Custom long non-coding RNA capture enhances detection sensitivity in different human sample types — **Source link** ⬈

Annelien Morlion, Celine Everaert, Justine Nuytens, Eva Hulstaert ...+3 more authors

Institutions: Ghent University, Ghent University Hospital

Related papers:

- High-Throughput Methods to Detect Long Non-Coding RNAs

- Bioinformatic Pipelines to Analyze lncRNAs RNAseq Data.

- Bioinformatics Pipeline to Analyze lncRNA Arrays.

- Preparation of highly multiplexed small RNA sequencing libraries

- Rapid microRNA Profiling on Encoded Gel Microparticles

Custom long non-coding RNA capture enhances detection sensitivity in different human sample types

Annelien Morlion[1,2]*, Celine Everaert[1,2]*, Justine Nuytens[1,2], Eva Hulstaert[1,2,3], Jo Vandesompele[1,2], Pieter Mestdagh[1,2]°

[1] OncoRNALab, Center for Medical Genetics, Department of Biomolecular Medicine, Ghent University, Ghent, Belgium

[2] Cancer Research Institute Ghent (CRIG), Ghent, Belgium

[3] Department of Dermatology, Ghent University Hospital, Ghent, Belgium

*joint first authors

°corresponding author: pieter.mestdagh@ugent.be

1 Abstract

2 Long non-coding RNAs (lncRNAs) are a heterogeneous group of transcripts that lack protein

3 coding potential and display regulatory functions in various cellular processes. As a result of

4 their cell- and cancer-specific expression patterns, lncRNAs have emerged as potential

5 diagnostic and therapeutic targets. The accurate characterization of lncRNAs in bulk

6 transcriptome data remains challenging due to their low abundance compared to protein

7 coding genes. To tackle this issue, we describe a unique short-read custom lncRNA capture

8 sequencing approach that relies on a comprehensive set of 565,878 capture probes for 49,372

9 human lncRNA genes. This custom lncRNA capture approach was evaluated on various sample

10 types ranging from artificial high-quality RNA mixtures to more challenging formalin-

11 fixed paraffin-embedded tissue and biofluid material. The custom enrichment approach

12 allows the detection of a more diverse repertoire of lncRNAs, with better reproducibility and

13 higher coverage compared to classic total RNA-sequencing.

14

15 Keywords

16 lncRNA, RNA sequencing, probes, lncRNome, RNA abundance, RNA expression, FFPE, biofluid

17

18 Introduction

19 While the majority of the human genome is actively transcribed into RNA transcripts, most of

20 these transcripts do not code for proteins (Djebali *et al.*, 2012). The non-coding RNA

21 transcripts longer than 200 nucleotides belong to the heterogeneous group of long non-

22 coding RNAs (lncRNAs), half of which are not poly-adenylated (Lorenzi *et al.*, 2019). These

23 lncRNAs are known to influence gene expression at both the transcriptional and post-

24 transcriptional level through a variety of mechanisms (Mercer *et al.*, 2009; Robinson *et al.*,

25 2020). Moreover, lncRNAs often show a particular cell- or cancer-type specific expression

26 pattern (Iyer *et al.*, 2015), which adds to their biomarker potential.

27 In the past, several high-throughput methods have been developed to profile the long non-

28 coding RNA transcriptome, study their structure or define their function (Cao *et al.*, 2019;

29 Turner *et al.*, 2019). Because of their generally low abundance compared to protein coding

30 genes, quantification of lncRNAs in bulk transcriptome data remains challenging. Enrichment

31 strategies favoring lncRNAs over the more abundant mRNAs could therefore result in more

32   lncRNAs being detected with a better transcript coverage, improving downstream analysis. A

33   promising method is RNA capture sequencing, a short-read sequencing method that can

34   enrich RNA targets of interest using oligonucleotide probes that are specifically designed to

35   tile the target sequences. These RNA capture sequencing technologies have mainly been

36   applied for deep sequencing of a selection of lncRNAs (Mercer *et al.*, 2014; Clark *et al.*, 2015).

37   Recently, the GENCODE consortium extended this method by applying long-read sequencing

38   after capturing about 14,470 lncRNAs genes to improve their structural annotation (RNA

39   Capture Long Seq, RNA CLS) (Lagarde *et al.*, 2017).

40   In this study, we describe a custom lncRNA capture sequencing approach that targets a very

41   comprehensive human lncRNome. This custom capture approach was evaluated on various

42   sample types ranging from high-quality RNA mixtures to more challenging formalin-

43   fixed paraffin-embedded (FFPE) tissue and biofluid material.

44

45   Material and methods

46   *Probe design*

47   Probes were designed against the highly confident set of LNCipedia 5.2 (hg19 genome build).

48   First, extended exons were created by concatenating each set of overlapping exons. For each

49   of these extended exons, probes of 120 nucleotides were tiled, resulting in (number of

50   nucleotides)-119 probes per concatenated exon. These exon tiling probes were mapped

51   against repeat regions and protein coding genes to filter out these that would capture off-

52   target fragments.

53   The resulting probe pool was extended with probes designed to capture both the Sequin and

54   ERCC spikes. These probes are 120-mers designed by tiling the spike-in sequences and,

55   inherent to the spike-design, these do not align to the human genome.

56   Further filtering was done by retaining the 120-mers with a GC content between 25-70%, a

57   GC-based Tm between 60-80 °C and a ΔG larger than -7 (calculated by UNAFold (version 3.8)

58   settings: hybrid-ss-min -E -n DNA -t 54 -T 54). The remaining probes underwent a selection

59   aimed at obtaining the minimal number of probes for an optimal coverage. In total, 565,878

60   probes against LNCipedia, 81,089 probes against novel genes (not discussed in this paper) and

61   2427 spike-in RNA probes were retained. Probes were synthesized by Twist Biosciences.

62

63   *Sample collection and RNA purification*

64   Sample collection was approved by the ethics committee of Ghent University Hospital, Ghent,

65   Belgium (#B670201734450 and #B670201733701) and written informed consent was

66   obtained from all donors. FFPE tissues were obtained from two colon cancer patients; the

67   biofluid samples (seminal and blood plasma) were collected from healthy donors.

68   *Platelet depleted blood plasma*

69   Venous blood from two healthy donors was drawn from an elbow vein after disinfection with

70   2% chlorhexidine in 70% alcohol. All blood draws were performed with a butterfly needle of

71   21 gauge (BD Vacutainer, Push Button Blood Collection Set, #367326, Becton Dickinson and

72   Company, NJ, USA) and blood was collected in 10 ml BD Vacutainer K2-EDTA tubes (#367525,

73   Becton Dickinson and Company, NJ, USA). The tubes were inverted 5 times and centrifuged

74   immediately after blood draw (15 min at 2500 g, room temperature, without brake). Per

75   donor, the upper plasma fractions were pipetted (leaving approximately 0.5 cm plasma above

76   the buffy coat) and pooled in a 15 ml tube. After gently inverting, the pooled plasma fraction

77   was centrifuged again (15 min at 2500 g, room temperature, without brake) and the upper

78   fraction was transferred to a new 15 ml tube, leaving approximately 0.5 cm plasma above

79   the separation. The resulting platelet depleted plasma was gently inverted, snap-frozen in

80   five aliquots (Safe-Lock cup DNA LoBind 2 ml PCR clean tubes, Eppendorf, #0030108078) and

81   stored at −80 °C. Platelets were counted and the degree of hemolysis was determined by

82   measuring levels of free hemoglobin by spectral analysis using a NanoDrop 1000

83   Spectrophotometer (Thermo Fisher Scientific). The entire plasma preparation protocol was

84   finished in two and a half hours. 200 µl was used for each RNA isolation.

