Charting extracellular transcriptomes in The Human Biofluid RNA Atlas

Eva Hulstaert (1,2,3), Annelien Morlion (1,2), Francisco Avila Cobos (1,2), Kimberly Verniers (1,2), Justine Nuytens (1,2), Eveline Vanden Eynde (1,2), Nurten Yigit (1,2), Jasper Anckaert (1,2), Anja Geerts (4), Pieter Hindryckx (4), Peggy Jacques (5,6), Guy Brusselle (7), Ken R. Bracke (7), Tania Maes (7), Thomas Malfait (7), Thierry Derveaux (8), Virginie Ninclaus (8), Caroline Van Cauwenbergh (8), Kristien Roelens (9), Ellen Roets (9), Dimitri Hemelsoet (10), Kelly Tilleman (11), Lieve Brochez (2,3), Scott Kuersten (12), Lukas Simon (13), Sebastian Karg (14), Alexandra Kautzky-Willers (15), Michael Leutner (15), Christa Nöhammer (16), Ondrej Slaby (17,18,19), Willinge Prins Roméo (20), Koster Jan (20), Steve Lefever (1,2), Gary P. Schroth (12), Jo Vandesompele (1,2)*, Pieter Mestdagh (1,2)*

1 Center for Medical Genetics, Department of Biomolecular Medicine, Ghent University, C. Heymanslaan 10, 9000, Ghent, Belgium
2 Cancer Research Institute Ghent (CRIG), Ghent University, C. Heymanslaan 10, 9000, Ghent, Belgium
3 Department of Dermatology, Ghent University Hospital, C. Heymanslaan 10, 9000, Ghent, Belgium
4 Department of Gastroenterology, Ghent University Hospital, C. Heymanslaan 10, 9000, Ghent, Belgium
5 Department of Rheumatology, Ghent University Hospital, C. Heymanslaan 10, 9000, Ghent, Belgium
6 VIB Inflammation Research Center, Ghent University, Technologiepark 71, 9052, Ghent, Belgium
7 Department of Respiratory Medicine, Ghent University Hospital, C. Heymanslaan 10, 9000, Ghent, Belgium
8 Department of Ophthalmology, Ghent University Hospital, C. Heymanslaan 10, 9000, Ghent, Belgium
9 Department of Obstetrics, Women's Clinic, Ghent University Hospital, C. Heymanslaan 10, 9000, Ghent, Belgium
10 Department of Neurology, Ghent University Hospital, C. Heymanslaan 10, 9000, Ghent, Belgium
11 Department of Reproductive Medicine, Ghent University Hospital, C. Heymanslaan 10, 9000, Ghent, Belgium
12 Illumina, San Diego, CA, USA
13 University of Texas Health Science Center, School of Biomedical Informatics, Center for Precision Health, Houston, TX, USA
14 Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany
15 Department of Internal Medicine III, Clinical Division of Endocrinology and Metabolism, Unit of Gender Medicine, Medical University of Vienna, Vienna, Austria
16 Austrian Institute of Technology, Center for Health and Bioresources, Molecular Diagnostics, Vienna, Austria
17 Masaryk Memorial Cancer Institute, Department of Comprehensive Cancer Care, Brno, Czech Republic
18 Department of Pathology, University Hospital Brno, Brno, Czech Republic
19 Central European Institute of Technology, Masaryk University, Brno, Czech Republic
20 Department of Oncogenomics, Amsterdam University Medical Centers (AUMC), University of Amsterdam, the Netherlands
* equally contributing authors

Abstract

Extracellular RNAs present in biofluids have emerged as potential biomarkers for disease. Where most studies focus on plasma or serum, other biofluids may contain more informative RNA molecules, depending on the type of disease. Here, we present an unprecedented atlas of messenger, circular and small RNA transcriptomes of a comprehensive collection of 20 different human biofluids. By means of synthetic spike-in controls, we compared RNA content across biofluids, revealing a more than 10 000-fold difference in RNA concentration. The circular RNA fraction is increased in nearly all biofluids compared to tissues. Each biofluid transcriptome is enriched for RNA molecules derived from specific tissues and cell types. In addition, a subset of biofluids, including stool, sweat, saliva and sputum, contains high levels of bacterial RNAs. Our atlas enables a more informed selection of the most relevant biofluid to monitor particular diseases. To verify the biomarker potential in these biofluids, four validation cohorts representing a broad spectrum of diseases were profiled, revealing numerous differential RNAs between case and control subjects. Taken together, our results reveal novel insights in the RNA content of human biofluids and may serve as a valuable resource for future biomarker studies. All spike-normalized data is publicly available in the R2 web portal and serve as a basis to further explore the RNA content in biofluids.

Keywords

RNA-sequencing, biofluids, circular RNA, messenger RNA, small RNA, biomarker, extracellular RNA, cell-free RNA

66 **Introduction**

67 Extracellular RNAs (exRNAs) in blood and other biofluids are emerging as potential
68 biomarkers for a wide range of diseases[1–6]. These so-called liquid biopsies may offer a non-
69 invasive alternative to tissue biopsies for both diagnosis and treatment response monitoring.
70 Previous studies have extensively profiled the small RNA content of several biofluids and
71 identified large differences in the small RNA content amongst different biofluids.[1–12] These
72 efforts were gathered by the NIH Extracellular RNA Communication Consortium in the
73 exRNA Atlas Resource (https://exrna-atlas.org).[8] Besides microRNAs (miRNAs), the most
74 studied small RNA biotype in biofluids, other small RNAs, such as piwi-interacting RNAs
75 (piRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), ribosomal RNAs
76 (rRNAs), transfer RNA fragments (tRNAs) and Y-RNAs have also been identified[5–7,9,12,13].
77 Weber et al.[13] was the first to compare the miRNA content in 12 different human biofluids
78 (pooled samples of plasma, saliva, tears, urine, amniotic fluid, colostrum, breast milk,
79 bronchial lavage fluid, cerebrospinal fluid, peritoneal fluid, pleural fluid and seminal plasma)
80 using reverse transcription quantitative polymerase chain reaction (RT-qPCR) of selected
81 miRNAs. Large variations in RNA concentration were observed among the different biofluids,
82 with the highest small RNA concentrations measured in breast milk and seminal fluid. Since
83 the advent of small RNA sequencing, other small RNA biotypes were characterized in various
84 biofluids, such as plasma, serum, stool, urine, amniotic fluid, bronchial lavage fluid, bile,
85 cerebrospinal fluid (CSF), saliva, seminal plasma and ovarian follicle fluid[5,7,9,9,12]. The
86 distribution of small RNA biotypes clearly varies across these biofluids, with a high
87 abundance of piRNAs and tRNAs reported in urine and a high abundance of Y-RNAs in
88 plasma[6,7,12]. Also non-human RNA sequences, mapping to bacterial genomes, were reported
89 in plasma, urine and saliva[6].
90 A systematic RNA-sequencing analysis of biofluids to explore the messenger RNAs (mRNA)
91 and circular RNA (circRNA) transcriptome is challenging due to low RNA concentration and
92 RNA fragmentation in biofluids. As such, most studies have explored the abundance of
93 individual mRNAs in one specific biofluid by RT-qPCR[14–20]. CircRNAs have been reported in
94 saliva[21], semen[22], blood[23] and urine[24,25]. Recently, the mRNA content of plasma and serum
95 has been investigated using dedicated sequencing approaches like Phospho-RNA-Seq,
96 SILVER-seq and SMARTer Stranded Total RNA-Seq method[26–29]. Studies comparing the small

97 RNA, mRNA and circRNA content in a wide range of human biofluids are currently lacking
98 and are essential to explore the biomarker potential of exRNAs.
99 The goal of the Human Biofluid RNA Atlas is to define the extracellular transcriptome across
100 a wide range of human biofluids (amniotic fluid, aqueous humor, ascites, bile, bronchial
101 lavage fluid, breast milk, cerebrospinal fluid, colostrum, gastric fluid, pancreatic cyst fluid,
102 plasma, saliva, seminal fluid, serum, sputum, stool, synovial fluid, sweat, tear fluid and urine)
103 and to assess biomarker potential in selected case-control cohorts. We used small RNA-
104 sequencing to quantify different small RNA species and present a dedicated mRNA-capture
105 sequencing workflow to simultaneously quantify mRNAs and circRNAs.
106 In the first phase of our study, small RNA sequencing and mRNA capture sequencing was
107 performed in a discovery cohort of 20 different biofluids (Fig. 1). The goal of this phase was
108 to assess the technical feasibility of the methodology and to generate a comprehensive set
109 of mRNAs, circRNAs and small RNAs in which the contributing tissues and cell types per
110 biofluid were assessed.
111 In the second phase of our study, we aimed to investigate the biological relevance of exRNAs
112 in various biofluids. Therefore, mRNA capture sequencing was applied to four different
113 case/control cohorts, each consisting of 16-24 samples (Fig. 1). These samples included
114 sputum samples from 8 patients with chronic obstructive pulmonary disease (COPD) versus
115 8 controls, urine samples from 12 bladder cancer patients versus 12 controls, CSF samples
116 from 12 glioblastoma patients versus 12 hydrocephalus patients and saliva samples from 12
117 diabetes mellitus patients versus 12 controls.
118 The resulting catalog of extracellular transcriptomes of 185 human samples can guide
119 researchers in the biomarker field to investigate other biofluids besides the well-studied
120 blood-derived ones and is a first step to more dedicated mRNA and circRNA profiling of
121 biofluids in larger cohorts.

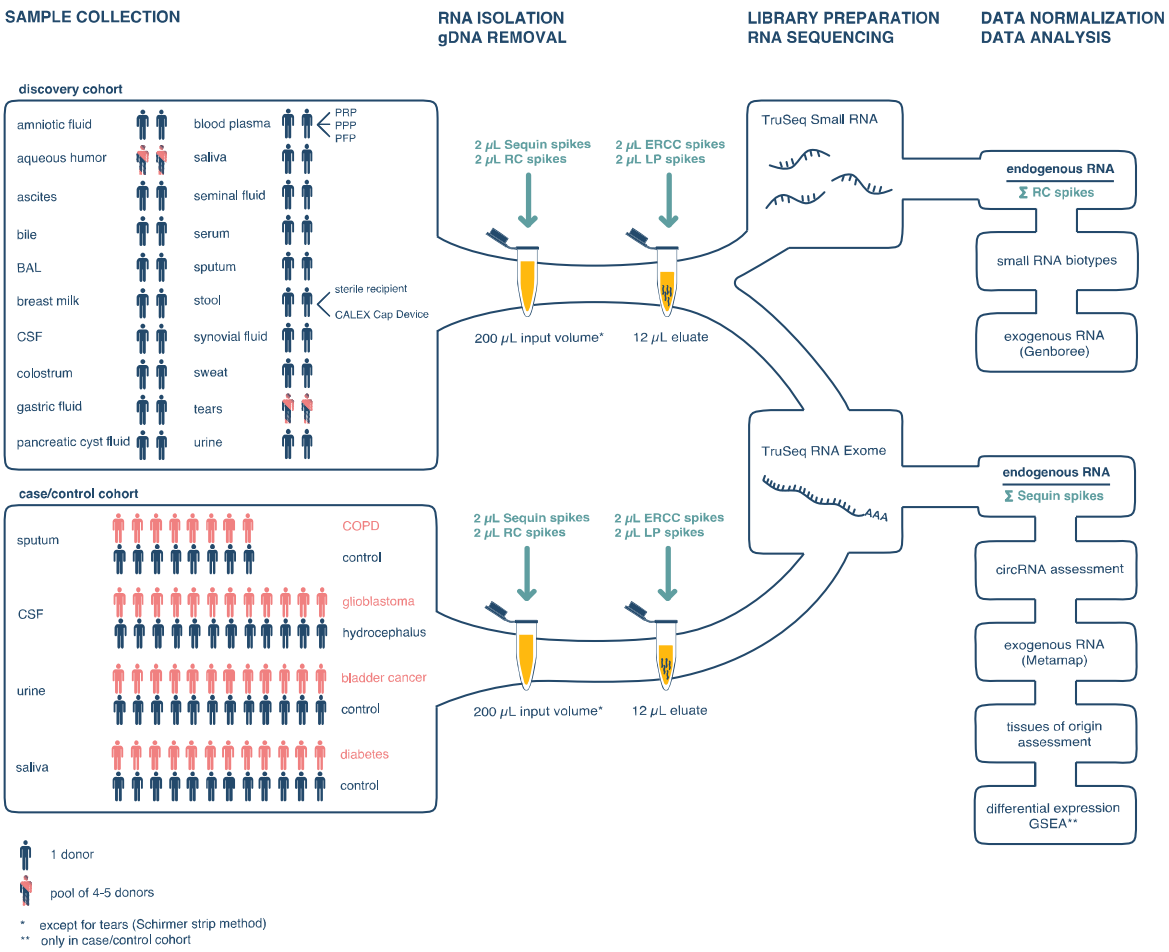


Fig.1 Study flow chart

In the discovery cohort, 20 different biofluids were collected in two donors or in a pool of 4-5 donors. In the case/control cohorts, selected biofluids (sputum, CSF, urine and saliva) were collected in 8-12 patients and an equal number of healthy controls. Both small RNA sequencing and mRNA capture sequencing were performed in the discovery cohort. In the case/control cohorts, mRNA capture sequencing was performed. To compare the RNA content across the different biofluids, the RC spikes and the Sequin spikes are used for normalization of small RNA and mRNA data, respectively.

