



HAL
open science

LncRNA analyses reveal increased levels of non-coding centromeric transcripts in hepatocellular carcinoma

Anamaria Necsulea, Philippe Veber, Tuyana Boldanova, Charlotte K Y Ng, Stefan Wieland, Markus H Heim

► **To cite this version:**

Anamaria Necsulea, Philippe Veber, Tuyana Boldanova, Charlotte K Y Ng, Stefan Wieland, et al.. LncRNA analyses reveal increased levels of non-coding centromeric transcripts in hepatocellular carcinoma. 2024. hal-04604260

HAL Id: hal-04604260

<https://cnrs.hal.science/hal-04604260v1>

Preprint submitted on 7 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **LncRNA analyses reveal increased levels of non-coding centromeric transcripts in**
2 **hepatocellular carcinoma**

3
4 Anamaria Necsulea^{1,*}, Philippe Veber¹, Tuyana Boldanova^{2,3}, Charlotte K Y Ng^{4,5}, Stefan
5 Wieland², Markus H Heim^{2,3,**}

6
7 ¹ Univ Lyon, Université Claude Bernard Lyon 1, CNRS, Laboratoire de Biométrie et Biologie
8 Évolutive, F-69100, Villeurbanne, France.

9 ² Department of Biomedicine, University Hospital Basel, University of Basel, Basel,
10 Switzerland

11 ³ Clarunis, University Center for Gastrointestinal and Liver Disease, Basel, Switzerland

12 ⁴ Institute of Pathology, University Hospital Basel, University of Basel, 4031 Basel, Switzerland

13 ⁵ Department for BioMedical Research (DBMR), Oncogenomics lab, University of Bern,
14 Switzerland

15
16 *Corresponding author. Tel: +33 4 72 43 35 82; E-mail: anamaria.necsulea@univ-lyon1.fr

17 **Corresponding author. Tel: +41 61 265 25 25; E-mail: markus.heim@unibas.ch

18
19
20
21 **Funding**

22 This work was supported by European Research Council Synergy grant 609883

23 Mechanisms of Evasive Resistance in Cancer (MERiC), by The Swiss Initiative in Systems
24 Biology grant SystemX – MERiC to M.H.H and by the Agence Nationale pour la Recherche
25 (ANR JCJC 2017 LncEvoSys). C.K.Y.N. is supported by the Swiss Cancer League (KFS-
26 4543-08-2018).

31 **Abstract**

32 The search for new biomarkers and drug targets for hepatocellular carcinoma (HCC) has
33 spurred an interest in long non-coding RNAs (lncRNAs), often proposed as oncogenes or
34 tumor suppressors. Furthermore, lncRNA expression patterns can bring insights into the
35 global de-regulation of cellular machineries in tumors. Here, we examine lncRNAs in a large
36 HCC cohort, comprising RNA-seq data from paired tumor and adjacent tissue biopsies from
37 114 patients. We find that numerous lncRNAs are differentially expressed between tumors
38 and adjacent tissues and between tumor progression stages. Although we find strong
39 differential expression for most lncRNAs previously associated with HCC, the expression
40 patterns of several prominent HCC-associated lncRNAs disagree with their previously
41 proposed roles. We examine the genomic characteristics of HCC-expressed lncRNAs and
42 reveal an enrichment for repetitive elements among the lncRNAs with the strongest
43 expression increases in advanced-stage tumors. This enrichment is particularly striking for
44 lncRNAs that overlap with satellite repeats, a major component of centromeres. Consistently,
45 we find increased non-coding RNA transcription from centromeres in tumors, in the majority
46 of patients, suggesting that aberrant centromere activation takes place in HCC.

47 **Introduction**

48 Following the realization that the human genome harbors thousands of non-coding RNA
49 genes (Carninci *et al*, 2005), many of which have important cellular functions (Mattick &
50 Makunin, 2006), a great deal of effort has been put into investigating the contributions of non-
51 coding RNAs to cancer biology (Gutschner & Diederichs, 2012). In particular, the roles of long
52 non-coding RNAs (lncRNAs) in cancer have been frequently scrutinized in the past decade.
53 This category of non-coding RNAs (defined simply as RNA molecules that lack protein-coding
54 capacity, at least 200 nucleotides long) comprises many transcripts with proven functions in
55 gene expression regulation, genome stability or nuclear architecture (Engreitz *et al*, 2016).
56 Numerous recent studies showed that lncRNA loci are part of the alterations that occur in
57 cancer cells (Yan *et al*, 2015). Thus, studying lncRNAs is perceived as a promising path
58 towards understanding the molecular mechanisms that underlie cancer onset. Ultimately,
59 lncRNAs may prove to be valuable in the diagnosis process, or serve as therapeutic targets.

60
61 The search for novel disease biomarkers and drug targets, including lncRNAs, is
62 understandably intensive for cancer types for which effective therapies are still lacking. This
63 is the case for hepatocellular carcinoma (HCC), which is a major cause of cancer-related
64 mortality world-wide (Yang & Roberts, 2010). As HCC is generally detected at late stages of
65 tumor progression, surgical treatment options are unavailable for the majority of patients
66 (Hartke *et al*, 2017). Several systemic therapies now exist, but they increase median patient

67 survival by less than 1 year (Finn *et al*, 2020). Thus, developing new treatments for HCC is
68 still an urgent need. With this aim, there has been extensive research aiming to identify the
69 genic and non-genic functional elements that are altered in HCC compared to the healthy liver.
70 Large-scale transcriptomics studies, comparing HCC samples with adjacent non-tumor tissue
71 or with normal liver samples, identified hundreds of differentially regulated protein-coding
72 genes and lncRNAs (Cui *et al*, 2017; Yang *et al*, 2017; Li *et al*, 2019; Jin *et al*, 2019; Unfried
73 *et al*, 2019). Some of the lncRNAs associated with HCC through genome-wide comparative
74 analyses were subject to further experimental investigations, aiming to elucidate their
75 mechanisms of action and the consequences of their differential regulation in tumors. For
76 some lncRNAs, there are now well-supported models for their behavior in HCC. This is the
77 case for example for *HOTTIP*, a lncRNA that is strongly up-regulated in HCC, and which likely
78 acts to enhance the expression of the neighboring genes by recruiting transcriptional co-
79 activators (Quagliata *et al*, 2014; Pradeepa *et al*, 2017; Quagliata *et al*, 2018). However, for
80 other lncRNAs experimental studies gave rise to conflicting results. For example, the *H19*
81 lncRNA (known as a parentally imprinted regulator of placenta growth (Keniry *et al*, 2012))
82 was alternatively proposed to act as a tumor suppressor (Hao *et al*, 1993; Yoshimizu *et al*,
83 2008; Schultheiss *et al*, 2017) or as an oncogene (Matouk *et al*, 2007; Zhou *et al*, 2019) in
84 various cancer types including HCC (Tietze & Kessler, 2020). Likewise, *MALAT1*, initially
85 described as an abundant lncRNA associated with the presence of metastases (Ji *et al*, 2003),
86 was first thought to promote tumor growth and invasion in breast cancer (Arun *et al*, 2016),
87 but is now believed to be a tumor suppressor (Kim *et al*, 2018, 1). In HCC, *MALAT1* was
88 mainly proposed to act as an oncogene (Hou *et al*, 2017, 1; Liu *et al*, 2019, 1; Chen *et al*,
89 2020), but there is no consensus on its mechanisms of action. This is also the case for most
90 of the lncRNAs that have been associated with HCC, although experimental data is
91 accumulating (Lanzafame *et al*, 2018). Thus, overall, the functions of lncRNAs in HCC and
92 other cancers are still poorly understood.

93

94 Although we are still far from developing therapies that target lncRNAs in HCC, in the more
95 immediate future, lncRNAs may prove to be useful as disease biomarkers, to help diagnose
96 HCC at an earlier stage and to better classify molecular subtypes of tumors. For this purpose,
97 large-scale transcriptomics comparisons that can identify differentially regulated lncRNAs in
98 tumor tissues are a valid approach, even in the absence of additional functional experiments.
99 Although such studies are abundant in the lncRNA literature, they are often restricted to small
100 cohorts of patients, thus potentially failing to reproduce the full extent of the molecular
101 heterogeneity of HCC (Boyault *et al*, 2007; Hoshida *et al*, 2009). Further work is still needed
102 to understand what part lncRNAs play in the molecular characteristics of HCC.

103

104 Studying lncRNA expression patterns in HCC and other cancers is also a means to better
105 understand the de-regulation of essential cellular machineries in tumors. Although many
106 lncRNAs have important biological roles (Engreitz *et al*, 2016), there is strong evidence that,
107 out of the tens of thousands of lncRNAs that are detected with sensitive transcriptome
108 sequencing approaches in human tissues (Pertea *et al*, 2018; Iyer *et al*, 2015), most may be
109 non-functional. This is indicated by their weak levels of evolutionary conservation (Necsulea
110 *et al*, 2014; Washietl *et al*, 2014) and by their expression patterns, which are often restricted
111 to tissues with open chromatin, permissive to spurious transcription (Soumillon *et al*, 2013;
112 Darbellay & Necsulea, 2020). Other evidence supporting non-functionality, or even a
113 deleterious effect of lncRNA transcription comes from their typical processing by the cellular
114 machinery. lncRNA transcripts are generally inefficiently spliced and poly-adenylated, and
115 are rapidly degraded by the RNA exosome (Melé *et al*, 2017; Schlackow *et al*, 2017). For
116 certain classes of lncRNAs, transcription is normally tightly repressed by chromatin-modifying
117 factors, and their de-repression leads to DNA replication stress and subsequently to cellular
118 senescence, due to an overlap with DNA replication origins (Nojima *et al*, 2018). It is not clear
119 yet to what extent similar principles apply to HCC and other cancers. However, the presence
120 of high lncRNA levels in cancer cells may be a sign of a global de-regulation of the molecular
121 machineries that normally keep deleterious transcription in check, even if individual lncRNAs
122 are not “oncogenes” *sensu stricto*. This further highlights the need for detailed investigations
123 of the patterns of lncRNA expression in cancer.

124

125 In this study, we set out to explore the patterns of lncRNA transcription in a large HCC cohort,
126 comprising paired tumor and adjacent tissue biopsies from 114 patients. Our work stands out
127 from previous efforts to characterize lncRNAs in HCC in several important ways. First, we take
128 advantage of an extensive transcriptome resource, which covers a wide range of tumor
129 progression stages and underlying liver diseases, and thus can provide a comprehensive
130 overview of transcriptional de-regulation during HCC development. Importantly, our
131 transcriptome dataset is derived from biopsies rather than tumor resections, and is thus likely
132 more faithful to the *in vivo* physiological status of the tumors. Second, we perform a meta-
133 analysis of the current literature on lncRNAs and HCC and we use our transcriptome collection
134 to critically re-evaluate previous claims regarding lncRNA expression patterns in HCC. We
135 can thus highlight the poor reproducibility of some prominent lncRNA-HCC associations.
136 Third, rather than attempting to propose new candidate oncogene or tumor suppressor
137 lncRNAs, we perform a detailed analysis of the genomic characteristics of de-regulated
138 lncRNAs. We thus reveal an increase in repetitive-element derived lncRNA expression in
139 tumor samples. In particular, we uncover a striking up-regulation of non-coding RNAs derived

140 from centromeric satellite repeats. We discuss the functional implications of this apparent
141 activation of centromeric chromatin in HCC tumors.

142 **Results**

143 Transcriptome dataset

144 We analyzed the patterns of protein-coding and lncRNA gene expression in a collection of
145 268 RNA-seq samples, derived from tumor and adjacent tissue biopsies from 114 HCC
146 patients (Figure 1a, Supplementary Table 1). This cohort comprises patients with different
147 underlying diseases, including hepatitis B and hepatitis C, alcoholic or non-alcoholic liver
148 diseases and cirrhosis (Supplementary Table 1). The Edmondson-Steiner differentiation
149 grade was recorded for each tumor sample (Supplementary Table 1). Biopsies were
150 performed during the diagnostic work-up of patients before therapy, and in 3 patients with
151 HCC recurrence after tumor resection (Supplementary Table 1). Our transcriptome data is not
152 restricted to poly-adenylated RNAs (Methods) and may thus better reflect the behavior of
153 lncRNA transcripts, which are inefficiently or not at all poly-adenylated (Schlackow *et al*, 2017).
154 With this dataset, we could analyze the expression patterns of 19,465 protein-coding genes
155 and 18,866 lncRNAs, including 7,959 lncRNAs detected *de novo* using our RNA-seq data
156 (Methods, Supplementary Dataset 1).

157 Global trends of gene expression variation in HCC tumors and adjacent tissue samples

158 We first aimed to evaluate broad patterns of gene expression variation among tumor and
159 adjacent tissue samples. To get a glimpse of the cellular composition changes that take place
160 in cancer tissue, we analyzed the expression patterns of liver cell type markers
161 (Supplementary Table 2), obtained from single cell transcriptomics data (MacParland *et al*,
162 2018). As expected, many of these markers display striking differences between tumor and
163 adjacent tissue samples, as well as among degrees of tumor differentiation (Figure 1b).
164 Hepatocyte markers (*PCK1*, *BCHE*, *ARG1*, *ALB*) are low in samples derived from
165 Edmondson-Steiner grade 4 tumors (Figure 1b). Immune cell markers (e.g., T cell markers
166 *PTPRC*, *NKG7*, *FCGR3A* or macrophage markers *CD52* and *CD68*) are generally expressed
167 at lower levels in tumor samples than in the adjacent tissue (Figure 1b). Overall, these patterns
168 confirm that the cellular environment is substantially different in HCC tumors compared to the
169 adjacent tissue, but also that there is considerable heterogeneity among tumors.

170

171 The molecular heterogeneity of HCC tumors is well illustrated by principal component
172 analyses (PCA) performed on protein-coding and lncRNA genes (Methods, Figure 1c,d).
173 However, although there is substantial variation among tumor samples, this gene expression
174 map is consistent with the histological classification. For both categories of genes the first axis
175 of the PCA separates samples with the highest Edmondson-Steiner grades and samples from

176 less advanced tumors and adjacent tissues (Figure 1c,d, Supplementary Figure 1a-d). The
177 second axis forms a gradient from adjacent tissue to the highest Edmondson-Steiner grades
178 (Figure 1c,d, Supplementary Figure 1a-d). Notably, paired biopsies do not cluster on the first
179 factorial map of the gene expression PCA, despite their shared genetic background. We
180 validated the sample pairing by evaluating the presence of shared alleles in exonic single
181 nucleotide polymorphisms that were reliably detected with our RNA-seq data (Methods). As
182 expected, samples stemming from the same patient are genetically very similar, in contrast to
183 samples derived from different patients (Supplementary Figure 1e).

184

185 In HCC, lncRNAs follow previously reported patterns: they are generally weakly expressed
186 and are thus detected in fewer samples than protein-coding genes (Supplementary Figure 2).
187 This trend is even stronger for *de novo* annotated lncRNAs (Supplementary Figure 2).

188

189 Differential expression of protein-coding genes and lncRNAs in HCC

190 We next tested for differential expression (DE) between paired tumor and adjacent tissue
191 biopsies and among tumors with different Edmondson-Steiner grades (Supplementary Table
192 3, Methods). We selected differentially expressed genes with a minimum fold change of 1.5
193 and maximum false discovery rate (FDR) of 1%. With these stringent settings, we found that
194 4,100 (21%) protein-coding genes and 3,315 (18%) lncRNAs were differentially expressed
195 between tumor and adjacent tissue biopsies. When comparing tumor samples grouped by
196 Edmondson-Steiner grade (grades 1 and 2 vs. grades 3 and 4), 2,537 (13%) protein-coding
197 genes and 2,065 (11%) lncRNAs were significantly differentially expressed. The distribution
198 of expression fold changes differs between the two categories of genes, with stronger positive
199 fold changes for lncRNAs for the latter analysis (Figure 2). Genes that were up-regulated in
200 tumors compared to adjacent tissues or in tumor samples with higher Edmondson-Steiner
201 grades were enriched in processes related to the cell cycle, to chromosome organization but
202 also to embryonic development (Figure 2, Supplementary Table 4). In contrast, downregulated
203 genes were enriched in metabolic processes characteristic of the healthy liver (Figure 2,
204 Supplementary Table 4). In addition, genes involved in immune response and in cell adhesion
205 are down-regulated in tumor samples compared to the adjacent tissue (Figure 2a,
206 Supplementary Table 4). There is substantial overlap between the sets of genes that are
207 differentially expressed in the two comparisons, with consistent directions of change, for both
208 protein-coding genes and lncRNAs (Supplementary Figure 3a,b).

