted in yellow in Fig. 2A, we analyzed the activating H3K27ac, H3K4me1, H3K4me3 and the
247 repressing H3K27me3 histone marks in H1-hESC (human embryonic stem cell line, light red), HepG2
248 (hepatoblastoma cell line, light blue), NHLF (lung fibroblasts, orange), K562 (lymphoblasts, brown),
249 NHEK (normal human epidermal keratinocytes, violet), and MCF7 (breast cancer, green) cell lines in
250 IGV from the ENCODE project (66). Unfortunately, for the hg19 assembly ENCODE did not include
251 H3K4me1 data in IGV. *GCK* showed H3K4me1 activating histone marks in K562 cells and to a lesser
252 extent also in HepG2 cells, whereas NHLF and NHEK cells harbored the repressive H3K27me3 mark
253 (Fig. 2A, left panel). MCF7 cells showed H3K27ac marks near the target region in exon 7. Furthermore,
254 the downstream regions overlapping exon 8-10 showed cell-type specific marks as indicated by
255 H3K4me1 in HepG2 and K562 cells, but H3K27me3 in H1-hESC, NHLF and NHEK cells. For *TM6SF2* (Fig.
256 2A, right panel) we found strong signals for the activating H3K4me1 histone modification in K562 cells,
257 but a H3K27me3 signals in NHEK and NHLF cells. HepG2 cells showed H3K4me1 and H3K4me3 marks
258 in the promoter region of *TM6SF2*. In conclusion, the analysis of ENCODE histone modifications
259 revealed cell type specific regulatory elements for both target regions.
260 To confirm these functions, we performed *in vitro* luciferase reporter assays for the
261 differentially methylated regions in *GCK* exon 7 and the *TM6SF2* intron 2-exon 3 boundary. To do this,
262 we cloned the regions of interest into two different CpG-free vectors upstream of the luciferase gene
263 and tested for their potential promoter and enhancer functions (Fig. 2B). We detected both, a
264 promoter (60.3±11.3-fold change, Welch two sample t-test, *p*=.0004, Fig. 2C) and enhancer (27.4±3.9-
265 fold change, Welch two sample t-test, *p*=.0001) function for the *GCK* region compared to the empty
266 vector. However, when the same region was methylated prior to the analysis, both functions were lost.
267 This trend was confirmed by TCGA data (61) of the nearest annotated CpG, where lower methylation

resulted in higher expression (Fig. 2D). In contrast, we did not find a promoter function (Fig. 2E) for the DMR located in *TM6SF2*, but the methylated construct within the CpG-free promoter vector showed a by 1.7±0.2-fold (Welch Two Sample t-test p=.004) elevated enhancer function. This was underlined by the correlation of the nearest annotated CpG and the expression of *TM6SF2* in TCGA data with higher methylation of the CpG in the regulatory region, resulting in higher expression (Fig. 2F).

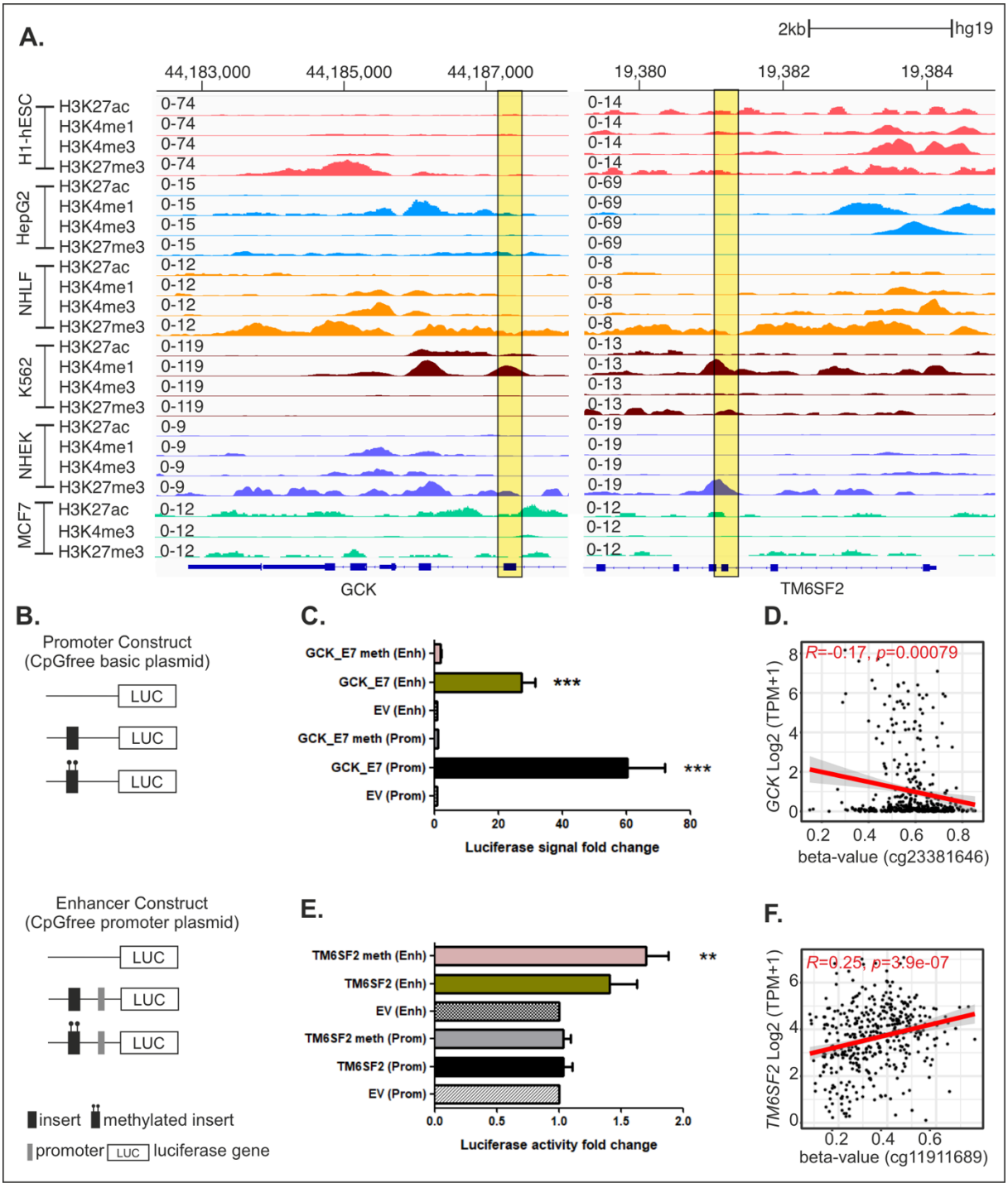


Figure 2. Regulatory roles of *GCK* and *TM6SF4* target regions. A. Activating (H3K27ac, H3K4me1 and H3K4me3) and repressing (H3K27me3) histone modifications obtained from ChIP-seq experiments from ENCODE for H1-hESC (stem cells, light red), HepG2 (hepatoblastoma, light blue), NHLF (healthy lung, orange), K562 (blood, brown), NHEK (Normal Human Epidermal Keratinocytes, violet) and MCF7 (breast cancer, green) cell lines (66) in *GCK* (left panel) and *TM6SF2* (right panel).

Yellow highlighted areas represent the pyrosequencing regions. Here we showed a cell-type specific pattern of activating and repressing histone marks. **B.** The CpG-free basic vector was used to test for a putative promoter function and the CpG-free promoter (containing a minimal promoter) vector to test for a putative enhancer function of the differentially methylated regions in a luciferase gene reporter assay in **C.** *GCK* exon 7 and **D.** *TM6SF2* intron 2-exon 3 boundary, in a methylated and unmethylated state and compared to EV (empty vector). **E.** and **F.** show the correlation of the nearest annotated CpG with gene expression in TCGA data (61). Pairwise comparison against empty vectors (Welch two sample t.test), **p*<.05, ***p*<.01, ****p*<.001.

To predict the potential neighboring targets of the *GCK* and *TM6SF2* regulatory regions, we analyzed Hi-C data in HepG2 cells (66) and from HCC patient material (67) using the 3D Genome Browser (68). *GCK* is located in a topological associated domain (TAD) of approximately 500 kb. When comparing the interaction frequencies of HepG2 cells and HCC decreased signal intensities were observed (Fig. 3A). Strikingly, within this TAD several genes are known for their association with metabolic diseases and cancer. *AEBP1* was upregulated in MASH (69) and its elevated expression predicted poor prognosis of liver fibrosis (70). *YKT6* was identified as metabolic syndrome risk gene in a genome-wide association study (GWAS) in the Taiwanese population (71) and the upregulation of *YKT6* transcription was associated with poor prognosis of HCC (72). *BLVRA* expression was upregulated in HCC (73). *URGCP* (also known as *URG4*) overexpression was related to tumor formation in mice (74), as well as in humans where it was found in several different cancer entities such as HCC, gastric and bladder cancer (75). Overexpression of *PGAM2* was found to downregulate glycolysis genes like phosphofructokinase or hexokinase 2 in heart tissue of mice (76). Two genes of the adjacent TAD were also linked to MetDs. The key cholesterol transport protein *NPC1L1* was identified as a risk factor for MAFLD (77). *CAMK2B* differential methylation was caused by high glucose treatment in Huh-7 cells, which were used for MAFLD and T2D models (78). By analyzing the potential contact points of the *GCK* DMP, we noticed that this region stays in contact with an enhancer region in *GCK* intron 1 (~chr7:44,157,293-44,158,172, hg38) and a proximal enhancer region between *MYL7* and *POLD2* (Fig. 3B). Of note, in a pan-cancer analysis, elevated POLD2 expression was found in advanced tumor stages and associated with poor prognosis, also for HCC (79).

TM6SF2 is located in a TAD together with *UPF1* (Fig. 3C), a gene involved in targeting a lncRNA *UCF1* implicated in cancer formation, and silencing of *UPF1* stimulated glycolysis in HCC (80). Furthermore, a genetic variant located in the gene *NCAN* found in the same TAD is associated with MAFLD (80). The interaction analysis showed that the *TM6SF2* DMP contacts a promoter-like region within intron 3 of *HPLN4* (~chr19:19,260,457, hg38) as well as intron 8 of *SUGP1* (~chr19:19,280,457, hg38), which was identified as regulator of cholesterol metabolism (81) (Fig. 3D).