BAL, bronchoalveolar lavage fluid; CSF, cerebrospinal fluid; PRP, platelet-rich plasma; PPP, platelet-poor plasma; PFP, platelet-free plasma

134 **Results**

135 ***RNA spike-in controls enable process control of the RNA sequencing workflow***

136 Synthetic spike-in RNA sequences are crucial to control the process from RNA isolation to
137 RNA sequencing, especially when working with challenging and low input material. We
138 applied 4 different mixes of synthetic RNA spike-in controls (in total 189 RNAs) as workflow
139 processing and normalization controls that enable direct comparison of the RNA profiles
140 across the different biofluids. Sequin and Small RNA extraction Control (RC) spikes were
141 added prior to RNA isolation whereas External RNA Control Consortium (ERCC) spikes and
142 small RNA Library Prep (LP) spikes were added to the RNA eluate prior to genomic DNA
143 (gDNA) removal (Fig. 1). Of note, every spike mix consists of multiple RNA molecules of
144 different lengths over a wide concentration range. Detailed information is provided in
145 Supplementary Note 1. Besides normalization, the spike-in controls enabled quality control
146 of the RNA extraction and library preparation steps in the workflow and relative
147 quantification of the RNA yield and concentration across the different biofluids.

148 First, the correlation between the expected and the observed relative quantities for all four
149 spike mixes can be used to assess quantitative linearity. In the discovery cohort, the
150 expected and the observed relative quantities for all four spike mixes were well correlated
151 (Pearson correlation coefficients range from 0.50 to 1.00 for Sequin spikes, 0.92 to 1.00 for
152 ERCC spikes, 0.44 to 0.98 for RC spikes and 0.40 to 0.96 for LP spikes). In some biofluids (e.g.
153 seminal plasma and tears), the sequencing coverage of spikes was low due to a high
154 concentration of endogenous RNA. Detailed information per sample is provided in
155 Supplementary Fig. 1.

156 The spike-in controls can also be used to assess the RNA isolation efficiency. The
157 Sequin/ERCC ratio and the RC/LP ratio reflect the relative mRNA and microRNA isolation
158 efficiency, respectively. A 170-fold and 104-fold difference in RNA isolation efficiency across
159 the samples was observed when assessing long and small RNAs, respectively (Supplementary
160 Fig. 2). These differences underline the challenges of working with heterogenous samples
161 and the importance of spike-in controls for proper data normalization and cross-sample
162 comparison of results.

163 Finally, the spikes can be utilized to normalize the endogenous RNA abundance data. In this
164 study, we applied a biofluid volume-based normalization by dividing the RNA reads

165 consumed by the endogenous transcripts by the sum of the Sequin spikes for mRNA data
166 and by the sum of the RC spikes for small RNA data. The spike-normalized data represent
167 relative abundance values of RNA molecules proportional to the input volume. Of note,
168 there is an inverse relationship between the number of spike-in RNA reads and the number
169 of endogenous RNA reads. As such, the ratio between the sum of the reads consumed by the
170 endogenous transcripts and the total number of spike-in reads is a relative measure for the
171 RNA concentration of the various samples.

172 ***Highly variable mRNA and small RNA content among biofluids in the discovery cohort***

173 Both small RNAs and mRNAs were quantified in each of the 20 biofluids in the discovery
174 cohort. Mapping rates varied substantially across the different biofluids (Fig. 2A). In general,
175 the proportion of mapped reads was higher for the mRNA capture sequencing data (further
176 referred to as mRNA data) than for the small RNA sequencing data, in line with the fact that
177 human mRNAs were enriched using biotinylated capture probes during the library
178 preparation. The fraction of mapped reads in the mRNA data ranged from 16% in stool to
179 97% in seminal plasma. Low mapping rates were observed in stool, in one of the bile
180 samples and in saliva. Mapping rates for samples in the case/control cohorts are in line with
181 these of the discovery cohort (Supplementary Fig. 3A). In the small RNA sequencing data, the
182 proportion of mapped reads ranged from ~7% in stool, saliva and CSF to 95% in platelet-rich
183 plasma (PRP).

184 A 10 000-fold difference in mRNA and small RNA concentration was observed between the
185 lowest concentrated fluids, i.e. platelet-free plasma, urine and CSF, and the highest
186 concentrated biofluids, i.e. tears, seminal plasma and bile (Fig. 2B). The generalizability of
187 the difference in mRNA concentration between highly concentrated biofluids (seminal
188 plasma) and lowly concentrated biofluids (CSF) was confirmed in additional samples
189 (Supplementary Fig. 3B). In the discovery cohort a 5547-fold difference in mRNA
190 concentration is observed between seminal plasma and CSF; in independent validation
191 samples, a similarly large 19 851-fold difference in mRNA concentration is observed between
192 both biofluids. In the discovery cohort, the mRNA and miRNA concentrations were
193 significantly correlated across biofluids (Pearson correlation coefficient 0.76, p-value = 8.5e-
194 10, Fig. 2D). Normalized abundance levels of exRNAs were significantly correlated between
195 biological replicates within each biofluid (Supplementary Fig. 4). The median Pearson
196 correlation coefficient of the mRNA and the small RNA data was 0.84 and 0.92, respectively.

While the mRNA and miRNA data was well correlated in most biofluids (e.g. tears, colostrum, saliva), correlation in other biofluids (e.g. bile, pancreatic cyst fluid) was poor. These biofluids are obtained with a more challenging collection method involving echo-endoscopy, impacting the reproducibility of collection and the correlation of the RNA content between biological replicates.

The likelihood of identifying RNA biomarkers in a given biofluid will not only depend on its relative RNA concentration, but also on its RNA diversity, here approximated by the fraction of read counts consumed by the top 10 most abundant mRNAs/miRNAs (Fig. 2C). In aqueous humor, the top 10 mRNAs represent up to 70% of all reads, indicating that this fluid does not contain a rich mRNA repertoire. In both PRP and PPP, about 50% of all reads go to the top 10 mRNAs. While amniotic fluid has a median RNA concentration, this fluid seems to contain a diverse mRNA profile, with only 7% of all reads going to the top 10 mRNAs. When looking into the miRNA data, the top 10 miRNAs represent more than 90% of all reads in PFP, urine and serum. BAL contains the most diverse miRNA repertoire, with 57% of all reads going to the top 10 miRNAs. Similar conclusions with respect to biofluid exRNA diversity can be drawn based on the number of miRNAs/mRNAs representing 50% of the counts (Supplementary Fig. 5). RNA diversity is also reflected by the number of detected exRNAs. The total number of mRNAs and miRNAs detected with at least 4 counts in both samples of the same biofluid ranged from 13 722 mRNAs in pancreatic cyst fluid to 107 mRNAs in aqueous humor and from 231 miRNAs in tears to 18 miRNAs in stool (Table 1).

Table 1 Number of mRNAs and miRNAs per biofluid.

The number of mRNAs and miRNAs with at least 4 unique read counts in both replicates is shown per biofluid.

biofluid	number of mRNAs
amniotic fluid	10 531
aqueous humor	107
ascites	5578
BAL	3565
bile	2279
breastmilk	11 607
colostrum	11 914
CSF	438
gastric fluid	9288
pancreatic cyst fluid	13 722
PFP	2699
PPP	4548

biofluid	number of miRNAs
amniotic fluid	119
aqueous humor	20
ascites	75
BAL	126
bile	45
breastmilk	213
colostrum	229
CSF	32
gastric fluid	21
pancreatic cyst fluid	129
PFP	95
PPP	113

PRP	5440
saliva	6353
seminal plasma	11 868
serum	4152
sputum	7738
stool	134
stool Calex	135
sweat	410
synovial fluid	1614
tears	13 366
urine	2094

PRP	192
saliva	110
seminal plasma	211
serum	122
sputum	91
stool	19
stool Calex	18
sweat	45
synovial fluid	122
tears	231
urine	41

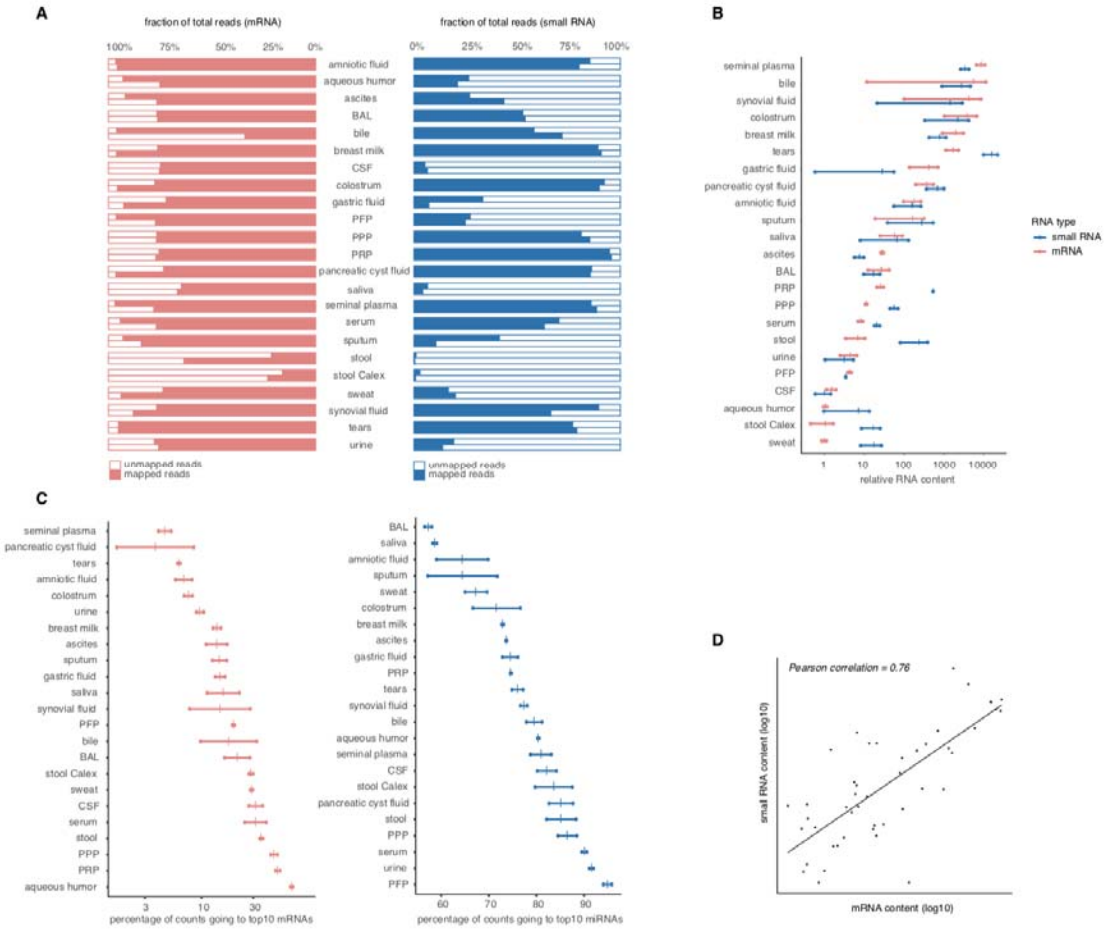


Fig. 2 mRNA and small RNA content varies across the 20 biofluids

(A) Percentage of the total read count mapping to the human transcriptome.