209

210 As expected given their involvement in essential cell cycle processes, protein-coding genes
211 that are up-regulated in tumors compared to the adjacent tissue or in the tumors with the

212 highest Edmondson-Steiner grades had significantly higher levels of evolutionary sequence
213 conservation than down-regulated genes (Wilcoxon rank sum test p-value $< 1e-10$ for the first
214 DE analysis, p-value 0.006 for the second DE analysis, Supplementary Figure 3c). For
215 lncRNAs, the increase in sequence conservation is only observed for those that are up-
216 regulated in the tumors compared to the adjacent tissue (Wilcoxon rank sum test, p-value 3.7
217 $e-4$, Supplementary Figure 3d). In contrast, lncRNAs that are up-regulated in tumors with
218 higher Edmondson-Steiner grades have slightly lower conservation scores than down-
219 regulated lncRNAs (Wilcoxon rank sum test, p-value 0.03, Supplementary Figure 3d).

220

221 We next verified whether the DE protein-coding genes and lncRNAs were isolated or clustered
222 in the genome. To do this, for each DE gene (defined as above) we verified the DE status for
223 neighboring genes, within a 50 kilobases (kb) window (Methods). We find that the proportion
224 of DE genes that have a DE neighbor with the same expression change direction is
225 significantly higher than expected by chance, for both protein-coding genes and lncRNAs
226 (randomization test, p-value < 0.01 , Supplementary Figure 4, Methods). In contrast, pairs of
227 neighboring genes with opposite DE orientation are significantly less frequent than expected
228 by chance (randomization test, p-value < 0.01 , Supplementary Figure 4). This pattern is
229 observed for both protein-coding and lncRNA genes and for both differential expression tests.

230

231 Finally, we also assessed the effect of other factors (namely, underlying liver disease,
232 presence of cirrhosis, sex of the patients) on gene expression patterns in HCC tumors. In
233 contrast with the large numbers of DE genes observed for the two comparisons described
234 above, only between 36 and 509 genes were significantly DE depending on one of these
235 factors (maximum FDR 0.01, minimum fold expression change 1.5, Supplementary Dataset
236 3). For the comparison between sexes, 180 genes were significantly DE, with the strongest
237 fold changes observed for genes located on sex chromosomes (Supplementary Dataset 3).

238

239 Expression patterns of prominent HCC-associated lncRNAs

240 We next aimed to evaluate the behavior of the most prominent HCC-associated lncRNAs in
241 our gene expression dataset. We performed a PubMed search with the key word
242 “hepatocellular carcinoma” in the article title, and parsed the abstracts of the resulting articles
243 to retrieve gene names or an unambiguous mention of lncRNAs as a class (Methods). We
244 found that the proportion of all HCC publications that mention lncRNAs increased rapidly in
245 the past decade, from 0 in 2010 to 6.3% in 2019 (Supplementary Figure 5a). In total, we could
246 find unambiguous citations for 262 lncRNAs, 160 (61%) of which were only mentioned in one
247 article (Supplementary Table 5, Supplementary Figure 5b). Only 29 lncRNAs were associated

248 with HCC in 5 or more articles. Expectedly, at the top of the list of highly-cited lncRNAs can
249 be found transcripts that are well known from other biological contexts, such as *MALAT1* (Ji
250 *et al*, 2003, 1), *H19* (Bartolomei *et al*, 1991), *HOTAIR* (Rinn *et al*, 2007) and *NEAT1*
251 (Hutchinson *et al*, 2007). The 5th highest-cited lncRNA is *HULC*, which was initially described
252 in the HCC context (Panzitt *et al*, 2007). Among the 262 HCC-associated lncRNAs, 98 (37%)
253 were significantly DE (maximum FDR 0.01 and minimum fold change 1.5) between tumor and
254 adjacent tissues, and 57 (22%) were significantly DE between tumor samples with
255 Edmondson-Steiner grades 1 and 2 and tumor samples with Edmondson-Steiner grades 3
256 and 4. These proportions are significantly higher than those observed for lncRNAs that are
257 not cited in the literature (17% and 11%, respectively, Chi-square test, p-value < 1e-10). In
258 total, 128 (49%) of the HCC-associated lncRNAs were significantly DE in at least one of the
259 tests; this proportion reached 81% with low stringency criteria (maximum FDR 0.1, no
260 minimum fold change).

261
262 We next examined the expression patterns of the 29 lncRNAs that were cited at least 5 times
263 in association with HCC (Figure 3). For this analysis, we set the maximum FDR at 0.01 as
264 described above, but we did not require a minimum fold expression change, to increase our
265 sensitivity. The great majority (90%) of these lncRNAs were significantly DE between tumors
266 and adjacent tissues, and 14 (48%) of them were also significantly DE between highly
267 differentiated (Edmondson-Steiner grades 1 and 2) and poorly differentiated tumors (grades 3
268 and 4). However, we observed several unexpected patterns among the best studied lncRNAs.
269 First, *MALAT1* was not significantly DE in neither one of the two analyses (Figure 3), despite
270 previous reports indicating its up-regulation in HCC tumors compared to adjacent tissues (Lin
271 *et al*, 2007; Lai *et al*, 2012). Importantly, this is not due to a lack of statistical power or due to
272 noisy expression, as *MALAT1* was expressed at high levels in all samples (Figure 3d).
273 Second, *HOTAIR* was overall very weakly expressed and not significantly DE in neither of the
274 two tests (FDR 0.046, Edmondson grades 1&2 against 3&4). Third, *NEAT1* was weakly but
275 significantly down-regulated in tumors compared to adjacent tissues, despite previous
276 evidence for up-regulation (Kou *et al*, 2020). For *HULC* (Panzitt *et al*, 2007), we confirmed the
277 previously reported up-regulation in tumor samples, but surprisingly, we found that it displayed
278 lower expression levels in advanced-stage tumors (Figure 3). In some cases, the results could
279 be explained by the distribution of tumor differentiation degrees among the tumor samples.
280 For example, *UCA1* is overall down-regulated in tumors compared to the adjacent tissue,
281 contrary to what was previously reported (Wang *et al*, 2015), but is expressed at higher levels
282 in samples with Edmondson-Steiner grades 3 and 4 (Figure 3). Some of the inconsistencies
283 observed between our DE analyses and previous reports, for the best-studied HCC-
284 associated lncRNAs, may also come from the distribution of patient characteristics, for

285 example underlying liver diseases, genetic background etc. However, out of the 29 tested
286 lncRNAs none showed significant expression differences between patients with different
287 underlying diseases (Supplementary Dataset 3). Only *XIST* was differently expressed
288 between sexes (Supplementary Dataset 3). We also did not observe any significant difference
289 between patients with or without cirrhosis (Supplementary Dataset 3).

290

291 *Increased repetitive sequence content in HCC-upregulated lncRNAs*

292 We next wanted to assess the genomic features of the lncRNAs that are significantly
293 differentially expressed in the two analyses described above. It was previously reported that
294 transposable elements that are repressed in healthy tissues can become active in cancer cells
295 (Burns, 2017). We thus analyzed the repetitive sequence content of differentially expressed
296 lncRNAs (Supplementary Table 6, Supplementary Dataset 4, Methods). The fraction of exonic
297 sequence covered by repeats was significantly higher for lncRNAs that were up-regulated in
298 tumors compared to adjacent tissues (median value 47%) than for down-regulated lncRNAs
299 (median value 40%, Wilcoxon rank sum test, p-value < 1e-10, Figure 4a). Likewise, in the DE
300 analysis comparing tumor samples with different Edmondson-Steiner grades, up-regulated
301 lncRNAs had significantly higher repetitive sequence content (median 48%) than down-
302 regulated lncRNAs (median value 43%, Wilcoxon rank sum test, p-value 1e-6, Figure 4a). For
303 protein-coding genes, the opposite trend was observed, with higher repetitive sequence
304 contents for down-regulated genes, in both DE analyses (Figure 4a). Among the most
305 abundant classes of repetitive elements, we found that this pattern was the strongest for
306 satellite repeats: for both DE analyses, up-regulated lncRNAs overlap significantly more
307 frequently with satellite repeats than down-regulated lncRNAs (Chi-square test, p-value 1e-4
308 for the first DE analysis, p-value 8e-5 for the second DE analysis, Figure 4b). Confirming this
309 observation, we found that lncRNAs that overlapped with satellite repeats had significantly
310 higher fold expression changes than lncRNAs without satellite repeats, for both DE analyses
311 (Wilcoxon rank sum test p-value 2e-6 for the first DE analysis, 0.02 for the second DE analysis,
312 Figure 4c). We also observed significantly higher fractions of exonic overlap with LTR repeats
313 for lncRNAs that are up-regulated in tumors with high Edmondson-Steiner grades, compared
314 to down-regulated lncRNAs (Supplementary Figure 6). However, for this repeat class there
315 was no significant difference between lncRNAs that are up-regulated or down-regulated
316 between tumors and adjacent tissues (Supplementary Figure 6).

317 *Up-regulation of centromeric non-coding RNAs and centromeric proteins in HCC*

318 Satellite repeats are a major functional component of centromeres (Hartley & O'Neill, 2019).
319 Following our observation that lncRNAs that overlap with satellite repeats tend to be
320 expressed at higher levels in tumors than in normal tissues, and in particular in advanced-

321 stage tumors, we performed a more direct examination of transcription in centromeric regions.
322 As these highly repetitive sequences can be difficult to capture with next generation
323 sequencing approaches, we first determined the centromeric regions that are mappable with
324 our RNA-seq data – that is, to which sequencing reads can be attributed unambiguously
325 (Methods). With the exception of the Y chromosome, which had a mappable length of 222 kb,
326 all centromeric regions had mappable lengths comprised between 1.2 Mb and 5.2 Mb
327 (Supplementary Figure 7a). We found 752 transcribed loci in centromeric regions, all but one
328 detected *de novo* with our RNA-seq data (Supplementary Dataset 5). In general, we could
329 detect at most 10 centromeric transcribed loci *per* chromosome (Supplementary Figure 7b).
330 However, we found large numbers of transcripts on chromosomes 2 and 18 (173 and 395
331 transcribed loci, respectively), as well as on chromosomes 1 and 19 (31 and 75 transcribed
332 loci, respectively). With the exception of an Ensembl-annotated pseudogene, these transcripts
333 were classified as non-coding, but only 243 passed all lncRNA filtering criteria (Supplementary
334 Dataset 5). The other non-coding transcripts were generally rejected from the lncRNA dataset
335 because they were too short (38% of the cases), they overlapped with unmappable regions
336 (14%), they had insufficient read coverage (4%), or because of a combination of these criteria.

337

338 We evaluated the abundance of centromeric transcripts by counting unambiguously mapped
339 RNA-seq for each chromosome and strand, normalized by dividing by the total unique read
340 count attributed to genes, for each sample (Methods, Supplementary Dataset 5). Most
341 centromeric RNA-seq reads were derived from chromosome 2, followed by chromosome 1
342 and 19 (Figure 5a). Chromosome 2 also stood out with respect to the differences between
343 tumor and adjacent tissue samples: on the reverse DNA strand, 94 patients (85%) had higher
344 transcript levels in tumors than in adjacent tissue samples (Figure 5b). We note that
345 transcription is not restricted to well-defined loci, but covers the entire centromeric region
346 (Figure 5c).

347

348 The degree of centromere transcript activation in tumor samples compared to adjacent tissue
349 samples varies considerably among patients (Figure 5b). To evaluate the determinants of
350 centromeric transcription variation, we analyzed the association between protein-coding gene
351 differential expression and centromeric transcript differential expression, across patients.
352 Specifically, for each patient, we computed the difference in TPM levels between tumors and
353 adjacent tissues, for each protein-coding gene; we also computed the difference in total
354 centromeric RPKM levels between tumors and adjacent tissues, and we correlated the two
355 sets of values across patients. Genes involved in mitotic cell cycle processes were often
356 positively associated with centromeric transcript activation levels (Supplementary Table 7,
357 gene ontology enrichment analysis presented in Supplementary Dataset 5). Among the genes

358 with the highest positive correlations with centromeric transcript activation levels were several
359 genes encoding centromeric proteins (*CENPJ*, *CENPF* and *CENPI*), the CENPA chaperone
360 *HJURP* (Hori *et al*, 2020), the *DNA2* nuclease/helicase that promotes centromeric DNA
361 replication (Li *et al*, 2018), etc. (Figure 5d, Supplementary Table 7). Interestingly, *CENPC*,
362 which is thought to repress alpha-satellite RNA levels (Bury *et al*, 2020), was negatively
363 associated with centromeric transcript activation levels (Figure 5d, Supplementary Table 7).

364

365 In addition to increased levels of centromeric non-coding RNAs in tumors, we also observed
366 a strong tendency for up-regulation for centromeric proteins (Supplementary Table 8). Out of
367 25 protein-coding genes annotated in Ensembl as “centromere proteins”, 20 (80%) were up-
368 regulated in the tumors compared to adjacent tissue and 13 (52%) were up-regulated in
369 tumors with Edmondson-Steiner grades 3&4 compared to tumors with Edmondson-Steiner
370 grades 1&2 (maximum FDR 0.01). At the top of the list, the genes coding for the histone variant
371 *CENPA* and for centromeric protein F (*CENPF*) were more than 4-fold over-expressed in
372 tumors compared to adjacent tissues (Supplementary Table 8). Confirming our previous
373 analysis, we also observed that *CENPC* was down-regulated in tumors compared with
374 adjacent tissues and in advanced-stage tumors compared to early-stage tumors
375 (Supplementary Table 8).

376

377 **Discussion**

378 *Protein-coding gene and lncRNA expression patterns in HCC*

379 With this analysis, our first aim was to investigate the gene expression alterations that
380 characterize HCC tumors. Compared to the numerous transcriptome collections that were
381 previously published in the HCC field, our dataset has the advantage of including a large
382 number of paired tumor and adjacent tissue samples, comprising a total of 268 samples from
383 114 patients. Importantly, the samples analyzed here are derived from biopsies, which are
384 likely to better reflect the situation *in vivo*, because they are devoid of changes induced by
385 hypoxia and hypoglycemia that occur in surgical resection specimens as a consequence of
386 segmental blood vessel occlusions during the operation. Moreover, our data includes both
387 poly-adenylated and non-poly-adenylated RNA species, which makes it better suited for the
388 study of inefficiently poly-adenylated lncRNAs (Schlackow *et al*, 2017).

389

390 We first explored the broad patterns of gene expression variation in our tumor and adjacent
391 tissue samples. By analyzing the expression patterns of molecular markers for the most
392 common cell types in the healthy liver (MacParland *et al*, 2018), we confirmed that HCC
393 tumors have very different cellular environments compared to adjacent tissue samples (Figure

394 1). In particular, immune cell populations appear to be diminished in the majority of tumors,
395 (Figure 1). Although these patterns are evidently better investigated with single-cell RNA-seq
396 data, these results confirm that our transcriptome collection reflects the cellular composition
397 changes that define the “tumor microenvironment” (Hanahan & Weinberg, 2011).

398

399 As expected, we found that gene expression patterns are in good agreement with the
400 histological classification of the tumor samples. For both protein-coding genes and lncRNAs,
401 tumor samples with Edmondson-Steiner grades 3 and 4 stand out from tumors with lower
402 grades and from adjacent tissue samples (Figure 1, Supplementary Figures 1). Other factors,
403 such as the underlying liver disease, the sex or the age of the patients, have comparatively
404 little effect on the overall gene expression variation. We thus focused on the protein-coding
405 genes and lncRNAs that are differentially expressed between paired tumor and adjacent
406 tissue samples, or between poorly differentiated tumors (Edmondson-Steiner grades 3 and 4)
407 and highly differentiated tumors (Edmondson-Steiner grades 1 and 2). We observed an over-
408 representation of biological processes associated with the cell cycle among genes that are
409 up-regulated in the tumors (Figure 2), which is expected given that cancer cells are rapidly
410 proliferating. Conversely, genes involved in the metabolic processes performed by the healthy
411 liver or in immune response tend to be down-regulated in the tumors (Figure 2).

