SUPPLEMENTARY MATERIALS

SUPPLEMENTARY METHODS

Sample collection and enrichment of EVs from human plasma, serum, and urine

For the prostate cancer dataset, human serum was collected using BD Vacutainer blood collection tubes (i.e., serum separation tubes). First, whole blood was centrifuged at 2,000g for 30 minutes at 4ºC followed by another centrifugation of the serum at 12,000g for 45 minutes at 4ºC to remove larger EVs (e.g. microvesicles and apoptotic bodies). The supernatant was carefully transferred to ultracentrifugation tubes (Beckman coulter, thick wall polypropylene tube, Cat. #355642) and ultracentrifuged for two rounds at 110,000g for 2 hours at 4ºC. The pellet was finally resuspend in 1 mL PBS and stored at -80C for further analysis. EV enrichment from human urine was performed with the same protocol.

For the HCC 'biomarker discovery' and 'biomarker validation' dataset, peripheral venous blood was collected in EDTA containing vacutainer (BD Vacutainer), stored on ice, and processed within 4 hours of collection. On the day of collection, we performed two centrifugation steps to separate plasma from other blood components and minimize cellular debris from our final isolate. First, whole blood was centrifuged at 1,600g for 10 minutes at 4ºC followed by another centrifugation of the plasma at 16,000g for 10 minutes at 4ºC to remove larger EVs (e.g. microvesicles and apoptotic bodies). The supernatant was then stored at -80°C until the ultracentrifugation was performed. For this, samples were thawed on ice and 0.5 - 1 mL of plasma was diluted in ~25 mL PBS and centrifuged at 120,000g for 2 hours at 4°C with a Type 50.2 Ti Fixed-Angle Titanium rotor (Beckman Coulter, k-factor = 69). Isolates were directly used for RNA extraction (see below) or resuspended in PBS and stored at -20°C until further analysis.

Characterization of EV-enriched isolates

Characterization procedures of our isolates were guided by recommendations from the International Society for Extracellular Vesicles (ISEV)[19]. After differential ultracentrifugation, the PBS-resuspended isolate was evaluated with transmission electron microscopy (TEM) in a Hitachi 7000 transmission electron microscope operating at 80 kV. Briefly, equal volumes of the isolate and 3% Glutaraldehyde were mixed and kept at room temperature for 1 hour. Two µl of osmium tetroxide was added to the mixture and incubated at room temperature for 1 hour. The solution was then transferred to formvar coated TEM grids and observed under the electron microscope. To estimate the size and concentration of the isolate, we conducted nanoparticle tracking analysis (NTA) on a NanoSight NS300 (Malvern Instruments Ltd, Malvern, UK) and analyzed the samples with the NTA 3.2 software (Malvern). For this, PBS-resuspended isolates were diluted 1:50 in PBS.

For immuno-labeling of the isolate, we performed Western Blotting for the intracellular marker TSG101 and Exoview™ analysis for colocalization of tetraspanins CD9, CD63, and CD81. For Western Blotting, we quantified protein concentration (Bradford assay, Biorad) and 20 μg of protein were separated by sodium dodecyl sulfate–polyacrylamide electrophoresis under reducing conditions and transferred to PVDF membranes (Life Technologies). Unspecific binding sites were blocked with 5% nonfat dry milk and membranes were incubated with mouse monoclonal TSG101 antibody (ab83, Abcam, RRID:AB_306450) at 4°C overnight followed by goat anti-mouse secondary antibody (A0447, Agilent Technologies) for 1 hour at room temperature. Chemiluminescence was detected using the ECL™ Prime Western Blotting System (RPN2232, GE Healthcare). The uncropped Western Blot image for TSG101 is displayed in **Supplementary Fig. S9**. Exoview™ experiments were carried out on an ExoView™ R100 imaging platform (NanoView Bioscience). With the Exoview™ Tetraspanin kit, 35 μl of PBS-resuspended isolate was incubated overnight on a microarray chip which has been functionalized with antibodies against CD9, CD63, CD81, plus IgG negative control to detect EVs expressing these surface markers. After washing off unbound particles, chips were stained with fluorescence-conjugated antibodies against CD9 (Alexa 647) or CD81 (Alexa 555) to identify subpopulations based on maker profiles. Analysis was done with the NanoViewer 2.4.5 (NanoView Bioscience).