85   *Seminal plasma*

86   Semen samples of healthy donors were produced by masturbation into a sterile container

87   and were allowed to liquefy for 30 min at 37 °C. Samples were centrifuged to remove

88   contaminating cells (10 min at 2000 g, room temperature, without brake) and stored at -80

89   °C within two hours after collection. 200 µl was used for each RNA isolation.

90   *Biofluid RNA purification*

91   RNA was isolated with the miRNeasy Serum/Plasma Kit (Qiagen, #217184) according to the

92   manufacturer's instructions. An input volume of 200 µL was used for all samples. Per 200 µL

93   biofluid input volume, 2 µL sequin spike-in controls (Garvan Institute of Medical Research)

94   were added before RNA isolation, in a 1/1300 000 dilution to blood plasma and in 1/1300

4

95   dilution to seminal plasma. Total RNA was eluted in 12 µL of RNAse-free water for (blood)

96   platelet depleted plasma, and in 20 µL of RNAse-free water for seminal plasma – in order to

97   adjust for viscosity. After RNA isolation, 2 µl External RNA Control Consortium (ERCC) spike-

98   in controls (ThermoFisher Scientific, #4456740) were added to the RNA isolation eluate of

99   blood plasma and seminal plasma in a dilution of 1/1000 000 and 1/1000, respectively. gDNA

100   heat-and-run removal was performed by adding 1 µl HL-dsDNase (ArcticZymes #70800-202,

101   2 U/µl) and 1.4 µl reaction buffer (ArcticZymes #66001) to the combination of 12 µl RNA

102   eluate and 2 µl ERCC spikes, followed by an incubation of 10 min at 37 °C and 5 min at 58 °C.

103   RNA was stored at -80 °C and only thawed on ice immediately before the start of the library

104   prep. Multiple freeze-thaw cycles did not occur. RNA obtained from three RNA isolations was

105   pooled per biofluid and per sample to avoid RNA isolation induced variation. This pooled RNA

106   was used as starting material for the different library preparations.

107   *FFPE*

108   Tumor RNA was isolated from five 10 µM sections of a formalin-fixed paraffin embedded

109   (FFPE) tissue block, applying macrodissection based on histopathological evaluation of

110   hematoxylin and eosin stained slides to select regions with high tumor cellularity. Within two

111   days after sectioning, the tissue sections were scraped into microcentrifuge tubes,

112   centrifuged for 5 min at 20,000 g, and deparaffinized in 320 µl Deparaffinization Solution

113   (Qiagen, #19093) for 3 min at 56 °C on a thermomixer (500 rpm). Samples were then cooled

114   to room temperature for 15 min. Subsequently, RNA was isolated using the miRNeasy FFPE

115   Kit (Qiagen, #217504), according to the manufacturer's protocol. gDNA heat-and-run removal

116   was performed by adding 1 µl HL-dsDNase (ArcticZymes #70800-202, 2 U/µl) and 0.68 µl

117   reaction buffer (ArcticZymes #66001) to 6.82 µl RNA (100 ng), followed by an incubation of

118   10 min at 37 °C and 5 min at 58 °C.

119   *MAQCA/B*

120   Two commercially available RNA samples, MAQCA and MAQCB, were used. MAQCA is the

121   Quantitative PCR Human Reference Total RNA (#750500, Agilent technologies), extracted

122   from cell lines representing different human tissues. MAQCB is FirstChoice Human Brain

123   Reference RNA (#AM7962, Life Technologies). gDNA heat-and-run removal was performed

124   on both RNA samples by adding 1 µl HL-dsDNase (ArcticZymes #70800-202, 2 U/µl) and 0.68

125   µl reaction buffer (ArcticZymes #66001) to 6.82 µl RNA (100 ng), followed by an incubation of

126   10 min at 37 °C and 5 min at 58 °C.

5

127   *Library preparation*

128   After RNA purification, four libraries were prepared for each sample: two technical replicates

129   for total RNA-seq and two technical replicates for custom lncRNA capture sequencing.

130   *SMARTer Stranded Total RNA library preparation*

131   Sequencing libraries were generated using SMARTer Stranded Total RNA-Seq Kit v2 - Pico

132   Input Mammalian (Takara Bio, #634413). The library preparation protocol started from 6 µL

133   eluate for the biofluid samples and 100 ng (or 10 ng) RNA for FFPE and MAQC. The

134   recommended amount of input RNA for SMARTer Stranded Total RNA sequencing is only up

135   to 10 ng while the capture method uses 100 ng. To make sure our analyses were not biased,

136   we decided to use the total RNA-seq method with 100 ng RNA input as well but also included

137   10 ng input samples. As shown in SFig 6 the results of 10 vs 100 ng RNA are similar. LncRNAs

138   that are only detected using one of the input amounts are mostly low abundant lncRNAs that

139   are just below the threshold. Compared to the manufacturer's protocol, the fragmentation

140   step was set to 2 min at 94 °C, hereafter the option to start from high-quality or partially

141   degraded RNA was used. During the final RNA seq library amplification, 16 PCR cycles were

142   used for the samples derived from platelet depleted (blood) plasma, 12 PCR cycles were used

143   for the other samples, and the cycles were followed by an extra 2 min at 68 °C before cooling

144   them down to 4 °C. Library quality control was performed with the Fragment Analyzer high

145   sense small fragment kit (Agilent Technologies, sizing range 50 bp-1000 bp). As Fragment

146   Analyzer profiles showed the presence of multiple adapter dimers, the final AMPure Bead

147   Purification step was repeated (17 µl AMPure beads added to each sample - 20 µl Tris Buffer

148   was used to resuspend the beads – and elution volume of 18 µl).

149   *Custom RNA capture library preparation*

150   Custom RNA capture-based libraries were prepared starting from 8.5 µL eluate for biofluid

151   samples and 100 ng RNA for FFPE and MAQCA/B using the TruSeq RNA Exome Library Prep

152   Kit (Illumina, USA). Library preparation happened according to the manufacturer's protocol

153   with some minor modifications. Fragmentation of RNA with the thermal cycler was set for 2

154   min at 94 °C (instead of 8) and incubation to synthesize first strand cDNA for 30 min at 16 °C

155   (instead of 60 min). After library validation with Fragment Analyzer (Agilent Technologies),

156   the Twist Human Core Exome EF Multiplex protocol (Twist Bioscience, San Francisco, USA)

157   was used starting with the pooling of amplified indexed libraries in sets of eight. One pool

158   consisted of MAQCA/B and seminal plasma libraries (with the required 187.5 ng per sample),

6

159 the other pool was a low-input pool containing the FFPE and (blood) plasma libraries (with

160 the available 20 ng per sample). Heated hybridization mix was added to the custom capture

161 probes without cooling down to room temperature in order to prevent the probes from

162 precipitating. After hybridization of probes with pools and binding to streptavidin beads, post

163 capture PCR amplification was performed at 8 cycles for the high-input pool and 12 cycles for

164 the low-input pool. After cleanup, the final libraries were validated with Fragment Analyzer

165 (Agilent Technologies).

166 *Sequencing*

167 Based on qPCR quantification with the KAPA Library Quantification Kit (Roche Diagnostics,

168 #KK4854), samples were pooled and loaded on NextSeq 500 with a loading concentration of

169 1.6 pM for the custom RNA capture libraries and 1.3 pM for the SMARTer Stranded Total RNA

170 libraries. Paired end sequencing was performed (2 x 75 nucleotides). Custom RNA capture

171 sequencing resulted in 168 million PE reads (median: 8.4 million PE reads/sample), SMARTer

172 Stranded Total RNA sequencing resulted in 110 million PE reads (median: 10.5 million PE

173 reads/sample). FASTQ data is currently being deposited in EGA.