412

413 Both protein-coding genes and lncRNAs contribute to the differential gene expression patterns
414 observed in HCC tumors (Figure 2). Differentially expressed protein-coding genes and
415 lncRNAs share many characteristics. For example, for both gene categories, we found that
416 genes that are up-regulated in tumors compared to adjacent tissue samples have significantly
417 higher levels of evolutionary sequence conservation than genes with the opposite expression
418 change (Supplementary Figure 3). This observation is consistent with the enrichment of cell
419 cycle functions among protein-coding genes that are up-regulated in the tumors, as these
420 genes have essential biological roles and are thus under strong constraint during evolution.
421 The increase in sequence conservation for lncRNAs that are up-regulated in tumors suggests
422 that these lncRNAs may also participate in essential cellular functions and contribute to
423 cellular proliferation. Another shared feature between protein-coding genes and lncRNA is the
424 presence of spatial clustering: differentially expressed genes are found in close proximity to
425 other differentially expressed genes with the same expression change direction significantly
426 more often than expected by chance (Supplementary Figure 4). This observation may be
427 explained by a tendency for co-regulation of neighboring lncRNA and protein-coding genes,
428 or may reflect the presence of large-scale structural variations (rearrangements, duplication
429 and deletions) in cancer cells, which can affect the expression patterns of multiple neighboring
430 genes (Spielmann *et al*, 2018). This finding also underlines the importance of evaluating the

431 broader genomic context when aiming to select candidate oncogenes, tumor suppressors, or
432 biomarkers: the most biologically relevant gene may be the neighbor of the gene initially
433 selected for validation.

434 Limited reproducibility of differential expression patterns for HCC-associated lncRNAs

435 In the past decade, the number of publications that discuss lncRNAs in the context of
436 hepatocellular carcinoma has increased exponentially (Supplementary Figure 5). lncRNAs
437 are often proposed as promising oncogenes or tumor suppressors, based on their patterns of
438 expression in tumors and healthy tissues. However, lncRNAs are weakly expressed and are
439 generally highly variable among tissues, cell types or individuals (Kornienko *et al*, 2016). Thus,
440 it is not clear to what extent the lncRNA expression patterns previously reported in the HCC
441 literature are reproducible with independent datasets. Here, we evaluated the expression
442 patterns of lncRNAs that were previously associated with HCC in our transcriptome collection.
443 The majority of these lncRNAs were strongly differentially expressed between paired tumors
444 and adjacent tissue samples or between groups of tumors with high or low differentiation.
445 However, we are still far from confirming differential expression patterns for all HCC-
446 associated lncRNAs, even when lowering the stringency of our criteria. Even when evaluating
447 the most prominent HCC-associated lncRNAs, which were cited by at least 5 publications, we
448 could not always recover the previously reported differential expression observations. This
449 was the case even for the three lncRNAs that were most frequently associated with HCC in
450 the literature: MALAT1, H19 and HOTAIR. The biological roles of these lncRNAs in cancer
451 were already controversial. For example, a recent study showed that MALAT1 suppresses
452 metastasis in breast cancer (Kim *et al*, 2018, 1), contrary to previous reports which proposed
453 that this lncRNA promotes metastasis (Arun *et al*, 2016). Likewise, H19 was alternatively
454 proposed as an oncogene (Matouk *et al*, 2007) or as a tumor suppressor (Yoshimizu *et al*,
455 2008). For HOTAIR, its role as a metastasis-promoting factor appears to be accepted in the
456 literature (Gupta *et al*, 2010). However, we note that the initially proposed function for this
457 lncRNA, namely a role in the regulation of *HOXD* genes during embryonic development (Rinn
458 *et al*, 2007), was refuted *in vivo* (Amândio *et al*, 2016). These examples illustrate the frailty of
459 some of the claims that are recurrently put forward regarding lncRNA functions, in cancer or
460 in other biological contexts, and again highlight the caution that should be exercised when
461 investigating lncRNAs.

462 Activated transcription of centromeric satellite repeats in HCC tumors

463 Transcriptome comparisons in HCC cohorts or in other cancer types generally aim to select
464 candidate oncogenes, tumor suppressors or biomarkers, that should be further verified
465 experimentally. As extensive functional validations were outside of the scope of our study, we
466 chose instead to analyze the genomic characteristics of the lncRNAs that were differentially

467 expressed in HCC tumors. We were thus able to detect an increase in the repetitive sequence
468 content of lncRNAs that were up-regulated in tumors compared to adjacent tissues, as well as
469 in poorly differentiated tumors compared to early stage tumors (Figure 4). Repetitive
470 sequences make up roughly half of the human genome (Lander *et al*, 2001). The high repeat
471 fraction observed for lncRNA exons, which is more than triple the fraction observed for protein-
472 coding gene exons (Figure 4), is likely due to the weak selective pressures that act on these
473 loci (Darbellay & Necsulea, 2020). However, the increase in repetitive sequence content for
474 tumor-upregulated lncRNAs cannot simply be explained by a lower proportion of functionally
475 constrained loci; on the contrary, average sequence conservation scores are higher for tumor-
476 upregulated lncRNAs than for tumor-downregulated lncRNAs (Supplementary Figure 3).
477 Moreover, we found that the over-representation of repetitive sequences in upregulated
478 lncRNAs does not affect all classes of repeats, but is strongest for satellite repeats (Figure 5).
479 This class of repeats is a major functional component of centromeres (Hartley & O'Neill, 2019).

480

481 Although centromeres were initially thought to be transcriptionally inert, it is now known that
482 they are transcribed into non-coding RNAs, which associate with centromeric chromatin and
483 potentially participate in kinetochore formation (Talbert & Henikoff, 2018). However, these
484 non-coding RNAs are generally weakly transcribed, and higher expression levels can lead to
485 impaired centromeric function (Bouzinba-Segard *et al*, 2006). Overexpression of centromeric
486 non-coding RNAs was previously reported in pancreatic cancers and in other types of
487 epithelial cancers (Ting *et al*, 2011). In mouse models of pancreatic cancers, it was shown
488 that overexpression of centromeric satellite repeats leads to increased DNA damage and
489 chromosomal instability, thereby accelerating tumor formation (Kishikawa *et al*, 2016, 2018).
490 Here, we reveal that centromeric non-coding RNAs are also aberrantly overexpressed in HCC.
491 This finding is supported by several lines of evidence. First, we showed that satellite repeats,
492 which are characteristic of centromeric regions, are over-represented in the exonic regions of
493 tumor-upregulated lncRNAs. Second, we directly quantified centromeric transcription, by
494 evaluating regions to which RNA-seq reads can be unambiguously attributed, despite the
495 repetitive sequence context. We thus showed that transcription stems from the entire length
496 of centromeric regions, rather than from well-defined non-coding RNA loci. Interestingly, all
497 chromosomes are not equal with respect to detectable centromeric transcription. The
498 centromere of chromosome 2 appears to be transcriptionally active in tumor samples for the
499 majority of patients (Figure 5). The mechanisms that underlie this over-representation of
500 chromosome 2 are unclear. This chromosome has a particular evolutionary history: it is
501 derived from a chromosome fusion event, which occurred after the divergence of human and
502 chimpanzee and which led to the loss of one of the two ancient centromeres (Chiatante *et al*,
503 2017). Although we verified that the over-representation of chromosome 2 is not simply due

504 to a better mappability of satellite repeats (Supplementary Figure 7), we cannot exclude other
505 technical issues that prevent us from detecting these highly repetitive transcripts from other
506 chromosomes.

507

508 The levels of centromeric non-coding RNA transcription were previously found to vary during
509 the cell cycle in mouse, with a peak in the G2/M phase (Ferri *et al*, 2009). Thus, our findings
510 may be partially explained by an over-representation of cells in the G2/M phase in tumor
511 samples compared to the adjacent tissue, expected given that cancerous cells are rapidly
512 proliferating. Indeed, our analysis revealed that genes involved in mitotic cell cycle processes
513 were positively associated with centromeric transcript up-regulation levels, across patients.
514 This included several genes encoding centromeric proteins (*CENPJ*, *CENPF* and *CENPI*)
515 (Figure 5d, Supplementary Table 7). Interestingly, the gene encoding centromere protein C
516 (*CENPC*) was negatively associated with centromeric transcript up-regulation levels across
517 patients, and was significantly down-regulated in tumors compared to adjacent tissues and in
518 advanced stage tumors (Figure 5d, Supplementary Tables 7-8). It was recently reported that
519 this protein acts to repress centromere-derived alpha-satellite RNA levels (Bury *et al*, 2020).
520 This observation could thus explain the up-regulation of centromeric transcripts in tumor
521 compared to adjacent tissue samples, which appears to occur in parallel with a down-
522 regulation of *CENPC* expression.

523

524 We also note that the ability to detect centromeric non-coding RNAs likely depends on the
525 methods used to generate RNA-seq data. Our transcriptome collection was generated from
526 ribo-depleted RNA samples, without enrichment for poly-adenylated RNA species (Methods).
527 Although it was reported that centromeric transcripts are poly-adenylated (Topp *et al*, 2004),
528 their subsequent processing into smaller RNA molecules (Talbert & Henikoff, 2018) may lead
529 to the loss of the polyA tail, thus hampering their detection in polyA-enriched RNA-seq data.
530 Furthermore, our RNA-seq data consists of relatively long reads (126-136 bp), which likely
531 increases our ability to unambiguously map RNA-seq reads to the genomic regions from which
532 they stem, even in the case of repetitive sequences.

533

534 To our knowledge, aberrant transcription of centromeric non-coding RNAs had not been
535 previously reported in HCC. Given that this phenomenon has been associated with tumor
536 formation in other types of cancer (Kishikawa *et al*, 2018), our observations are highly relevant
537 for the search for oncogenic factors driving hepatocellular carcinoma, and thus warrant further
538 investigations.

539

540

541 **Methods**

542 Biological sample collection

543 The analyses presented in this manuscript were performed on carcinoma and adjacent liver
544 tissue biopsies obtained from 114 patients. Human tissues were obtained from patients
545 undergoing diagnostic liver biopsy at the University Hospital Basel. Written informed consent
546 was obtained from all patients. The study was approved by the ethics committee of the
547 northwestern part of Switzerland (Protocol Number EKNZ 2014-099). The samples analysed
548 here were derived from pre-treatment biopsies, with the exception of 3 patients, for which
549 samples were collected after tumor resection (Supplementary Table 1). We recorded the sex,
550 age at the time of biopsy and underlying liver diseases for each patient (Supplementary Table
551 1). We also recorded the percentage of tumor tissue in the biopsies and the Edmondson-
552 Steiner grades of the tumors (Supplementary Table 1). Multiple tumor and adjacent tissue
553 biopsies were collected for 26 and 3 patients, respectively. In total, we analysed 268 samples,
554 corresponding to 117 adjacent tissue and 151 tumor biopsies.

555 RNA extraction and library preparation

556 We extracted RNA and DNA from tissue biopsies using the ZR-Duet DNA/RNA MiniPrep Plus
557 kit (Zymo Research, catalog number D7003). We performed the in-column DNase I treatment
558 as specified in the kit to remove residual DNA from the RNA fraction. We prepared RNA-seq
559 libraries using the Illumina TruSeq stranded RNA protocol, without polyA selection. We
560 depleted ribosomal RNA using the Ribo-Zero Gold kit from Illumina. We generated single-end
561 reads, 126 or 136 nucleotides (nt) long (Supplementary Table 1).

562 RNA-seq data processing

563 We aligned the RNA-seq reads on the genome using HISAT2 (Kim *et al*, 2015, 2) version
564 2.0.5. We used the primary assembly of the human genome version GRCh38 (hg38),
565 downloaded from Ensembl (Cunningham *et al*, 2019). We built the HISAT2 genome index
566 using additional splice site information from Ensembl release 97, as well as from the CHES
567 (Pertea *et al*, 2018) and MiTranscriptome (Iyer *et al*, 2015) transcript assemblies. We
568 extracted unambiguously mapped reads based on the NH tag from HISAT2 reported
569 alignments. To evaluate the prevalence of strand errors during library preparation, we
570 identified introns with GT-AG and GC-AG splice sites, supported by spliced RNA-seq reads
571 aligned on at least 8 nucleotides on each neighboring exon and with a maximum mismatch
572 frequency of 2%. We then compared the strand inferred based on splice site information with
573 the strand inferred based on the read alignment orientation and on the library type. All libraries
574 had strand error rates below 2.5% (Supplementary Table 1). The presence of contradictory
575 strand assignments was used as a red flag in our lncRNA filtering procedure (see below).

576 Single nucleotide polymorphism analysis

577 We verified that samples derived from the same patient were correctly paired by assessing
578 their genetic similarity, using RNA-seq information alone. To do this, we first scanned the
579 RNA-seq alignments to detect putative single nucleotide polymorphisms (SNPs). We used a
580 a pipeline combining tools from GATK (Van der Auwera *et al*, 2013) version 4.1.9.8 and Picard
581 (<http://broadinstitute.github.io/picard/>) version 2.18.7. Briefly, we analyzed non-duplicated
582 aligned RNA-seq reads, re-calibrated the alignment quality around known variants from
583 dbSNP (Sherry *et al*, 2001) release 151 and called variants with a minimum base quality score
584 threshold of 20. We combined the detected SNPs across all samples and filtered them to keep
585 only positions found in dbSNP and in exonic regions, excluding repetitive sequences. For all
586 resulting SNPs, we counted the number of reads supporting each allele using the
587 ASEReadCounter tool. We kept biallelic SNPs supported by at least 50 reads. To allow for
588 sequencing or mapping errors, SNPs were considered to be heterozygous if the estimated
589 allele frequency was between 0.1 and 0.9, and homozygous if the allele frequency was equal
590 to 0 or 1. For all pairs of samples, we computed the proportion of SNPs with shared alleles
591 out of all biallelic SNPs. We compared this measure of genetic similarity between pairs of
592 samples derived from the same patient or from different patients (Supplementary Figure 1).
593 We also evaluated the proportion of heterozygous SNPs out of all detected SNPs on
594 autosomes and on sex chromosomes, for each sample. We excluded the pseudo-autosomal
595 regions from sex chromosomes. For one male patient (identifier 42), we observed high levels
596 of heterozygosity on the X chromosome and high *Xist* expression levels, for both tumor and
597 adjacent tissue biopsies. This patient was excluded from differential expression analyses
598 (Supplementary Table 1).

599 Evaluation of genomic DNA contamination

600 To assess the amount of genomic DNA contamination, we evaluated the RNA-seq read
601 coverage on repeat-masked intergenic regions, on both forward and reverse strands. As
602 genuinely transcribed regions are generally strongly biased in favor of one strand, we
603 computed the number of regions that had relatively symmetric strand distribution, *i.e.* for which
604 the absolute value of the $(\text{forward-reverse})/(\text{forward+reverse})$ coverage ratio was below 0.5.
605 We then computed for each sample the proportion of intergenic regions with symmetric
606 coverage, out of all intergenic regions with RNA-seq coverage. We considered that samples
607 with more than 5% symmetrically transcribed intergenic regions had significant DNA
608 contamination. These samples were excluded from differential expression analyses
609 (Supplementary Table 1).

610 Identification of “mappable” and “unmappable” genomic regions

611 To determine whether RNA-seq reads can be correctly traced back to their genomic region of
612 origin, we performed a “mappability” analysis. To do this, we generated single-end sequencing
613 reads with the same lengths as in our data (126 and 136 nt) from sliding genomic windows
614 with 5 nt step. Reads were generated with perfect sequence quality and no mismatches. We
615 aligned these reads on the genome using HISAT2 with the same parameters as for the real
616 RNA-seq data. Genomic regions to which simulated reads were mapped back unambiguously
617 and on their entire length were said to be mappable. We defined unmappable regions by
618 subtracting mappable intervals from full-length chromosomes.

619 Transcript assembly

620 We performed a genome- and transcriptome-guided transcript assembly with StringTie
621 (Pertea *et al*, 2015) release 2.1.2. We used as an input the unambiguously mapped reads
622 obtained with HISAT2, combined across all samples. We used annotations from Ensembl
623 (Cunningham *et al*, 2019) release 99, excluding read-through transcripts, as a guide for the
624 assembly. We ran StringTie separately for each chromosome and strand; unassembled
625 contigs and the mitochondrion were excluded. We filtered the StringTie output to discard
626 artefactual antisense transcripts stemming from library preparation errors. To do this, we
627 computed the sense and antisense exonic read coverage for each transcript and kept only
628 those transcripts which had a sense/antisense ratio of at least 5% in at least one sample. We
629 also removed transcripts that contained splice junctions with contradictory strand assignments
630 based on the splice site (GT-AG or GC-AG) and on the read alignment and library type. We
631 combined Ensembl 99 and filtered StringTie transcript annotations by adding to the Ensembl
632 reference those *de novo* annotated transcripts which had exonic overlap with at most 1
633 Ensembl-annotated gene. Ensembl-annotated transcripts were not altered, with the exception
634 of read-through transcripts (defined as transcripts that overlap with more than one multi-exonic
635 gene), which were discarded. LncRNAs that overlapped with annotated microRNAs were
636 annotated separately from the miRNA products.