RNA extraction, small library preparation and next-generation sequencing

For the prostate cancer dataset, total RNA was extracted from the serum/urine bump fraction (nanoDLD, serum only), UC isolates, or bulk tissue using the Total Exosome and Protein Isolation Kit (Invitrogen 4478545) by following the protocol. For the HCC biomarker discovery and biomarker validation datasets, RNA was extracted from the UC isolate on the same day of ultracentrifugation using the miRNeasy Plasma/Serum kit (Qiagen) according to the manufacturer's recommendations including the spike-in *C. elegans* miR-39 miRNA mimic and stored at -80°C until further use. RNA quantitation and quality was assessed on a 2100 Bioanalyzer Instrument (Agilent) with the RNA 6000 Pico Kit (Agilent). Indexed Illumina Small RNA libraries were prepared with the SMARTer® smRNA-Seq Kit (Clontech Laboratories, Inc.) and sequenced on an Illumina HiSeq 4000 (prostate cancer dataset) or HiSeq2500 (liver cancer dataset) platform.

Trimming

The SMARTer™ smRNA-Seq kit yields reads are flanked on the 5’ end by a leading triad of three bases from SMARTer™ template switching activity, and on the 3’ end by the Illumina adapter and extra bases from the oligo dT (which are exactly 15 bp in length). We used Cutadapt[40]

to remove the first 3 nucleotides of all reads, specify the homopolymer adapter sequence AAAAAAAAAA to remove along with any sequence 3' of it, and finally discard all reads that are smaller than 15 bp long after these filters are applied. The exact command used, as recommended by the (strand-sensitive) SMARTer™ smRNA-Seq kit, is

```
cutadapt -m 15 -u 3 -a AAAAAAAAAA input.fastq > output.fastq
```

Therefore our set of initial small RNAs are at least 15 bp long, and are trimmed from positions 1-3 and also from the oligo dT 3' through to the adapter. We note in passing that although template switching at low frequencies can add more than 3 nucleotides to the 5' end, we did not trim any further on the 5' end.

Alignment and multiple small RNA mapping

All samples' smRNA adapters were trimmed as above. With the aim of identifying the transcriptional site of origin for each small RNA-seq (small RNA-seq) read, we aligned these small RNA-seq data to the GRCh38 reference genome. In order to deal with the preponderance of multiply mapping small RNA-seq reads (i.e., a single read that maps with equal confidence to m >= 2 genomic regions), we adopted strategies based on those of ERANGE 4.0a[41], SiLoCO[42–44], which quantify expression of multiply mapped reads as proportional to the number of uniquely mapped reads in the vicinity. It has been extensively shown by Johnson et al.[44] that compared to simply ignoring multiply mapped reads (minimizing sensitivity but maximizing specificity) or randomly assigning them (maximizing sensitivity, minimizing specificity, the default bowtie setting), local weighting of multiply mapped reads (m < 50) by genomic context leads to improved small RNA-seq alignment performance. Briefly, the procedure is as follows:

a) For each trimmed fastq file, using bowtie[45], determine all *best*-matched alignments to genomic reference GRCh38 for each small RNA-seq read and discard those that multiply map > 50 times;

```
bowtie -q -v 1 -S -a -m 50 –-best --strata <hg38_genome>  < <input.fastq> > <output.sam>
```

b) Read-sort the sam files (using samtools sort) and then finally merge the sorted sam files (using samtools merge) (samtools, RRID:SCR_005240). Count the number of uniquely mapped reads in 50 bp bins across the reference genome, with bin start coordinates defined by the 5' edge of the uniquely mapped read. For every multiply mapped read (m < 50) the