In summary, both *in silico* data for histone marks and HiC data, as well as the luciferase reporter assay provide evidence that *GCK* exon 7 and the *TM6SF2* intron 2-exon 3 boundary harbor important regulatory elements, which are not only related to metabolic diseases, but also could play roles in tumor formation and prognosis. Furthermore, our analysis points out that these regions contact

neighboring genes and that changes in their methylation patterns potentially also affect the interactions with these elements. This strongly indicates that changes in DNA methylation at these regulatory elements is not only associated with MAFLD but could also be linked to the enhanced cancer risk in MetD patients corroborated also by the changes in the HiC data within these TADs.

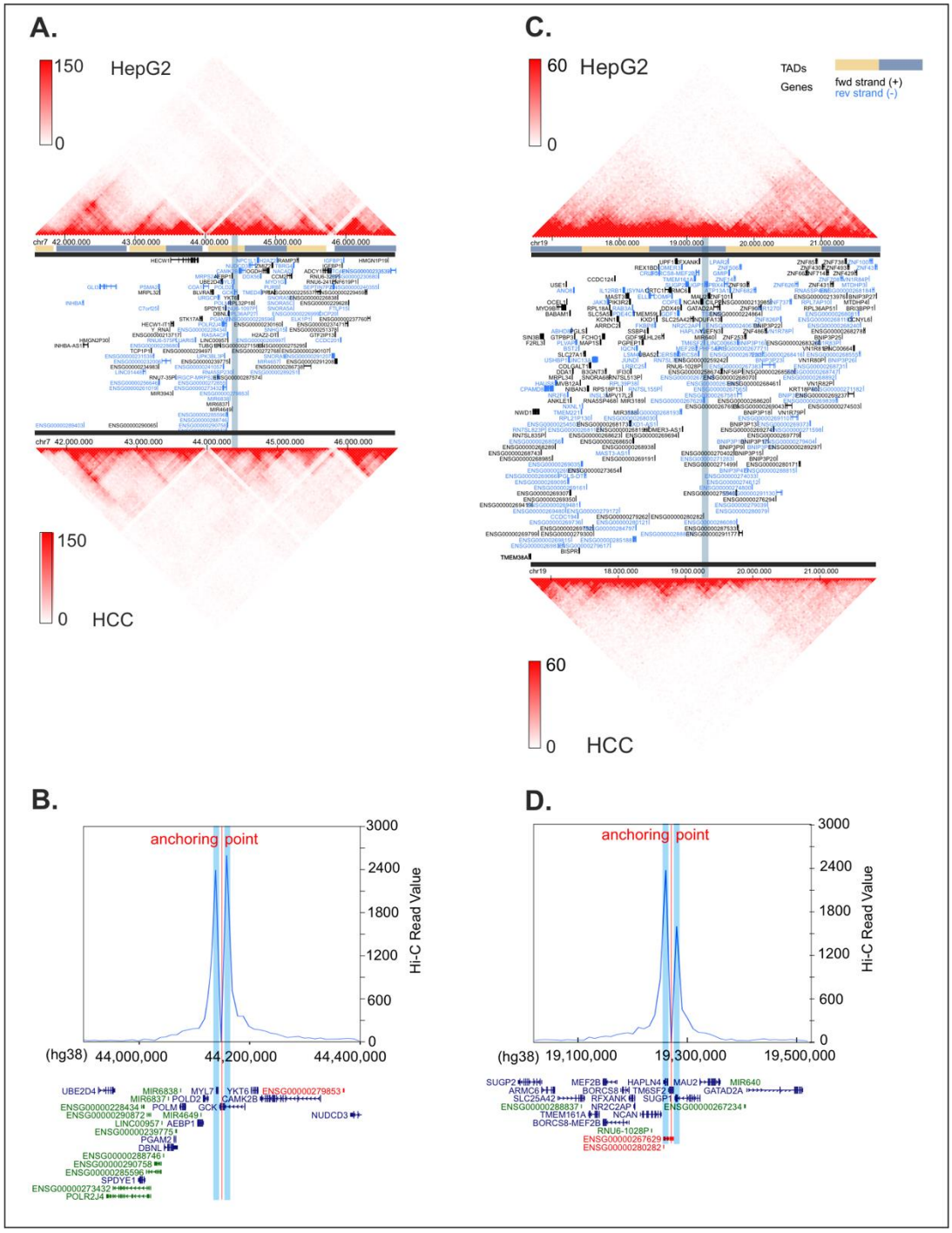


Figure 3. HiC data analysis in HepG2 cells and in HCC tumor material visualized in 3D genome browser (68). A. TAD formation on chromosome 7 around the *GCK* DMP in HepG2 (upper panel) (66) and HCC tumor tissue (lower panel) (67). **B.** Analysis of potential contact points of the regulatory region (red vertical line) shows interactions (blue highlighted areas) with *GCK* intron 1 and an intergenic region between *MYL7* and *POLD2* near an enhancer region. **C.** TAD formation on chromosome 19 around TM6SF2 in HepG2 (upper panel) (66) and HCC tumor tissue (lower panel) (67). **D.** Analysis of potential contact points of the

327 regulatory region (red vertical line) shows interactions (blue highlighted areas) with *SUGP1* and *HAPLN4*. Chromosome
328 annotations are given in the newer assembly hg38.

329 **DMPs detected in *GCK* and *TM6SF2* are able to form non-B DNA structures**

330 G-rich regions with a high GC-skew can form non-B DNA structures such as G4s (82) and R-
331 loops (83). To investigate if the detected DMPs in these two regulatory regions are located in non-B
332 DNA structures, we retrieved information about R-loop forming sequences (RLFS) (83) and putative G4
333 motifs (PQS) from pqsfinder (84). The *GCK* DMR does not harbor a predicted RLFS, albeit this exon is
334 overlapped by the annotated antisense lncRNA LOC105375258, but we found several G4 motifs (Fig.
335 4A left panel). All four CpG positions analyzed in *GCK* exon 7 are located in a G4 motif, with the top
336 differentially methylated CpG identified in all four MetD groups located in the longest loop of the G4
337 (Fig. 4A). Furthermore, exons 8-10 harbor both RLFS and PQS (Fig. 4A left panel). The *TM6SF2* DMR
338 contained a RLFS as well as PQS (Fig. 4A middle panel). These predicted motifs are not found in an
339 intronic control region at *PARD3B* (Fig. 4A right panel). The differentially methylated CpGs 4-6 were
340 located directly within the G4 motif in *TM6SF2* intron 2-exon 3 boundary and CpG 1-3 flanked the G4
341 sequence motif (Fig. 4A).
342 *In vitro* G4 formation was verified by analyzing published G4-sequencing (G4-Seq) maps based
343 on a polymerase stop assay in a buffer containing the G4 stabilizer pyridostatin (PDS) in HEK293T cells
344 (85) (Fig. 4A, black track). Strong signal intensities in *GCK* and *TM6SF2,* but not in *PARD3B*, confirm the
345 ability of the predicted motifs to form DNA secondary structures. As G4s are not necessarily formed in
346 living cells and their formation can vary strongly in different cell types, we analyzed published G4 ChIP-
347 Seq data in HEK293 cells (86). Especially, *GCK* shows enriched signal intensities compared to its input
348 control over the whole region spanning from exon 7 to exon 10 (Fig. 4A left panel), whereas *TM6SF2*
349 showed a more discreate peak for replicate 1 but not for replicate 2 (Fig. 4A middle panel). Since
350 antibody-based methods have to be taken with care, we aimed to verify these data with an alternative
351 method. Therefore, we updated a chemical method based on permanganate/S1 nuclease footprinting
352 (87), but linked it with direct adapter ligation and called it PDAL-Seq (permanganate/S1 nuclease
353 footprinting with direct adapter ligation and sequencing) (54). Since PDAL-Seq is a footprinting
354 method, we expected low or absent read frequency directly at the non-B DNA structure but increased
355 read frequency at the non-B DNA flanking sites. We did not detect a direct overlap of footprints for
356 HEK293 cells rather than a slight upstream shift. Only in MCF7 cell the footprint was more obvious (Fig.
357 4A left panel, light blue highlighted box). Of note, footprints detected in *GCK* exon 8-10 were more
358 pronounced and overlapped well with the strong G4-Seq signals (Fig. 4A left panel, light red highlighted
359 box), while the signals from the antibody-based G4 ChIP-Seq method were less clear (Fig. 4A, red
360 tracks). One explanation for the difference in G4 formation could be that for exon 8-10 a RLFS was

361 predicted, while this was absent for our target region in exon 7 and in general R-loops and G4s stabilize
362 each other. In agreement to *GCK*, PDAL-Seq data confirmed G4 formation at the *TM6SF2* intron 2-exon
363 3 boundary due to the presence of a footprint in the tracks (Fig. 4A middle panel, blue tracks, light blue
364 highlighted box). As G4s are important regulators of alternative splicing events (88), this signal directly
365 at the exon-intron boundary could be involved in alternative splicing events. In contrast, the intronic
366 region of *PARD3B* used as negative control did not contain a G4 motif, nor did it form G4s *in vitro* or in
367 living cells (Fig. 4A right panel).

368 Next, we analyzed EGR1 ChIP-Seq data in K562 from ENCODE (66). Unfortunately, there were
369 no data available in HEK293 cells. Strong EGR1 signals were detected co-occurring with the G4-Seq
370 peak and the footprinting valley of PDAL-Seq. Currently we do not know, if non-B DNAs are formed in
371 K562 cells, but bioinformatic studies predict EGR1 as a potential G4 binding protein (89). Nevertheless,
372 also the negative G4 region in *PARD3B* showed a EGR1 binding site indicating that EGR1 does not only
373 target G4 regions.