637 Protein-coding potential of newly assembled transcripts

638 We used the PhyloCSF (Mudge *et al*, 2019) codon substitution frequency score to evaluate
639 the protein-coding potential of newly assembled transcripts. To do this, we overlapped exonic
640 coordinates with protein-coding regions predicted by PhyloCSF, in all possible reading frames.
641 Transcripts were said to be potentially protein-coding if they overlapped with a PhyloCSF
642 protein-coding region on at least 150 nt. Due to the nature of the genetic code, some
643 substitutions are synonymous on both DNA strands, which can generate artefactually high
644 PhyloCSF scores on the antisense strand of protein-coding regions. We thus required that the
645 overlap with PhyloCSF regions be higher on the sense strand than on the antisense strand of

646 the transcripts. We also evaluated the similarity between lncRNA sequences and known
647 proteins and protein domains, using DIAMOND (Buchfink *et al*, 2015) against SwissProt
648 (UniProt Consortium, 2019) and Pfam (El-Gebali *et al*, 2019). We retained SwissProt entries
649 with high confidence scores (1 to 3) and the Pfam-A subset of Pfam. We searched for hits on
650 repeat-masked cDNA sequences with the “blastx” flavor of DIAMOND and we required a
651 maximum e-value of 0.01. Transcripts were said to be potentially protein-coding if they had
652 similarity with a known protein or protein domain on at least 150 nt, with at least 40% sequence
653 identity. Genes were said to be potentially protein-coding if at least one of their isoforms was
654 predicted as protein-coding with either method.

655 LncRNA dataset

656 We established a lncRNA dataset by combining lncRNAs annotated in Ensembl (gene biotype
657 “lncRNA”) and transcribed loci annotated with StringTie that passed several filters: no protein-
658 coding potential, evaluated as described above; minimum exonic length of 200 nt for multi-
659 exonic loci and 500 nt for mono-exonic loci; at most 5% exonic length overlap with unmappable
660 genomic regions; no overlap with Ensembl-annotated protein-coding genes on the same
661 strand; at least 5000 nt away from protein-coding gene exons; at most 25% exonic length
662 overlap with RNA repeats; at most 10% exonic length overlap with retrogenes (coordinates
663 downloaded from the UCSC Genome Browser database (Casper *et al*, 2018)). We also
664 required transcribed loci to be supported by at least 100 RNA-seq reads. lncRNA annotations
665 are provided in Supplementary Dataset 1 online.

666 Literature search for HCC-associated lncRNAs

667 We searched for articles in PubMed with the key word “hepatocellular carcinoma” in the article
668 title. We retrieved the article abstract, title, journal and publication date. We searched for gene
669 names in the abstract, based on a list of common gene names and synonyms in the Ensembl
670 database. We excluded gene names that were ambiguous and matched with common terms
671 in the HCC literature (e.g., MRI, TACE, etc). We also checked if articles contained general
672 references to lncRNAs as a class, based on the “long non-coding RNA” and “lncRNA”
673 keywords, with spelling variations (e.g. “noncoding” instead of “non-coding”, “lincRNA” instead
674 of “lncRNA”, etc.).

675 Gene expression estimation

676 We evaluated gene expression values with Kallisto (Bray *et al*, 2016) release 0.46.1 (patch by
677 P.V. to correct bootstrap estimates). We obtained effective read counts and transcript *per*
678 million (TPM) values for each isoform and obtained gene-level TPM values using tximport
679 (Soneson *et al*, 2015) in R. We performed an additional normalization across samples, with a
680 previously-proposed median-scaling approach based on the 100 genes that vary least in terms
681 of expression ranks among samples (Brawand *et al*, 2011). This approach was applied on

682 gene-level TPM values. For most gene expression analyses, we used log2-transformed TPM
683 values, adding an offset of 1 (TPM to $\log_2(\text{TPM}+1)$). As a control, we also evaluated unique
684 read counts per gene using featureCounts from Rsubread (Liao *et al*, 2019). Expression
685 estimation analyses were performed on the full set of detected transcribed loci, including
686 protein-coding genes, lncRNAs and other types of genes. Gene expression levels are
687 provided in Supplementary Dataset 2 online.

688 Principal component analyses

689 We performed principal component analyses using the dudi.pca function in the ade4 library in
690 R (Dray & Dufour, 2007). We used log2-transformed TPM levels, for all protein-coding and
691 lncRNA genes or for each gene type separately. We enabled variable centering but not scaling
692 and kept 5 axes.

693 Differential expression analyses

694 We used DESeq2 release 1.28.0 and txlImport (Soneson *et al*, 2015) release 1.16.1 in R to
695 assess differential expression, based on Kallisto-estimated effective read counts *per*
696 transcript. We performed all differential expression analyses on the combined set of protein-
697 coding and lncRNA genes. Given that the number of biopsies varied among patients, we first
698 selected one pair of tumor and adjacent tissue samples *per* patient, to ensure patients
699 contributed equally to DE results. For patients where biopsies were done before and after
700 tumor resection, we selected the biopsies obtained before resection. For one patient, an
701 adjacent tissue biopsy was performed before onset of HCC; we excluded it from DE analyses.
702 For all other cases where multiple biopsies were available, we selected the sample with the
703 largest number of uniquely mapped reads for each tissue type. We tested for differential
704 expression between pairs of tumors and adjacent tissues by fitting a model that explains gene
705 expression variation as a function of two factors, the tissue type and the patient of origin. We
706 then evaluated the difference between tumors and adjacent tissues with a Wald test
707 contrasting the two tissue types and estimated the effect size with the “apeglm” shrinkage
708 method (Zhu *et al*, 2019). We repeated this analysis separately for males and females. We
709 also tested for differences in gene expression among tumor samples, depending on the patient
710 sex Edmondson-Steiner grade, presence or absence of hepatitis C, hepatitis B, cirrhosis,
711 alcoholic liver disease or non-alcoholic liver disease. To do this, we fitted an additive model
712 including all these factors and then evaluated the effect of each factor by contrasting its levels
713 with a Wald test, using the “apeglm” shrinkage method to estimate the effect size (Zhu *et al*,
714 2019). For the Edmondson-Steiner grade, we contrasted grades 1 and 2 against grades 3 and
715 4. We performed a preliminary test for an age effect, but as there were no significantly DE
716 genes this factor was not included in the model. Differential expression analyses were

717 performed only on protein-coding and lncRNA genes. Results are provided in Supplementary
718 Dataset 3 online.

719 Gene ontology analyses

720 We performed gene ontology enrichment analyses with GOrilla (Eden *et al*, 2009), contrasting
721 the lists of up- or down-regulated protein-coding genes in each test with a background set
722 consisting of protein-coding genes expressed in those samples. To define the background set,
723 we evaluated the minimum expression level (DESeq2-normalized read counts) of differentially
724 expressed genes and selected genes that had higher or equal expression levels. For the
725 analysis of the association between protein-coding gene expression and centromeric
726 transcript levels, across patients, we analyzed the gene ontology enrichment in a single
727 ranked list, comparing genes at the top of the list (with high, positive correlation coefficients)
728 to genes at the bottom of the list (with low, negative correlation coefficients).

729 Cell type marker analyses

730 We analyzed the expression patterns of common markers for the most frequent cell types in
731 the liver from a single cell RNA-seq study (MacParland *et al*, 2018). We computed a Z-score
732 matrix from the log2-transformed normalized TPM values across samples.

733 Sequence conservation analyses

734 We downloaded PhastCons (Siepel *et al*, 2005) sequence conservation scores, computed on
735 a multiple genome alignment on human and 29 other mammalian species, from the UCSC
736 Genome Browser (Casper *et al*, 2018). We computed average PhastCons scores on exonic
737 regions and splice sites. For loci that overlapped with other genes, we also computed average
738 scores on non-overlapping exonic regions. Results are provided in Supplementary Dataset 4
739 online.

740 Repetitive sequence analyses

741 We downloaded repetitive element coordinates predicted with RepeatMasker (Smit *et al*,
742 2003) from the UCSC Genome Browser (Casper *et al*, 2018). We overlapped the exonic
743 coordinates of all protein-coding and lncRNA loci with repetitive elements and we analyzed
744 the exonic fraction covered by each repeat class. Results are provided in Supplementary
745 Dataset 4 online.

746 Centromeric transcription analyses

747 We downloaded centromeric region coordinates from the UCSC Genome Browser (Casper *et*
748 *al*, 2018). We determined the mappable regions within each centromere as described above,
749 by discarding regions deemed unmappable for 126 or 136 nt read lengths. We counted the
750 number of unambiguously mapped reads that could be attributed to each mappable
751 centromeric region, on each DNA strand, using featureCounts in the Rsubread R package

752 (Liao *et al*, 2019). We computed normalized expression values (RPKM) by dividing the read
753 counts by the mappable region length (expressed in kilobases) and by the number of million
754 mapped reads, counted on the gene models annotated in Ensembl or detected with StringTie.
755 We extracted centromeric proteins based on the “centromere” keyword in the Ensembl gene
756 description. We used the same set of samples selected for differential expression analyses.
757 Results are provided in Supplementary Dataset 5 online. Centromeric transcription read
758 coverage tracks are available online for all patients: [http://pbil.univ-](http://pbil.univ-lyon1.fr/members/necsulea/MERIC_IncRNAs/)
759 [lyon1.fr/members/necsulea/MERIC_IncRNAs/](http://pbil.univ-lyon1.fr/members/necsulea/MERIC_IncRNAs/) .

760 *Co-expression between centromeric transcript levels and protein-coding gene expression*

761 To evaluate the determinants of centromeric transcription variation, we estimated the
762 correlation between protein-coding gene expression and centromeric transcript levels across
763 patients. Specifically, for each patient we estimated the difference in centromeric RPKM
764 between tumor and adjacent tissue samples, using the samples selected for differential
765 expression analyses. In parallel, we computed the difference in gene TPM between tumor and
766 adjacent tissue samples, for all protein-coding genes. We then computed Spearman’s
767 correlation coefficients for each protein-coding gene, using the values described above for all
768 patients. Results are presented in Supplementary Table 7 and in Supplementary Dataset 5
769 online.

770 *Data and code availability*

771 The sequencing data used in this project was submitted to the European Genome-Phenome
772 Archive under the accession number EGAS00001004976. Supplementary datasets
773 containing all the information needed to reproduce the results are available at the address:
774 http://pbil.univ-lyon1.fr/members/necsulea/MERIC_IncRNAs/ . Scripts are available in GitLab:
775 <https://gitlab.in2p3.fr/anamaria.necsulea/meric> .

776

777 **Acknowledgements**

778 This work was supported by European Research Council Synergy grant 609883
779 Mechanisms of Evasive Resistance in Cancer (MERiC), by The Swiss Initiative in Systems
780 Biology grant SystemX – MERiC to M.H.H and by the Agence Nationale pour la Recherche
781 (ANR JCJC 2017 LncEvoSys). C.K.Y.N. is supported by the Swiss Cancer League (KFS-
782 4543-08-2018). This work was performed using the computing facilities of the CC
783 LBBE/PRABI. We would also like to thank the French Institute of Bioinformatics (IFB, ANR-
784 11-INBS-0013) for providing storage and computing resources on its national life science
785 Cloud.

786

787

788

789

790

791 **Figure legends**

792 Figure 1. Global expression patterns in HCC tumors and adjacent tissue samples.

793 a. Numbers of patients and RNA-seq samples included in our study. All samples are derived
794 from pre-treatment biopsies.

795 b. Heatmap representing relative expression levels (log₂-transformed TPM values, divided by
796 the maximum value across samples), for 36 markers of the most common cell types in the
797 healthy liver (MacParland *et al*, 2018).

798 c. Scatter plot representing the first factorial map of a principal component analysis, performed
799 on log₂-transformed TPM values for protein-coding genes. Each dot represents one sample.
800 Colors represent sample types (adjacent tissue in grey, tumor samples colored according to
801 the Edmonson-Steiner grade).

802 d. Same as c, for lncRNAs.

803

804 Figure 2. Differentially expressed genes between HCC tumors and adjacent tissue samples.

805 a. Density plot of the log₂ fold expression change, for genes that are significantly differentially
806 expressed (maximum FDR 0.01) between paired tumor and adjacent tissue samples
807 (Methods). Red: protein-coding genes; blue: lncRNAs. The dotted vertical lines mark an
808 expression change threshold of 1.5. The numbers of genes that pass the FDR and minimum
809 fold change thresholds are shown at the top of the plot. The main enriched gene ontology
810 categories for up-regulated and down-regulated genes are shown below the plot (Methods).

811 b. Same as a, for the analysis comparing tumor samples with different stages (Edmondson-
812 Steiner grades 1 and 2 vs. Edmondson-Steiner groups 3 and 4).

813

814 Figure 3. Differential expression patterns for prominent HCC-associated lncRNAs.

815 a. Distribution of patient characteristics for the 151 tumor samples analyzed in this study. ALD:
816 alcoholic liver disease.

817 b. Distribution of the difference in expression levels between tumor and adjacent tissue
818 samples, across patients, for the 29 lncRNAs that are cited in at least 5 HCC publications. The
819 black line shows a density plot of the ratio (TPM tumor – TPM adjacent tissue)/(TPM tumor +
820 TPM adjacent tissue), computed for each patient. Only samples used for the differential
821 expression analyses were considered. The vertical red line represents the median value.

822 c. Presence/absence and direction of significant expression changes between paired tumor
823 and adjacent tissue biopsies. Upward arrows indicate up-regulation in tumor samples,
824 downward arrows indicate down-regulation in tumor samples, with a maximum false FDR of
825 0.01 (no fold change requirement). Gray arrows represent marginally significant changes
826 (FDR < 0.1, no fold change requirement).

827 d. Expression levels (log2-transformed TPM) for tumor samples, for the 29 lncRNAs that are
828 cited in at least 5 HCC publications. Samples are colored depending on the Edmondson-
829 Steiner grade.

830 e. Same as c, for the differential expression analysis comparing tumor samples with
831 Edmondson-Steiner grades 3 and 4, *versus* tumor samples with Edmondson-Steiner grades
832 1 and 2.

833

834 Figure 4. Over-representation of satellite repeats among tumor-upregulated lncRNAs.

835 a. Boxplots of the percentage of exonic sequences covered by repetitive sequences, for
836 protein-coding genes (red) and lncRNAs (blue). We display separately genes that are
837 differentially expressed (maximum FDR 0.01, minimum fold expression change 1.5) in tumors
838 compared to adjacent tissues, and in tumors with Edmondson-Steiner grades 3 and 4
839 compared to tumors with Edmondson-Steiner grades 1 and 2. Horizontal segments represent
840 median values; notches represent 95% confidence intervals for the median; dashed segments
841 extend to 1.5 times the inter-quartile range.

842 b. Percentage of genes that have exonic overlap with satellite repeats, for protein coding
843 genes (red) and lncRNAs (blue). As in a, we display separately genes that show significant
844 expression differences in our two main DE analyses.

845 c. Distribution of the log2 fold expression changes in our two main DE analyses, for lncRNAs
846 that overlap with satellite repeats (dark blue) or not (light blue). Only lncRNAs that are show
847 significant differences (maximum FDR 0.01, no minimum fold change requirement) are shown.

848

849 Figure 5. Increased centromeric transcription in tumors compared to adjacent tissue samples.

850 a. Dot chart representing the median normalized expression levels (RPKM) for centromeric
851 regions, across samples, for each chromosome and strand. Red: transcripts on the forward
852 DNA strand, blue: reverse strand. The bars represent the 95% confidence intervals.

853 b. Density plot of the RPKM difference between tumor and adjacent tissue, across patients,
854 for the three chromosome/strand combinations with highest RPKM levels (chromosome 2
855 reverse, 1 reverse and 19 forward strand).

856 c. Top: representation of the regions considered to be unambiguously mappable (Methods),
857 for the chromosome 2 centromere. Next panels: unique read coverage distribution on the
858 chromosome 2 centromere, forward and reverse strands, for one patient (identifier 13). The
859 read coverage was normalized for each sample based on the number of million mapped reads
860 attributed to genes.