local count of uniquely mapped reads, in each 50 bp bin, is computed across all alignment coordinates and converted to fractions of the sum of total counts across those locations. These fractions are the probabilities for placement of the multiply-mapped read in a particular alignment position, as drawn from a normal distribution. In cases where there are no proximal uniquely mapped reads recorded near the multiply-mapped loci, those loci will be randomly chosen if there are 3 or fewer choices. Otherwise they are discarded. Uniquely mapped or assigned (guided or m < 3) multiply mapped reads are called primary alignments.

c) Following Johnson et al.[44,46], we assign the following auxiliary CIGAR codes to the alignments in the final merged and sorted SAM/BAM file to reflect the uniqueness and mapping status of each alignment.

1. XY:Z:N -- Unmapped with zero valid alignments
2. XY:Z:M -- Unmapped with too many alignments: alignments > 50
3. XY:Z:O -- Unmapped because no guidance was possible (> 3 choices with no proximal unique-mappers)
4. XY:Z:U -- Uniquely mapped
5. XY:Z:R -- Multi-mapped with primary alignment chosen randomly (no guidance, <= 3 choices)
6. XY:Z:P -- Multi-mapped with primary alignment chosen guided by unique-weighting. Record probability of uniquely-weighted assignment

d) The total tallies of these counts across the two cohorts (Prostate smRC training cohort, n = 41), (HCC smRC biomarker discovery cohort, n = 15) are given below. Primary alignments are all placed reads (CIGAR codes U + R + P), and percentages in brackets are given in terms of total number of input reads retained after cutadapt adapter trimming (**Supplementary Table S6**).

Deconvolution analysis

EV carrier deconvolution analysis was performed as a post-processing step to the standard exceRpt pipeline[47], which was applied to the entire HCC smRC discovery dataset (n=15). The output of exceRpt is collated (using mergePipelineRuns.R from https://github.com/rkitchen/exceRpt) to form summary data of count matrices for key annotated, noncoding RNA biotypes (piRNA, circRNA, miRNA, tRNA counts), aggregated QC data, adapter sequence data, and diagnostic plots. At this point we applied their deconvolution algorithm on the

summarized data. Briefly, this consists of two key stages: In the first stage, constituent cargo profiles are estimated using a modified version of a methylation deconvolution technique in Onuchic et al.[48]. Next, deconvolution is performed using the Read Counts or RPM sample profiles from the exRNA Atlas and the per-sample proportion enrichments of each profile are estimated.

smRC definition, properties, exRNA-specific smRCs

Details on trimming and alignment can be found in the supplementary methods. Clusters of primary alignments that are genomically localized – as expected biologically from localized small RNA precursors giving rise to multiple small RNA mature products -- can be defined by a simple moving average smoothing window via minimum coverage and primary alignment spacing (padding) constraints as shown in **Fig. 3A**. Essentially, all regions of the genome that are tiled edge-to-edge by small RNA-seq reads within the padding constraint are filtered for those that contain at least the number of small RNA-seq reads specified by the minimum coverage threshold. Increasing the minimal coverage threshold decreases sensitivity (oversmooths) while increasing the padding decreases specificity. For the prostate cancer 'smRC characterization' cohort, we set the minimum coverage threshold to just over 1 rpm (read per million primary alignments) (U + P + R), which works out to imply that at least 248 small RNA-seq reads must be contained within the cluster, while for the 'HCC biomarker discovery' cohort, we set the same rpm threshold and obtained 205 reads. In both cohorts we empirically set the padding to be approximately three quarters of the small RNA-seq read length, 75 bp.