174 *Sequencing data quality control*

175 The SMARTer Stranded Total RNA seq libraries were trimmed using cutadapt (v.1.16) to

176 remove 3 nucleotides of the 5' end of read 2 (Martin, 2011). Reads with a low a base calling

177 accuracy (< 99% in at least 80% of the bases in both mates) were discarded. To enable a fair

178 comparison, we started data-analysis from an equal number of reads by downsampling to the

179 minimum available paired-end reads per sample type (rounded to half a million): 6.5 million

180 for FFPE, 7.5 million for MAQCA/B, 6 million for seminal plasma, 3 million for platelet-

181 depleted (blood) plasma. Downsampling was done with Seqtk (v1.3) (Li, 2021). Next, read

182 duplicates were removed with Clumpify (BBMap v.38.26, standard settings) using the

183 following specifications: paired-end mode, 2 substitutions allowed, kmersize of 31, and 20

184 passes (Bushnell, 2021). For duplicate removal, only the first 60 nucleotides of both reads

185 were considered to account for the sequencing quality drop at the end of the reads. Full-

186 length read sequences were retrieved after duplicate removal for further quantification.

187 *Quantification of Ensembl and LNCipedia genes*

188 Strand-specific transcript-level quantification of the deduplicated FASTQ files was performed

189 with Kallisto (v.0.44.0) in –rf-stranded mode (Bray *et al.*, 2016). Quantification was performed

190 with two references. The first one is a custom Ensembl v75 reference where lncRNAs are only

7

191   taken from LNCipedia 5.2 (high-confidence set) (Volders *et al.*, 2019; Yates *et al.*, 2020). This

192   reference was used to design the custom probes. The second reference is only based on a

193   more recent version of Ensembl v91.

194   Further processing was done with R (v.4.0.3) making use of tidyverse (v.1.3.0). A count

195   threshold for filtering low abundant genes was set based on an analysis of single positive

196   genes in technical replicates (Mestdagh *et al.*, 2014). Single positives are genes with a zero

197   count value in one replicate and a non-zero value in the other one. After applying a threshold

198   of 10 counts, at least 95% of the single positives are removed (SFig 3).

199

200   Results

201   In brief, 565,878 lncRNA capture probes of 120 nucleotides in length were designed against

202   the high confidence set of LNCipedia v5.2 (Volders *et al.*, 2019) that comprises 107,039

203   transcripts belonging to 49,372 lncRNA genes. This probe set targets 45,284 lncRNA genes or

204   91.72% of the LNCipedia high confidence set. The median number of probes designed per

205   lncRNA is 5 (SFig 1a), ranging from 1 up to 1675 probes for lnc-TBC1D22A-4 (with a length of

206   152,544 bp). The selected probe designs have a median GC of 43.33%, a Tm of 72.42 °C and

207   ΔG of -2.8 (SFig 1b,c,d).

208   The custom lncRNA capture approach was applied to RNA from four different human sample

209   types: high-quality RNA (artificial RNA mixture from human cell lines, MAQCA, and human

210   brain reference RNA, MAQCB (Shi *et al.*, 2006)), formalin-fixed paraffin embedded colon

211   tissue samples (FFPE), platelet-depleted blood plasma and seminal plasma. Each sample was

212   also profiled with a total RNA-sequencing workflow representing the gold standard for

213   quantification of both polyadenylated and non-polyadenylated lncRNAs.

214   We observed a clear enrichment of the lncRNA fraction with custom lncRNA capture

215   compared to total RNA-seq when mapping reads to a LNCipedia transcriptome reference. Up

216   to 75% of mapped reads in the custom capture method are derived from lncRNAs (Fig 1a),

217   which is a 3.5-fold enrichment compared to total RNA-seq for FFPE, 4-fold for high quality

218   MAQCA/B RNA and 8.5-fold for seminal plasma. This enrichment was also observed when

219   aligning reads to a less comprehensive lncRNA reference (the Ensembl v91 reference),

220   although the level of enrichment was lower (SFig 2a). In blood plasma, only a small fraction

221   of reads aligned to lncRNAs for both the custom lncRNA capture and total RNA-seq method,

222   resulting in the detection of just a few hundred lncRNAs (data not shown). In FFPE, the

223   fraction of reads mapping to ribosomal RNA was higher in total RNA-seq (38% and 52% for

224   donor 1 and 2, respectively) compared to custom capture (10% and 20% for donor 1 and 2,

225   respectively) (Fig 1a). In other sample types, the lower fraction of lncRNA reads in total RNA

226   sequencing compared to custom capture sequencing is almost exclusively compensated by a

227   higher fraction of protein coding RNA (mRNA) reads.

228   After downsampling to the same number of reads, we applied a minimal coverage of 10

229   counts to select for lncRNAs that are reproducibly detected (SFig 3) and compared detection

230   sensitivity between both methods. Although both methods were able to detect several

231   thousands of lncRNAs, the custom capture method on average resulted in two times more

232   uniquely detected lncRNAs compared to total RNA-seq (Fig 1b). The maximum number of

233   detected lncRNAs with the custom capture approach was 8186 for FFPE, 11,238 for

234   MAQCA/B, and 6910 for seminal plasma. As expected, the majority of lncRNAs detected in all

235   total RNA-seq replicates were also detected in all custom capture replicates: 87%-91% of

236   lncRNAs based on LNCipedia reference (Fig 1c); 83%-91% based on Ensembl reference (SFig

237   2c). More importantly, custom capture enabled the detection of several thousands of

238   additional lncRNAs (59%-61% of all lncRNAs reproducibly detected by custom capture were

239   not detected by total RNA-seq), illustrating the sensitivity of this procedure (Fig 1c & SFig 2c).

240   Expression abundance analysis revealed that these uniquely detected lncRNAs are generally

241   less abundant compared to lncRNAs detected by both methods (Fig 1d & SFig 2d).

242   Next, we evaluated reproducibility based on absolute log2 fold changes of lncRNA abundance

243   between technical replicates (ideally, these fold changes are close to zero). As shown in Fig 2

244   and SFig 4, we observed a higher fraction of lncRNAs with a log fold change close to zero in

245   the custom capture approach compared to the total RNA-seq approach, indicating a better

246   reproducibility for the custom capture approach. Only the total RNA-seq data of seminal

247   plasma from donor 1 showed better reproducibility (SFig 4e), yet the custom approach in

248   general still had lower fold changes between technical replicates (Kolmogorov-Smirnov test

249   p-value < 0.001). Note that seminal plasma from donor 1 also resulted in a lower number of

250   unique lncRNAs than that of donor 2 (Fig 1b).

251   We also compared transcript coverage of lncRNAs that were detected with both approaches

252   by looking at their TPM distributions. In general, coverage was higher in the custom capture

253   approach than in total RNA-seq (Fig 2 & SFig 4). Median TPM values for the custom capture

254 and total RNA-seq approach, respectively, were 8.2 and 2.0 TPM in FFPE, 9.7 and 2.5 TPM in

255 MAQCA/B, and 16.1 and 2.3 TPM in seminal plasma. In terms of gene body coverage, both

256 methods covered the entirety of the lncRNA body with an expected lower coverage towards

257 the 5' and 3' ends. The custom capture sequencing, however, showed a more pronounced

258 reduction in coverage towards the 3' end of the lncRNAs compared to total RNA-seq (SFig 5).