(B) Relative RNA concentration per biofluid; every dot represents the relative RNA concentration in one sample, every vertical mark indicates the mean per biofluid.

(C) The diversity of the RNA content expressed as fraction of read counts consumed by the top 10 most abundant mRNAs/miRNAs. Only genes with at least 4 unique reads

231 *are taken into account. Every dot represents the fraction in one sample, every vertical*
232 *mark indicates the mean percentage per biofluid.*
233 *(D) Correlation between the small RNA and the mRNA relative concentration. The*
234 *Pearson correlation coefficient is 0.76 (p-value = 8.58 x 10^{-10}). The correlation*
235 *coefficients is calculated on log10 transformed data.*
236 *BAL, bronchoalveolar lavage fluid; CSF, cerebrospinal fluid; PRP, platelet-rich plasma;*
237 *PPP, platelet-poor plasma; PFP, platelet-free plasma*
238
239 ***The distribution of small RNA biotypes varies across the different biofluids***
240 The distribution of small RNA biotypes shows distinct patterns among the 20 different
241 biofluids (Fig. 3). The exceptionally high percentage of miscellaneous RNAs (mainly Y-RNAs)
242 observed in blood-derived fluids is in line with a previous study[12] and with the Y-RNA
243 function in platelets. The fraction of reads mapping to miRNAs is lower than 15% in all
244 samples but platelet-free plasma and one synovial fluid sample. Tears, bile and amniotic
245 fluid have the highest fraction of tRNA fragments while saliva has the highest fraction of
246 piRNAs. The rRNA fraction is higher than 15% in all samples but tears, aqueous fluid and the
247 three plasma fractions. The majority of these reads map to the 45S ribosomal RNA
248 transcript. The not annotated read fraction contains uniquely mapped reads that could not
249 be classified in one of the small RNA biotypes. These reads most likely originate from
250 degraded longer RNAs, such as mRNAs and long non-coding RNAs.

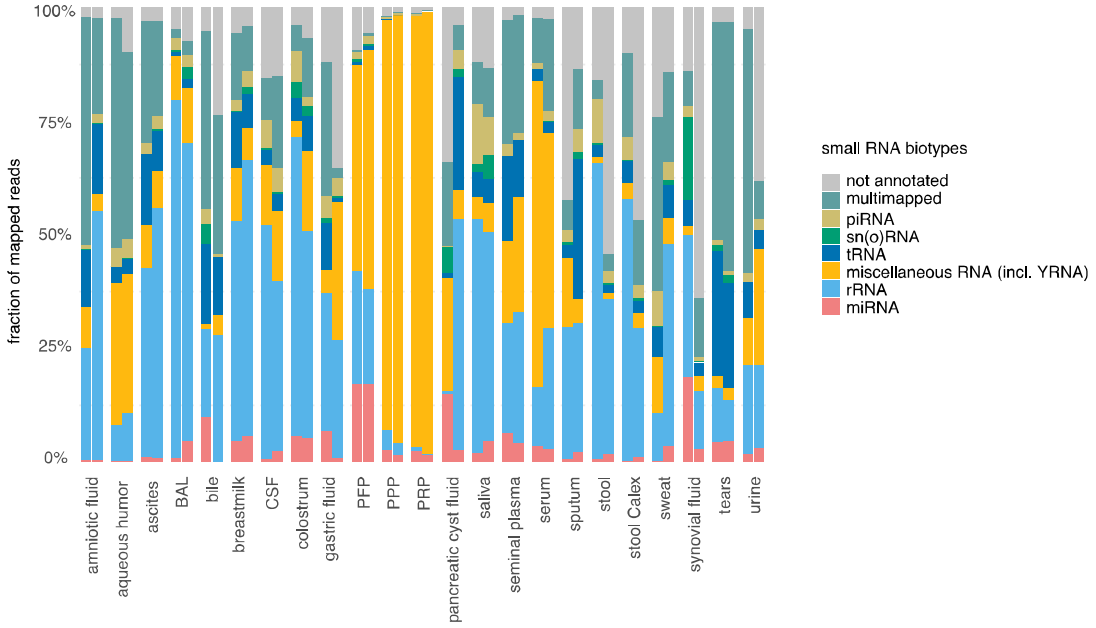
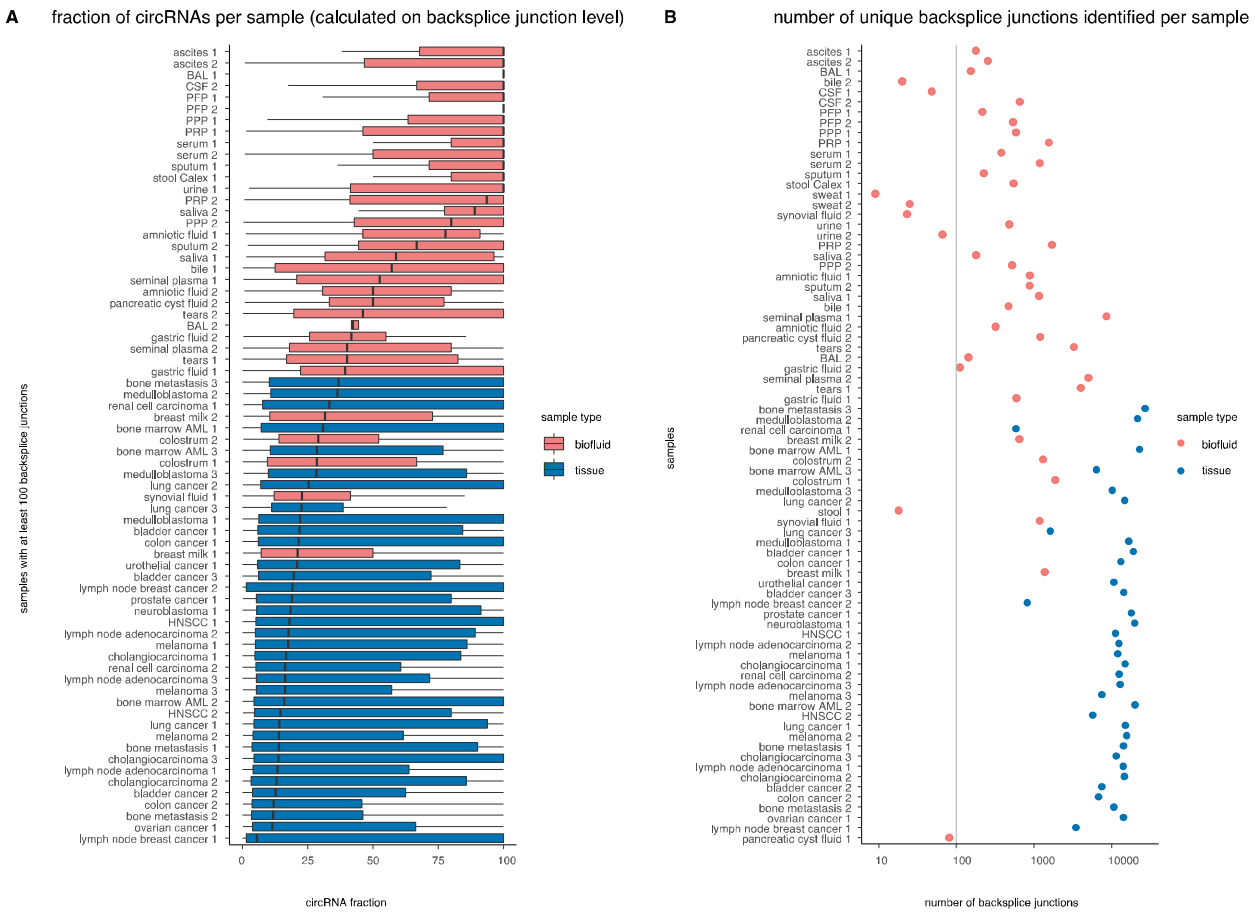


Fig. 3 Distinct small RNA biotype patterns are present across the different biofluids

The fraction of reads that align to small RNA biotypes are shown per biofluid. Only mapped reads of the small RNA sequencing data are taken into account. BAL, bronchoalveolar lavage fluid; CSF, cerebrospinal fluid; miRNA: microRNA; *PFP, platelet-free plasma; PPP, platelet-poor plasma; PRP, platelet-rich plasma;* piRNAs: piwi-interacting RNA; sn(o)RNAs: small nuclear and nucleolar RNAs; tRNAs: transfer RNA.

Circular RNAs are enriched in biofluids compared to tissues

CircRNAs are produced from unspliced RNA through a process called back-splicing where a downstream 5' donor binds to an upstream 3' acceptor. CircRNAs are resistant to endogenous exonucleases that target free 5' or 3' terminal ends. As a result, circRNAs are highly stable and have extended half-lives compared to linear mRNAs.[30] CircRNAs have been reported to be present in numerous human tissues[24] and in a few biofluids such as saliva[21], blood[31], semen[22] and urine[24,25]. A direct comparison of the circRNA read fraction between biofluids and tissues is currently lacking in literature. We compared the circRNA fraction, for genes that produce both linear and circular transcripts, identified through mRNA capture sequencing of the 20 biofluids in this study with the circRNA fraction identified in mRNA capture sequencing of 36 cancerous tissue types obtained from the MiOncoCirc Database[24]. While more unique backsplice junctions were identified in tissues compared to biofluids, in



line with the higher RNA concentration in tissues (Fig. 4B), the circRNA read fraction is clearly higher in biofluid exRNA compared to cellular RNA (Fig. 4A). The median circRNA read fraction in biofluids is 84.4%, which is significantly higher than the median circRNA read fraction in tissues of 17.5% (Mann-Whitney-U test, two-sided, p-value = 5.36×10^{-12}). For genes that produce both linear and circular transcripts, the stable circRNAs are more abundant than the linear mRNAs in biofluids, while it is the other way around in tissues.

We used two different methods to define the circRNA read fraction (see *"Circular RNA detection"* in methods; Supplementary Fig. 6): one based on individual backsplice junctions (shown in Fig. 4) and another method based on backsplice junctions aggregated at gene-level (Supplementary Fig. 7). Both methods clearly point towards a substantial enrichment of circRNAs in biofluids.

Fig. 4 CircRNAs are enriched in biofluids compared to tissues

284 *(A) The circRNA fraction, calculated at the backsplice junction level, is plotted per sample*
285 *and is higher in cell-free biofluid RNA than in tissue RNA. Only samples with at least*
286 *100 backsplice junctions are plotted.*
287 *(B) The number of unique backsplice junctions per sample is higher in tissues compared*
288 *to biofluids, in line with the higher input concentration of RNA into the library prep.*
289 *AML, acute myeloid leukemia; BAL, bronchoalveolar lavage fluid; CSF, cerebrospinal fluid;*
290 *HNSCC: head and neck squamous-cell carcinoma; PFP, platelet-free plasma; PPP, platelet-*
291 *poor plasma; PRP, platelet-rich plasma*
292
293 ***Assessment of exogenous RNA in human biofluids***
294 Two dedicated pipelines were used for the non-trivial assessment of the presence of
295 microbial or viral RNA in human biofluid extracellular RNA. Overall, the fraction of bacterial
296 reads is higher in small RNA sequencing data than in the mRNA data, in line with the
297 unbiased nature of small RNA sequencing and the targeted hybrid capture enrichment using
298 probes against human RNA during the mRNA capture library preparation. Stool (both
299 collection methods), sweat, saliva and sputum are among the biofluids with the highest
300 fraction of bacterial RNA in both the small RNA sequencing data and the mRNA data. The
301 percentage of bacterial reads in mRNA data and in small RNA data are significantly
302 correlated across biofluids (Pearson correlation coefficient 0.78, p-value = 1.94e-10).
303 Bacterial reads in aqueous humor and CSF, two fluids with very low endogenous RNA
304 content that were collected in a sterile setting (and thus presumed to be sterile), most likely
305 reflect background contamination during the workflow[32]. To illustrate the biological
306 relevance of the bacterial signal, we looked into reads mapping to *Campylobacter concisus*, a
307 gram-negative bacterium that is known to primarily colonize the human oral cavity, with
308 some strains translocated to the intestinal tract[33]. We confirm the selective presence of
309 reads mapping to *Campylobacter concisus* in saliva in both the small RNA and the mRNA
310 data(Fig. 5B). In all samples and for both the small RNA and the mRNA data, the percentage
311 of the total reads that maps to viral transcriptomes is less than 1%.
312