374 Finally, G4 folding was confirmed by native DNA polyacrylamide (PAA) gel electrophoresis (Fig.
375 4B). The *GCK* and *TM6SF2* target regions and two controls, *MYC* as a strong G4-folding positive region
376 (90) and an enhancer region of *ESR1* as a G4-negative region (91), were PCR amplified and separated
377 on a native 10% DNA PAA gel. Strikingly, both *GCK* and *TM6SF2*, as well as *MYC* showed significant
378 shifted bands after the G4 folding procedure with potassium and heat induction (Fig. 4B, labeled with
379 K), and even stronger after G4-stabilization with PDS (Fig. 4B, labeled with PDS), compared to the
380 unfolded controls only containing Tris-buffer (Fig. 4B, labeled with U). This effect was not detected for
381 the negative control region *ESR1*. In agreement with the native gel electrophoresis, CD experiments of
382 the positive G4 folding region *MYC*, and our target regions in *GCK* and *TM6SF2*, showed characteristic
383 CD spectra for parallel G4s with a negative peak at 240 nm and a positive peak at 260 nm (92) (Fig. 4C).
384 This was not observed for the negative region in *ESR1*, as already confirmed in the DNA PAA gel.

385 To summarize, we identified G4 formation at the *GCK* exon 7 and *TM6SF2* intron 2-exon 3
386 boundary with five different methods: published G4-Seq and G4 ChIP-Seq data, PDAL-Seq, native PAA
387 gel electrophoresis and CD spectra. The newly established PDAL-Seq approach emerged as a valid
388 alternative method to detect non-B DNA formation. Taken together, the DNA methylation data
389 obtained from patients´ blood, analysis of histone marks and HiC data, *in vitro* luciferase assays and
390 the identification of non-B forming sites and their overlap with EGR1 binding, highlighted that non-B
391 formation overlaps with sites of methylation instability within regulatory elements potentially
392 affecting important cellular processes causing dysregulation of the cell's transcription patterns or
393 alternative splicing processes.

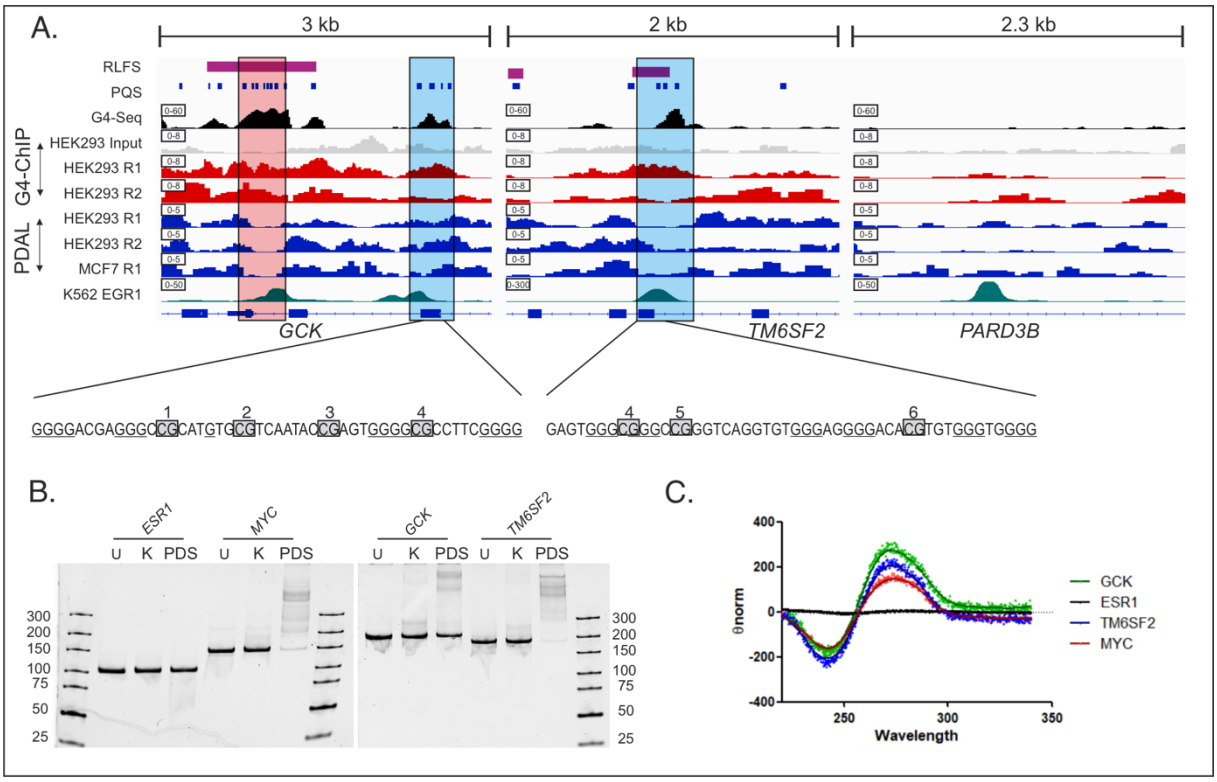


Figure 4. Non-B DNA formation in *GCK* and *TM6SF2*: A. Shows *GCK* (chr7: 44,184,901-44,187,919), *TM6SF2* (chr19: 19,380,358-19,382,362) and the G4-negative region *PARD3B* (chr2:205,780,593-205,782,891); Data from different sources are shown in individual rows: Pink rectangles indicate R-loop forming sequences (RLFS) calculated by Quantitative Model of R-loop Forming Sequence (RLFS) finder (QmRLFS-finder) (83). Blue rectangles represent putative G4 forming motifs (PQS) obtained from pqsfinder (84), Black track: *in vitro* G4 formation stabilized with pyridostatin (PDS) in HEK293T cells (85); Grey track: ChIP input control (86); Red tracks: G4-ChIP-Seq tracks in HEK293 cells (2 replicates) (86); Blue tracks: PDAL-Seq replicates 1 and 2 of HEK293 and MCF7 cells, Olive tracks: transcription factor EGR1 binding in K546 cells (66). The sequences within *GCK* and *TM6SF2* forming a G4 structure are detailed below. G-tracks are underlined, grey boxes indicate CpG positions. Blue highlighted areas: identified differentially methylated regions; light red highlighted area: a second regulatory region in *GCK,* which can form G4s. **B.** Native DNA PAA gels identifies *in vitro* G4 formation in the G4 control region in *MYC, GCK* exon 7 and *TM6SF2* intron 2-exon 3, but not for *ESR1* (negative control). PCR products are loaded as unfolded (50 mM Tris-HCl pH 8.3) (U), folded with 150 mM KCl and heat induction (K), and folded with 150 mM KCl, heat induction and 10 µM PDS (PDS). **C.** CD spectra of G4 negative region *ESR1*, G4 positive region *MYC*, and the two target regions in *GCK* and *TM6SF2*. *MYC, GCK* and *TM6SF2* showed characteristic CD spectra for parallel G4s with a negative peak at 240 nm and a positive peak at 260 nm in a 50mM Tris-HCl pH 8.3 and 150 mM KCl buffer.

411 ## 3. Discussion

412 Metabolic diseases often arise due to a complex interplay of genetic and epigenetic processes.
413 Whereas the genetic risk is predefined by the base composition, the epigenetic landscape is flexible
414 and heavily influenced by the individuals´ environment. To assess epigenetic changes in obesity and
415 metabolic diseases, we analyzed DNA methylation patterns in intragenic (exonic and intron-exon
416 boundary) regions of two key metabolism genes *GCK* and *TM6SF2* overlapping annotated lncRNAs in
417 five different groups: lean- and obese healthy subjects, lean and obese MAFLD, as well as patients
418 suffering from advanced MetDs. Significant 5mC changes were detected for both target genes. While
419 *GCK* exon 7 was significantly hypomethylated in all groups especially at CpG 3 compared to lean healthy
420 subjects, we observed a diverse pattern for the *TM6SF2* intron 2-exon 3 boundary. Lean- and obese
421 MAFLD subjects showed a hypomethylation, whereas hypermethylation was detected in advanced
422 MetD. This pattern likely mirrors the elevated blood glucose levels observed also as HbA_{1c}>7% within
423 the advanced MetD group, potentially leading to hypermethylation. Since increasing chronological age
424 is one major risk factor for developing metabolic diseases (reviewed in (55)), our findings of a
425 significant age correlation with DNA methylation within *GCK* and *TM6SF2* were expected. In a previous
426 study, Xu et al. already showed a relation between DNA methylation of a CpG island in exon 9 and 10
427 of *GCK* (hg19, chr7: 44,184,771-44,185,695) with age in 36 coronary heart disease patients and 36
428 matched controls (93), but to the best of our knowledge, there are currently no data available for
429 *TM6SF2*. Furthermore, our results clearly showed significant differences in DNA methylation in
430 individual disease groups, and this was also confirmed in liver cancer versus adjacent normal tissue
431 from hepatocellular carcinoma patients obtained from *TCGA* data (61). Albeit the trend was reversed
432 for *TM6SF2* with hypermethylation in aMetD but hypomethylation in HCC, the observation of opposing
433 directions of differential methylation in blood versus the affected tissue is not unusual. In our latest
434 epigenome-wide association study, a regulatory region upstream of NHL repeat containing 1 (*NHLRC1*)
435 was found to be hypervariable, and hypermethylated in blood of lung cancer patients, but
436 hypomethylated in lung cancer versus normal adjacent tissue (94).

437 Especially if differential methylation occurs within regulatory elements, effects on gene
438 expression patterns are logical consequences. By investigating the target regions in luciferase reporter
439 assays, we detected a promoter and enhancer function for *GCK* exon 7, which was abolished by DNA
440 methylation. Furthermore, the *TM6SF2* target region showed an enhancer function only in a
441 methylated state. Although transcriptome analysis is not available for our cohort, the comparisons
442 with recent MAFLD gene expression studies (62) and TCGA data (61) revealed that expression of *GCK*
443 and *TM6SF2* as well as of their overlapping lncRNA transcripts was altered in disease states. Of note,
444 the expression of the lncRNA ENSG00000092054 annotated to *TM6SF2* is prognostic for the survival of

445 intrahepatic cholangiocarcinoma (64) as well as for AML (65). Nevertheless, the decreased expression
446 of *GCK* in HCC is also caused by its replacement by hexokinase 2 in tumor cells (95).