861 d. Boxplots representing the distribution of gene TPM differences between tumor and adjacent
862 tissues, for three classes of patients defined based on the degree of centromeric transcript
863 “activation” in tumors compared to adjacent tissues. The first class comprises 23 patients for

864 which the difference in RPKM values for total centromeric transcripts between tumors and
865 adjacent tissues is below 0 RPKM; the second class comprises 61 patients for which the
866 difference is between 0 and 50 RPKM; the third class comprises 26 patients for which the
867 difference is above 50 RPKM. We display the 6 genes mentioned in the text: *CENPJ*, *HJURP*,
868 *CENPF*, *DNA2*, *CENPI*, *CENPC*. P-values correspond to Kruskal-Wallis non-parametric tests,
869 for differences among the three classes of patients.
870

871 **Supplementary figure legends**

872 Supplementary Figure 1. Sample clustering based on gene expression and genetic similarity.

873 a. Boxplots representing the distribution of sample coordinates on principal component 1, for
874 the PCA performed on protein-coding genes (displayed in figure 1). Samples are grouped
875 depending on tissue types. Gray: adjacent tissue samples; yellow to red: tumors grouped by
876 Edmondson-Steiner grade. Horizontal segments represent median values; notches represent
877 95% confidence intervals for the median; dashed segments extend to 1.5 times the inter-
878 quartile range.

879 b. Same as a, for principal component 2.

880 c. Same as a, for the PCA performed on lncRNAs (displayed in figure 1).

881 d. Same as c, for principal component 2.

882 e. Distribution of the proportion of shared alleles for pairs of samples, for single nucleotide
883 polymorphisms detected with our RNA-seq data (Methods). Red: distribution observed for
884 pairs of samples derived from different patients; black: distribution observed from pairs of
885 samples derived from the same patient.

886

887 Supplementary Figure 2. Expression patterns of protein-coding genes and lncRNAs in HCC
888 samples.

889 a. Distribution of the maximum expression level (log₂-transformed TPM, maximum observed
890 across samples) for protein-coding genes (red), previously known lncRNAs (dark blue) and
891 newly annotated lncRNAs (light blue). The dotted vertical line represents the TPM = 1
892 threshold. Numbers of genes above the threshold are shown in the figure legend.

893 b. Histogram of the number of samples in which the expression level is above the TPM = 1
894 threshold, for the three categories of genes described in a.

895

896 Supplementary Figure 3. Differential expression patterns in HCC samples.

897 a. Comparison between the log₂ expression fold changes observed for our two main
898 differential expression analyses (tumors vs. adjacent tissue samples, tumors with
899 Edmondson-Steiner grades 3 and 4 vs. tumors with Edmondson-Steiner grades 1 and 2), for
900 protein-coding genes. We show only genes that were significantly DE with a maximum FDR
901 of 0.01 and a fold expression change above 1.5 in at least one of the two analyses. Green:
902 genes with consistent expression changes in the both analyses; red: genes with opposite
903 expression changes; orange: genes that are significantly DE only in the first DE analysis;
904 purple: genes that are significantly DE only in the second DE analysis.

905 b. Same as a, for lncRNAs.

906 c. Distribution of sequence conservation scores for exonic regions (Methods), for protein-
907 coding genes. Genes that are up-regulated or down-regulated in our two main DE analyses

908 are shown separately. The dot represents the median conservation score, the vertical
909 segments represents the 95% confidence interval for the median.

910 d. Same as c, for lncRNAs.

911

912 Supplementary Figure 4. Genomic clustering of differentially expressed genes.

913 a. Proportion of differentially expressed genes (maximum FDR 0.01, minimum fold expression
914 change 1.5), in the comparison between paired tumors and adjacent samples, that have
915 another differentially expressed gene within a 50kb distance. Red dots represent the values
916 observed for protein-coding genes, blue dots for lncRNAs. The gray dots and vertical intervals
917 represent the average and the 95% confidence intervals for the random expectation, obtained
918 through simulations (Methods). The direction of the expression change required for the focus
919 gene and the neighboring gene is displayed below the plot.

920 b. Same as a, for the comparison between tumors with Edmondson-Steiner grades 3 and 4
921 vs. tumors with Edmondson-Steiner grades 1 and 2.

922

923 Supplementary Figure 5. Growing interest for lncRNAs in the HCC field.

924 a. Bar plot of the fraction of publications that mention lncRNAs and HCC, from 2009 to 2019.
925 The bars represent the percentage of publications that mention lncRNAs, out of the total
926 number of HCC publications. The numbers of publications that mention lncRNAs are shown
927 above the bars.

928 b. Histogram of the number of publications that cite each lncRNA in the context of HCC.
929 lncRNAs that are cited in 5 or more publications are indicated in the plot.

930

931 Supplementary Figure 6. Increased repetitive sequence content in tumor-upregulated
932 lncRNAs.

933 Percentage of genes that have exonic overlap with major classes of repeats, for protein coding
934 genes (red) and lncRNAs (blue). We display separately genes that show significant expression
935 differences in our two main DE analyses. Significantly different proportions (Chi-square test,
936 p-value <0.05) are marked by an asterisk.

937

938 Supplementary Figure 7. Centromeric transcription characteristics.

939 a. Bar plot representing the total mappable length of centromeric regions, for each
940 chromosome (Methods).

941 b. Bar plot representing the number of transcribed loci found in centromeric regions, annotated
942 with our RNA-seq data.

943

944 **Supplementary tables**

945 Supplementary Table 1. Description of the 268 RNA-seq samples in our transcriptome
946 collection.

947 Supplementary Table 2. List of cell type-specific markers for the most abundant cells in the
948 healthy liver.

949 Supplementary Table 3. Results of our two main differential expression analyses for protein-
950 coding genes and lncRNAs.

951 Supplementary Table 4. Gene ontology enrichment for differentially expressed protein-coding
952 genes.

953 Supplementary Table 5. Number of HCC-related articles that mention each protein-coding and
954 lncRNA genes.

955 Supplementary Table 6. Statistics for the overlap with different classes of repetitive elements.

956 Supplementary Table 7. Correlation between protein-coding gene expression and centromeric
957 transcript levels across patients.

958 Supplementary Table 8. Results of our two main differential expression analyses for protein-
959 coding genes involved in centromere functions.

960

961 **Supplementary datasets**

962 Supplementary Dataset 1. Genome annotation used in this analysis, obtained by combining
963 annotations from Ensembl 99 and gene models detected *de novo* with our RNA-seq data.

964 Supplementary Dataset 2. Gene expression data.

965 Supplementary Dataset 3. Full results of the differential expression analyses.

966 Supplementary Dataset 4. Evolutionary sequence conservation and repetitive element overlap
967 statistics.

968 Supplementary Dataset 5. Mappable region coordinates and expression estimates for
969 centromeric regions.

970

971

972 **References**

- 973 Amândio AR, Necsulea A, Joye E, Mascrez B & Duboule D (2016) Hotair is dispensible for
 974 mouse development. *PLoS Genet* 12: e1006232
- 975 Arun G, Diermeier S, Akerman M, Chang K-C, Wilkinson JE, Hearn S, Kim Y, MacLeod AR,
 976 Krainer AR, Norton L, *et al* (2016) Differentiation of mammary tumors and reduction in
 977 metastasis upon Malat1 lncRNA loss. *Genes Dev* 30: 34–51
- 978 Bartolomei MS, Zemel S & Tilghman SM (1991) Parental imprinting of the mouse H19 gene.
 979 *Nature* 351: 153–155
- 980 Bouzinba-Segard H, Guais A & Francastel C (2006) Accumulation of small murine minor
 981 satellite transcripts leads to impaired centromeric architecture and function. *Proc Natl*
 982 *Acad Sci USA* 103: 8709–8714
- 983 Boyault S, Rickman DS, de Reyniès A, Balabaud C, Rebouissou S, Jeannot E, Hérault A,
 984 Saric J, Belghiti J, Franco D, *et al* (2007) Transcriptome classification of HCC is related
 985 to gene alterations and to new therapeutic targets. *Hepatology* 45: 42–52
- 986 Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A,
 987 Aximu-Petri A, Kircher M, *et al* (2011) The evolution of gene expression levels in
 988 mammalian organs. *Nature* 478: 343–348
- 989 Bray NL, Pimentel H, Melsted P & Pachter L (2016) Near-optimal probabilistic RNA-seq
 990 quantification. *Nat Biotechnol* 34: 525–527
- 991 Buchfink B, Xie C & Huson DH (2015) Fast and sensitive protein alignment using DIAMOND.
 992 *Nat Methods* 12: 59–60
- 993 Burns KH (2017) Transposable elements in cancer. *Nature Reviews Cancer* 17: 415–424
- 994 Bury L, Moodie B, Ly J, McKay LS, Miga KH & Cheeseman IM (2020) Alpha-satellite RNA
 995 transcripts are repressed by centromere-nucleolus associations. *Elife* 9
- 996 Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T,
 997 Lenhard B, Wells C, *et al* (2005) The transcriptional landscape of the mammalian
 998 genome. *Science* 309: 1559–1563
- 999 Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM,
 1000 Lee BT, Karolchik D, *et al* (2018) The UCSC Genome Browser database: 2018 update.
 1001 *Nucleic Acids Res* 46: D762–D769
- 1002 Chen S, Wang G, Tao K, Cai K, Wu K, Ye L, Bai J, Yin Y, Wang J, Shuai X, *et al* (2020) Long
 1003 noncoding RNA metastasis-associated lung adenocarcinoma transcript 1 cooperates
 1004 with enhancer of zeste homolog 2 to promote hepatocellular carcinoma development by
 1005 modulating the microRNA-22/Snail family transcriptional repressor 1 axis. *Cancer Sci*
 1006 111: 1582–1595
- 1007 Chiatante G, Giannuzzi G, Calabrese FM, Eichler EE & Ventura M (2017) Centromere Destiny
 1008 in Dicentric Chromosomes: New Insights from the Evolution of Human Chromosome 2
 1009 Ancestral Centromeric Region. *Mol Biol Evol* 34: 1669–1681
- 1010 Cui H, Zhang Y, Zhang Q, Chen W, Zhao H & Liang J (2017) A comprehensive genome-wide
 1011 analysis of long noncoding RNA expression profile in hepatocellular carcinoma. *Cancer*
 1012 *Med* 6: 2932–2941
- 1013 Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J,
 1014 Billis K, Boddu S, *et al* (2019) Ensembl 2019. *Nucleic Acids Res* 47: D745–D751
- 1015 Darbellay F & Necsulea A (2020) Comparative transcriptomics analyses across species,
 1016 organs, and developmental stages reveal functionally constrained lncRNAs. *Mol Biol*
 1017 *Evol* 37: 240–259
- 1018 Dray S & Dufour A (2007) The ade4 package: implementing the duality diagram for ecologists.
 1019 *Journal of Statistical Software* 22: 1–20
- 1020 Eden E, Navon R, Steinfeld I, Lipson D & Yakhini Z (2009) GOrilla: a tool for discovery and
 1021 visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10: 48
- 1022 El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ,
 1023 Salazar GA, Smart A, *et al* (2019) The Pfam protein families database in 2019. *Nucleic*
 1024 *Acids Res* 47: D427–D432

1025 Engreitz JM, Ollikainen N & Guttman M (2016) Long non-coding RNAs: spatial amplifiers that
1026 control nuclear structure and gene expression. *Nat Rev Mol Cell Biol* 17: 756–770

1027 Ferri F, Bouzinba-Segard H, Velasco G, Hubé F & Francastel C (2009) Non-coding murine
1028 centromeric transcripts associate with and potentiate Aurora B kinase. *Nucleic Acids Res*
1029 37: 5071–5080

1030 Finn RS, Qin S, Ikeda M, Galle PR, Ducreux M, Kim T-Y, Kudo M, Breder V, Merle P, Kaseb
1031 AO, *et al* (2020) Atezolizumab plus bevacizumab in unresectable hepatocellular
1032 carcinoma. *N Engl J Med* 382: 1894–1905

1033 Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai M-C, Hung T, Argani P,
1034 Rinn JL, *et al* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to
1035 promote cancer metastasis. *Nature* 464: 1071–1076

1036 Gutschner T & Diederichs S (2012) The hallmarks of cancer: a long non-coding RNA point of
1037 view. *RNA Biol* 9: 703–719

1038 Hanahan D & Weinberg RA (2011) Hallmarks of cancer: the next generation. *Cell* 144: 646–
1039 674

1040 Hao Y, Crenshaw T, Moulton T, Newcomb E & Tycko B (1993) Tumour-suppressor activity of
1041 H19 RNA. *Nature* 365: 764–767

1042 Hartke J, Johnson M & Ghabril M (2017) The diagnosis and treatment of hepatocellular
1043 carcinoma. *Semin Diagn Pathol* 34: 153–159

1044 Hartley G & O’Neill RJ (2019) Centromere repeats: hidden gems of the genome. *Genes*
1045 (*Basel*) 10

1046 Hori T, Cao J, Nishimura K, Ariyoshi M, Arimura Y, Kurumizaka H & Fukagawa T (2020)
1047 Essentiality of CENP-A depends on its binding mode to HJURP. *Cell Rep* 33: 108388

1048 Hoshida Y, Nijman SMB, Kobayashi M, Chan JA, Brunet J-P, Chiang DY, Villanueva A, Newell
1049 P, Ikeda K, Hashimoto M, *et al* (2009) Integrative transcriptome analysis reveals common
1050 molecular subclasses of human hepatocellular carcinoma. *Cancer Res* 69: 7385–7392

1051 Hou Z, Xu X, Zhou L, Fu X, Tao S, Zhou J, Tan D & Liu S (2017) The long non-coding RNA
1052 MALAT1 promotes the migration and invasion of hepatocellular carcinoma by sponging
1053 miR-204 and releasing SIRT1. *Tumour Biol* 39: 1010428317718135

1054 Hutchinson JN, Ensminger AW, Clemson CM, Lynch CR, Lawrence JB & Chess A (2007) A
1055 screen for nuclear transcripts identifies two linked noncoding RNAs associated with
1056 SC35 splicing domains. *BMC Genomics* 8: 39

1057 Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans
1058 JR, Zhao S, *et al* (2015) The landscape of long noncoding RNAs in the human
1059 transcriptome. *Nat Genet* 47: 199–208

1060 Ji P, Diederichs S, Wang W, Böing S, Metzger R, Schneider PM, Tidow N, Brandt B, Buerger
1061 H, Bulk E, *et al* (2003) MALAT-1, a novel noncoding RNA, and thymosin beta4 predict
1062 metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 22: 8031–
1063 8041

1064 Jin Y, Lee WY, Toh ST, Tennakoon C, Toh HC, Chow PK-H, Chung AY-F, Chong SS, Ooi LL-
1065 P-J, Sung W-K, *et al* (2019) Comprehensive analysis of transcriptome profiles in
1066 hepatocellular carcinoma. *J Transl Med* 17: 273

1067 Keniry A, Oxley D, Monnier P, Kyba M, Dandolo L, Smits G & Reik W (2012) The H19 lincRNA
1068 is a developmental reservoir of miR-675 that suppresses growth and Igf1r. *Nat Cell Biol*
1069 14: 659–665

1070 Kim D, Langmead B & Salzberg SL (2015) HISAT: a fast spliced aligner with low memory
1071 requirements. *Nat Methods* 12: 357–360

1072 Kim J, Piao H-L, Kim B-J, Yao F, Han Z, Wang Y, Xiao Z, Siverly AN, Lawhon SE, Ton BN, *et*
1073 *al* (2018) Long noncoding RNA MALAT1 suppresses breast cancer metastasis. *Nat*
1074 *Genet* 50: 1705–1715

1075 Kishikawa T, Otsuka M, Suzuki T, Seimiya T, Sekiba K, Ishibashi R, Tanaka E, Ohno M,
1076 Yamagami M & Koike K (2018) Satellite RNA increases DNA damage and accelerates
1077 tumor formation in mouse models of pancreatic cancer. *Mol Cancer Res* 16: 1255–1262

1078 Kishikawa T, Otsuka M, Yoshikawa T, Ohno M, Ijichi H & Koike K (2016) Satellite RNAs
1079 promote pancreatic oncogenic processes via the dysfunction of YBX1. *Nat Commun* 7:
1080 13006

1081 Kornienko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, Cleary C, Kralovics R,
1082 Pauler FM & Barlow DP (2016) Long non-coding RNAs display higher natural expression
1083 variation than protein-coding genes in healthy humans. *Genome Biol* 17: 14

1084 Kou J-T, Ma J, Zhu J-Q, Xu W-L, Liu Z, Zhang X-X, Xu J-M, Li H, Li X-L & He Q (2020) LncRNA
1085 NEAT1 regulates proliferation, apoptosis and invasion of liver cancer. *Eur Rev Med*
1086 *Pharmacol Sci* 24: 4152–4160

1087 Lai M, Yang Z, Zhou L, Zhu Q, Xie H, Zhang F, Wu L, Chen L & Zheng S (2012) Long non-
1088 coding RNA MALAT-1 overexpression predicts tumor recurrence of hepatocellular
1089 carcinoma after liver transplantation. *Med Oncol* 29: 1810–1816