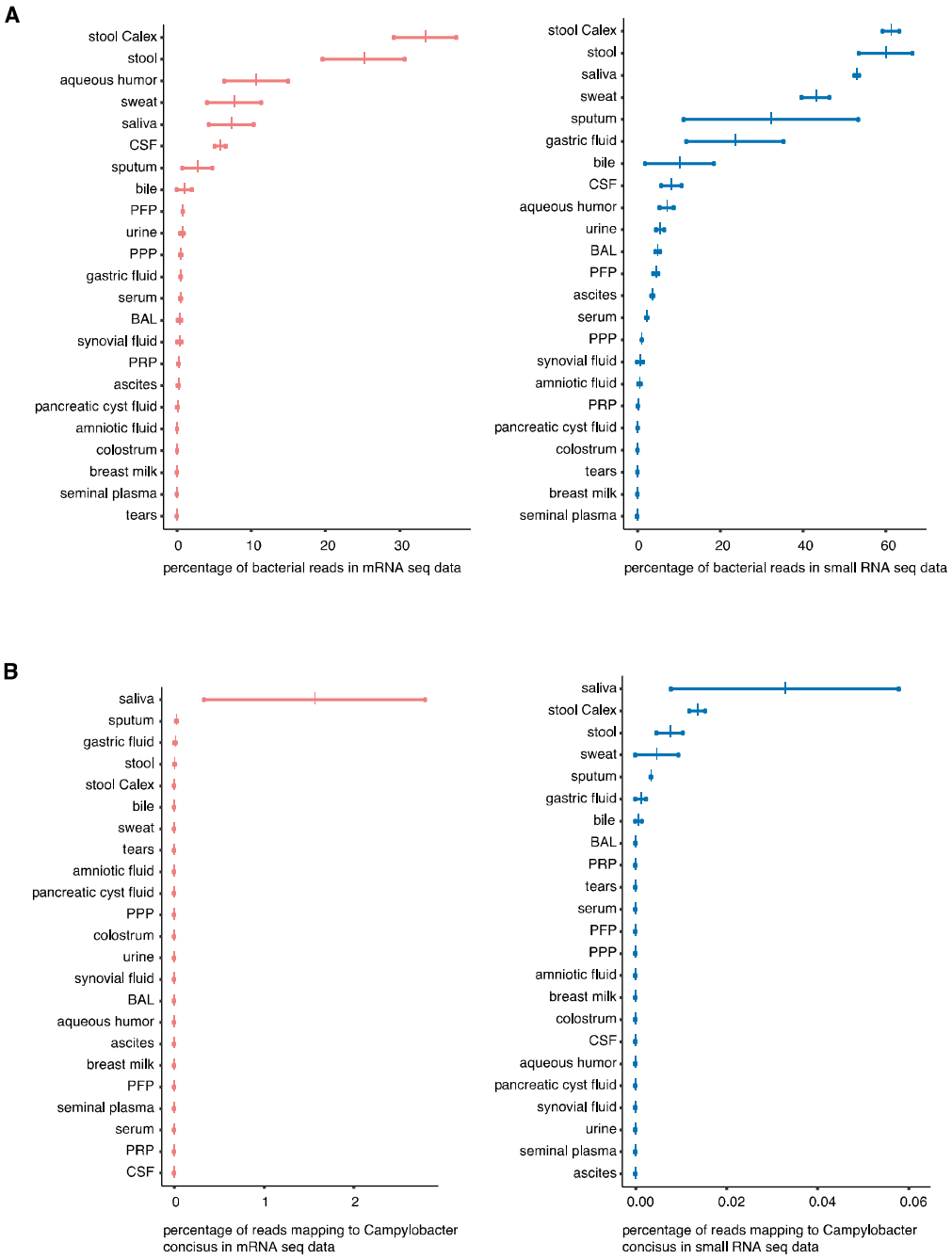


Fig. 5. Reads mapping to bacterial genomes

(A) Percentage of reads mapping to bacteria in mRNA data (pink) and in small RNA sequencing data (blue).

(B) Percentage of reads mapping to Campylobacter concisus in mRNA data (pink) and in small RNA sequencing data (blue). Campylobacter concisus is known to be present in saliva.

320 ***Assessment of the tissues of origin and deconvolution of pancreatic cyst fluid***

321 Gaining insights in tissue contribution to biofluid RNA profiles may guide the selection of the
322 most appropriate biofluid to investigate a given disease. To define tissues that specifically
323 contribute RNA molecules to individual biofluids, we explored the relationship between
324 extracellular mRNA levels and tissue or cell type specific mRNA signatures. The heatmap in
325 Fig. 6A highlights the relative contribution of tissues and cell types to a specific biofluid
326 compared to the other biofluids. More detailed results per biofluid are shown in
327 Supplementary Fig. 8. The results of this analysis were validated in an independent sample
328 cohort for CSF, saliva, sputum, seminal plasma and urine (Supplementary Fig. 3C). As
329 expected, prostate tissue RNA markers are more abundant in urine and in seminal plasma
330 than in any other biofluid. Both sputum and saliva contain mRNAs specific for trachea and
331 esophagus. In amniotic fluid, markers for esophagus, small intestine, colon and lung are
332 more abundant than the other tissues and cell types, probably reflecting organs that actively
333 shed RNA (at the gestational age of sampling) into the amniotic cavity. These data strongly
334 suggest that biofluid mRNA levels, at least to some degree, reflect intracellular mRNA levels
335 from cells that produce or transport the fluid. To further investigate the origin of biofluid
336 RNA at the cellular level, we applied computational deconvolution of the pancreatic cyst
337 fluid RNA profiles using single cell RNA sequencing data from 10 pancreatic cell types[34]. Fig.
338 6B reveals that pancreatic cyst fluid 1 consists of 45% of activated stellate cells and 43% of
339 endothelial cells, while pancreatic cyst fluid 2 mainly consists of quiescent stellate cells
340 (38%), endothelial cells (31%) and acinar cells (19%).

341

342

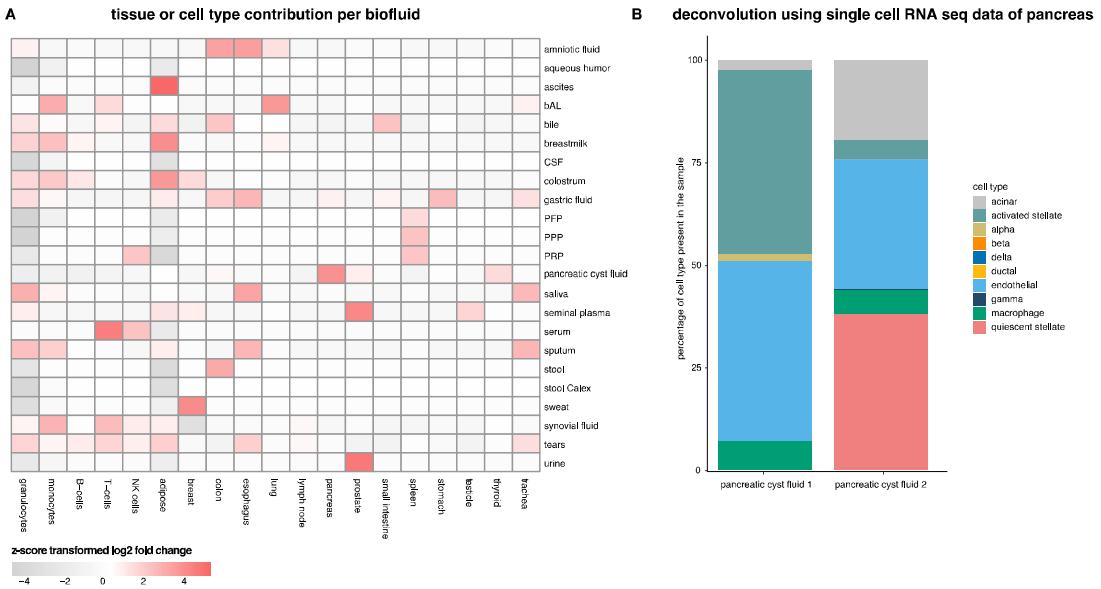


Fig. 6 Identification of the tissues of origin per biofluid and deconvolution of pancreatic cyst fluid

(A) Assessment of the tissues of origin in the biofluids of the discovery cohort.

Heatmap showing tissues and cell types that contribute more specifically to a certain biofluid compared to the other biofluids. Rows depict the biofluids of the discovery cohort and the columns are the tissues or cell types for which markers were selected based on the RNA Atlas[35]. For visualization purposes, only tissues and cell types with a z-score transformed log2 fold change ≥ |1| in at least one biofluid are shown.

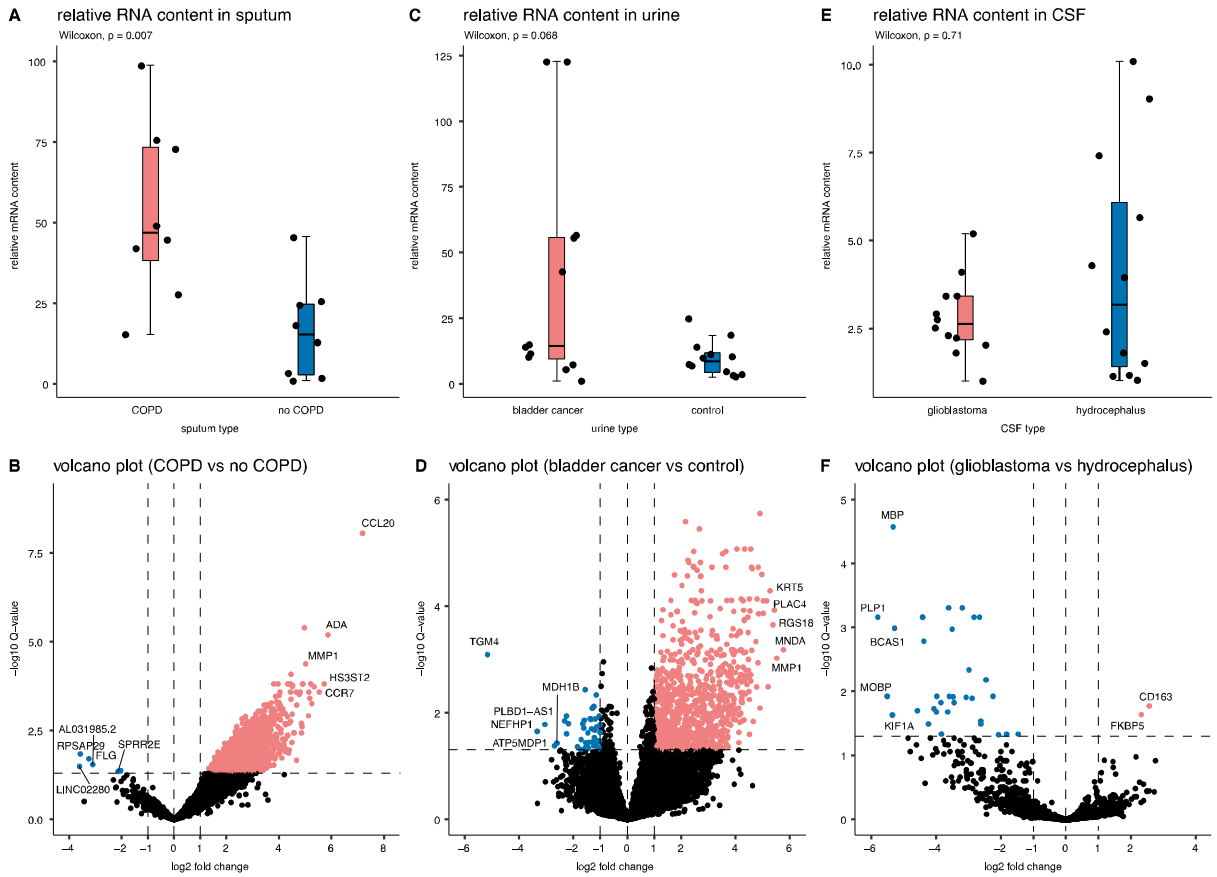
(B) Composition of pancreatic cyst fluid samples based on deconvolution using sequencing data from 10 pancreatic cell types.