447 Next, we studied the occurrence of non-B DNA structures. Although we selected our target
448 region based on the annotation and expression of lncRNAs, R-loop formation was predicted only for
449 the *TM6SF2* intron 2-exon 3 boundary. Unlike *GCK* exon 7, which lacks RLFS, the upstream exon 8-10
450 potentially can form them, although a lncRNA is not annotated. Both target regions harbor several PQS
451 motifs likely forming *in vivo* G-quadruplexes. Experimental validation of non-B DNA formation was
452 conducted using several techniques, including published G4-Seq and ChIP-Seq data, PDAL-Seq, native
453 DNA PAA gel electrophoresis, and CD spectroscopy. These analyses confirmed the formation of G4
454 structures at the predicted sites in GCK exon 7 and TM6SF2 intron 2-exon 3 boundary, but also in the
455 upstream *GCK* exon 8-10 region. Additionally, EGR1 binding sites were identified, suggesting a
456 regulatory role for G4 structures in these regions. Strikingly, the differentially methylated CpGs are
457 located directly within G4 forming sequences pointing out the important roles of G4s causing
458 methylation instability. While we did not analyze DNA methylation in the second G4 cluster of *GCK* in
459 exon 8-10, Xu et al. detected significant hypomethylation in exons 9-10 in patients suffering from
460 coronary heart disease (49.77 ± 6.43%) compared to controls (54.47 ± 7.65%, p=.018) (93), reflecting
461 the same trend as in our study. G4 signals are stronger in regions with predicted RLFS and PQS motifs
462 by showing a more pronounced footprint compared to regions harboring PQS alone. Notably, since we
463 have used cell lines for our study, and within these cells the expression of the lncRNAs could differ
464 from liver tissue from MetD patients, probably the G4 signals are slightly different. Nevertheless,
465 differential methylation is detected in all three clusters (GCK exon 7, GCK exon 9-10 by (93) and *TM6SF2*
466 intron 2-exon 3 boundary) pointing out the important role of the G-quadruplex and probably the R-
467 loop as a stabilizer of the DNA secondary structures. Of particular note, PDAL-Seq emerges as a robust
468 alternative method for detecting non-B DNA structures, providing complementary evidence to the
469 other experimental approaches.

470 Based on our results and extensive literature research, we hypothesize that the pathologically
471 increased concentrations of glucose and fatty acids in obesity and metabolic diseases lead to
472 reformations or topological changes of the non-B DNA landscape (Fig.5). Since G4s are known as sites
473 for interactions and bound by DNMTs, other TFs as well as splicing factors (38), this in turn likely causes
474 changes in DNA methylation patterns, which we have observed. Ultimately, this could impact gene
475 expression or alternative splicing events and finally promotes disease formation and progression. First
476 of all, G4s are fairly sensitive to changes in concentrations of ions like potassium (96), or of the pH (97),
477 but strikingly a recent preprint by Guneri and colleagues observed G-quadruplex and i-motif formation
478 (non-B DNA structure formed opposite of a G4 on the C-rich strand) in the promoter of the insulin gene
479 (98). Its expression is dependent on the formation of these non-B DNA structures. Moreover, it has

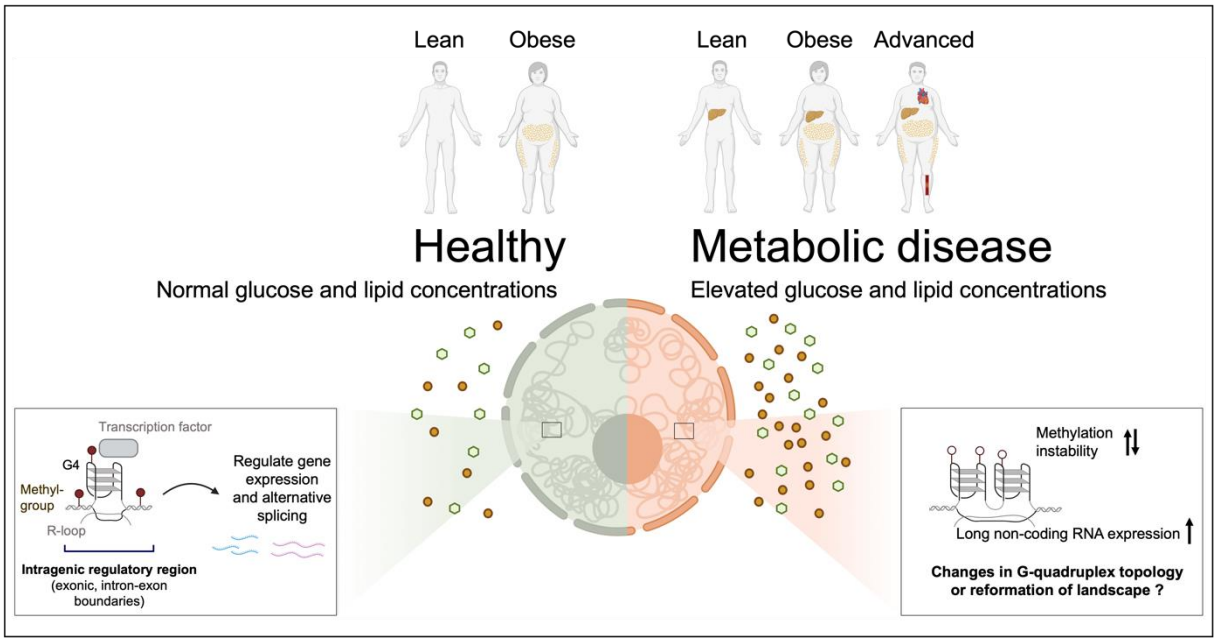
480 been shown that pathologically high glucose concentrations formed a stable G4 and lead to
481 significantly increased expression compared to normal glucose concentrations within a luciferase
482 reporter assay (98). The link between G4s and cancer is well known. While the number of G4s naturally
483 decreases during cell differentiation (99), G4s are increased in tumor tissue (100) and in the human
484 keratinocytes (HaCaT) cancer cell line versus normal human epidermal keratinocytes NHEK (101).
485 Evidence that G4s are involved in regulatory processes emerges also from the observation, that they
486 function as TF binding hubs (37, 38). Amongst their top binders splice factors (38) and DNA
487 methyltransferases were found (26, 39). Mao et al. hypothesized that G4 structures sequester DNMTs
488 and direct DNA methylation away from regulatory elements (26). Hence topological changes caused
489 by pathological and long-term elevated glucose or fatty acid concentrations as observed in obesity and
490 metabolic diseases most likely lead to the observed differential methylation.

491 Additionally, a recent publication has shown that the cell-type specific transcriptome is shaped
492 by G4s (102), fitting to our observation of cell type specific regulatory elements in *GCK* and *TM6SF2*.
493 One of these TFs is EGR1, which recognizes a G-rich motif (103). Bioinformatic studies have shown that
494 EGR1 belongs to a set of potential G4 binding proteins (89). The analysis of published EGR1 ChIP-Seq
495 data in K562 showed that both target regions are strongly bound by EGR1. Therefore, it is reasonable
496 to assume that changes in the G4 formation, or its topology could change the binding of EGR1 to these
497 regulatory elements, and in consequence also the transcription of other genes, since HiC data clearly
498 show several contact points for these regulatory elements.

499 Finally, G4s have been shown to control alternative splicing processes and they are enriched
500 within approximately 100 bp next to the 3´ or 5´ splice sites (88). The splicing factor U2AF1, which
501 recognizes the 3´splice site of a gene, is amongst the top ten candidates associated with G4s (38). Since
502 we noticed that G4s flank the alternative splicing region intron 2 to intron 3, which spans the
503 alternatively retained introns in a *TM6SF2* transcript like the lncRNA ENSG00000092054, changes in
504 the G4 landscape could also trigger the enhanced formation of this lncRNA. This is underlined by the
505 observation of Wang et al, that retained introns often co-occur with a high level of G4 formation (104).

506 Although further research is necessary to get a deeper understanding of the mechanisms involved,
507 we were able to connect detected DMPs in metabolic risk genes to G-quadruplex formation in a patient
508 cohort suffering from metabolism dysfunction-associated fatty liver disease and type II diabetes, for
509 the very first time. Unfortunately, we are currently lacking G4 formation data in human liver biopsies,
510 since this type of material is rare. Although the assessment of anthropometric and blood parameters,
511 abdominal MRI and liver biopsies are the current standard for MAFLD diagnosis, liver biopsies cannot
512 be applied on a broad scale in asymptomatic obese or lean-MAFLD subjects for ethical concerns or low
513 patient acceptance. In conclusion, since metabolic diseases are linked to an elevated cancer risk, and
514 aberrant DNA methylation as well as elevated non-B DNA formation occurs during cancer formation,

515 these mechanisms could point to a common mechanism of metabolic diseases and cancer
516 development. Understanding this intricate connection is paramount in unraveling the complexities of
517 metabolic diseases, marking a significant step forward in research.
518



519
520 **Figure 5: Altered non-B DNA formation and lncRNA expression in patients suffering from metabolic diseases cause**
521 **methylation instability and altered gene expression patterns or alternative splicing mechanisms.**

522 ## 4. Methods

523 **Study cohort demographics**

524 We designed a nested cohort study by selecting 148 blood samples from Caucasian origin collected
525 at the University Hospital Salzburg, Austria, which fit into one of five groups according to the criteria
526 below. The local ethics committee (Ethikkommission des Landes Salzburg No. 415-E/1521/6-2012, and
527 an ethics approval by the Paracelsus Medical University Salzburg no 01-2022) approved the study and
528 informed consent was obtained from all participants. Alanine transaminase (ALT) and aspartate
529 transaminase (AST) levels ≤40 U/L for women and ≤50 U/L for men, with or without increased levels
530 of γ-glutamyl transpeptidase (GGT), were considered as normal liver tests. Routine clinical parameters
531 were determined during sampling at the outpatient clinics of the University Hospital Salzburg, Austria.
532 (1) Lean healthy subjects (LH, n=24): body mass index (BMI) ≤25.0 kg/m^2, normal liver tests,
533 normal ultrasound and no components of MetS according to ATPIII criteria (105)
534 (2) Obese healthy subjects (OH, n=31): metabolic healthy obese (MHO) BMI ≥30.0 kg/m^2, normal
535 ultrasound and normal liver tests, and ≤2 components of MetS (obesity and one additional
536 metabolic syndrome defining criterion allowed and glycated hemoglobin (HbA_{1c})<6.0%;

537 (3) Lean MAFLD subjects (LM, n=27): metabolic unhealthy normal weight (MUNW), BMI ≤25.0
538 kg/m^2, ultrasound evidence of a fatty liver, with or without elevated liver tests, HbA_{1c}<7%;
539 (4) Obese MAFLD subjects (OM, n=32): BMIs ≥30.0 kg/m^2, unequivocal fatty liver ultrasound, with
540 or without elevated liver tests, HbA_{1c}<7%;
541 (5) Advanced MetD (aMetD, n=34): Elevated fibrosis-4 (FIB4) score>2.67 and/or established T2DM
542 with elevated HbA_{1c}>7%.