259 Finally, we wanted to further assess the relevance of the custom capture approach for

260 biological or clinical applications. We evaluated the abundance of previously described

261 prostate-cancer related lncRNAs (Helsmoortel et al., 2018) in seminal plasma samples

262 between both methods. As shown in Fig 3, coverage of detected lncRNAs is consistently

263 higher with custom capture sequencing than with total RNA-seq. In total, 16 prostate-cancer

264 related lncRNAs were detected above threshold in at least one sample. While none of those

265 lncRNAs were exclusively detected by total RNA-seq, five lncRNAs (LINC01564, lnc-HNF1A-1,

266 lnc-SPATA31A6-6, PCA3, and PCAT7) were detected by custom capture sequencing only. The

267 custom capture counts of these lncRNAs ranged from 11 to 40 when taking the mean of both

268 technical replicates of donor 2. Yet, these lncRNAs (except LINC01564) did not reach the

269 detection threshold in custom capture sequencing samples of donor 1. For the 11 lncRNAs

270 that were detected with both methods, custom capture sequencing resulted in 2 to 15 times

271 more counts compared to total RNA-seq with an average fold change of 6. This increased

272 sensitivity could greatly benefit biomarker research.

273 In summary, these findings demonstrate the added value of our custom lncRNA capture

274 method for applications aimed at establishing a more complete lncRNA expression landscape.

275

276 Discussion

277 An extensive enrichment combined with a higher coverage of lncRNAs may further improve

278 our understanding of lncRNA association to various conditions or phenotypes. Studies aiming

279 to identify lncRNA biomarkers could equally benefit from these advantages. We have

280 demonstrated a superior performance of custom lncRNA capture sequencing compared to

281 classic total RNA-sequencing, across different sample types. Fewer reads are consumed by

282 RNA biotypes other than lncRNAs, which results in a better lncRNA coverage. Interestingly,

283 we also observed a better lncRNA detection reproducibility between technical replicates for

284 the custom capture compared to total RNA-seq (Fig 2). Deeper sequencing could event

285 further improve the performance.

286 The custom capture method, however, did not outperform the total RNA-sequencing method

287 in platelet-depleted blood plasma. In these samples, both methods only detected a few

288 hundred lncRNAs. This observation is in line with the fact that the extracellular mRNA

289 concentration in this sample type is low (Hulstaert *et al.*, 2020). Additionally, the blood plasma

290 samples were not sequenced at high depth (3 million paired-end reads before duplicate

291 removal), suggesting that results may improve when generating more reads.

292 For 10 to 25% of lncRNAs that were uniquely detected with total RNA-seq, no probes were

293 present in the custom capture probe set because of a failure to satisfy probe design

294 requirements. About half of the lncRNAs with at least one custom probe were still detected

295 in some of the capture libraries but failed to reach the threshold in other replicates (and

296 where therefore labeled as undetected in these libraries). While data was downsampled to

297 the same read depth, increasing sequencing depth may solve this discrepancy. Some of the

298 lncRNAs did not have probes complementary to the transcript regions that were detected

299 with total RNA-seq. Incorporating additional probes against those regions could further

300 improve performance, although this would require loosening probe design criteria, which

301 may result in more non-specific hybridization and off-target capture. For the remaining

302 lncRNAs, further optimization of the probe designs may be required to enable proper capture.

303 Note that the custom capture library preparation is considerably more expensive than total

304 RNA-sequencing. The price difference is mainly driven by the large custom probe set. Of note,

305 probe cost could be substantially reduced when offered off-the-shelf or by transitioning from

306 a discovery phase to a validation phase, including only those probes that target lncRNAs of

307 interest. In this study, the stranded TruSeq RNA Exome Library Prep Kit was used for the

308 custom capture approach, but other library prep methods would work too.

309 Taken together, we demonstrated that lncRNA capture sequencing is able to detect a more

310 diverse repertoire of lncRNAs compared to standard total RNA sequencing, and increases

311 coverage as well as reproducibility in both high-quality high input as well as fragmented

312 and/or low input RNA samples.

313

327   Conflict of interest

328   The authors declare no conflicts of interest.

329

330   References

331   Bray,N.L. *et al.* (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**,

332   525–527.

333   Bushnell,B. (2021) BBMap. *SourceForge*.

334   Cao,M. *et al.* (2019) Genome-wide methods for investigating long noncoding RNAs. *Biomed.*

335   *Pharmacother.*, **111**, 395–401.

336   Clark,M.B. *et al.* (2015) Quantitative gene profiling of long noncoding RNAs with targeted RNA

337   sequencing. *Nat. Methods*, **12**, 339–342.

338   Djebali,S. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

339   Helsmoortel,H. *et al.* (2018) Detecting long non-coding RNA biomarkers in prostate cancer

340   liquid biopsies: Hype or hope? *Non-Coding RNA Res.*, **3**, 64–74.

341   Hulstaert,E. *et al.* (2020) Charting Extracellular Transcriptomes in The Human Biofluid RNA

342   Atlas. *Cell Rep.*, **33**, 108552.

343   Iyer,M.K. *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome.
344   *Nat. Genet.*, **47**, 199–208.

345   Lagarde,J. *et al.* (2017) High-throughput annotation of full-length long noncoding RNAs with
346   Capture Long-Read Sequencing. *Nat. Genet.*, **49**, 1731–1740.

347   Li,H. (2021) lh3/seqtk.

348   Lorenzi,L. *et al.* (2019) The RNA Atlas, a single nucleotide resolution map of the human
349   transcriptome. *bioRxiv*, 807529.

350   Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing
351   reads. *EMBnet.journal*, **17**, 10–12.

352   Mercer,T.R. *et al.* (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, **10**,
353   155–159.

354   Mercer,T.R. *et al.* (2014) Targeted sequencing for gene discovery and quantification using
355   RNA CaptureSeq. *Nat. Protoc.*, **9**, 989–1009.

356   Mestdagh,P. *et al.* (2014) Evaluation of quantitative miRNA expression platforms in the
357   microRNA quality control (miRQC) study. *Nat. Methods*, **11**, 809–815.

358   Robinson,E.K. *et al.* (2020) The how and why of lncRNA function: An innate immune
359   perspective. *Biochim. Biophys. Acta Gene Regul. Mech.*, **1863**, 194419.

360   Shi,L. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and
361   intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, **24**,
362   1151–1161.

363   Turner,A.W. *et al.* (2019) Multi-Omics Approaches to Study Long Non-coding RNA Function in
364   Atherosclerosis. *Front. Cardiovasc. Med.*, **6**.

365   Volders,P.-J. *et al.* (2019) LNCipedia 5: towards a reference set of human long non-coding
366   RNAs. *Nucleic Acids Res.*, **47**, D135–D139.