Biomarker potential of mRNA in sputum, urine, CSF and saliva in selected case/control cohorts

Additional biofluid samples were collected in patients with a specific disease or in healthy controls to investigate potential biologically relevant differences in mRNA content between both groups. Sequin RNA spikes were used for biofluid volume-based data normalization. Strikingly, the relative RNA concentration in sputum of COPD patients was higher than in non-COPD patients, probably reflecting the high turnover of immune cells during the state of chronic inflammation (Fig. 7A). Differential expression analysis revealed 5513 and 6 mRNAs that were significantly up- and downregulated, respectively, in sputum from COPD patients

364 compared to healthy controls (Fig. 7B). CCL20, the most differential mRNA, showed a 146-
365 fold upregulation in COPD patients compared to healthy donors. This potent chemokine
366 attracting dendritic cells has previously been linked to the pathogenesis of COPD[36,37]. ADA
367 and MMP1, also among the most differential mRNAs, have also been associated with the
368 pathogenesis of COPD[38–40]. To verify the RNA-seq findings, 8/8 of the most differentially
369 abundant mRNAs were validated by RT-qPCR (Supplementary Fig. 9A-B).
370 In contrast to COPD, the relative RNA content is comparable in urine from bladder cancer
371 patients and healthy volunteers, in CSF from glioblastoma patients and hydrocephalus
372 patient, and in saliva from diabetes patients and healthy volunteers (Fig. 7C/E,
373 Supplementary Fig. 10). A higher RNA yield in CSF from glioblastoma patients compared to
374 CSF from healthy controls has been reported by Saugstad et al.[41], however the collection
375 method of CSF differed between both groups and it is therefore not possible to assess
376 whether the reported difference in RNA yield between both groups is due to the different
377 CSF collection site (lumbar puncture versus craniotomy) or due to the neurological disease.
378 In urine from patients with a muscle invaded bladder cancer, 529 mRNAs and 9 mRNAs were
379 significantly upregulated and downregulated, respectively, compared to urine from healthy
380 volunteers (Fig. 7D). Some of the upregulated mRNAs, such as MDK, SLC2A1, GPRC5A, KRT17
381 and KRT5, have been reported in urine and were suggested as biomarker for the accurate
382 detection and classification of bladder cancer[42–45]. In CSF from glioblastoma patients, only 2
383 mRNAs are significantly upregulated compared to CSF from hydrocephalus patients. CD163,
384 one of the upregulated genes in glioblastoma, has been linked with glioblastoma
385 pathogenesis[46]. In saliva from diabetes patients and saliva from healthy volunteers, no
386 differentially expressed genes could be identified. A list with differentially expressed genes
387 in all case/control cohorts can be found in Supplementary Data 5.
388 Differential abundance analysis was performed for circular RNAs as well, but in none of the
389 case/control cohorts differentially abundant circRNAs could be detected (data not shown).
390 As circular RNAs can only be identified based on their backsplice junction, the read coverage
391 is generally (too) low for biomarker discovery based on mRNA capture sequencing data.
392 When applying a similar strategy for mRNAs by looking at the reads of only one “linear only”
393 junction per gene (outside every detected back-splice junction) a significantly lower number
394 of differentially abundant mRNAs were detected (sputum: 13 out of 5519 mRNAs; urine: 0

395 out of 538 mRNAs; CSF: 0 out of 35 mRNAs). These results strongly suggest that a dedicated
396 circRNA enrichment strategies may be needed to assess circRNA biomarker potential.
397 To validate the identification of the 10 most abundant circRNAs detected by mRNA capture
398 sequencing in sputum, an orthogonal validation by RT-qPCR of the backsplice sequen region
399 was performed. For 9 of the 10 circRNAs, the RNA-sequencing results could be validated.
400 (Supplementary Fig. 9C)



401
402 ***Fig. 7 Relative RNA concentration and volcano plot in case/control cohorts***

403 *Top: Boxplots of relative mRNA content, bottom: Volcano plots of differentially expressed*
404 *mRNAs (q<0.05; pink up; blue down in patient vs. control) with labeling of up to 5 most*
405 *differential genes. (A) Sputum from COPD patients (n = 8) compared to sputum from healthy*
406 *donors (n = 8; Wilcoxon rank test, two-sided, p = 0.007); (B) 5513 and 6 mRNAs up and down,*
407 *respectively in COPD samples. (C) Urine from bladder cancer patients (n = 12) compared to*
408 *urine from healthy donors (n = 12; Wilcoxon signed-rank test, two-sided, p = 0.068). (D) 529*
409 *and 9 mRNAs up and down, respectively in bladder cancer samples. (E) CSF from*
410 *glioblastoma cancer patients (n = 12) compared to CSF from hydrocephalus patients (n = 12);*

411 *Wilcoxon signed-rank test, two-sided, p = 0.71); (D) 2 and 33 mRNAs up and down,*
412 *respectively in glioblastoma samples*

413 **Discussion**

414 By applying two complementary RNA-sequencing technologies on 20 different biofluids, we
415 assembled the most comprehensive human biofluid transcriptome, covering small RNAs,
416 mRNAs and circRNAs. Until now, most efforts to investigate and compare the RNA content
417 within biofluids focused on small RNA sequencing, most likely because of technical
418 limitations and unawareness of the abundance of extracellular mRNA (fragments)[5–7,9,12,13].
419 The availability of both small RNA sequencing data and mRNA data allows a more in-depth
420 characterization of the human transcriptome in biofluids. To our knowledge, this is the first
421 study reporting on the mRNA content, generated through a dedicated mRNA enrichment
422 sequencing method, in tear fluid, amniotic fluid, aqueous humor, bile, bronchial lavage fluid,
423 gastric fluid, saliva, seminal plasma, synovial fluid, sweat and urine. Selected mRNAs were
424 previously studied by means of RT-qPCR in amniotic fluid[14], pancreatic cyst fluid[15,18], seminal
425 plasma[16], sputum[17], stool[19] and in extracellular vesicles isolated from cell-free urine[20]. In
426 saliva, selected mRNAs were detected using microarrays[47]. We have demonstrated that it is
427 technically feasible to generate mRNA data from low input biofluid samples. This is expected
428 to accelerate biomarker research in these fluids. Further efforts to profile and share the
429 mRNA and circRNA content in larger sample cohorts of biofluids, comparable to the exRNA
430 Atlas Resource for small RNAs, are necessary to move this scientific field forward.[8]
431 Our small RNA results confirm previous studies observing high miRNA concentration in
432 tears[13], low mapping rates in CSF[5,48] and low miRNA concentration in cell-free urine[12]. A
433 direct comparison of the absolute number of detected miRNAs, mRNAs and circRNAs
434 detected per sample in our study with the numbers in published literature is hampered by
435 the fact that the absolute read count is dependent on the input volume of the biofluids, the
436 RNA isolation kit and library preparation method used, the sequencing depth and data-
437 analysis settings (e.g. mapping without mismatches, filtering of the data). In addition,
438 different pre-analytical variables when preparing the biofluid samples may also affect the
439 sequencing results. However, on a higher level, we can look into the most abundant miRNAs
440 detected in specific biofluids. The majority of the 10 most abundant miRNAs detected in 9

441 specific biofluids reported by Godoy et al. are also detected amongst the most abundant
442 miRNAs in the samples from the discovery cohort (Supplementary Data 10)[5].
443 We compared the mRNA results of the discovery cohort with these of the case/control
444 cohorts. Mapping rates for samples in the discovery cohort are in the same range for saliva,
445 sputum and seminal plasma. The mapping rates for CSF and urine are about 15% higher in
446 the case/control cohorts compared to the discovery cohort. These differences may be due to
447 different pre-analytical variables between both cohorts (collection tube, centrifugation
448 speed and the portion of urine collected (Supplementary Fig. 3A; Supplemental material and
449 methods).
450 In the discovery cohort on average 53% of all small RNA reads in saliva can be traced to
451 bacteria, perfectly in line with the average of 45.5% reads mapping to bacteria reported by
452 Yeri et al.[6] Aqueous humor and CSF, although collected in a sterile setting and presumed to
453 be sterile, contain up to 11% of reads mapping to bacteria, in line with previous studies[5,48].
454 However, bacterial cultures of our two CSF samples were negative. As both CSF and aqueous
455 humor display a very low relative RNA content, the exogenous sequences may represent
456 bacterial contaminants introduced during the sample processing workflow. Contaminants
457 can derive from contaminated spin columns used during RNA purification[32], enzymes
458 produced in microorganisms [49], or various environmental sources[50]. Such contaminant
459 signals are likely underrepresented in samples with high concentration of endogenous
460 exRNAs.
461 Although we collected a broad range of biofluids, only two samples per biofluid were
462 studied, limiting our ability to assess donor variability. The input volume for the RNA
463 isolations in all biofluids was set to 200 µL and a volume-based comparison of the RNA
464 content was made among the biofluids. We did not explore if higher input volumes would
465 result in higher RNA yields in biofluids where this could have been possible (e.g. urine). We
466 also note that the results in Table 1 are impacted by biofluid input volume in the RNA
467 purification, RNA input in the sequencing library prep, and the sequencing depth.
468 Biofluid data normalization with synthetic spike-in controls is a unique and powerful
469 approach and reflects more accurately the biological situation compared to classic
470 normalization approaches where global differences on overall abundance are neutralized.
471 For instance, the relative mRNA concentration in sputum from COPD patients is higher than
472 in sputum from healthy donors. Typically, RNA sequencing data is subsampled or normalized

473 based on the library size before performing a differential expression analysis, resulting in an
474 artificially more balanced volcano plot, an overcorrection of the biological situation and a
475 loss of information, which is not the case when the data is normalized based on spike-in
476 controls.
477 Our results highlighting tissues and cell types that contribute more specifically to a certain
478 biofluid compared to the other biofluids (Fig. 6A) can be used as a roadmap to formulate
479 hypotheses when initiating biomarker research. Not surprisingly, the RNA signal from
480 prostate is reflected in urine and seminal plasma. Both fluids can be collected in a non-
481 invasive way and may be of value to investigate further in prostate cancer patients. Of
482 interest, the mRNA concentration in seminal plasma is 1000-fold higher than in urine and
483 seminal plasma contains more unique mRNAs compared to urine, suggesting that the
484 biomarker potential of seminal plasma is higher. However, one should also be cautious in
485 interpreting the tissue enrichment results: while the RNA signal of breast seems relatively
486 enriched in sweat, this biofluid has the lowest RNA concentration. The limited number of
487 detected mRNAs in sweat show overlap with mRNAs related to secretion (MCL1 gene,
488 SCGB2A2 gene, SCGB1D2 gene) that also appear as markers in breast tissue.
489 The pancreatic tissue RNA signal appears to be enriched in pancreatic cyst fluid and a
490 different cell type composition is observed when both samples are deconvoluted using
491 single cell RNA sequencing data of pancreatic cell types (Fig. 6B). Pancreatic cyst fluid was
492 collected in these donors to investigate a cystic lesion in the pancreas. The routine
493 cytological analysis of these fluid samples was inconclusive at the moment of sample
494 collection. By following up both patients, we discovered that the first patient developed a
495 walled off necrosis collection after necrotizing pancreatitis. The incipient high fraction of
496 activated stellate cells in the first cyst fluid sample may have been an indication pointing
497 towards the inflammation and necrosis that finally occurred. The second patient was
498 diagnosed with a side-branch intra papillary mucinous neoplasia, probably reflected by the
499 relative high fraction of acinar cells. Pancreatic cysts are often detected on abdominal
500 imaging, resulting in a diagnostic and treatment dilemma. Furthermore, pancreatic cysts
501 represent a broad group of lesions, ranging from benign to malignant entities. The main
502 challenge in their management is to accurately predict the malignant potential and to
503 determine the risk to benefit of a surgical resection[51]. Our results show that the cellular
504 contribution to the RNA content of pancreatic cyst fluids can be estimated through

505 deconvolution and that these results may be associated with clinical phenotypes. Larger
506 cohorts are necessary to investigate the clinical potential of this approach and pancreatic
507 tumor cells may also need to be added to the reference set with single cell RNA sequencing
508 data to improve the accuracy of the prediction.
509 In addition to linear mRNA transcripts, we also explored the circular RNA content in
510 biofluids. CircRNAs are a growing class of non-coding RNAs and a promising RNA biotype to
511 investigate in the liquid biopsy setting, as they are presumed to be less prone to degradation
512 compared to linear forms[52]. The circRNA fraction in tissues has previously been reported and
513 is in line with our findings[53]. In our study, we demonstrated that for genes that produce both
514 circRNAs and linear mRNAs, the circRNAs are more abundant than the linear forms in
515 biofluids. Further assessment of the biomarker potential of circRNAs in biofluids require
516 dedicated library preparation methods with circRNA enrichment.
517 In conclusion, The Human Biofluid RNA Atlas provides a systematic and comprehensive
518 comparison of the extracellular RNA content in 20 different human biofluids. The results
519 presented here may serve as a valuable resource for future biomarker studies.
520
521
522

523 **Material and methods**

524 ***Donor material, collection and biofluid preparation procedure***

525 Sample collection for the discovery cohort and sputum collection for the case/control cohort
526 was approved by the ethics committee of Ghent University Hospital, Ghent, Belgium (no.
527 B670201734450) and written informed consent was obtained from all donors according to
528 the Helsinki declaration. Breast milk, colostrum, plasma, serum, sputum, seminal plasma,
529 sweat, stool, tears and urine were obtained in healthy volunteers. All other biofluids were
530 collected from non-oncological patients.