543 **Sequence specific DNA methylation analysis by pyrosequencing**

544 DNA was extracted from 200 µL whole blood with Promega ReliaPrep Blood gDNA Miniprep
545 System (Cat #A5081) according to the manufacturer´s guidelines. Bisulfite conversion was performed
546 with 500 ng of genomic DNA (gDNA) input material using the Zymo Research Bisulfite Conversion Kit
547 (Cat #D5001). DNA was eluted with 60 µL M-Elution buffer (~8 ng/µL). PCR was performed with 2 µL
548 (~16 ng) or 4 µL (~32 ng) of BT-converted DNA with HotStar Taq Polymerase (Qiagen), depending on
549 the genomic region. Pyrosequencing was performed using 20 µL PCR product and 7.5 or 9 µM
550 sequencing primer. Detailed assay design, primer sequences, PCR mastermixes, annealing
551 temperatures and sequencing conditions can be found in the Supplement (Material & Methods section
552 1.1). Data analysis was performed with Pyromark Q24 Advanced software, R version 4.3.1 (106) and R
553 studio (107).

554 **TCGA data and ENCODE histone mark and TF binding data analysis**

555 DNA methylation and gene expression data were obtained from tumor versus normal tissue
556 from patients suffering from hepatocellular carcinoma (HCC) via the Shiny Methylation Analysis
557 Resource Tool (SMART), an interactive tool for analyzing *The Cancer Genome Atlas* (TCGA) data (108).
558 *The Encyclopedia of DNA Elements* (ENCODE) data for H3K27ac, H3K4me1 and H3K4me3 from
559 the liver hepatoblastoma cell line HepG2 and the lymphoblast cell line K562 (66) were downloaded
560 and visualized using the Integrative Genome Viewer (IGV) (109). Early growth response 1 (EGR1)
561 chromatin immunoprecipitation and sequencing (ChIP-Seq) data from K562 cells were obtained from
562 the ENCODE database (66).

563 **Analyzing contact points in HiC datasets**

564 Genome organization and chromatin interaction data (HiC) were obtained from HepG2 cells
565 via ENCODE (66), and for HCC material (67), and visualized with 3D Genome Browser (68). To detect
566 potential interactions with surrounding genes, we used the coordinates chr7: 44,187,355 in hg19

567 (chr7:44,147,756; hg38) for *GCK* exon 7 and chr19:19,381,266 in hg19 (chr19:19,270,457; hg38) for
568 *TM6SF2* intron 2-exon 3 for virtual 4C analysis using HiC data in the 3D Genome Browser (68).

569 **Testing for regulatory functions using the luciferase reporter assay**

570 To verify the regulatory role of the DMPs identified in *GCK* exon 7 and *TM6SF2* intron 2-exon
571 3 boundary, the regions were cloned into pCpG-free vectors (Invivogen). The pCpG-free basic vector
572 was used to detect potential promoter activity and the pCpG-free promoter vector contains a minimal
573 promoter to test for enhancer activity of the regions of interest. Since these vectors do not contain
574 any CpG dinucleotides in the backbone, *in vitro* DNA methylation can be performed with the *M.SssI*
575 enzyme, which then selectively methylates only the inserted human *GCK* or *TM6SF2* region and allows
576 to selectively study the effect of 5mC in the insert on luciferase gene expression.
577 Briefly, DNA was isolated from HepG2 cells and the regions were amplified with PCR primers
578 containing restriction enzyme (RE) cutting sites for *NsiI* (New England Biolabs, NEB) and *BamHI*/*HindIII*
579 (NEB) at the 5´ends of the forward and reverse primers, respectively. Each 500 ng of PCR amplicons
580 and plasmid were double-digested at 37°C for 20 min and cleaned up with Promega DNA cleanup and
581 concentration system (Cat# A2893). Primer sequences, restriction enzymes and PCR conditions can be
582 found in the Supplement (Material & Methods section 1.2).
583 The ligation reaction was performed in a 1:5 plasmid to insert copy number ratio using T4 DNA
584 ligase (NEB) at 16°C overnight. 5 µL of the ligated construct were transferred to chemically competent
585 *E.coli* GT115 cells (*pir* mutant strain, Invivogen) and spread on lysogeny broth (LB) plates containing
586 zeocin (50 µL/ml; Invivogen). Singe colonies were picked and incubated in 3 ml LB-zeocin medium
587 overnight. Plasmid was extracted with Promega PureYield Plasmid Midiprep System and successful
588 cloning was verified by Sanger Sequencing at LGC genomics (Berlin). 1 µg of plasmid was incubated
589 with 8 U *M.SssI* (NEB) at 37°C for 4 h for *in vitro* DNA methylation. The reaction was heat deactivated
590 at 65°C for 20 min, followed by plasmid cleanup with Promega ReliaPrep DNA Clean-Up and
591 Concentration kit.
592 24 h prior to transfection, 0.15 x 10^5 HepG2 cells per well were seeded in 24 well plates with
593 DMEM low glucose medium + 10% fetal bovine serum without antibiotics. Lipofectamin 3000 (Thermo
594 Scientific, Cat # L3000001) was used to transfect 500 ng plasmid and 25 ng pGL4 SV40 background
595 plasmid (firefly luciferase) and incubated overnight. After cell lysis, chemiluminescent read-out was
596 detected on a Tecan Spark plate reader (Tecan). Renilla readouts were normalized to the firefly
597 luciferase and finally, data were expressed as fold-change normalized against the empty vector.

598 **Non-B DNA detection with permanganate/S1 nuclease footprinting with direct adapter ligation and**
599 **sequencing (PDAL-Seq)**

600 PDAL-Seq was performed according to the protocol described in (54). Further details can be
601 found in the Supplement (Material & Methods Section 1.3). In short, 5-7 x 10^6 cells were washed two
602 times with 1x PBS without Ca^{2+} and Mg^{2+} and incubated in 1 mL low salt buffer (15 mM Tris-HCl pH 7.5,
603 15 mM NaCl, 60 mM KCl, 300 mM sucrose, 0.5 mM EGTA) for 1 min. 670 µL of freshly prepared $KMnO_4$
604 stock solution (40 mM for HEK293, 70mM for MCF7) were added and incubated at 37°C in the dark for
605 80 sec. Subsequently, the reaction was stopped with 2.67 mL stop solution (50 mM EDTA, 700 mM β-
606 mercaptoethanol, 1% SDS) and 79.5 µL proteinase K (20 mg/ml stock solution) were added. After
607 overnight incubation at 37°C, DNA was extracted with phenol chloroform isoamylalcohol (PCI) and
608 precipitated with 2 M NH_4OAc and 100% ethanol. After resuspending the pellet in 1 mL of 10 mM Tris
609 HCl pH 8.0 buffer, RNA was digested by adding 2.5 µL of 20 mg/mL RNase A for 1 h followed by another
610 PCI extraction. Free DNA ends were blocked with 500 U Terminal Transferase, 120 µM cordycepin-5´-
611 triphosphate and 5 mM $CoCl_2$ in 500 µL total reaction volume, followed by a PCI extraction and two
612 ethanol precipitations to avoid cordycepin carry over. 40 µg DNA were digested with 20 U S1 nuclease
613 (0.5 U/1 µg DNA) in a total reaction volume of 600 µL at 37°C for 20 min. The reaction was stopped by
614 adding 10 mM EDTA. After PCI extraction, end repair using T4 DNA polymerase was performed for 15
615 min at 12°C, followed by PCI extraction. 500 nM biotinylated P5 adapter were linked with 2000 U T4
616 DNA ligase in a final volume of 100 µL at 16°C overnight. To remove unligated P5 adapters, Agencourt
617 SPRI bead cleanup was performed with 0.9 volumes of beads and eluted in 132 µL 1x TE buffer (10 mM
618 Tris HCl pH 8.0, 1mM EDTA). Sonication was performed with an Covaris M220 focused-ultrasonicator
619 in 130 µL Covaris microTube-AFA Fiber Pre-Slit Snap Cap (Cat #520045) tubes with the parameters:
620 peak incidence power: 75, duty factor: 10, cycles per burst: 200 and treatment time: 80 sec, to reach
621 a target peak of approximately 400 bp. Fragments were bound to streptavidin coated magnetic beads
622 (Dynabeads® M-280 kilobaseBINDER kit, Cat #60101) and carefully washed with 10 mM Tris-HCl pH
623 7.5. The beads were then resuspended in 100 µL end repair mix (NEBNext End Repair Module, Cat
624 #E6050) and incubated at 20°C for 30 min. After washing with 10 mM Tris-HCl pH 7.5, P7 adapter
625 ligation (1 µM adapters, 400 U T4 DNA ligase) was performed on a wheel tube rotator at 20°C for 4 h.
626 Beads were washed with 10 mM Tris-HCl pH 7.5 and resuspended in 100 µL PCR mix and PCR was
627 performed. Detailed information about PCR reaction and primers are provided in the Supplement
628 (Material and Methods section 1.3). SPRI bead purification removed unbound adapters and primer
629 dimers and the quality of the resulting libraries was checked using an Agilent Fragment Analyzer.
630 Library concentration was determined with KAPA Library Quantification Kit (Cat #KK4835) and
631 sequenced on a NovaSeq platform with approximately 80 million reads. After demultiplexing the raw
632 reads and quality control, reads were aligned to the human reference genome hg19 with the BWA-

633 MEM aligner (110). Subsequent filtering removed low-quality reads (mapping quality MQ <30) and
634 ENCODE blacklisted regions (66). BamCoverage files were generated with a 50 bp bin size and
635 normalized to 1× coverage and visualized in IGV (109). Data analysis was performed using the open-
636 source platform Galaxy (111).