1090 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle
1091 M, FitzHugh W, *et al* (2001) Initial sequencing and analysis of the human genome. *Nature*
1092 409: 860–921

1093 Lanzafame M, Bianco G, Terracciano LM, Ng CKY & Piscuoglio S (2018) The role of long
1094 non-coding RNAs in hepatocarcinogenesis. *Int J Mol Sci* 19

1095 Li G, Shi H, Wang X, Wang B, Qu Q, Geng H & Sun H (2019) Identification of diagnostic long
1096 non-coding RNA biomarkers in patients with hepatocellular carcinoma. *Mol Med Rep* 20:
1097 1121–1130

1098 Li Z, Liu B, Jin W, Wu X, Zhou M, Liu VZ, Goel A, Shen Z, Zheng L & Shen B (2018) hDNA2
1099 nuclease/helicase promotes centromeric DNA replication and genome stability. *EMBO J*
1100 37

1101 Liao Y, Smyth GK & Shi W (2019) The R package Rsubread is easier, faster, cheaper and
1102 better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res*

1103 Lin R, Maeda S, Liu C, Karin M & Edgington TS (2007) A large noncoding RNA is a marker
1104 for murine hepatocellular carcinomas and a spectrum of human carcinomas. *Oncogene*
1105 26: 851–858

1106 Liu S, Qiu J, He G, Liang Y, Wang L, Liu C & Pan H (2019) LncRNA MALAT1 acts as a miR-
1107 125a-3p sponge to regulate FOXM1 expression and promote hepatocellular carcinoma
1108 progression. *J Cancer* 10: 6649–6659

1109 MacParland SA, Liu JC, Ma X-Z, Innes BT, Bartczak AM, Gage BK, Manuel J, Khuu N,
1110 Echeverri J, Linares I, *et al* (2018) Single cell RNA sequencing of human liver reveals
1111 distinct intrahepatic macrophage populations. *Nat Commun* 9: 4383

1112 Matouk IJ, DeGroot N, Mezan S, Ayesh S, Abu-lail R, Hochberg A & Galun E (2007) The H19
1113 non-coding RNA is essential for human tumor growth. *PLoS ONE* 2: e845

1114 Mattick JS & Makunin IV (2006) Non-coding RNA. *Hum Mol Genet* 15 Spec No 1: R17-29

1115 Melé M, Mattioli K, Mallard W, Shechner DM, Gerhardinger C & Rinn JL (2017) Chromatin
1116 environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs.
1117 *Genome Res* 27: 27–37

1118 Mudge JM, Jungreis I, Hunt T, Gonzalez JM, Wright JC, Kay M, Davidson C, Fitzgerald S,
1119 Seal R, Tweedie S, *et al* (2019) Discovery of high-confidence human protein-coding
1120 genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci. *Genome*
1121 *Res*

1122 Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Grutzner F & Kaessmann H (2014)
1123 The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505:
1124 635–640

1125 Nojima T, Tellier M, Foxwell J, Ribeiro de Almeida C, Tan-Wong SM, Dhir S, Dujardin G, Dhir
1126 A, Murphy S & Proudfoot NJ (2018) Deregulated expression of mammalian lncRNA
1127 through loss of SPT6 induces R-loop formation, replication stress, and cellular
1128 senescence. *Mol Cell* 72: 970-984.e7

1129 Panzitt K, Tschernatsch MMO, Guelly C, Moustafa T, Stradner M, Strohmaier HM, Buck CR,
1130 Denk H, Schroeder R, Trauner M, *et al* (2007) Characterization of HULC, a novel gene
1131 with striking up-regulation in hepatocellular carcinoma, as noncoding RNA.
1132 *Gastroenterology* 132: 330–342

1133 Perteza M, Perteza GM, Antonescu CM, Chang T-C, Mendell JT & Salzberg SL (2015) StringTie
1134 enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*
1135 33: 290–295

1136 Perteza M, Shumate A, Perteza G, Varabyou A, Breitwieser FP, Chang Y-C, Madugundu AK,
1137 Pandey A & Salzberg SL (2018) CHESS: a new human gene catalog curated from
1138 thousands of large-scale RNA sequencing experiments reveals extensive transcriptional
1139 noise. *Genome Biol* 19: 208

1140 Pradeepa MM, McKenna F, Taylor GCA, Bengani H, Grimes GR, Wood AJ, Bhatia S &
1141 Bickmore WA (2017) Psp1/p52 regulates posterior Hoxa genes through activation of
1142 lncRNA Hottip. *PLoS Genet* 13: e1006677

1143 Quagliata L, Matter MS, Piscuoglio S, Arabi L, Ruiz C, Procino A, Kovac M, Moretti F,
1144 Makowska Z, Boldanova T, *et al* (2014) Long noncoding RNA HOTTIP/HOXA13
1145 expression is associated with disease progression and predicts outcome in
1146 hepatocellular carcinoma patients. *Hepatology* 59: 911–923

1147 Quagliata L, Quintavalle C, Lanzafame M, Matter MS, Novello C, di Tommaso L, Pressiani T,
1148 Rimassa L, Tornillo L, Roncalli M, *et al* (2018) High expression of HOXA13 correlates
1149 with poorly differentiated hepatocellular carcinomas and modulates sorafenib response
1150 in in vitro models. *Lab Invest* 98: 95–105

1151 Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Bruggmann SA, Goodnough LH, Helms JA,
1152 Farnham PJ, Segal E, *et al* (2007) Functional demarcation of active and silent chromatin
1153 domains in human HOX loci by noncoding RNAs. *Cell* 129: 1311–1323

1154 Schlackow M, Nojima T, Gomes T, Dhir A, Carmo-Fonseca M & Proudfoot NJ (2017)
1155 Distinctive patterns of transcription and RNA processing for human lincRNAs. *Mol Cell*
1156 65: 25–38

1157 Schultheiss CS, Laggai S, Czepukojc B, Hussein UK, List M, Barghash A, Tierling S, Hosseini
1158 K, Golob-Schwarzl N, Pokorny J, *et al* (2017) The long non-coding RNA H19 suppresses
1159 carcinogenesis and chemoresistance in hepatocellular carcinoma. *Cell Stress* 1: 37–54

1160 Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM & Sirotkin K (2001) dbSNP:
1161 the NCBI database of genetic variation. *Nucleic Acids Res* 29: 308–311

1162 Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth
1163 J, Hillier LW, Richards S, *et al* (2005) Evolutionarily conserved elements in vertebrate,
1164 insect, worm, and yeast genomes. *Genome Res* 15: 1034–50

1165 Smit AFA, Hubley R & Green P (2003) RepeatMasker Open-4.0.

1166 Sonesson C, Love MI & Robinson MD (2015) Differential analyses for RNA-seq: transcript-level
1167 estimates improve gene-level inferences. *F1000Res* 4: 1521

1168 Soumillon M, Necseulea A, Weier M, Brawand D, Zhang X, Gu H, Barthès P, Kokkinaki M, Nef
1169 S, Gnirke A, *et al* (2013) Cellular source and mechanisms of high transcriptome
1170 complexity in the mammalian testis. *Cell Rep* 3: 2179–2190

1171 Spielmann M, Lupiáñez DG & Mundlos S (2018) Structural variation in the 3D genome. *Nat*
1172 *Rev Genet* 19: 453–467

1173 Talbert PB & Henikoff S (2018) Transcribing centromeres: noncoding RNAs and kinetochore
1174 assembly. *Trends Genet* 34: 587–599

1175 Tietze L & Kessler SM (2020) The good, the bad, the question-H19 in hepatocellular
1176 carcinoma. *Cancers (Basel)* 12

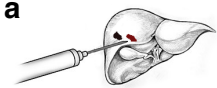
1177 Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, Coffman EJ, Contino G,
1178 Deshpande V, Iafrate AJ, Letovsky S, *et al* (2011) Aberrant overexpression of satellite
1179 repeats in pancreatic and other epithelial cancers. *Science* 331: 593–596

1180 Topp CN, Zhong CX & Dawe RK (2004) Centromere-encoded RNAs are integral components
1181 of the maize kinetochore. *Proc Natl Acad Sci USA* 101: 15986–15991

1182 Unfried JP, Serrano G, Suárez B, Sangro P, Ferretti V, Prior C, Boix L, Bruix J, Sangro B,
1183 Segura V, *et al* (2019) Identification of coding and long non-coding RNAs differentially
1184 expressed in tumors and preferentially expressed in healthy tissues. *Cancer Res*

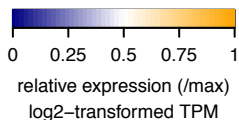
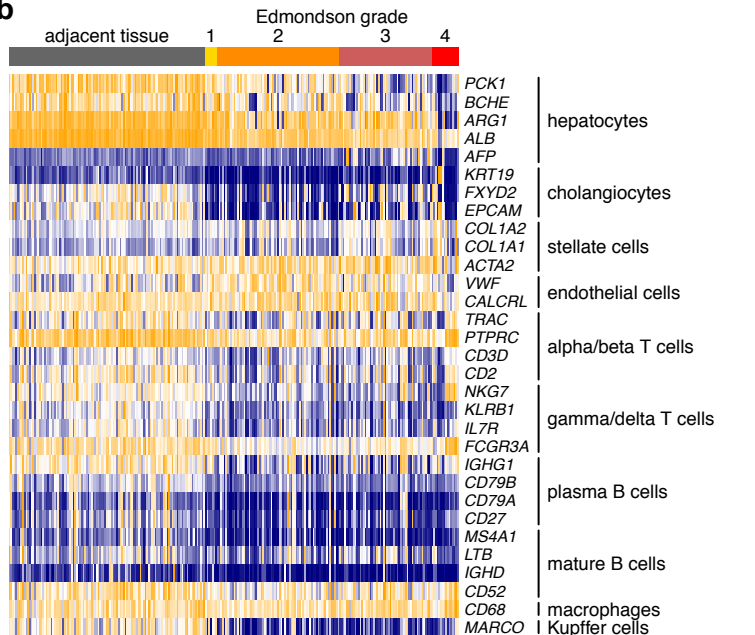
1185 UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*
1186 47: D506–D515

1187 Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan
1188 T, Shakir K, Roazen D, Thibault J, *et al* (2013) From FastQ data to high confidence
1189 variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc*
1190 *Bioinformatics* 43: 11.10.1-11.10.33
1191 Wang F, Ying H-Q, He B-S, Pan Y-Q, Deng Q-W, Sun H-L, Chen J, Liu X & Wang S-K (2015)
1192 Upregulated lncRNA-UCA1 contributes to progression of hepatocellular carcinoma
1193 through inhibition of miR-216b and activation of FGFR1/ERK signaling pathway.
1194 *Oncotarget* 6: 7899–7917
1195 Washietl S, Kellis M & Garber M (2014) Evolutionary dynamics and tissue specificity of human
1196 long noncoding RNAs in six mammals. *Genome Res* 24: 616–28
1197 Yan X, Hu Z, Feng Y, Hu X, Yuan J, Zhao SD, Zhang Y, Yang L, Shan W, He Q, *et al* (2015)
1198 Comprehensive genomic characterization of long non-coding RNAs across human
1199 cancers. *Cancer Cell* 28: 529–540
1200 Yang JD & Roberts LR (2010) Hepatocellular carcinoma: A global view. *Nat Rev Gastroenterol*
1201 *Hepatol* 7: 448–458
1202 Yang Y, Chen L, Gu J, Zhang H, Yuan J, Lian Q, Lv G, Wang S, Wu Y, Yang Y-CT, *et al*
1203 (2017) Recurrently deregulated lncRNAs in hepatocellular carcinoma. *Nature*
1204 *Communications* 8: 14421
1205 Yoshimizu T, Miroglio A, Ripoche M-A, Gabory A, Vernucci M, Riccio A, Colnot S, Godard C,
1206 Terris B, Jammes H, *et al* (2008) The H19 locus acts in vivo as a tumor suppressor. *Proc*
1207 *Natl Acad Sci USA* 105: 12417–12422
1208 Zhou Y, Fan R-G, Qin C-L, Jia J, Wu X-D & Zha W-Z (2019) LncRNA-H19 activates
1209 CDC42/PAK1 pathway to promote cell proliferation, migration and invasion by targeting
1210 miR-15b in hepatocellular carcinoma. *Genomics* 111: 1862–1872
1211 Zhu A, Ibrahim JG & Love MI (2019) Heavy-tailed prior distributions for sequence count data:
1212 removing the noise and preserving large differences. *Bioinformatics* 35: 2084–2092
1213



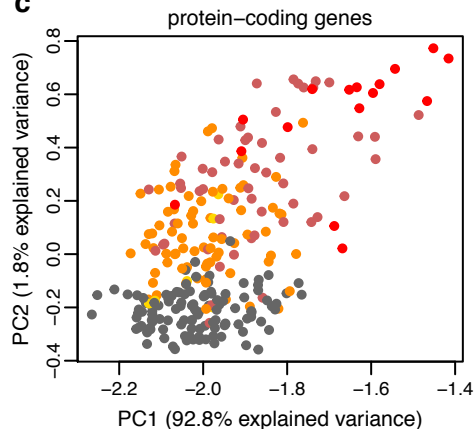
114 patients
151 tumor biopsies
117 adjacent tissue biopsies

b

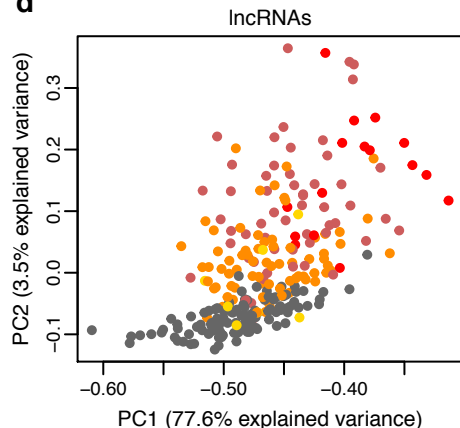


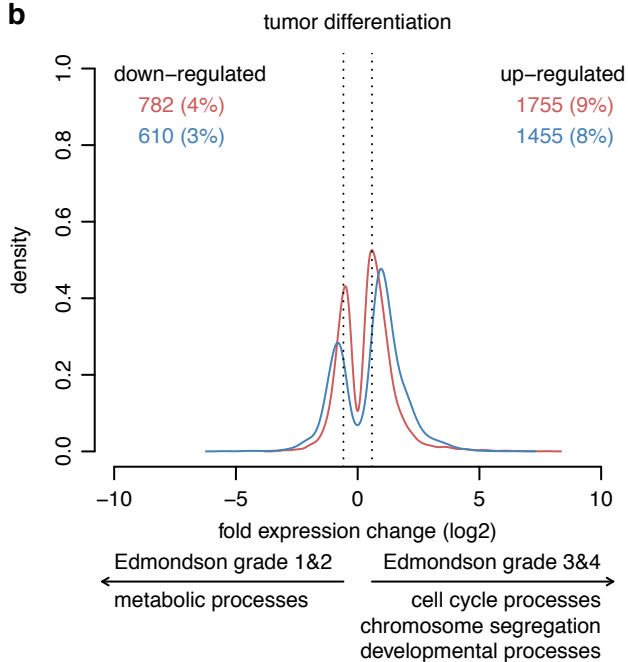
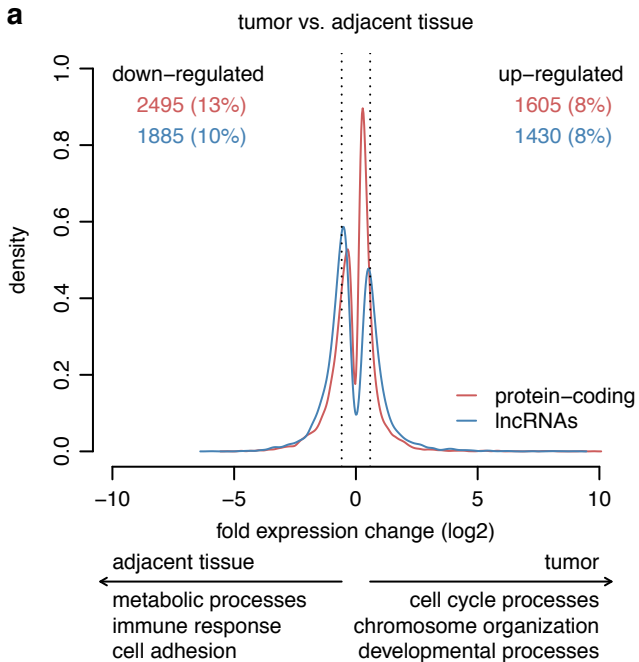
- adjacent tissue (n=117)
- Edmondson grade 1 (n=7)
- Edmondson grade 2 (n=73)
- Edmondson grade 3 (n=55)
- Edmondson grade 4 (n=16)