531 The collection of two case series of each 12 cases and 12 control samples was approved by
532 the Masaryk Memorial Cancer Institute, Brno, Czech Republic (no. 14-08-27-01 and no.
533 MOU190814). Urine was collected in healthy donors and muscle-invasive bladder cancer
534 patients; CSF was collected in hydrocephalus patients and glioblastoma patients.

535 Collection of saliva samples in 12 healthy donors and in patients with diabetes mellitus for
536 the case/control cohort was approved by the ethics committee of the Medical University of
537 Vienna, Vienna, Austria (no. 2197/2015). Written informed consent was obtained from all
538 donors. The demographic and clinical patient information is provided in Supplementary
539 Table 1. Detailed information on the sample collection per biofluid is provided in
540 Supplementary Note 2. All samples, except tear fluid, plasma and serum, were centrifuged at
541 2000 g (rcf) for 10 minutes without brake at room temperature. All samples were processed
542 within 2 hours after collection. The cell-free supernatant was carefully pipetted into 2 mL
543 LoBind tubes (Eppendorf LoBind microcentrifuge tubes, Z666556-250EA) and stored at -80
544 °C.

545 ***RNA isolation and gDNA removal***

546 *RNA isolation from all biofluids, except tears*

547 In the discovery cohort, two RNA isolations per biofluid and per sample were simultaneously
548 performed by two researchers (E.V.E. and E.H.). In the end, RNA obtained from both RNA
549 isolations was pooled per biofluid and per sample and this pooled RNA was used as starting
550 material for both library preparations. Hence, small RNA and mRNA capture sequencing on
551 the discovery cohort were performed on the same batch of RNA. In the case/control cohorts,
552 one RNA isolation was performed per sample and the RNA was used as starting material for
553 mRNA capture sequencing.

554 RNA was isolated with the miRNeasy Serum/Plasma Kit (Qiagen, Hilden, Germany, 217184)
555 according to the manufacturer's instructions. An input volume of 200 µL was used for all
556 samples, except for tear fluid, and total RNA was eluted in 12 µL of RNAse-free water. Tear
557 fluid was collected with Schirmer strips and RNA was isolated directly from the strips (see
558 further). Per 200 µL biofluid input volume, 2 µL Sequin spike-in controls (Garvan Institute of
559 Medical Research) and 2 µl RNA extraction Control (RC) spike-ins (Integrated DNA
560 Technologies)[54] were added to the lysate for TruSeq RNA Exome Library Prep sequencing
561 and TruSeq Small RNA Library Prep sequencing, respectively. Details on the spike-in controls
562 are available in the Supplementary Note 1.
563 Briefly, 2 µl External RNA Control Consortium (ERCC) spike-in controls (ThermoFisher
564 Scientific, Waltham, MA, USA, 4456740), 2 µl Library Prep Control (LP) spike-ins (Integrated
565 DNA Technologies)[55], 1 µl HL-dsDNase and 1.6 µl reaction buffer were added to 12 µl RNA
566 eluate, and incubated for 10 min at 37 °C, followed by 5 min at 55 °C. Per biofluid and per
567 donor the RNA after gDNA removal was pooled. RNA was stored at -80 °C and only thawed
568 on ice immediately before the start of the library prep. Multiple freeze/thaw cycles did not
569 occur.
570 *RNA isolation from tear fluid*
571 Tear fluid was collected in 8 healthy donors with Schirmer strips (2 strips per eye per donor),
572 as previously described[56,57]. RNA was isolated within two hours after tear collection with the
573 miRNeasy Serum/Plasma Kit (Qiagen, Hilden, Germany, 217184), starting from one 2 mL
574 tube containing each 4 Schirmer strips. The same reagent volumes as suggested by the
575 manufacturer for a 200 µL input volume were used. Throughout the RNA isolation protocol,
576 the two final RNA samples each result from 4 tear fluid samples (each containing the 4 strips
577 of a single donor) that were pooled in a two-step method. First, the upper aqueous phase of
578 two tear fluid samples was put together (in step 8 of the RNA isolation protocol). Second,
579 the RNA eluate of these two samples was pooled into the final RNA that was used as input
580 for the library prep (in step 15 of the RNA isolation protocol).
581 ***TruSeq RNA Exome library prep sequencing***
582 Messenger RNA capture based libraries were prepared starting from 8.5 µL DNase treated
583 and spike-in supplemented RNA eluate using the TruSeq RNA Exome Library Prep Kit
584 (Illumina, San Diego, CA, USA). Each sample underwent individual enrichment according to
585 the manufacturer's protocol. The quality and yield of the prepared libraries were assessed

586 using a high sensitivity Small DNA Fragment Analysis Kit (Agilent Technologies, Santa Clara,
587 CA, USA) according to manufacturer's instructions. The libraries were quantified using qPCR
588 with the KAPA Library Quantification Kit (Roche Diagnostics, Diegem, Belgium, KK4854)
589 according to manufacturer's instructions. Based on the qPCR results, equimolar library pools
590 were prepared.
591 Paired-end sequencing was performed on a NextSeq 500 instrument using a high output v2
592 kit (Illumina, San Diego, CA, USA) with a read length of 75 nucleotides to an average
593 sequencing depth of 11 million read pairs in the discovery cohort, 16.8 million read pairs in
594 the sputum case/control cohorts, 15.4 million read pairs in the urine case/control cohort, 15
595 million read pairs in the CSF case/control cohort and 18.8 million read pairs in the saliva
596 case/control cohort. Samples from the discovery cohort were randomly assigned over two
597 pools and sequenced with a loading concentration of 1.2 pM (5% PhiX) and 1.6 pM (5%
598 PhiX), respectively. Urine, CSF and saliva samples from the case/control cohorts were loaded
599 in 3 separate runs at 2 pM (2% PhiX) and sputum samples from the case/control cohorts
600 were loaded at 1.6 pM (5% PhiX).
601 ***TruSeq Small RNA library prep sequencing***
602 Small RNA libraries were prepared starting from 5 µL DNase treated and spike-in
603 supplemented RNA eluate using a TruSeq Small RNA Library Prep Kit (Illumina, San Diego,
604 CA, USA) according to the manufacturer's protocol with two minor modifications(1). The
605 RNA 3' adapter (RA3) and the RNA 5' adapter (RA5) were 4-fold diluted with RNase-free
606 water(2) and the number of PCR cycles was increased to 16.
607 First, a volume-based pool of all 46 samples of the discovery cohort was sequenced. After
608 PCR amplification, quality of libraries was assessed using a high sensitivity Small DNA
609 Fragment Analysis Kit (Agilent Technologies, Santa Clara, CA, USA) according to
610 manufacturer's instructions. Size selection of the pooled samples was performed using 3%
611 agarose dye-free marker H cassettes on a Pippin Prep (Sage Science, Beverly, MA, USA)
612 following manufacturer's instructions with a specified collection size range of 125–163 bp.
613 Libraries were further purified and concentrated by ethanol precipitation, resuspended in
614 10 µl of 10 mM tris-HCl (pH = 8.5) and quantified using qPCR with the KAPA Library
615 Quantification Kit (Roche Diagnostics, Diegem, Belgium, KK4854) according to
616 manufacturer's instructions. The pooled library was quality controlled via sequencing at a
617 concentration of 1.7 pM with 35% PhiX on a NextSeq 500 using a mid-output v2 kit (single-

618 end 75 nucleotides, Illumina, San Diego, CA, USA), resulting in an average sequencing depth
619 of 1 million reads, ranging from 3341 reads to 14 million reads. Twenty-three samples with
620 less than 200 000 reads were assigned to a low concentrated pool, 23 samples with more
621 than 17 million reads were assigned to a highly concentrated pool. Based on the read
622 numbers from the mid output run, two new equimolar pools were prepared, purified and
623 quantified as described higher. Both re-pooled libraries were then sequenced at a final
624 concentration of 1.7 pM with 25% PhiX on a NextSeq 500 using a high output v2 kit (single-
625 end, 75 nucleotides, Illumina, San Diego, CA, USA), resulting in an average sequencing depth
626 of 9 million reads (range 817 469 – 41.7 million reads).

627 ***RT-qPCR***

628 To validate findings observed in the RNA sequencing data, we performed a targeted mRNA
629 and circRNA expression profiling with RT-qPCR for 8 differentially expressed mRNAs in
630 sputum (COPD versus healthy control) and for the 10 most abundant circRNAs in sputum. As
631 reference RNAs for normalization purposes, we selected Sequin spikes stably detected in all
632 samples based on the available RNA sequencing data. The assays to measure mRNA, circRNA
633 and Sequin spike expression were custom designed using primerXL[58] (Supplementary Data 9)
634 and purchased from Integrated DNA Technologies, Inc. (Coralville, USA).

635 For cDNA synthesis, 5 µl of total RNA was reverse transcribed using the iScript Advanced
636 cDNA Synthesis Kit (BioRad, USA) in a 10 µL volume. 5 µL of cDNA was pre-amplified in a 12-
637 cycle PCR reaction using the Sso Advanced PreAmp Supermix (Bio-Rad, USA) in a 50 µL
638 reaction. Pre-amplified cDNA was diluted (1:8) and 2 µL was used as input for a 45-cycle
639 qPCR reaction, quantifying 8 mRNAs and 10 circRNAs of interest with the SsoAdvanced™
640 Universal SYBR Green Supermix (BioRad, USA). All reactions were performed in 384-well
641 plates on the LightCycler480 instrument (Roche) in a 5 µL reaction volume using 250 nM
642 primer concentrations. Cq-values were determined with the LightCycler®480 Software
643 (release 1.5.0, Roche) with the "Abs Quant/2nd Derivative Max" method.

644 The geNorm analysis to select the optimal number of reference targets was performed using
645 Biogazelle's qbase+ software (www.qbaseplus.com) using log2-transformed RNA count data.
646 We observed medium reference target stability (average geNorm M ≤ 1.0) with an optimal
647 number of reference targets in this experimental situation of two (geNorm V < 0.15 when
648 comparing a normalization factor based on the two or three most stable targets). As such,

649 the optimal normalization factor can be calculated as the geometric mean of reference
650 targets R2_150 and R2_65. These Sequin spike RNAs were considered as reference RNAs.

651 ***Data analysis***

652 *Processing TruSeq RNA Exome sequencing data*

653 Read quality was assessed by running FastQC (v0.11.5) on the FASTQ files and reads shorter
654 than 35 nucleotides and with a quality (phred) score < 30 were removed. The reads were
655 mapped with STAR (v2.6.0). Mapped reads were annotated by matching genomic
656 coordinates of each read with genomic locations of mRNAs (obtained from UCSC
657 GRCh38/hg38 and Ensembl, v91) or by matching the spike-in sequences. Picard (v2.18.5) was
658 used for duplicate removal. HTSeq (v0.9.1) was used for quantification of PCR deduplicated
659 reads. A cut-off for filtering noisy genes was set based on historic data to remove noisy
660 genes. Using a threshold of 4 counts, at least 95% of the single positive replicate values are
661 filtered out. A table with the read count of mRNAs per sample is provided in Supplementary
662 Data 6.