367   Yates,A.D. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.

368

369   Figures

370
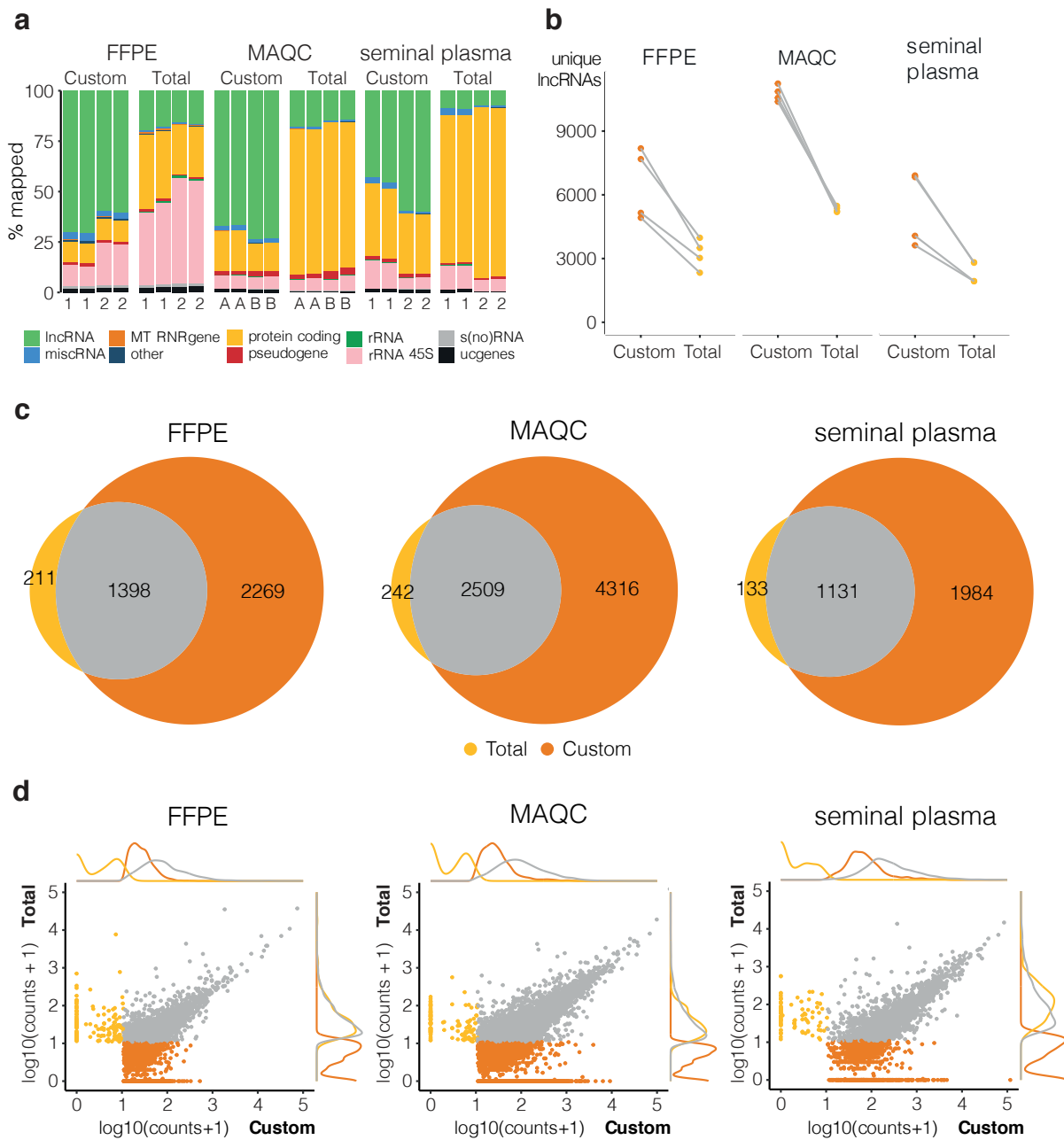
Fig 1: Custom capture sequencing (Custom) is able to detect more lncRNAs than total RNA-sequencing (Total). Quantification based on combined reference of Ensembl and LNCipedia. a: RNA biotype distribution plot of mapped reads where 1 and 2 indicate the two different donors and A and B refer to MAQCA and MAQCB, respectively (lncRNAs: high-confidence lncRNAs based on LNCipedia 5.2; miscRNA: miscellaneous RNA, non-coding RNA that cannot be classified; MT RNR gene: mitochondrially encoded ribosomal RNAs; protein coding: protein coding RNA transcripts; pseudogene; rRNA (45S): (45S) ribosomal RNA; s(no)RNA: small nuclear/nucleolar RNA; ucgenes: unannotated cancer genes; other: T cell receptor genes,

14

379   Immunoglobulin genes, TEC (To be Experimentally Confirmed) - regions with EST clusters that

380   have polyA features that could indicate the presence of protein coding genes, vaultRNA -

381   short non coding RNA genes that form part of the vault ribonucleoprotein complex;

382   microRNAs; ribozymes); b: number of unique lncRNAs with at least 10 counts (filter

383   threshold), data points from same donor or MAQC type are linked (grey lines); c: overlap

384   between lncRNAs that are detected above threshold in all replicates of a certain library prep

385   method, plots made with eulerr package (v6.1.0) in R; d: correlation and density plots of

386   overlapping (grey) and specific lncRNAs for custom capture (orange) and total RNA-

387   sequencing (yellow); lncRNAs below count threshold in both methods were left out.
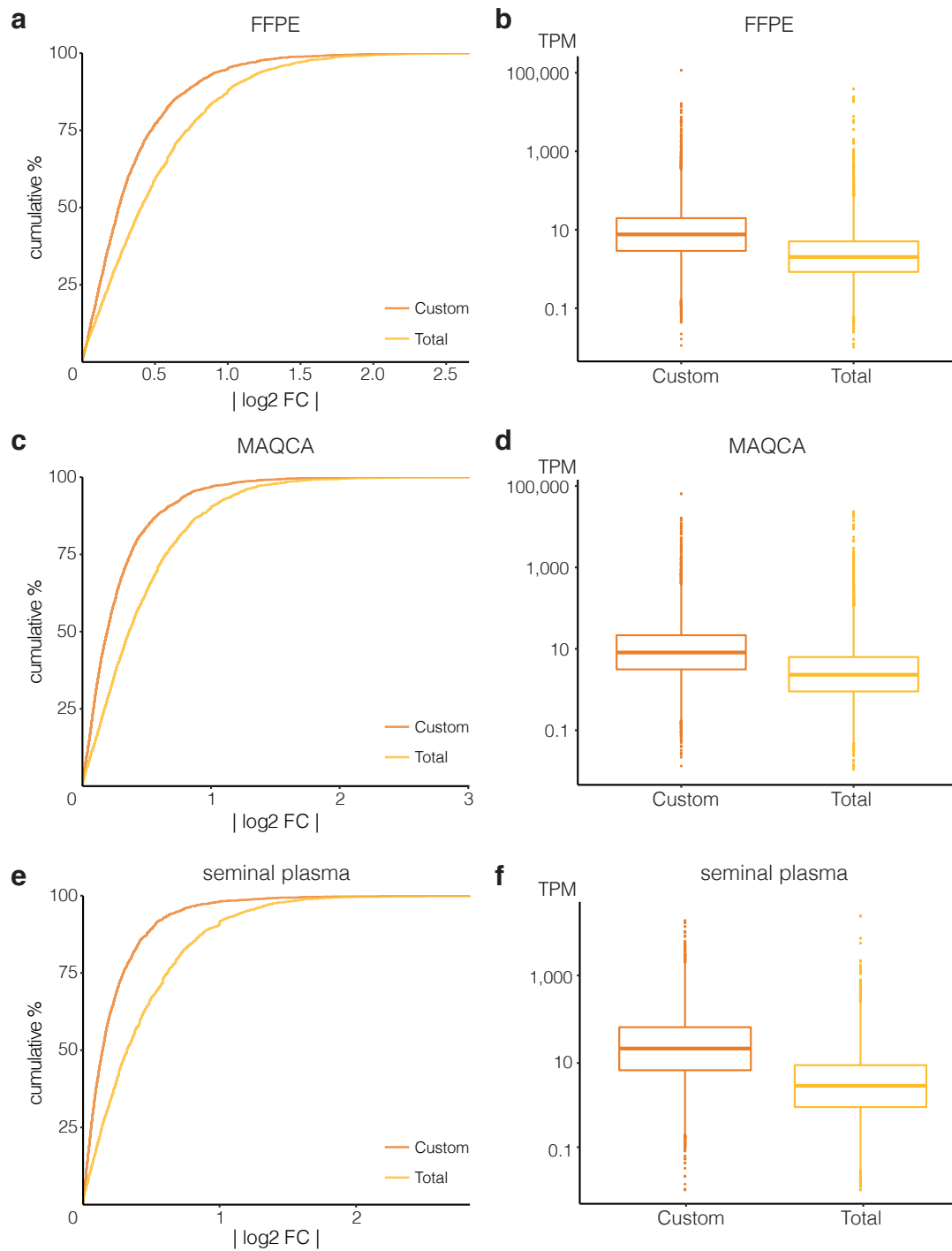
Fig 2: Custom capture seq (Custom) has a higher lncRNA count reproducibility and coverage than total RNA-seq (Total). Cumulative distributions of absolute log2 fold changes (log2 FC) between lncRNA counts in the two technical replicates are shown for (a) FFPE from donor 1, (c) MAQCA, and (e) seminal plasma from donor 2. Kolmogorov–Smirnov tests each time showed significant difference in distributions between Total and Custom (p-value < 0.001). Boxplot of corresponding transcripts per million (TPM) values of these lncRNAs are shown in (b) for FFPE, (d) for MAQCA, and (f) for seminal plasma.
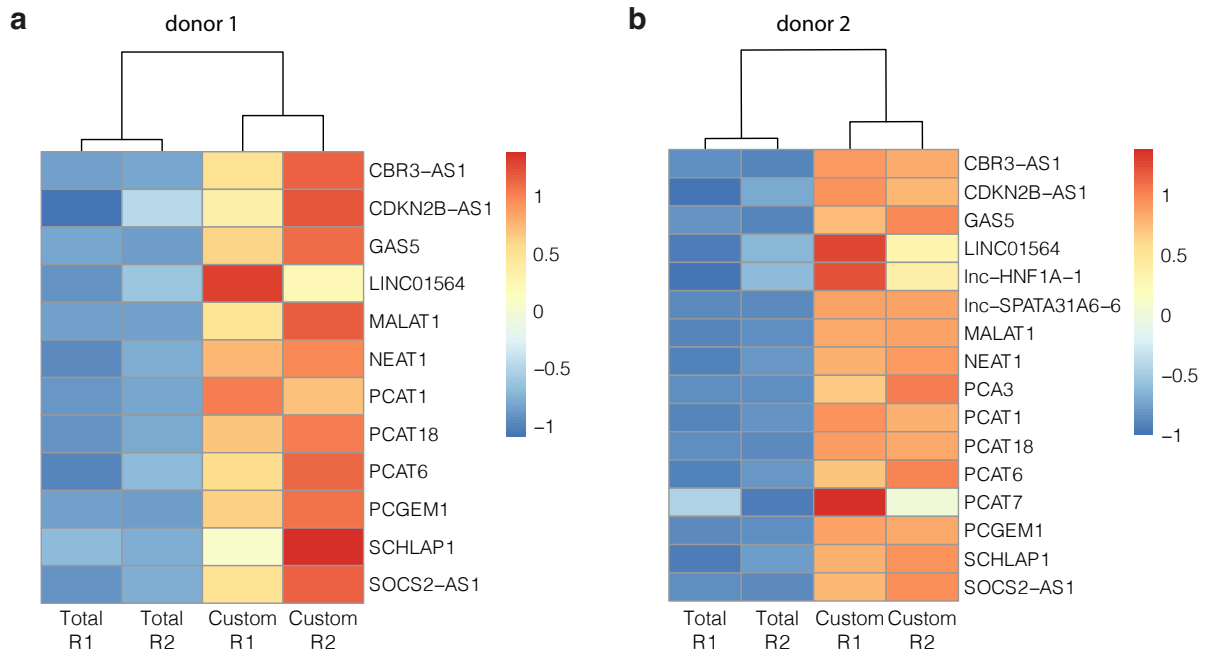
Fig 3: Higher coverage for prostate-cancer related lncRNAs with custom capture (Custom) than total RNA-sequencing (Total). Heatmaps based on z-score transformed lncRNA counts of seminal plasma samples from donor 1 (a) and donor 2 (b), respectively. Per donor, only lncRNAs detected above count threshold (10 counts) in at least one replicate were considered. A higher z-score (orange/red) indicates relatively more coverage. Complete clustering of samples based on Euclidean distance. R1: technical replicate 1; R2: technical replicate 2.