637 **Studying *in vitro* G4 formation with native DNA polyacrylamide gels and circular dicroism spectra**

638 Target regions harboring the DMPs in *GCK* and *TM6SF2*, a positive control region in *MYC*
639 containing a predicted G4 motif (90), and a negative control region within a regulatory region of the
640 estrogen receptor substrate 1 (*ESR1)* without a predicted G4 motif (91) were PCR-amplified with
641 HotStar Taq Polymerase (Qiagen). The PCR fragments were cleaned-up with Promega ReliaPrep DNA
642 concentration kit and eluted in dH_2O. The amplicons were then diluted to a final concentration of 20
643 ng/µL in 50 mM Tris-HCl (pH 8.3). To enhance G4 folding, we prepared the same dilutions and added
644 150 mM KCl or 150 mM KCl and 10 µM pyridostatin (PDS) to the Tris-HCl buffer. The fragments were
645 then heated to 95°C and cooled gradually down to 25°C in a thermocycler overnight. DNA fragments
646 were loaded onto a native 10% DNA polyacrylamide (PAA) gel using 1× TBE buffer at 68 V (8 V/cm^2) for
647 2h. DNA PAA gels were post-stained with GelRed (Biotinum Cat #41003, 7 µL GelRed mixed with 15 ml
648 dH_2O) for 5 min and visualized with a BioRad ChemiDoc station. The detailed protocol can be found in
649 the Supplement (Material & Methods Section 1.4).
650 Circular dichroism (CD) spectra were obtained the same PCR amplicons, except for *ESR1*. This region
651 did not result in a sufficient amount of PCR product. Therefore, we used a 20 bp oligo within the *ESR1*
652 region in a concentration of 1 µM as negative control. All amplicons, including the *ESR1* oligo, were
653 reconstituted in 50 mM Tris-HCl (pH 8.3) and 150 mM KCl in a final volume of 200 µL. CD spectra were
654 recorded using the Chirascan™ Plus CD Spectrophotometer instrument (Applied Photophysics) in a 0.5
655 cm path cuvette at 25°C. Spectra were scanned from 340 to 220 nm., with a step size of 0.2 nm and a
656 bandwidth of 1 nm. Four spectra per sample were averaged. After blank correction, the CD ellipticity
657 θ in millidegrees was normalized using the relation $\theta norm = \theta / (32980 \times c \times l)$, where c is the DNA
658 concentration in mol/L and l is the path length in cm.

659 **Statistical analyses**

660 All statistical analyses were performed using R version 4.3.1 (106) and R Studio (107).
661 Significant changes in DNA methylation in all groups were detected using analysis of variance (ANOVA)
662 with $p<.05$. For individual group comparisons, Student´s t-test ($p<.05$) was applied. Welch´s two sample
663 t-tests were performed to calculate significant differences in luciferase reporter assays ($p<.05$).

664 **Author contribution**

665 A. L. established PDAL Seq, designed assays and conceptualized the study; A. L., V. E., A. O., Z.
666 L. and E. S. performed experiments, B. P. and E. A. collected and grouped the samples and gave
667 substantial advice and intellectual input to the study; A. L. wrote the manuscript, A. L., E.A. and A.R.
668 revised and edited the manuscript. All authors read and approved the final manuscript.

669 **Funding**

670 A.L was supported by a *Starter Grant* awarded by the Austrian Diabetes Association (ÖDG).
671 This research was in part funded by the County of Salzburg, Cancer Cluster Salzburg [grant number
672 20102-P1601064-FPR].

673 **Acknowledgments**

674 We want to thank the Austrian Diabetes Association (ÖDG) very warmly for their support as
675 well as all patients who contributed their blood samples to the study. Furthermore, we thank Kateryna
676 Makova and Matthias Weissensteiner for plenty of fruitful discussions and their input on G-quadruplex
677 structures.

678 **Data accessibility**

679 PDAL-Seq data are available at the Sequence Read Archive (SRA) (112) under the BioProject
680 number PRJNA1088738.

681 **Conflicts of Interest**

682 A. L., V. E., A. O., Z. L., E. S., B. P., E. A. and A. R. do not have any conflicts of interest.

683 **Abbreviations**

5mC	5-methylcytosine
ChIP-Seq	Chromatin immunoprecipitation and sequencing
DMP	Differentially methylated position
DMR	Differentially methylated region
G4	G-quadruplex
GCK	Glucokinase
HiC/4C	Chromatin conformation capture
MAFLD	Metabolism dysfunction-associated fatty liver disease
MetD	Metabolic disease

MetS	Metabolic syndrome
Non-B DNA	Non-canonical DNA structure
PDAL-Seq	Permanganate/S1 nuclease footprinting with direct adapter ligation
PQS	Putative G4 forming sequence
R-loop	RNA:DNA hybrid
RLFS	R-loop forming sequence
T2DM	Type 2 diabetes mellitus
TF	Transcription factor
TM6SF2	Transmembrane 6 superfamily 2

686 **References**

687

688 1. WHO. Obesity and overweight, Fact Sheets, https://www.who.int/en/news-room/fact-
689 sheets/detail/obesity-and-overweight 2020 [updated 01.04.2020.
690 2. Koenen M, Hill MA, Cohen P, Sowers JR. Obesity, Adipose Tissue and Vascular Dysfunction.
691 Circ Res. 2021;128(7):951-68.
692 3. Collaboration NCDRF. Worldwide trends in body-mass index, underweight, overweight, and
693 obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in
694 128.9 million children, adolescents, and adults. Lancet. 2017;390(10113):2627-42.
695 4. Samblas M, Milagro FI, Martinez A. DNA methylation markers in obesity, metabolic
696 syndrome, and weight loss. Epigenetics. 2019;14(5):421-44.
697 5. Fabbrini E, Sullivan S, Klein S. Obesity and nonalcoholic fatty liver disease: biochemical,
698 metabolic, and clinical implications. Hepatology. 2010;51(2):679-89.
699 6. Yamamura S, Eslam M, Kawaguchi T, Tsutsumi T, Nakano D, Yoshinaga S, et al. MAFLD
700 identifies patients with significant hepatic fibrosis better than NAFLD. Liver Int. 2020;40(12):3018-30.
701 7. Afshin A, Reitsma MB, Murray CJL. Health Effects of Overweight and Obesity in 195
702 Countries. N Engl J Med. 2017;377(15):1496-7.
703 8. Chen Y, Liu L, Wang X, Wang J, Yan Z, Cheng J, et al. Body mass index and risk of gastric
704 cancer: a meta-analysis of a population with more than ten million from 24 prospective studies.
705 Cancer Epidemiol Biomarkers Prev. 2013;22(8):1395-408.
706 9. Genkinger JM, Spiegelman D, Anderson KE, Bernstein L, van den Brandt PA, Calle EE, et al. A
707 pooled analysis of 14 cohort studies of anthropometric factors and pancreatic cancer risk. Int J
708 Cancer. 2011;129(7):1708-17.
709 10. Wang Y, Wang B, Shen F, Fan J, Cao H. Body mass index and risk of primary liver cancer: a
710 meta-analysis of prospective studies. Oncologist. 2012;17(11):1461-8.
711 11. Ma Y, Yang Y, Wang F, Zhang P, Shi C, Zou Y, et al. Obesity and risk of colorectal cancer: a
712 systematic review of prospective studies. PLoS One. 2013;8(1):e53916.
713 12. Carreras-Torres R, Johansson M, Haycock PC, Wade KH, Relton CL, Martin RM, et al. Obesity,
714 metabolic factors and risk of different histological types of lung cancer: A Mendelian randomization
715 study. PLoS One. 2017;12(6):e0177875.
716 13. Li M, Cao SM, Dimou N, Wu L, Li JB, Yang J. Association of Metabolic Syndrome With Risk of
717 Lung Cancer: A Population-Based Prospective Cohort Study. Chest. 2023.
718 14. Grasselli E, Canesi L, Portincasa P, Voci A, Vergani L, Demori I. Models of non-Alcoholic Fatty
719 Liver Disease and Potential Translational Value: the Effects of 3,5-L-diiodothyronine. Ann Hepatol.
720 2017;16(5):707-19.
721 15. Younossi Z, Anstee QM, Marietti M, Hardy T, Henry L, Eslam M, et al. Global burden of NAFLD
722 and NASH: trends, predictions, risk factors and prevention. Nat Rev Gastroenterol Hepatol.
723 2018;15(1):11-20.
724 16. Roden M. Mechanisms of Disease: hepatic steatosis in type 2 diabetes--pathogenesis and
725 clinical relevance. Nat Clin Pract Endocrinol Metab. 2006;2(6):335-48.
726 17. Gaggini M, Morelli M, Buzzigoli E, DeFronzo RA, Bugianesi E, Gastaldelli A. Non-alcoholic fatty
727 liver disease (NAFLD) and its connection with insulin resistance, dyslipidemia, atherosclerosis and
728 coronary heart disease. Nutrients. 2013;5(5):1544-60.
729 18. Geltinger F, Schartel L, Wiederstein M, Tevini J, Aigner E, Felder TK, et al. Friend or Foe: Lipid
730 Droplets as Organelles for Protein and Lipid Storage in Cellular Stress Response, Aging and Disease.
731 Molecules. 2020;25(21).
732 19. Chalasani N, Younossi Z, Lavine JE, Diehl AM, Brunt EM, Cusi K, et al. The diagnosis and
733 management of non-alcoholic fatty liver disease: practice Guideline by the American Association for
734 the Study of Liver Diseases, American College of Gastroenterology, and the American
735 Gastroenterological Association. Hepatology. 2012;55(6):2005-23.