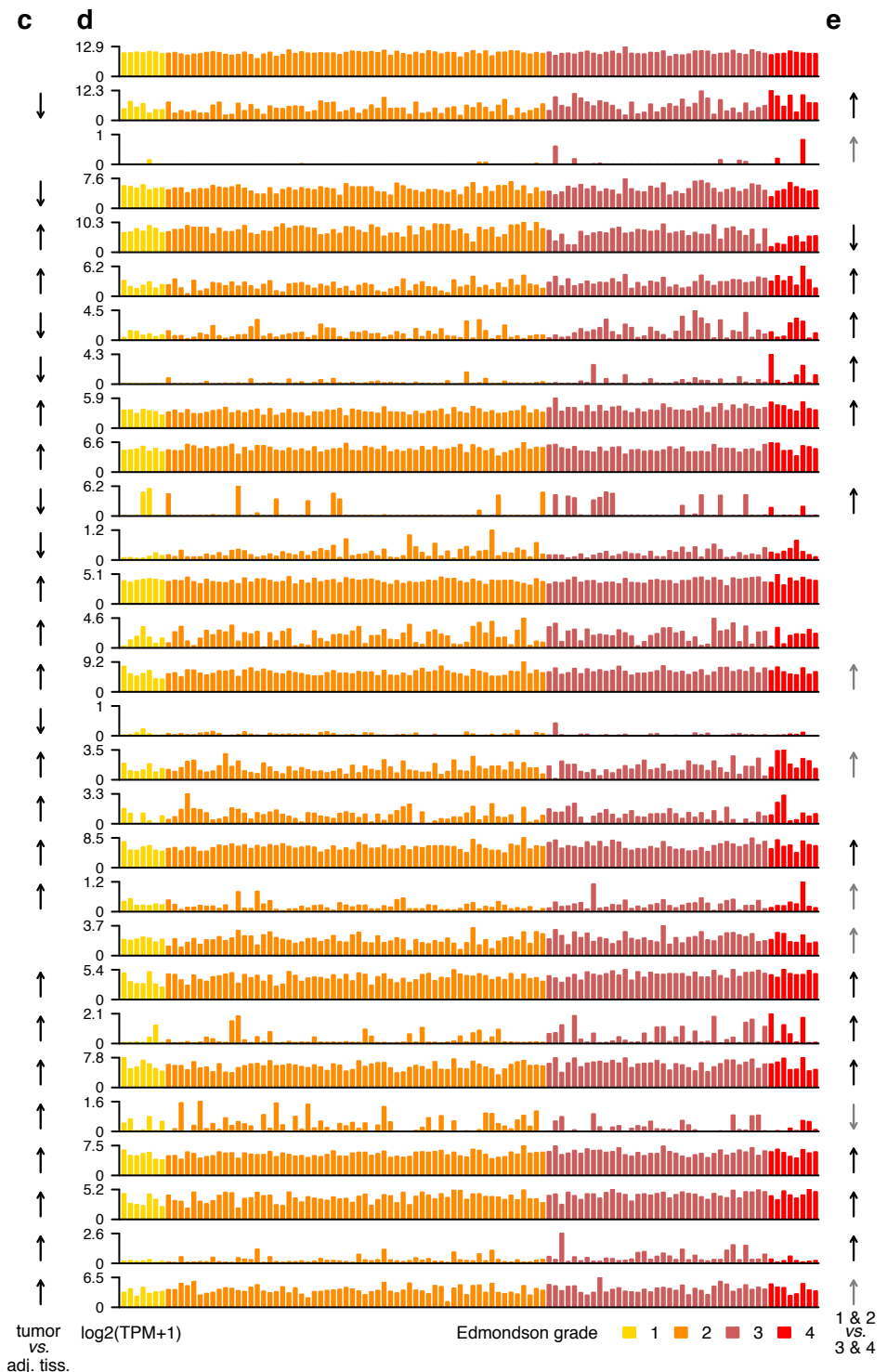
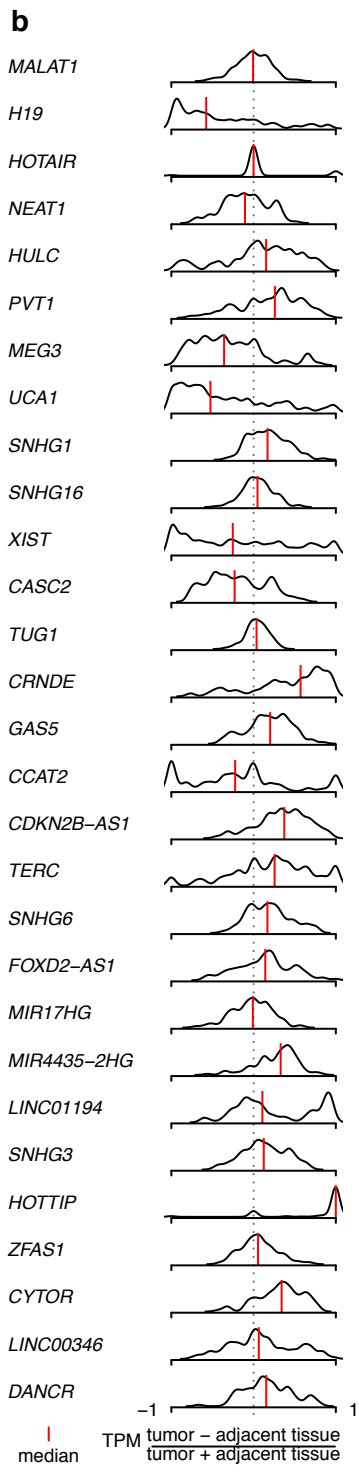
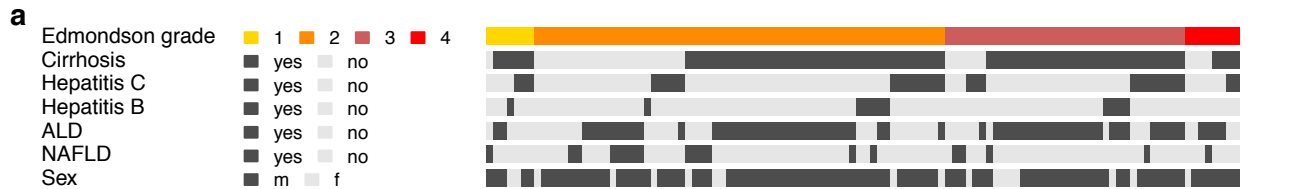
c

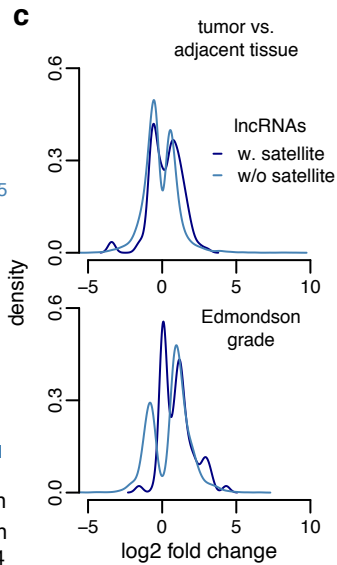
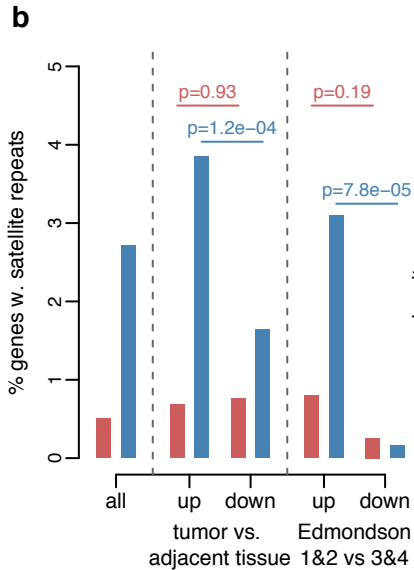
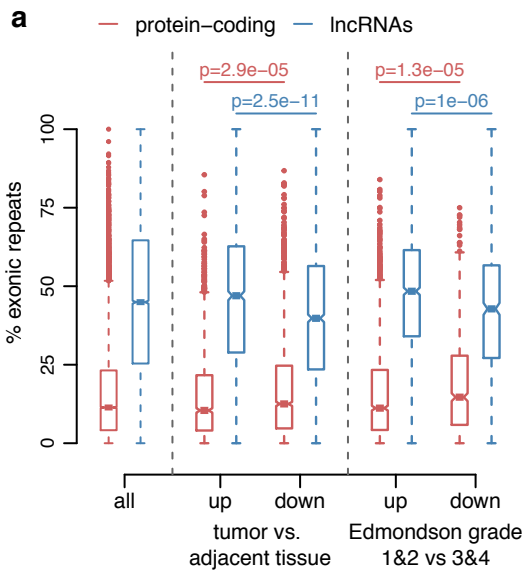


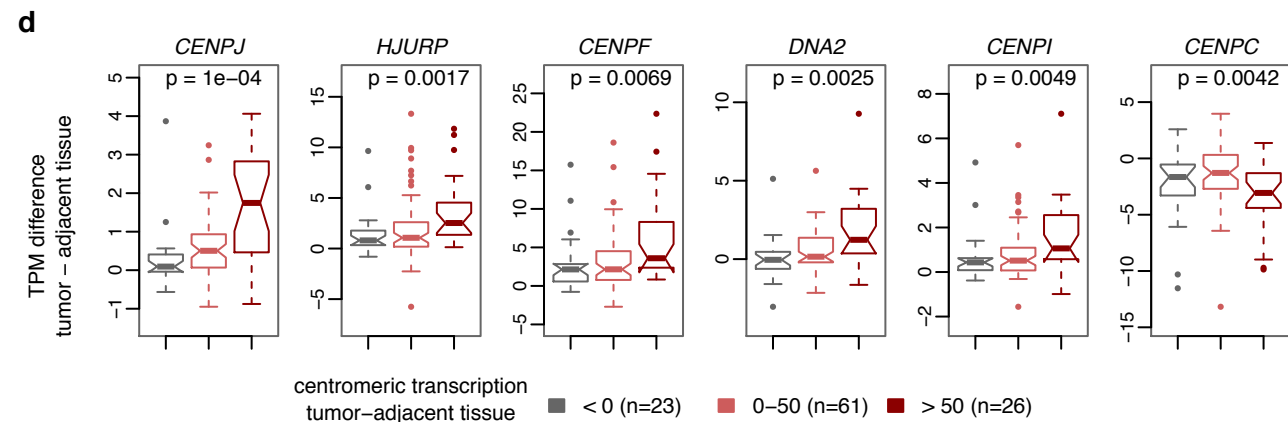
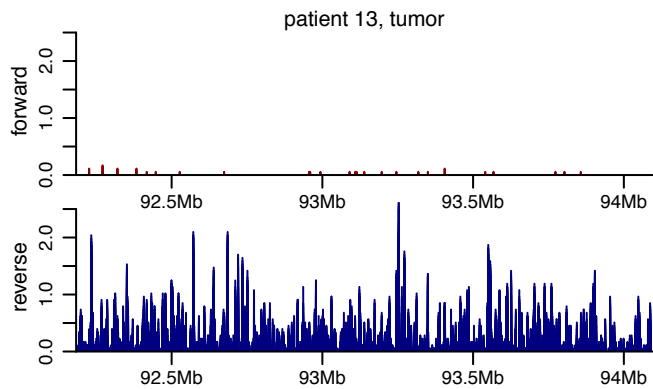
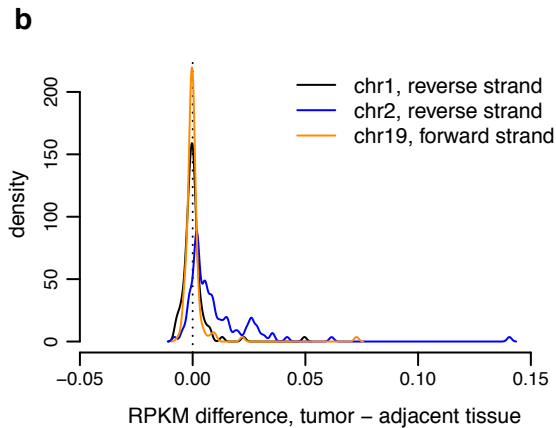
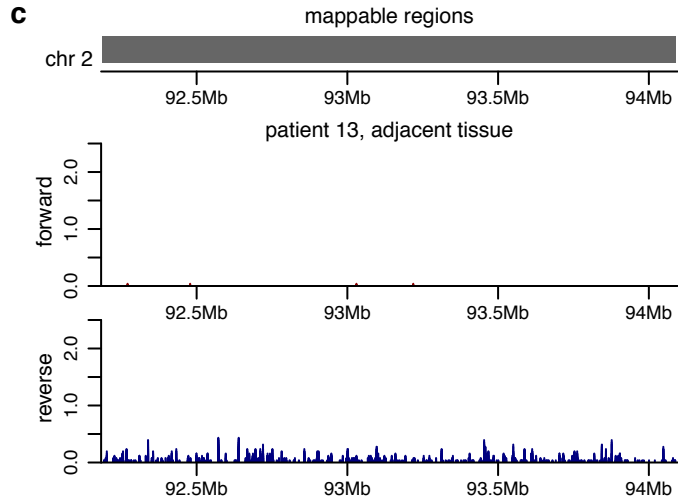
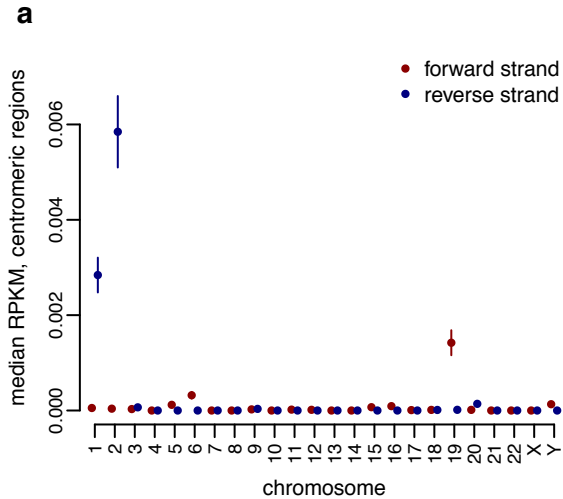
d