405   Supplemental Figures
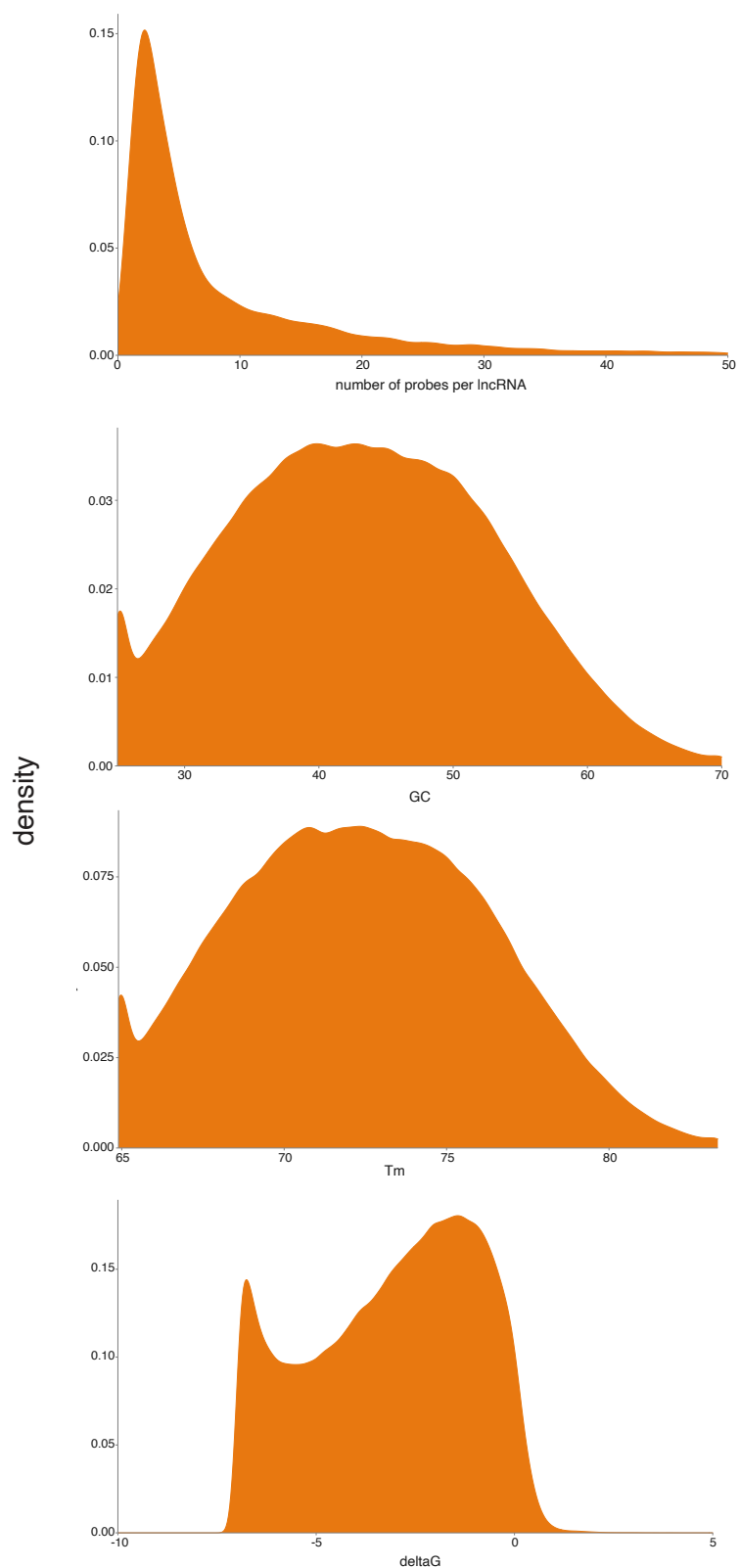
406

407   **SFig 1: Distributions of number of probes per gene, GC%, melting temperature and ΔG of**