736 20. Ma J, Nano J, Ding J, Zheng Y, Hennein R, Liu C, et al. A Peripheral Blood DNA Methylation
737 Signature of Hepatic Fat Reveals a Potential Causal Pathway for Nonalcoholic Fatty Liver Disease.
738 Diabetes. 2019;68(5):1073-83.
739 21. Gerhard GS, Malenica I, Llaci L, Chu X, Petrick AT, Still CD, et al. Differentially methylated loci
740 in NAFLD cirrhosis are associated with key signaling pathways. Clin Epigenetics. 2018;10(1):93.
741 22. Ahrens M, Ammerpohl O, von Schonfels W, Kolarova J, Bens S, Itzel T, et al. DNA methylation
742 analysis in nonalcoholic fatty liver disease suggests distinct disease-specific and remodeling
743 signatures after bariatric surgery. Cell Metab. 2013;18(2):296-302.
744 23. Kirchner H, Sinha I, Gao H, Ruby MA, Schonke M, Lindvall JM, et al. Altered DNA methylation
745 of glycolytic and lipogenic genes in liver from obese and type 2 diabetic patients. Mol Metab.
746 2016;5(3):171-83.
747 24. Nano J, Ghanbari M, Wang W, de Vries PS, Dhana K, Muka T, et al. Epigenome-Wide
748 Association Study Identifies Methylation Sites Associated With Liver Enzymes and Hepatic Steatosis.
749 Gastroenterology. 2017;153(4):1096-106 e2.
750 25. Rauchhaus J, Robinson J, Monti L, Di Antonio M. G-quadruplexes Mark Sites of Methylation
751 Instability Associated with Ageing and Cancer. Genes (Basel). 2022;13(9).
752 26. Mao SQ, Ghanbarian AT, Spiegel J, Martinez Cuesta S, Beraldi D, Di Antonio M, et al. DNA G-
753 quadruplex structures mold the DNA methylome. Nat Struct Mol Biol. 2018;25(10):951-7.
754 27. Williamson JR, Raghuraman MK, Cech TR. Monovalent cation-induced structure of telomeric
755 DNA: the G-quartet model. Cell. 1989;59(5):871-80.
756 28. Sen D, Gilbert W. Formation of parallel four-stranded complexes by guanine-rich motifs in
757 DNA and its implications for meiosis. Nature. 1988;334(6180):364-6.
758 29. Haider S, Parkinson GN, Neidle S. Crystal structure of the potassium form of an Oxytricha
759 nova G-quadruplex. J Mol Biol. 2002;320(2):189-200.
760 30. Frasson I, Pirota V, Richter SN, Doria F. Multimeric G-quadruplexes: A review on their
761 biological roles and targeting. Int J Biol Macromol. 2022;204:89-102.
762 31. Varshney D, Spiegel J, Zyner K, Tannahill D, Balasubramanian S. The regulation and functions
763 of DNA and RNA G-quadruplexes. Nat Rev Mol Cell Biol. 2020;21(8):459-74.
764 32. Makova KD, Weissensteiner MH. Noncanonical DNA structures are drivers of genome
765 evolution. Trends Genet. 2023;39(2):109-24.
766 33. Duquette ML, Handa P, Vincent JA, Taylor AF, Maizels N. Intracellular transcription of G-rich
767 DNAs induces formation of G-loops, novel structures containing G4 DNA. Genes Dev.
768 2004;18(13):1618-29.
769 34. Niehrs C, Luke B. Regulatory R-loops as facilitators of gene expression and genome stability.
770 Nat Rev Mol Cell Biol. 2020;21(3):167-78.
771 35. Robinson J, Raguseo F, Nuccio SP, Liano D, Di Antonio M. DNA G-quadruplex structures: more
772 than simple roadblocks to transcription? Nucleic Acids Res. 2021;49(15):8419-31.
773 36. Wulfridge P, Yan Q, Rell N, Doherty J, Jacobson S, Offley S, et al. G-quadruplexes associated
774 with R-loops promote CTCF binding. Mol Cell. 2023;83(17):3064-79 e5.
775 37. Raiber EA, Kranaster R, Lam E, Nikan M, Balasubramanian S. A non-canonical DNA structure is
776 a binding motif for the transcription factor SP1 in vitro. Nucleic Acids Res. 2012;40(4):1499-508.
777 38. Spiegel J, Cuesta SM, Adhikari S, Hansel-Hertsch R, Tannahill D, Balasubramanian S. G-
778 quadruplexes are transcription factor binding hubs in human chromatin. Genome Biol.
779 2021;22(1):117.
780 39. Smith SS, Kan JL, Baker DJ, Kaplan BE, Dembek P. Recognition of unusual DNA structures by
781 human DNA (cytosine-5)methyltransferase. J Mol Biol. 1991;217(1):39-51.
782 40. Osbak KK, Colclough K, Saint-Martin C, Beer NL, Bellanne-Chantelot C, Ellard S, et al. Update
783 on mutations in glucokinase (GCK), which cause maturity-onset diabetes of the young, permanent
784 neonatal diabetes, and hyperinsulinemic hypoglycemia. Hum Mutat. 2009;30(11):1512-26.
785 41. Thuesen ACB, Jensen RT, Maagensen H, Kristiansen MR, Sorensen HT, Vaag A, et al.
786 Identification of pathogenic GCK variants in patients with common type 2 diabetes can lead to
787 discontinuation of pharmacological treatment. Mol Genet Metab Rep. 2023;35:100972.

788 42. Hattersley AT, Turner RC, Permutt MA, Patel P, Tanizawa Y, Chiu KC, et al. Linkage of type 2
789 diabetes to the glucokinase gene. Lancet. 1992;339(8805):1307-10.
790 43. Liu J, Fu H, Kang F, Ning G, Ni Q, Wang W, et al. beta-Cell glucokinase expression was
791 increased in type 2 diabetes subjects with better glycemic control. J Diabetes. 2023;15(5):409-18.
792 44. Kozlitina J, Smagris E, Stender S, Nordestgaard BG, Zhou HH, Tybjaerg-Hansen A, et al.
793 Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic
794 fatty liver disease. Nat Genet. 2014;46(4):352-6.
795 45. Liu YL, Reeves HL, Burt AD, Tiniakos D, McPherson S, Leathart JB, et al. TM6SF2 rs58542926
796 influences hepatic fibrosis progression in patients with non-alcoholic fatty liver disease. Nat
797 Commun. 2014;5:4309.
798 46. Luo F, Oldoni F, Das A. TM6SF2: A Novel Genetic Player in Nonalcoholic Fatty Liver and
799 Cardiovascular Disease. Hepatol Commun. 2022;6(3):448-60.
800 47. Milano M, Aghemo A, Mancina RM, Fischer J, Dongiovanni P, De Nicola S, et al.
801 Transmembrane 6 superfamily member 2 gene E167K variant impacts on steatosis and liver damage
802 in chronic hepatitis C patients. Hepatology. 2015;62(1):111-7.
803 48. Holmen OL, Zhang H, Fan Y, Hovelson DH, Schmidt EM, Zhou W, et al. Systematic evaluation
804 of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and
805 myocardial infarction risk. Nat Genet. 2014;46(4):345-51.
806 49. Xie Z, Xie T, Liu J, Zhang Q, Xiao X. Glucokinase Inactivation Ameliorates Lipid Accumulation
807 and Exerts Favorable Effects on Lipid Metabolism in Hepatocytes. Int J Mol Sci. 2023;24(5).
808 50. Matschinsky FM, Wilson DF. The Central Role of Glucokinase in Glucose Homeostasis: A
809 Perspective 50 Years After Demonstrating the Presence of the Enzyme in Islets of Langerhans. Front
810 Physiol. 2019;10:148.
811 51. Peter A, Stefan N, Cegan A, Walenta M, Wagner S, Konigsrainer A, et al. Hepatic glucokinase
812 expression is associated with lipogenesis and fatty liver in humans. J Clin Endocrinol Metab.
813 2011;96(7):E1126-30.
814 52. Li TT, Li TH, Peng J, He B, Liu LS, Wei DH, et al. TM6SF2: A novel target for plasma lipid
815 regulation. Atherosclerosis. 2018;268:170-6.
816 53. Mahdessian H, Taxiarchis A, Popov S, Silveira A, Franco-Cereceda A, Hamsten A, et al.
817 TM6SF2 is a regulator of liver fat metabolism influencing triglyceride secretion and hepatic lipid
818 droplet content. Proc Natl Acad Sci U S A. 2014;111(24):8913-8.
819 54. Lahnsteiner A, Craig SJC, Kamali K, Weissensteiner B, McGrath B, Risch A, et al. In vivo
820 detection of DNA secondary structures using Permanganate/S1 Footprinting with Direct Adapter
821 Ligation and Sequencing (PDAL-Seq). Methods in Enzymology. 2024.
822 55. Bresgen N, Kovacs M, Lahnsteiner A, Felder TK, Rinnerthaler M. The Janus-Faced Role of Lipid
823 Droplets in Aging: Insights from the Cellular Perspective. Biomolecules. 2023;13(6).
824 56. ENSEMBL. https://www.ensembl.org/index.html 2013 [
825 57. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome
826 browser at UCSC. Genome Res. 2002;12(6):996-1006.
827 58. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol.
828 2013;14(10):R115.
829 59. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation
830 profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49(2):359-67.
831 60. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, et al. An epigenetic biomarker
832 of aging for lifespan and healthspan. Aging (Albany NY). 2018;10(4):573-91.
833 61. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger
834 BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-20.
835 62. Ilieva M, Dao J, Miller HE, Madsen JH, Bishop AJR, Kauppinen S, et al. Systematic Analysis of
836 Long Non-Coding RNA Genes in Nonalcoholic Fatty Liver Disease. Noncoding RNA. 2022;8(4).
837 63. Fagerberg L, Hallstrom BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of
838 the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-
839 based proteomics. Mol Cell Proteomics. 2014;13(2):397-406.