663 *Processing TruSeq Small RNA sequencing data*

664 Adaptor trimming was performed using Cutadapt (v1.8.1) with a maximum error rate of
665 0.15. Reads shorter than 15 nts and those in which no adaptor was found were discarded.
666 For quality control the FASTX-Toolkit (v0.0.14) was used, a minimum quality score of 20 in at
667 least 80% of nucleotides was applied as a cutoff. The reads were mapped with Bowtie
668 (v1.1.2) without allowing mismatches. Mapped reads were annotated by matching genomic
669 coordinates of each read with genomic locations of miRNAs (obtained from miRBase, v22)
670 and other small RNAs (obtained from UCSC GRCh38/hg38 and Ensembl, v91) or by matching
671 the spike-in sequences. Reads assigned as "not annotated" represent uniquely mapped
672 reads that could not be classified in one of the small RNA biotype groups. As for the mRNA
673 data, genes with fewer than 4 counts were filtered out. A table with the read count of
674 miRNAs per sample is provided in Supplementary Data 7.

675 *Exogenous RNA characterization*

676 The exogenous RNA content in the mRNA data was assessed using the MetaMap pipeline[59].
677 Briefly, all reads were mapped to the human reference genome (hg38) using STAR (v2.5.2)[60].
678 Unmapped reads were subsequently subjected to metagenomic classification using CLARK-S
679 (v1.2.3)[61]. Reads were summed across all bacterial species.

680 The exogenous RNA content in the small RNA data was assessed using the exceRpt small
681 RNA-seq pipeline (v4.6.2) in the Genboree workbench with default settings[62]. Briefly, after
682 adapter trimming, read quality was assessed by FASTQC (v0.11.2). A minimum quality score
683 of 20 in at least 80% of nucleotides was applied as cutoff. The minimum read length after
684 adapter trimming was set to 18 nucleotides. Reads were first mapped to the custom spike-in
685 sequences using Bowtie2 (v2.2.6), followed by mapping the unmapped reads with STAR
686 (v2.4.2a) to UniVec contaminants and human ribosomal (rRNA) sequences to exclude them
687 before mapping (also with STAR) to the following databases: miRbase (v21), gtRNAdb,
688 piRNABank, GeneCode version 24 (hg38) and circBase (version last updated in July 2017). A
689 single mismatch was allowed during mapping to the human genome. Unmapped reads were
690 then mapped with STAR to exogenous miRNAs and rRNAs. In the end, the remaining
691 unmapped reads were mapped to the genomes of all sequenced species in Ensembl and
692 NCBI. No mismatches were allowed during exogenous alignment. Raw read counts obtained
693 from the Genboree workbench were further analyzed in R (v3.5.1) making use of tidyverse
694 (v1.2.1).

695 *Circular RNA detection and circular/linear ratio determination*

696 Only TruSeq RNA Exome reads passing quality control (base calling accuracy of ≥ 99% in at
697 least 80% of the nucleotides in both mates of a pair) were included in this analysis. Clumpify
698 dedupe (v38.26) was used to remove duplicates in paired-end mode (2 allowed
699 substitutions, kmer size of 31 and 20 passes). We used a two-step mapping strategy to
700 identify forward splice (further referred to as linear) junction reads and backsplice junction
701 reads. First, reads were aligned with TopHat2 (v2.1.0) to the GRCh38/hg38 reference
702 genome (Ensembl, v91). Micro-exons were included, a minimum anchor length of 6
703 nucleotides was required, and up to two mismatches in the anchor region were allowed. The
704 resulting output contains linear junction information. Secondly, unmapped reads from the
705 first mapping strategy were realigned with TopHat2 (v2.1.0) to the same reference, but this
706 time with the fusion search option that can align reads to potential fusion transcripts.
707 Processing the fusion search output with CIRCexplorer2 parse (v2.3.3) results in backsplice
708 junction information. Junction read counts obtained with the mapping strategies described
709 above were used as a measure for the relative level of linear and circular RNA in each
710 sample. Only genes with at least one detected backsplice junction were considered.
711 Junctions that could be part of both linear and circular transcripts (ambiguous junctions)

712 were filtered out. As there is currently no consensus on how to calculate the circular to
713 linear ratio (CIRC/LIN), we decided to calculate the ratio in two different ways
714 (Supplementary Fig. 8). The circRNA fraction is defined as 100*CIRC/(CIRC+LIN). The first
715 method (referred to as "backsplice junction-level method") zooms in on each particular
716 backsplice junction. CIRC was defined as the backsplice junction read count of one particular
717 backsplice junction. LIN was defined as the average read count of all junctions flanking the
718 backsplice junction of interest. The second method (referred to as "gene-level method")
719 considers all backsplice junctions within a given gene. CIRC was defined as the average
720 number of backsplice junction reads for a given gene. LIN was defined as the average
721 number of linear junction reads for a given gene. For both methods, CIRC > 3 was used as a
722 cut-off for filtering noisy backsplice junctions. To enable a comparison of the circular/linear
723 genic ratios in biofluids with those of tissues, the mRNA capture sequencing FASTQ files of
724 16 cancerous tissue types (34 samples in total) were downloaded from the MiOncoCirc
725 database (dbGaP Study Accession phs000673.v3.p1)[24]. A list with the downloaded samples is
726 attached in Supplementary Table 2. A table with the read count of backsplice junctions per
727 sample is provided in Supplementary Data 8.

728 *Assessment of tissue and cell contribution to biofluid exRNA*

729 Using total RNA-sequencing data from 27 normal human tissue types and 5 immune cell
730 types from peripheral blood from the RNA Atlas[35], we created gene sets containing marker
731 genes for each individual entity (Supplementary Data 4). We removed redundant tissues and
732 cell types from the original RNA Atlas (e.g. granulocytes and monocytes were present twice;
733 brain was kept and specific brain sub-regions such as cerebellum, frontal cortex, occipital
734 cortex and parietal cortex were removed) and we used genes where at least one tissue or
735 cell type had expression values greater or equal to 1 TPM normalized counts. A gene was
736 considered to be a marker if its abundance was at least 5 times higher in the most abundant
737 sample compared to the others. For the final analysis, only tissues and cell types with at
738 least 3 markers were included, resulting in 26 tissues and 5 immune cell types.

739 Gene abundance read counts from the biofluids were normalized using Sequin spikes as size
740 factors in DESeq2 (v1.22.2). For all marker genes within each gene set, we computed the
741 log2 fold changes between the median read count of a biofluid sample pair versus the
742 median read count of all other biofluids. The median log2 fold change of all markers in a
743 gene set was selected, followed by z-score transformation over all biofluids (Fig. 7). For

744 visualization purposes, only tissues and cell types with a z-score ≥ |1| in at least one biofluid
745 were used.

746 *Cellular deconvolution of pancreatic cyst fluid samples*

747 To build the reference matrix for the computational deconvolution of pancreatic cyst fluid
748 samples, single cell RNA sequencing data of 10 pancreatic cell types[34] was processed with
749 the statistical programming language R (v3.6.0). For each gene, the mean count across all
750 individual cells from each cell type was computed. Next, this reference matrix was
751 normalized using the trimmed means of M values (TMM) with the edgeR package
752 (v3.26.4)[63,64]. Limma-voom (v3.40.2)[65] was used for subsequent differential gene expression
753 analysis and those genes with an absolute fold change greater or equal to 2 and an adjusted
754 p-value < 0.05 (Benjamini-Hochberg) were retained as markers[66]. Finally, using these
755 markers and both the pancreatic cyst fluid samples and the reference matrix described
756 above, the cell type proportions were obtained through computational deconvolution using
757 non-negative least squares (nnls package; v1.4)[67,68].

758 *Differential expression analysis in case/control cohorts*

759 Further processing of the count tables was done with R (v3.5.1) making use of tidyverse
760 (v1.2.1). Gene expression read counts from the biofluids were normalized using Sequin
761 spikes as size factors in DESeq2 (v1.20.0)[69]. To assess the biological signal in the case/control
762 cohorts, we performed differential expression analysis between the patients and control
763 groups using DESeq2 (v1.20.0). Genes were considered differentially expressed when the
764 absolute log2 fold change > 1 and at q < 0.05.

765

766 ***Data availability***

767 The raw RNA-sequencing data have been deposited at the European Genome-phenome
768 Archive (EGA) under accession number EGAS00001003917. All spike-normalized sequencing
769 data can be readily explored in the interactive web-based application R2: Genomics analysis
770 and visualization platform (http://r2.amc.nl), and via a dedicated accessible portal
771 (http://r2platform.com/HumanBiofluidRNAAtlas). This portal allows the analysis and
772 visualization of mRNA, circRNA and miRNA abundance, as illustrated in Supplementary Fig.
773 11. All samples can be used for correlation, principle component, and gene set enrichment
774 analyses, and many more. All other data are available within the article and supplementary
775 information.

776

777 ***Code availability***

778 The R scripts to reproduce the analyses and plots reported in this paper are available from
779 the corresponding authors upon request.

780

781 **Acknowledgments**

782 E.H. is a recipient of a grant of the Fund for Scientific Research Flanders (FWO). F.A.C. and
783 A.M. are supported by Special Research Fund (BOF) scholarship of Ghent University
784 (BOF.DOC.2017.0026.01, BOF.DOC.2019.0047.01). A.G. is senior clinical researcher of the
785 Fund for Scientific Research Flanders (FWO; 1805718N). This research is partly funded by
786 "RNA-MAGIC", "LNCCA" Concerted Research Actions of Ghent University (BOF19/GOA/008,
787 BOF16/GOA/023), "Kom Op Tegen Kanker" and by the Czech Ministry of Health grants
788 (NV18-03-00398, NV18-03-00360). We thank Tim Mercer for providing the Sequin spikes.
789 The results published here are in part based upon data generated by the Clinical Sequencing
790 Exploratory Research (CSER) consortium established by the NHGRI. Funding support was
791 provided through cooperative agreements with the NHGRI and NCI through grant numbers
792 U01 HG006508 (Exploring Cancer Medicine for Sarcoma and Rare Cancers). Information
793 about CSER and the investigators and institutions who comprise the CSER consortium can be
794 found at http://www.genome.gov/27546194.

795

796 **Contributions**

797 J.V. and P.M. conceived and supervised the project; E.H., K.V., N.Y., E.V.E., J.N. designed and
798 performed the experiments; E.H., A.M., J.A. and F.A.C. analyzed the data; L.S. and S.K.
799 performed analysis using the MetaMap pipeline; G.S. and S.K. contributed technical support
800 and resources; E.H., A.G., P.H., P.J., G.B., K.B., T.M., T.D., V.N., C.V.C., K.R., E.R., D.H., K.T.,
801 O.S., C.N. collected samples; S.L. designed RT-qPCR primers; E.H., P.M and J.V. wrote the
802 paper; J.K. developed dedicated tools to analyze RNA atlas data and results and
803 implemented them in the online portal R2. All authors contributed to manuscript editing
804 and approved the final draft.