408   **the selected probes.**

409
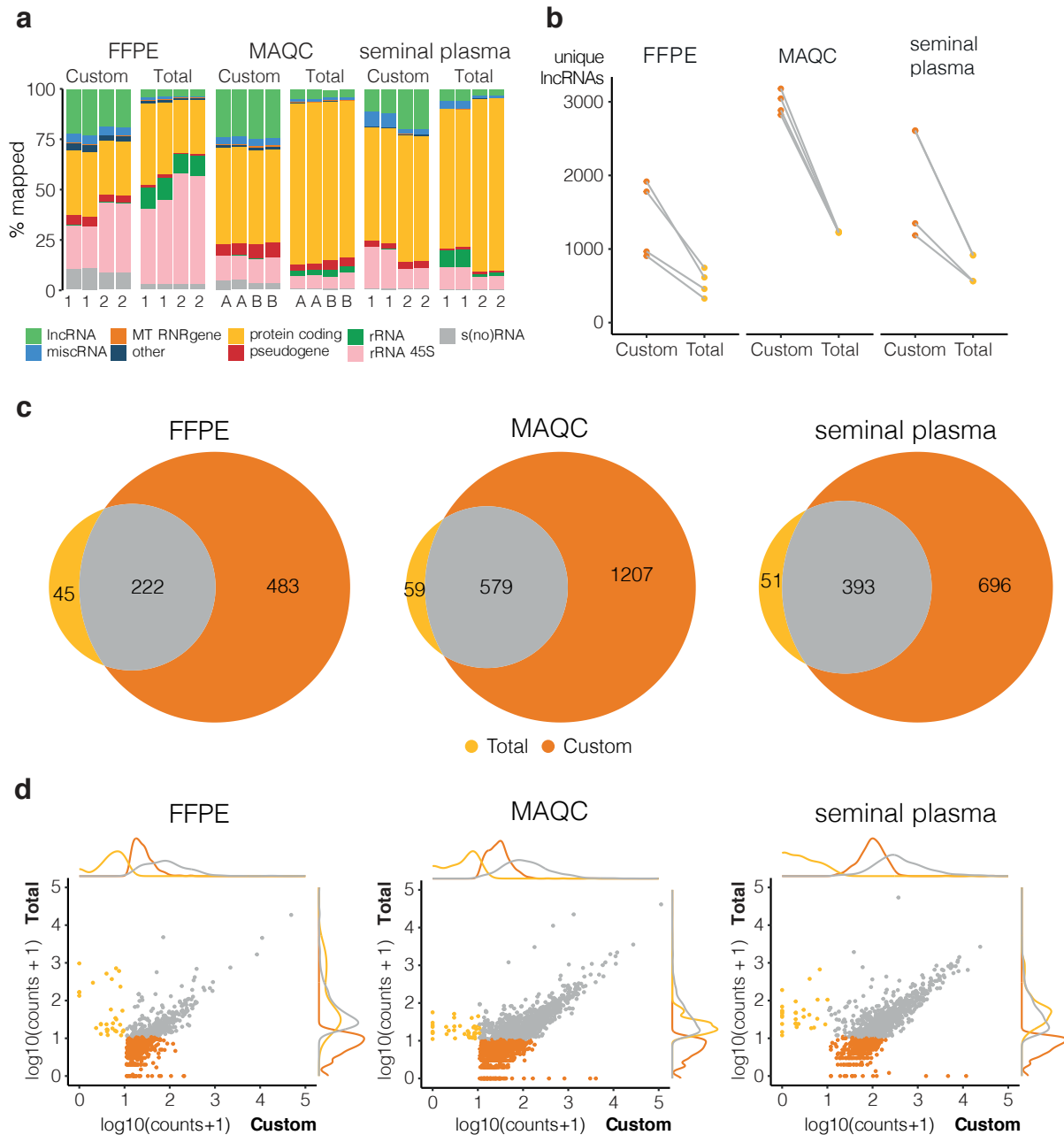
SFig 2: Custom capture sequencing (Custom) is able to detect more lncRNAs than total RNA-sequencing (Total). Quantification based on Ensembl v91 reference. a: RNA biotype distribution plot of mapped reads where 1 and 2 indicate the two different donors and A and B refer to MAQCA and MAQCB, respectively (lncRNAs: Ensembl lncRNAs; miscRNA: miscellaneous RNA, non-coding RNA that cannot be classified; MT RNR gene: mitochondrially encoded ribosomal RNAs; protein coding: protein coding RNA transcripts; pseudogene; rRNA (45S): (45S) ribosomal RNA; s(no)RNA: small nuclear/nucleolar RNA; ucgenes: unannotated cancer genes; other: T cell receptor genes, Immunoglobulin genes, TEC (To be Experimentally Confirmed) - regions with EST clusters that have polyA features that could indicate the

19

419   presence of protein coding genes, vaultRNA - short non coding RNA genes that form part of

420   the vault ribonucleoprotein complex; microRNAs; ribozymes); b: number of unique lncRNAs

421   with at least 10 counts (filter threshold), data points from same donor are linked (grey lines);

422   c: overlap between lncRNAs that are detected above threshold in all replicates of a certain

423   library prep method, plots made with eulerr package (v6.1.0) in R; d: correlation and density

424   plots of overlapping (grey) and specific lncRNAs for custom capture (orange) and total RNA-

425   sequencing (yellow); lncRNAs below count threshold in both methods were left out.

426
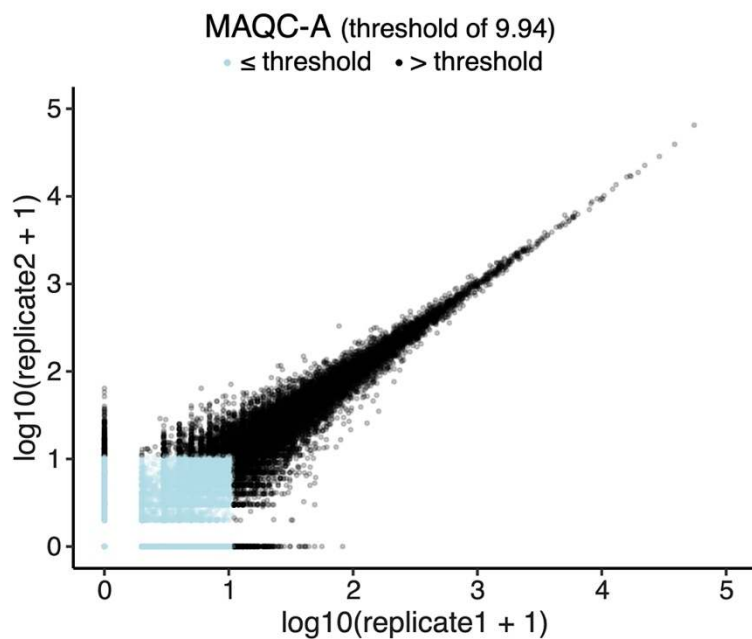
427

428   **SFig 3: Reproducibility threshold is determined based on elimination of at least 95% of**

429   **single positives between technical (library preparation) replicates.** Single positives are

430   detected (at least 1 count) in one replicate and not detected in the other replicate. Here,

431   example of replicate correlation of MAQCA with total RNA sequencing is shown. Count

432   threshold that filters out 95% of single positives is 9.94 (kallisto quantification leads to decimal

433   counts). Data points that will be filtered with this threshold are in blue. Slope of linear model