In order to quantify the peakiness of the distribution of reads within the moving average window we define a tiling complexity (C) measure as the percentage of the total window coverage comprised of unique read sequences. Low percentages ($C < 0.1$) indicate a highly peaked distribution, while high percentages ($C > 0.9$) reflect a relatively more uniform distribution. We note in passing that a similar measure could be defined for the dominant read strandedness of the window (+, -, mixed). In addition, within the window we identify the peak coverage and its consensus sequence.

We define small RNA clusters (smRC) as regions passing the above coverage and padding filters possessing total expression (computed in log2CPM units), complexity, and peak consensus sequences

smRC := (chr:start-end; log2CPM, complexity, peak_consensus_sequence,…)

and annotate the genome with smRCs. The ellipsis indicates that other important parameters can be added (average strandedness, peakiness, smRC length, etc). For each smRC, we calculate the relative contribution of each sample to the total coverage by tallying a raw read count matrix with row dimension equal to the number of smRCs and column dimension equal the number of samples. The total number of smRCs in the prostate cancer 'smRC characterization' cohort is 40,879 (while for the 'HCC biomarker discovery' cohort there are 229,677 smRCs), and their length ranges from 15 bp to 25 kbp in both datasets (see **Supplementary Fig. S2A** for length distribution within prostate cancer 'smRC characterization' cohort). Furthermore, as demonstrated by **Fig. 3B and Supplementary Fig. S3B**, capture of smRCs in general is relatively robust across different EV isolation technologies (nanoDLD and ultracentrifugation, UC), with over 80% of well-expressed smRCs captured by both techniques (**Supplementary Fig. 3A**).

Apart from being about as numerous as annotated genes, smRCs possess a standard overdispersed count (heteroscedastic) mean-variance profile across all samples, as shown in **Supplementary Fig. 2B** for the prostate cancer 'smRC characterization' cohort. Their variance profile tracks key axes of variation in the dataset, as shown in **Supplementary Fig. 2C** and in the PCA plot in **Supplementary Fig. 2D**, which demonstrates the first two principal components of smRC expression demarcate clear separation across exRNA *versus* cellular RNA origin and biofluid (serum/urine), respectively, implying that their differential expression may be tractably analyzed via standard techniques employing explicit parametric negative binomial estimations (e.g., DESeq2[49]) and nonparametric transformation models (e.g., voom/limma[50]). Normalizing the counts of all samples in the prostate cancer 'smRC characterization' cohort by library size and filtering for minimum expression (cpm > 5) across at least 3 samples, i.e., imposing

$$\text{rowSums}(\text{cpm}(\text{DGE\_smRC}) > 5) > 2$$

we obtain 34,297 well-expressed smRCs, where DGE_smRC is the smRC count matrix. The flexibility and interpretability of linear modeling can then be utilized to construct well-defined contrasts in expression and test associated null hypotheses. Leveraging the matched cellular and exRNA expression smRC profiles in the prostate cancer 'smRC characterization' cohort, we tested the expression null hypothesis (between exRNA and cellular RNA isolate sample smRCs)

$$H0 := \text{exRNA} - \text{cellular} = 0$$

using the standard voom/limma workflow[50]. Any smRCs rejecting this null hypothesis with positive (negative) logFC are defined as exRNA (cellular) specific in the same patient. Of all the well-expressed smRCs, 25,771 (14,458 cellular, 11,313 exRNA) had a BH-adjusted p-value (FDR)

< 0.05 to reject the null hypothesis as shown in the volcano plot of **Fig. 3C**. Plotting the tiling complexity and peakiness of these significant exRNA- and cellular-specific smRCs, and annotating by logFC, the striking difference between exRNA and cellular smRCs becomes manifest. exRNA-associated (cellular-associated) smRCs have low (high) complexity and dominant peaks. Furthermore, examining the maximum value logFC for all significant smRCs as a function of the length of the smRC peak consensus sequence in **Supplementary Fig. 2E**, we see that smRCs with shorter peak consensus sequences (smRC length < 26 bp) tend to be exRNA-derived while those with longer sequences (26 bp < smRC length < 97 bp) are preferentially expressed in cellular smRCs. Similarly, **Supplementary Fig. 2F** demonstrates that shorter smRCs, which tend to be better expressed in exRNA compared to cellular samples, also tend to rely more heavily on multimapping (XY:Z:P and XY:Z:R) reads. We therefore define exRNA-associated smRCs

exRNA_smRC := (chr:start-end; log2CPM, complexity ~ 0, short peak_consensus_sequence,…)

as the subset of well-expressed smRCs that are short, peaky, relatively low tiling complexity, and have relatively higher average reliance on multi-mapped reads.