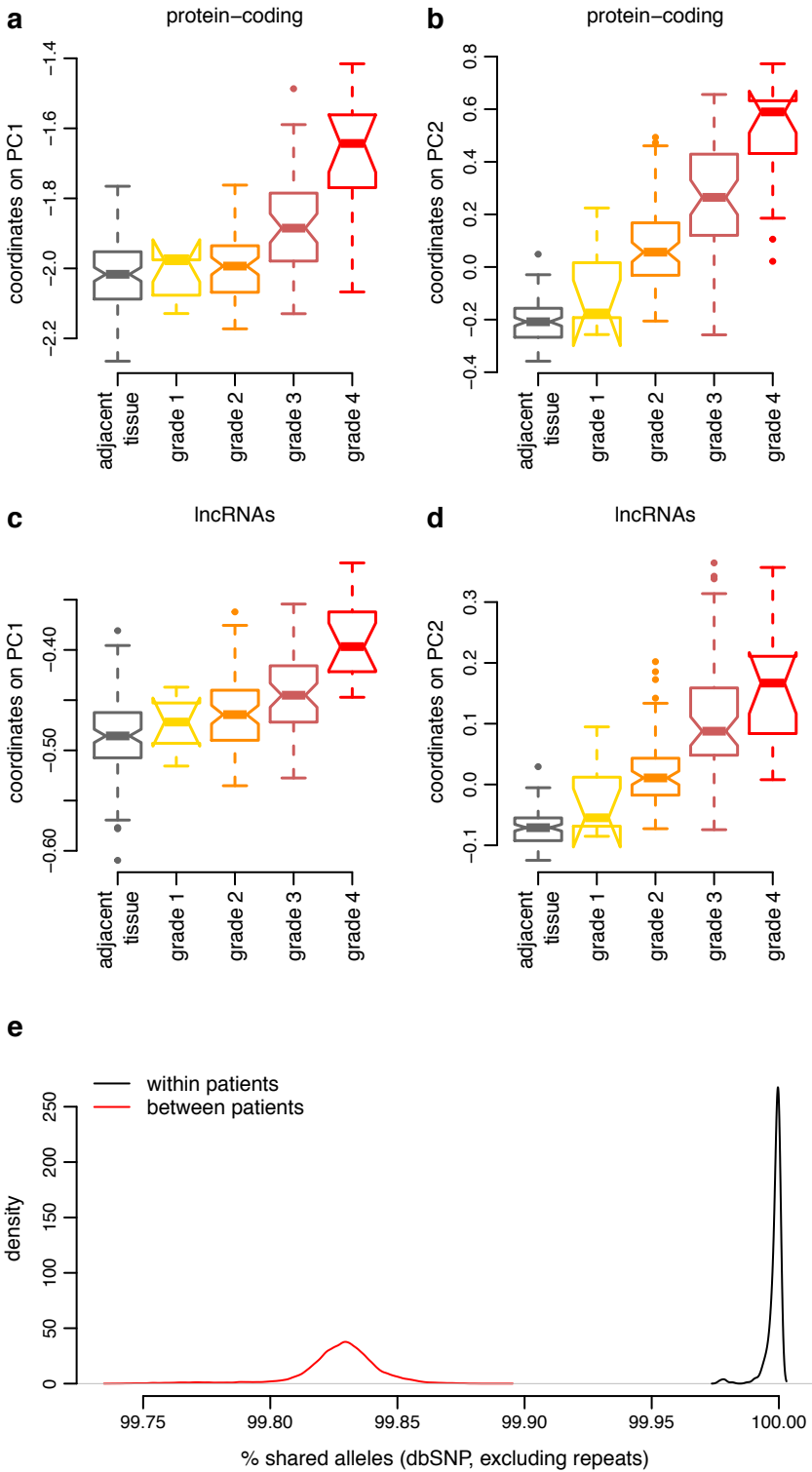


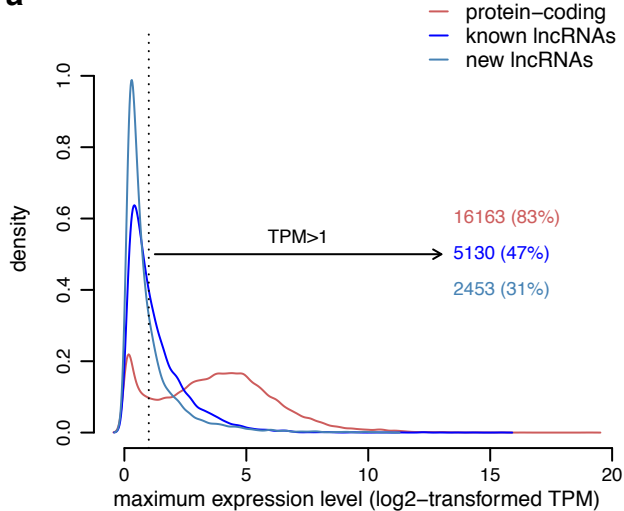
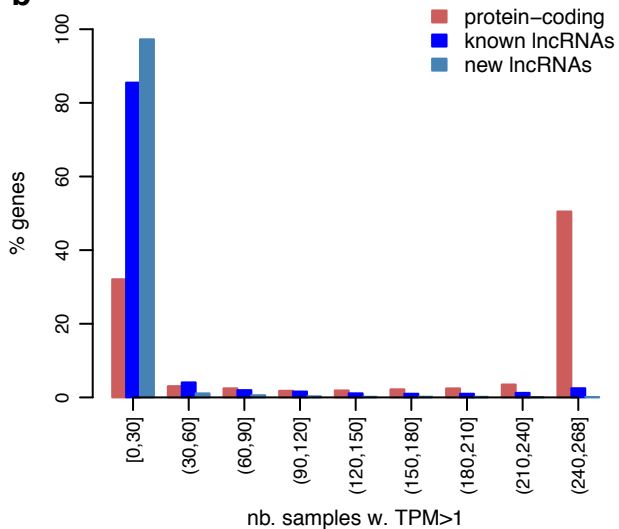


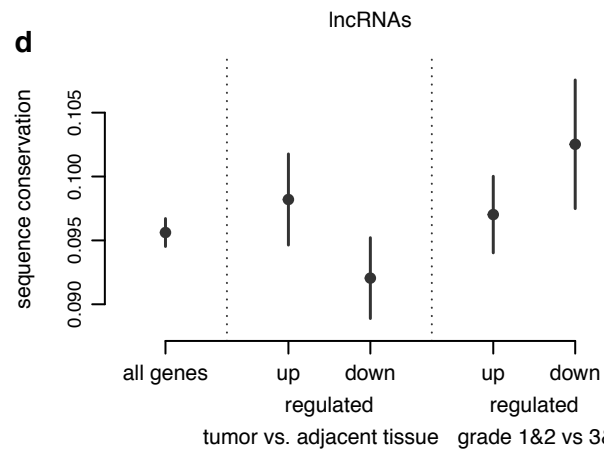
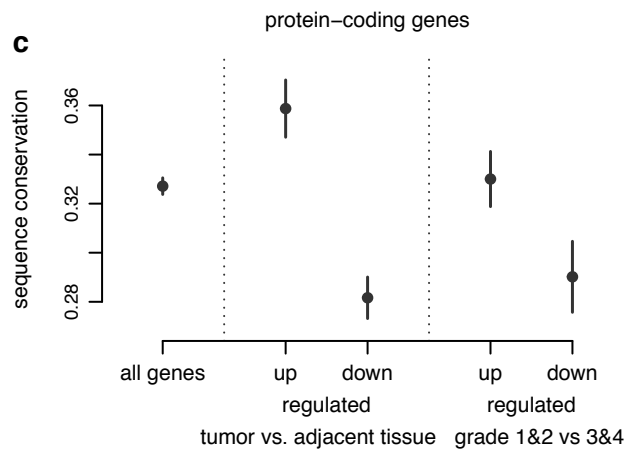
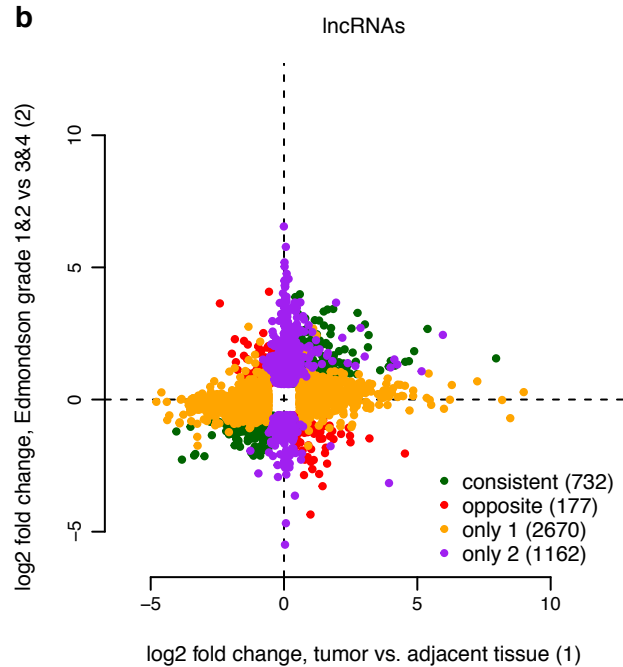
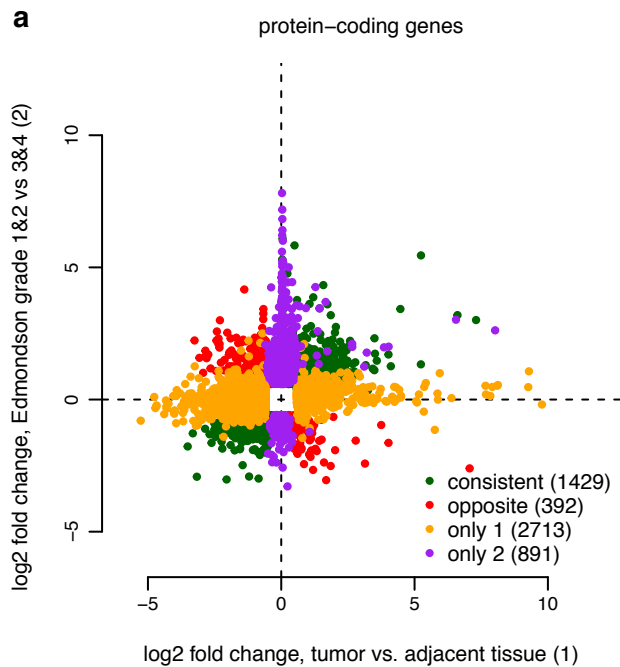


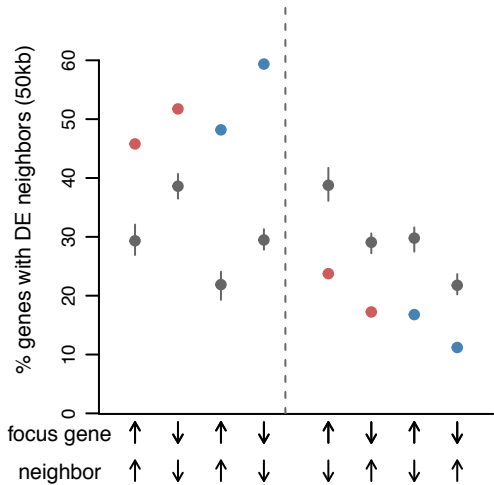
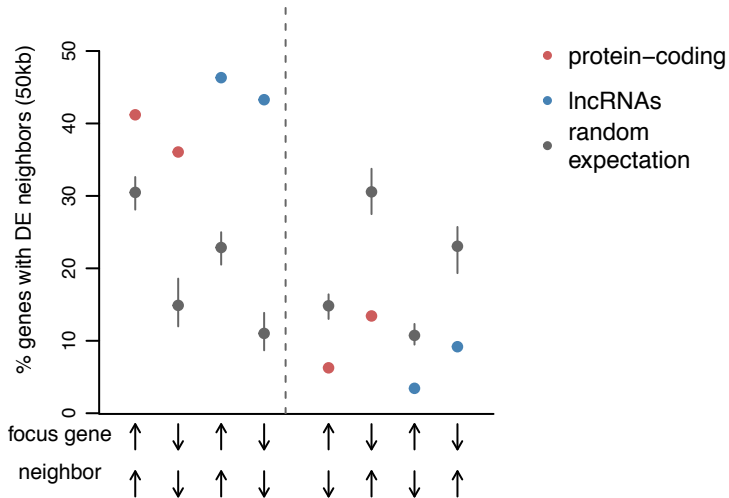


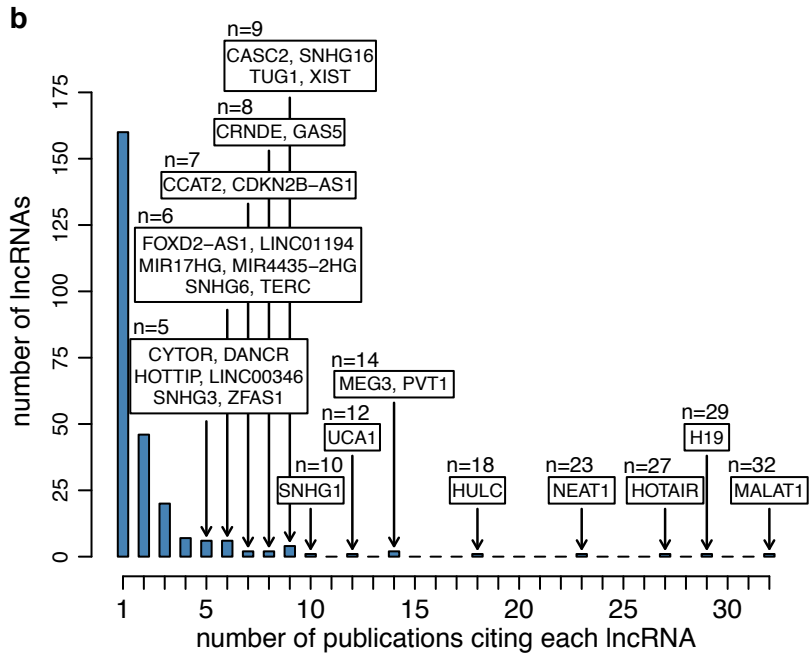
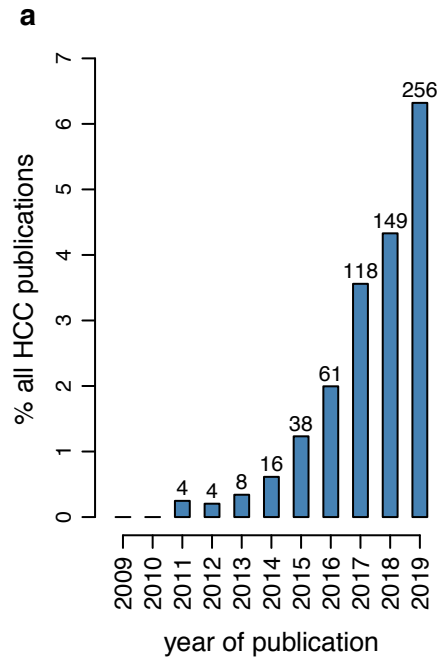


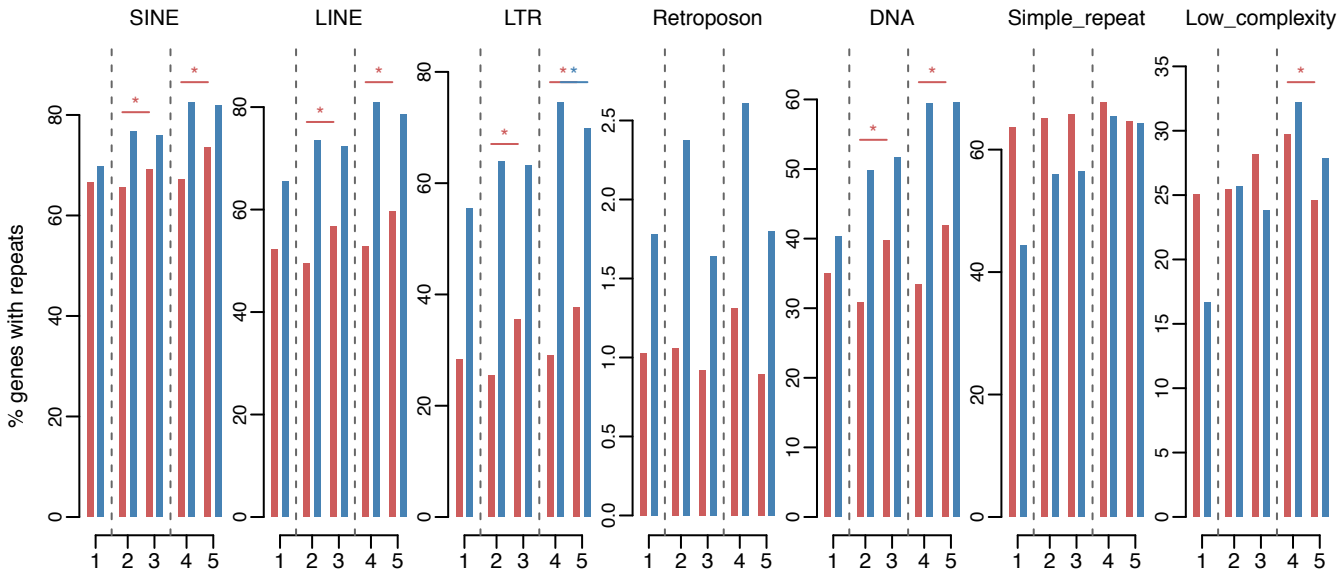


a**b**



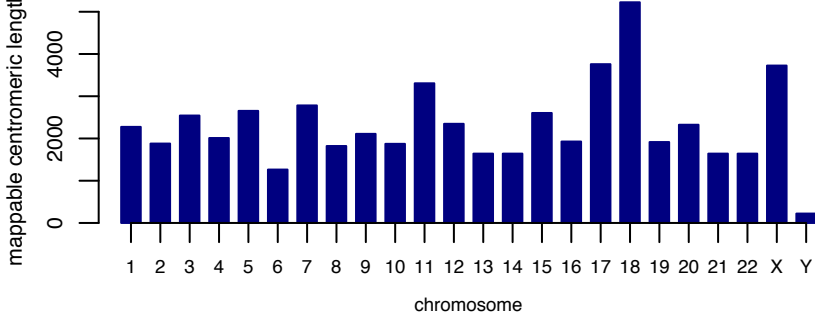
a tumor vs. adjacent tissue**b** Edmondson grade 1&2 vs. 3&4





1: all genes 2: up-regulated tumor vs. adjacent tissue 3: down-regulated tumor vs. adjacent tissue
 4: up-regulated grade 1&2 vs 3&4 5: down-regulated grade 1&2 vs 3&4

■ protein-coding
 ■ lncRNAs

a**b**