840 64. Zou W, Wang Z, Wang F, Li L, Liu R, Hu M. A metabolism-related 4-lncRNA prognostic
841 signature and corresponding mechanisms in intrahepatic cholangiocarcinoma. BMC Cancer.
842 2021;21(1):608.
843 65. Pan JQ, Zhang YQ, Wang JH, Xu P, Wang W. lncRNA co-expression network model for the
844 prognostic analysis of acute myeloid leukemia. Int J Mol Med. 2017;39(3):663-71.
845 66. ENCODE. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the
846 human genome. Nature. 2012;489(7414):57-74.
847 67. Xu Z, Lee DS, Chandran S, Le VT, Bump R, Yasis J, et al. Structural variants drive context-
848 dependent oncogene activation in cancer. Nature. 2022;612(7940):564-72.
849 68. Wang Y, Song F, Zhang B, Zhang L, Xu J, Kuang D, et al. The 3D Genome Browser: a web-based
850 browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol.
851 2018;19(1):151.
852 69. Lou Y, Chen YD, Sun FR, Shi JP, Song Y, Yang J. Potential Regulators Driving the Transition in
853 Nonalcoholic Fatty Liver Disease: a Stage-Based View. Cell Physiol Biochem. 2017;41(1):239-51.
854 70. Zhang W, Li YJ, Zhang N, Chen SY, Tong XF, Wang BQ, et al. Fibroblast-specific adipocyte
855 enhancer binding protein 1 is a potential pathological trigger and prognostic marker for liver fibrosis
856 independent of etiology. J Dig Dis. 2023;24(10):550-61.
857 71. Ho CY, Lee JI, Huang SP, Chen SC, Geng JH. A Genome-Wide Association Study of Metabolic
858 Syndrome in the Taiwanese Population. Nutrients. 2023;16(1).
859 72. Xu JZ, Jiang JJ, Xu HJ, Sun XD, Liu ZC, Hu ZM. High expression of YKT6 associated with
860 progression and poor prognosis of hepatocellular carcinoma. Scand J Gastroenterol.
861 2021;56(11):1349-54.
862 73. Kubickova KN, Subhanova I, Konickova R, Matousova L, Urbanek P, Parobkova H, et al.
863 Predictive role BLVRA mRNA expression in hepatocellular cancer. Ann Hepatol. 2016;15(6):881-7.
864 74. Tufan NL, Lian Z, Liu J, Pan J, Arbuthnot P, Kew M, et al. Hepatitis Bx antigen stimulates
865 expression of a novel cellular gene, URG4, that promotes hepatocellular growth and survival.
866 Neoplasia. 2002;4(4):355-68.
867 75. Dodurga Y, Secme M, Lale Satiroglu-Tufan N. A novel oncogene URG4/URGCP and its role in
868 cancer. Gene. 2018;668:12-7.
869 76. Okuda J, Niizuma S, Shioi T, Kato T, Inuzuka Y, Kawashima T, et al. Persistent overexpression
870 of phosphoglycerate mutase, a glycolytic enzyme, modifies energy metabolism and reduces stress
871 resistance of heart in mice. PLoS One. 2013;8(8):e72173.
872 77. Toyoda Y, Takada T, Umezawa M, Tomura F, Yamanashi Y, Takeda K, et al. Identification of
873 hepatic NPC1L1 as an NAFLD risk factor evidenced by ezetimibe-mediated steatosis prevention and
874 recovery. FASEB Bioadv. 2019;1(5):283-95.
875 78. Wang Y, Chen L, Pandak WM, Heuman D, Hylemon PB, Ren S. High Glucose Induces Lipid
876 Accumulation via 25-Hydroxycholesterol DNA-CpG Methylation. iScience. 2020;23(5):101102.
877 79. Cong F, Long J, Liu J, Deng Z, Yan B, Liang C, et al. An integrative analysis revealing POLD2 as a
878 tumor suppressive immune protein and prognostic biomarker in pan-cancer. Front Genet.
879 2022;13:877468.
880 80. Zhou Y, Li Y, Wang N, Li X, Zheng J, Ge L. UPF1 inhibits the hepatocellular carcinoma
881 progression by targeting long non-coding RNA UCA1. Sci Rep. 2019;9(1):6652.
882 81. Kim MJ, Yu CY, Theusch E, Naidoo D, Stevens K, Kuang YL, et al. SUGP1 is a novel regulator of
883 cholesterol metabolism. Hum Mol Genet. 2016;25(14):3106-16.
884 82. Burge S, Parkinson GN, Hazel P, Todd AK, Neidle S. Quadruplex DNA: sequence, topology and
885 structure. Nucleic Acids Res. 2006;34(19):5402-15.
886 83. Jenjaroenpun P, Wongsurawat T, Yenamandra SP, Kuznetsov VA. QmRLFS-finder: a model,
887 web server and stand-alone tool for prediction and analysis of R-loop forming sequences. Nucleic
888 Acids Res. 2015;43(20):10081.
889 84. Hon J, Martinek T, Zendulka J, Lexa M. pqsfinder: an exhaustive and imperfection-tolerant
890 search tool for potential quadruplex-forming sequences in R. Bioinformatics. 2017;33(21):3373-9.

891 85. Marsico G, Chambers VS, Sahakyan AB, McCauley P, Boutell JM, Antonio MD, et al. Whole
892 genome experimental maps of DNA G-quadruplexes in multiple species. Nucleic Acids Res.
893 2019;47(8):3862-74.
894 86. Li C, Wang H, Yin Z, Fang P, Xiao R, Xiang Y, et al. Ligand-induced native G-quadruplex
895 stabilization impairs transcription initiation. Genome Res. 2021;31(9):1546-60.
896 87. Kouzine F, Wojtowicz D, Baranello L, Yamane A, Nelson S, Resch W, et al. Permanganate/S1
897 Nuclease Footprinting Reveals Non-B DNA Structures with Regulatory Potential across a Mammalian
898 Genome. Cell Syst. 2017;4(3):344-56 e7.
899 88. Georgakopoulos-Soares I, Parada GE, Wong HY, Medhi R, Furlan G, Munita R, et al.
900 Alternative splicing modulation by G-quadruplexes. Nat Commun. 2022;13(1):2404.
901 89. Brazda V, Cerven J, Bartas M, Mikyskova N, Coufal J, Pecinka P. The Amino Acid Composition
902 of Quadruplex Binding Proteins Reveals a Shared Motif and Predicts New Potential Quadruplex
903 Interactors. Molecules. 2018;23(9).
904 90. Zheng KW, Chen Z, Hao YH, Tan Z. Molecular crowding creates an essential environment for
905 the formation of stable G-quadruplexes in long double-stranded DNA. Nucleic Acids Res.
906 2010;38(1):327-38.
907 91. Lam EY, Beraldi D, Tannahill D, Balasubramanian S. G-quadruplex structures are stable and
908 detectable in human genomic DNA. Nat Commun. 2013;4:1796.
909 92. Monsen RC, DeLeeuw LW, Dean WL, Gray RD, Chakravarthy S, Hopkins JB, et al. Long
910 promoter sequences form higher-order G-quadruplexes: an integrative structural biology study of c-
911 Myc, k-Ras and c-Kit promoter sequences. Nucleic Acids Res. 2022;50(7):4127-47.
912 93. Xu L, Zheng D, Wang L, Jiang D, Liu H, Xu L, et al. GCK gene-body hypomethylation is
913 associated with the risk of coronary heart disease. Biomed Res Int. 2014;2014:151723.
914 94. Faltus C, Lahnsteiner A, Barrdahl M, Assenov Y, Husing A, Bogatyrova O, et al. Identification
915 of NHLRC1 as a Novel AKT Activator from a Lung Cancer Epigenome-Wide Association Study (EWAS).
916 Int J Mol Sci. 2022;23(18).
917 95. Perrin-Cocon L, Vidalain PO, Jacquemin C, Aublin-Gex A, Olmstead K, Panthu B, et al. A
918 hexokinase isoenzyme switch in human liver cancer cells promotes lipogenesis and enhances innate
919 immunity. Commun Biol. 2021;4(1):217.
920 96. Fujii T, Podbevsek P, Plavec J, Sugimoto N. Effects of metal ions and cosolutes on G-
921 quadruplex topology. J Inorg Biochem. 2017;166:190-8.
922 97. Nishio M, Tsukakoshi K, Ikebukuro K. G-quadruplex: Flexible conformational changes by
923 cations, pH, crowding and its applications to biosensing. Biosens Bioelectron. 2021;178:113030.
924 98. Guneri D, Alexandrou E, Omari KE, Dvořáková Z, Chikhale RV, Pike D, et al. Structural Insights
925 into Regulation of Insulin Expression Involving i-Motif DNA Structures in the Insulin-Linked
926 Polymorphic Region. bioRxiv. 2023:2023.06.01.543149.
927 99. Zyner KG, Simeone A, Flynn SM, Doyle C, Marsico G, Adhikari S, et al. G-quadruplex DNA
928 structures in human stem cells and differentiation. Nat Commun. 2022;13(1):142.
929 100. Biffi G, Tannahill D, Miller J, Howat WJ, Balasubramanian S. Elevated levels of G-quadruplex
930 formation in human stomach and liver cancer tissues. PLoS One. 2014;9(7):e102711.
931 101. Hansel-Hertsch R, Beraldi D, Lensing SV, Marsico G, Zyner K, Parry A, et al. G-quadruplex
932 structures mark human regulatory chromatin. Nat Genet. 2016;48(10):1267-72.
933 102. Lago S, Nadai M, Cernilogar FM, Kazerani M, Dominiguez Moreno H, Schotta G, et al.
934 Promoter G-quadruplexes and transcription factors cooperate to shape the cell type-specific
935 transcriptome. Nat Commun. 2021;12(1):3885.
936 103. Cao X, Mahendran R, Guy GR, Tan YH. Detection and characterization of cellular EGR-1
937 binding to its recognition site. J Biol Chem. 1993;268(23):16949-57.
938 104. Wang Q, Conlon EG, Manley JL, Rio DC. Widespread intron retention impairs protein
939 homeostasis in C9orf72 ALS brains. Genome Res. 2020;30(12):1705-15.
940 105. Grundy SM, Cleeman JI, Daniels SR, Donato KA, Eckel RH, Franklin BA, et al. Diagnosis and
941 management of the metabolic syndrome: an American Heart Association/National Heart, Lung, and
942 Blood Institute scientific statement: Executive Summary. Crit Pathw Cardiol. 2005;4(4):198-203.
943 106. Team RC. R: A Language and Environment for Statistical Computing. 2023.

107. Team RS. RStudio: Integrated Development for R. 2020.
108. Li Y, Ge D, Lu C. The SMART App: an interactive web application for comprehensive DNA methylation analysis and visualization. Epigenetics Chromatin. 2019;12(1):71.
109. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24-6.
110. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
111. Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. Nucleic Acids Res. 2020;48(W1):W395-W402.
112. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. The sequence read archive. Nucleic Acids Res. 2011;39(Database issue):D19-21.