HCC smRC biomarker selection

Using exactly the same methodology as outlined in **smRC definition and properties** in the supplementary material, we computed HCC-specific smRCs from the 'HCC biomarker discovery' cohort by testing the null hypothesis

H1 := early HCC – CLD = 0

and adjusting for etiology, age, gender, and sequencing batch. Only 269 smRCs (41 upregulated, 228 downregulated) were significantly associated (FDR < 0.05) with HCC. We focused on well-expressed smRCs which were over-expressed in HCC compared to chronic liver disease, and applied the following filtering criteria using the intuition of the prostate cancer 'smRC characterization' cohort:

a) Filter to the top 90th percentile of positive logFC for significant smRCs, which is 2.51, which leaves (29/41);
b) Order the remaining smRCs in descending order average expression (log2cpm) across all samples;

c) Remove all smRCs with complexity higher than mean of candidates, all those with larger size than the mean, and all those with higher than median smRC consensus sequence length; (6/41)
d) Remove the smRCs whose peak consensus sequences have significant overlap with repeat elements (LINEs, SINEs, low complexity sequence, etc), in order to aid RT-qPCR validation; (5/41)
e) Select the top 4 smRCs;

Supplementary Table S5 displays the relevant information from the top four selected smRCs. Subsequent RT-qPCR validation revealed that smRC_125851 had relatively poor discriminatory power between HCC and CLD, so it was removed. The remaining three were profiled via RT-qPCR in the early 'HCC biomarker validation' cohort and subsequently used to create an early HCC risk function using penalized logistic regression.

Reverse transcriptase quantitative polymerase chain reaction (RT-qPCR)

We designed custom TaqMan® Small RNA Assays to target our candidate smRCs (ThermoFisher, **Supplementary Table S4+S5)** and purchased a catalog TaqMan® miRNA Assay against cel-miR-39-3p (ThermoFisher) to target the spike-in miRNA mimic which was used during the exRNA extraction for normalization purposes. Three µl of extracted exRNA were used for reverse transcription (RT) to cDNA with the conventional TaqMan™ MicroRNA Reverse Transcription Kit (ThermoFisher) and target-specific RT primers, followed by quantitative real-time PCR according to the manufacturer's protocol. For our 3-smRC signature, raw ct values of smRCs were corrected against ct values of the spike-in (ΔCt) and normalized to the average ΔCt of all controls (ΔΔCt). Overall, the turnaround time from blood sampling to final test results can be achieved in less than 12 hours.

smRC overlap with known RNA biotypes

We next investigated if well-expressed exRNA and cellular smRCs preferentially capture (enclose) any key known RNA biotypes, as we would expect with both exRNA and cellular smRCs for miRNA for example, and to what extent they do so across all key biotypes. Indeed, for a specific RNA biotype we first computed the smRC capture percentage (i.e., whether or not the smRC completely or only partially enclosed the RNA biotype). Then, for a particular smRC capture percentage, we asked how frequent a particular RNA biotype was among all biotypes. **Supplementary Fig. S4A** shows the relative breakdown of RNA biotypes at several extremal