805

806

807 **References**

808 1. Weiland, M., Gao, X.-H., Zhou, L. & Mi, Q.-S. Small RNAs have a large impact: circulating
809 microRNAs as biomarkers for human diseases. *RNA Biol.* **9**, 850–859 (2012).
810 2. Max, K. E. A. *et al.* Human plasma and serum extracellular small RNA reference profiles and their
811 clinical utility. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E5334–E5343 (2018).
812 3. Freedman, J. E. *et al.* Diverse human extracellular RNAs are widely detected in human plasma.
813 *Nat. Commun.* **7**, 11106 (2016).
814 4. Yuan, T. *et al.* Plasma extracellular RNA profiles in healthy and cancer patients. *Sci. Rep.* **6**, 19413
815 (2016).
816 5. Godoy, P. M. *et al.* Large Differences in Small RNA Composition Between Human Biofluids. *Cell*
817 *Rep.* **25**, 1346–1358 (2018).
818 6. Yeri, A. *et al.* Total Extracellular Small RNA Profiles from Plasma, Saliva, and Urine of Healthy
819 Subjects. *Sci. Rep.* **7**, 44061 (2017).
820 7. Ferrero, G. *et al.* Small non-coding RNA profiling in human biofluids and surrogate tissues from
821 healthy individuals: description of the diverse and most represented species. *Oncotarget* **9**,
822 (2018).
823 8. Murillo, O. D. *et al.* exRNA Atlas Analysis Reveals Distinct Extracellular RNA Cargo Types and
824 Their Carriers Present across Human Biofluids. *Cell* **177**, 463-477.e15 (2019).
825 9. Srinivasan, S. *et al.* Small RNA Sequencing across Diverse Biofluids Identifies Optimal Methods for
826 exRNA Isolation. *Cell* **177**, 446-462.e16 (2019).
827 10. Fehlmann, T., Ludwig, N., Backes, C., Meese, E. & Keller, A. Distribution of microRNA biomarker
828 candidates in solid tissues and body fluids. *RNA Biol.* **13**, 1084–1088 (2016).
829 11. Umu, S. U. *et al.* A comprehensive profile of circulating RNAs in human serum. *RNA Biol.* **15**, 242–
830 250 (2018).
831 12. El-Mogy, M. *et al.* Diversity and signature of small RNA in different bodily fluids using next
832 generation sequencing. *BMC Genomics* **19**, 408 (2018).
833 13. Weber, J. A. *et al.* The microRNA spectrum in 12 body fluids. *Clin. Chem.* **56**, 1733–1741 (2010).
834 14. Welch, R. A., Shaw, M. K. & Welch, K. C. Amniotic fluid LPCAT1 mRNA correlates with the
835 lamellar body count. *J. Perinat. Med.* **44**, 531–532 (2016).
836 15. Marzioni, M. *et al.* PDX-1 mRNA expression in endoscopic ultrasound-guided fine needle
837 cytoaspirate: perspectives in the diagnosis of pancreatic cancer. *Dig. Liver Dis. Off. J. Ital. Soc.*
838 *Gastroenterol. Ital. Assoc. Study Liver* **47**, 138–143 (2015).
839 16. Tian, Y., Li, L., Zhang, F. & Xu, J. Seminal plasma HSPA2 mRNA content is associated with semen
840 quality. *J. Assist. Reprod. Genet.* **33**, 1079–1084 (2016).
841 17. Oreo, K. M. *et al.* Sputum ADAM8 expression is increased in severe asthma and COPD. *Clin. Exp.*
842 *Allergy J. Br. Soc. Allergy Clin. Immunol.* **44**, 342–352 (2014).
843 18. Maker, A. V. *et al.* Cyst Fluid Biosignature to Predict Intraductal Papillary Mucinous Neoplasms of
844 the Pancreas with High Malignant Potential. *J. Am. Coll. Surg.* **228**, 721–729 (2019).
845 19. Herring, E., Kanaoka, S., Tremblay, E. & Beaulieu, J.-F. A Stool Multitarget mRNA Assay for the
846 Detection of Colorectal Neoplasms. *Methods Mol. Biol. Clifton NJ* **1765**, 217–227 (2018).
847 20. Bazzell, B. G. *et al.* Human Urinary mRNA as a Biomarker of Cardiovascular Disease. *Circ.*
848 *Genomic Precis. Med.* **11**, e002213 (2018).
849 21. Bahn, J. H. *et al.* The landscape of microRNA, Piwi-interacting RNA, and circular RNA in human
850 saliva. *Clin. Chem.* **61**, 221–230 (2015).
851 22. Liu, B. *et al.* Characterization of tissue-specific biomarkers with the expression of circRNAs in
852 forensically relevant body fluids. *Int. J. Legal Med.* **133**, 1321–1331 (2019).
853 23. Li, S. *et al.* exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes.
854 *Nucleic Acids Res.* **46**, D106–D112 (2018).
855 24. Vo, J. N. *et al.* The Landscape of Circular RNA in Cancer. *Cell* **176**, 869-881.e13 (2019).

856 25. Kölling, M. *et al.* Circular RNAs in Urine of Kidney Transplant Patients with Acute T Cell-Mediated
857 Allograft Rejection. *Clin. Chem.* (2019) doi:10.1373/clinchem.2019.305854.
858 26. Giraldez, M. D. *et al.* Phospho-RNA-seq: a modified small RNA-seq method that reveals
859 circulating mRNA and lncRNA fragments as potential biomarkers in human plasma. *EMBO J.* **38**,
860 (2019).
861 27. Zhou, Z. *et al.* Extracellular RNA in a single droplet of human serum reflects physiologic and
862 disease states. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 19200–19208 (2019).
863 28. Everaert, C. *et al.* Performance assessment of total RNA sequencing of human biofluids and
864 extracellular vesicles. *Sci. Rep.* **9**, (2019).
865 29. Metzenmacher *et al.* Plasma Next Generation Sequencing and Droplet Digital-qPCR-Based
866 Quantification of Circulating Cell-Free RNA for Noninvasive Early Detection of Cancer. *Cancers*
867 **12**, 353 (2020).
868 30. Li, X., Yang, L. & Chen, L.-L. The Biogenesis, Functions, and Challenges of Circular RNAs. *Mol. Cell*
869 **71**, 428–442 (2018).
870 31. Memczak, S., Papavasileiou, P., Peters, O. & Rajewsky, N. Identification and Characterization of
871 Circular RNAs As a New Class of Putative Biomarkers in Human Blood. *PloS One* **10**, e0141214
872 (2015).
873 32. Heintz-Buschart, A. *et al.* Small RNA profiling of low biomass samples: identification and removal
874 of contaminants. *BMC Biol.* **16**, 52 (2018).
875 33. Liu, F., Ma, R., Wang, Y. & Zhang, L. The Clinical Importance of Campylobacter concisus and
876 Other Human Hosted Campylobacter Species. *Front. Cell. Infect. Microbiol.* **8**, 243 (2018).
877 34. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals
878 Inter- and Intra-cell Population Structure. *Cell Syst.* **3**, 346-360.e4 (2016).
879 35. Lorenzi, L. *et al.* The RNA Atlas, a single nucleotide resolution map of the human transcriptome.
880 *bioRxiv* (2019) doi:10.1101/807529.
881 36. Demedts, I. K. *et al.* Accumulation of dendritic cells and increased CCL20 levels in the airways of
882 patients with chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **175**, 998–
883 1005 (2007).
884 37. Bracke, K. R. *et al.* Cigarette smoke-induced pulmonary inflammation and emphysema are
885 attenuated in CCR6-deficient mice. *J. Immunol. Baltim. Md 1950* **177**, 4350–4359 (2006).
886 38. Karmouty-Quintana, H. *et al.* Adenosine A2B receptor and hyaluronan modulate pulmonary
887 hypertension associated with chronic obstructive pulmonary disease. *Am. J. Respir. Cell Mol.*
888 *Biol.* **49**, 1038–1047 (2013).
889 39. Singh Patidar, B. *et al.* Adenosine Metabolism in COPD: A Study on Adenosine Levels, 5'-
890 Nucleotidase, Adenosine Deaminase and Its Isoenzymes Activity in Serum, Lymphocytes and
891 Erythrocytes. *COPD* **15**, 559–571 (2018).
892 40. Stankovic, M., Nikolic, A., Nagorni-Obradovic, L., Petrovic-Stanojevic, N. & Radojkovic, D. Gene-
893 Gene Interactions Between Glutathione S-Transferase M1 and Matrix Metalloproteinases 1, 9,
894 and 12 in Chronic Obstructive Pulmonary Disease in Serbians. *COPD* **14**, 581–589 (2017).
895 41. Saugstad, J. A. *et al.* Analysis of extracellular RNA in cerebrospinal fluid. *J. Extracell. Vesicles* **6**,
896 1317577 (2017).
897 42. Holyoake, A. *et al.* Development of a multiplex RNA urine test for the detection and stratification
898 of transitional cell carcinoma of the bladder. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **14**,
899 742–749 (2008).
900 43. Eckstein, M. *et al.* mRNA-Expression of KRT5 and KRT20 Defines Distinct Prognostic Subgroups of
901 Muscle-Invasive Urothelial Bladder Cancer Correlating with Histological Variants. *Int. J. Mol. Sci.*
902 **19**, (2018).
903 44. Murakami, T. *et al.* Bladder cancer detection by urinary extracellular vesicle mRNA analysis.
904 *Oncotarget* **9**, 32810–32821 (2018).
905 45. Lin, H., Zhou, Q., Wu, W. & Ma, Y. Midkine Is a Potential Urinary Biomarker for Non-Invasive
906 Detection of Bladder Cancer with Microscopic Hematuria. *OncoTargets Ther.* **Volume 12**, 11765–
907 11775 (2020).

908 46. Chen, T. *et al.* CD163, a novel therapeutic target, regulates the proliferation and stemness of
909 glioma cells via casein kinase 2. *Oncogene* **38**, 1183–1199 (2019).
910 47. Zhang, L. *et al.* Salivary transcriptomic biomarkers for detection of resectable pancreatic cancer.
911 *Gastroenterology* **138**, 949-957.e1–7 (2010).
912 48. Waller, R. *et al.* Small RNA Sequencing of Sporadic Amyotrophic Lateral Sclerosis Cerebrospinal
913 Fluid Reveals Differentially Expressed miRNAs Related to Neural and Glial Activity. *Front.*
914 *Neurosci.* **11**, (2018).
915 49. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based
916 microbiome analyses. *BMC Biol.* **12**, (2014).
917 50. Strong, M. J. *et al.* Microbial contamination in next generation sequencing: implications for
918 sequence-based analysis of clinical samples. *PLoS Pathog.* **10**, e1004437 (2014).
919 51. Farrell, J. J. Pancreatic Cysts and Guidelines. *Dig. Dis. Sci.* **62**, 1827–1839 (2017).
920 52. Jeck, W. R. & Sharpless, N. E. Detecting and characterizing circular RNAs. *Nat. Biotechnol.* **32**,
921 453–461 (2014).
922 53. Guo, J. U., Agarwal, V., Guo, H. & Bartel, D. P. Expanded identification and characterization of
923 mammalian circular RNAs. 14 (2014).
924 54. Locati, M. D. *et al.* Improving small RNA-seq by using a synthetic spike-in set for size-range
925 quality control together with a set for data normalization. *Nucleic Acids Res.* **43**, e89 (2015).
926 55. Hafner, M. *et al.* RNA-ligase-dependent biases in miRNA representation in deep-sequenced small
927 RNA cDNA libraries. *RNA N. Y. N* **17**, 1697–1712 (2011).
928 56. Green-Church, K. B., Nichols, K. K., Kleinholz, N. M., Zhang, L. & Nichols, J. J. Investigation of the
929 human tear film proteome using multiple proteomic approaches. *Mol. Vis.* **14**, 456–470 (2008).
930 57. Pieragostino, D. *et al.* Tear Film Steroid Profiling in Dry Eye Disease by Liquid Chromatography
931 Tandem Mass Spectrometry. *Int. J. Mol. Sci.* **18**, 1349 (2017).
932 58. Lefever, S. *et al.* High-throughput PCR assay design for targeted resequencing using primerXL.
933 *BMC Bioinformatics* **18**, 400 (2017).
934 59. Simon, L. M. *et al.* MetaMap: an atlas of metatranscriptomic reads in human disease-related
935 RNA-seq data. *GigaScience* **7**, (2018).
936 60. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21
937 (2013).
938 61. Ounit, R. & Lonardi, S. Higher classification sensitivity of short metagenomic reads with CLARK-S.
939 *Bioinforma. Oxf. Engl.* **32**, 3823–3825 (2016).
940 62. Rozowsky, J. *et al.* exceRpt: A Comprehensive Analytic Platform for Extracellular RNA Profiling.
941 *Cell Syst.* **8**, 352-357.e3 (2019).
942 63. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis
943 of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
944 64. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential
945 expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* **26**, 139–140 (2010).
946 65. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and
947 microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
948 66. Benjamini, Yoav, H., Yosef. Controlling the False Discovery Rate: A Practical and Powerful
949 Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*
950 **57**, 289–300 (1995).
951 67. Mullen, K. M. & Van Stokkum, I. H. M. nnls: The Lawson-Hanson algorithm for non-negative lease
952 squares (NNLS). R package version 1.4. https://CRAN.R-project.org/package=nnls. (2012).
953 68. Avila Cobos, F., Vandesompele, J., Mestdagh, P. & De Preter, K. Computational deconvolution of
954 transcriptomics data from mixed cell populations. *Bioinforma. Oxf. Engl.* **34**, 1969–1979 (2018).
955 69. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for
956 RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
957