434   is 0.969, pearson correlation is 0.999.
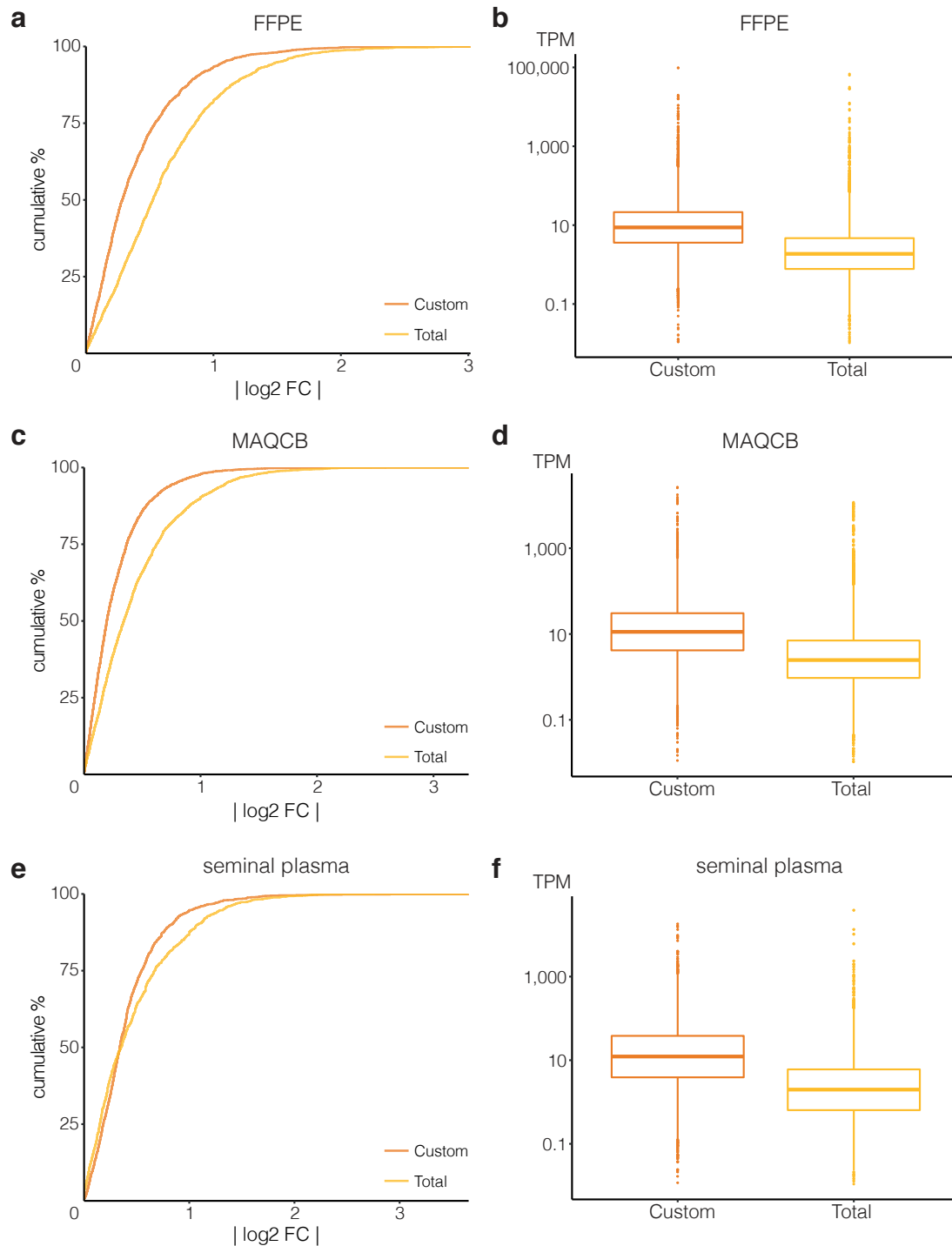
**SFig 4: Custom capture seq (Custom) has a higher lncRNA count reproducibility and coverage than SMARTer Stranded Total RNA seq (Total).** Cumulative distribution of absolute log2 fold changes between lncRNA counts in the two technical replicates are shown for (a) FFPE from donor 2, (c) MAQCB, and (e) seminal plasma from donor 1. Kolmogorov–Smirnov tests each time showed significant difference in distributions between Total and Custom (p-value < 0.001). Boxplot of corresponding transcripts per million (TPM) values of these lncRNAs are shown in (b) for FFPE, (d) for MAQCB, and (f) for seminal plasma.
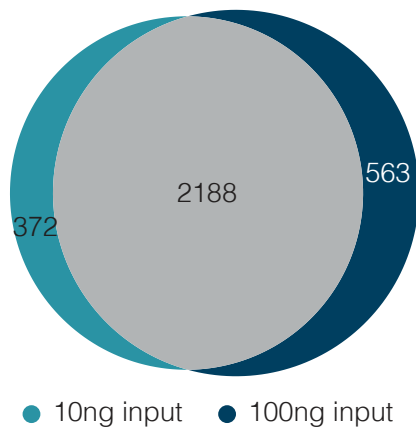
443
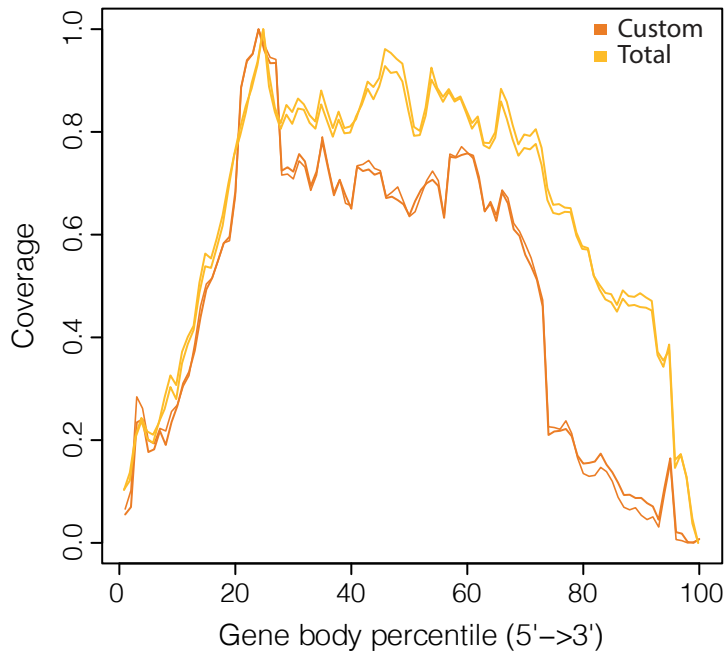
SFig 5: Distribution of gene body coverage shows quite stable coverage for majority of gene (lncRNA) body but drops towards the 5' and 3' ends. Distribution shown for both technical replicates of MAQCA. Custom: custom capture sequencing; Total: SMARTer Stranded Total RNA sequencing.
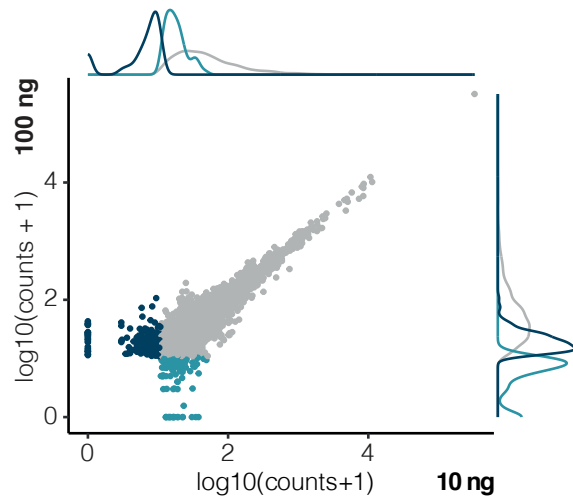
448

449

SFig 6: Impact of 10 vs 100 ng RNA input on SMARTer Stranded Total RNA sequencing results is limited. a: overlap between lncRNAs detected (> 10 counts) in MAQC with 10 vs 100 ng RNA input for library preparation; b: corresponding count correlation and density plot (counts shown for one technical replicate of MAQCA).

22