points of the smRC capture percentage (1%, 70%, 100%), where plainly miRNA, snoRNA, snRNA, and other small RNA are preferentially *completely* captured (i.e., they are the dominant RNA biotypes with capture overlap ~ 1) by smRCs compared to mRNA, which are dominantly *grazed* (i.e., protein coding biotype is dominant for capture overlap << 1). In other words, as expected, when a smRC completely or mostly encloses a known RNA biotype, it is mostly likely a small RNA and very unlikely a protein-coding RNA. Indeed, plotting the RNA biotype frequency across all exRNA and cellular smRC capture overlap percentages separately for mRNA, lincRNA, miRNA, and snoRNA, yields **Supplementary Fig. S4B**. We find that exRNA smRCs dominantly *partially capture (graze)* mRNA at most to about 25% of the mRNA transcript, and never capture more, while cellular smRCs tend to overlap more protein-coding mRNA and can actually completely enclose mRNA. Similarly, using **Supplementary Fig. S4B**, one can conclude miRNA are preferentially completely enclosed by both exRNAEV and cellular smRCs at the same rate, at most 50% of a lncRNA is captured by an exRNA smRC, and snoRNAs are preferentially completely enclosed by cellular smRCs compared to exRNA smRCs. Taken together, when smRCs do enclose known RNA biotypes they can either do so predominantly partially (as with mRNA and lncRNA) or predominantly completely (as with miRNA, snoRNA, and other small RNA), with key differences in the statistics observed between exRNA and cellular smRCs. Finally, one can ask if these overlap properties are principally driven by the number and relative size distributions of exRNA and cellular smRCs (as opposed to a genuine property of small RNA accumulation in exRNA and cells). Randomly generating genomic regions with the same number of regions, and exRNA and cellular smRC size distributions (masking for repeat regions and centromeres), we repeat the above overlap computations and use a Kolmogorov-Smirnov test to assess if the underlying distributions of overlaps and capture percentages are the same within sampling noise. It turns out that all pairwise (x = smRC, y = random) Kolmogorov-Smirnov tests with two sided alternative hypotheses are highly significant, especially for lncRNAs, indicating that the exRNA and RNA biotype specific overlap patterns are not solely attributable to the size distributions (or number) of smRCs. As **Supplementary Table S3** illustrate for the two separate one-sided KS tests, interesting trends emerge: for mRNA, both exRNA and cellular smRCs tend to overlap more exons than expected by random simulation; for lncRNA, exRNA smRCs overlap than expected more while cellular smRCs overlap much less; for miRNA, both exRNA and cellular smRCs overlap far more than expected by chance; for snoRNA, cellular smRCs overlap far more than expected while exRNA smRCs have slightly more evidence for relative depletion.

In summary, exRNA smRCs overlap known RNA biotypes in a non-random fashion, and when they completely or almost completely enclose a biotype it is overwhelmingly likely to be a

known *small* RNA biotype, as opposed to similar but distinct trends for cellular smRCs. To aid in interpretation and comparison, **Supplementary Fig. S4C** also includes the simulated fractional overlap curves. However, as **Supplementary Fig. S4A** demonstrates a significant fraction of exRNA smRCs are well-expressed from unannotated genomic regions.

Prostate smRC consensus sequence motifs

In the absence of functional data, we speculate that like other small RNA, exRNA small RNA payloads are in complex with RNA binding proteins (RBPs), or may bear vestigial evidence of exRNA related packing by RBPs. Using MEME[51], we investigated if the exRNA smRC peak consensus sequences had any evidence of being enriched in ungapped motif sequences that in turn had homology to known RBP motifs. Parsimoniously, we assumed that each peak sequence contains at most **one** occurrence of a motif, but likely none. We also assumed that if nucleotide frequency biases exist there would be only single-nucleotide biases (as opposed to dimer biases such as GC content, or even higher order biases), and only searched for motifs between 3 and 6 nucleotides long, rejecting all those that had a sufficiently high E-value (probability of being found randomly). This amounts to running an instance of a zeroth order Hidden Markov Model in the zoops (zero or one per sequence) setting of MEME on a fasta file of exRNA-specific smRCs, which we took to be those with positive logFC in the null H0 and FDR < 0.001:

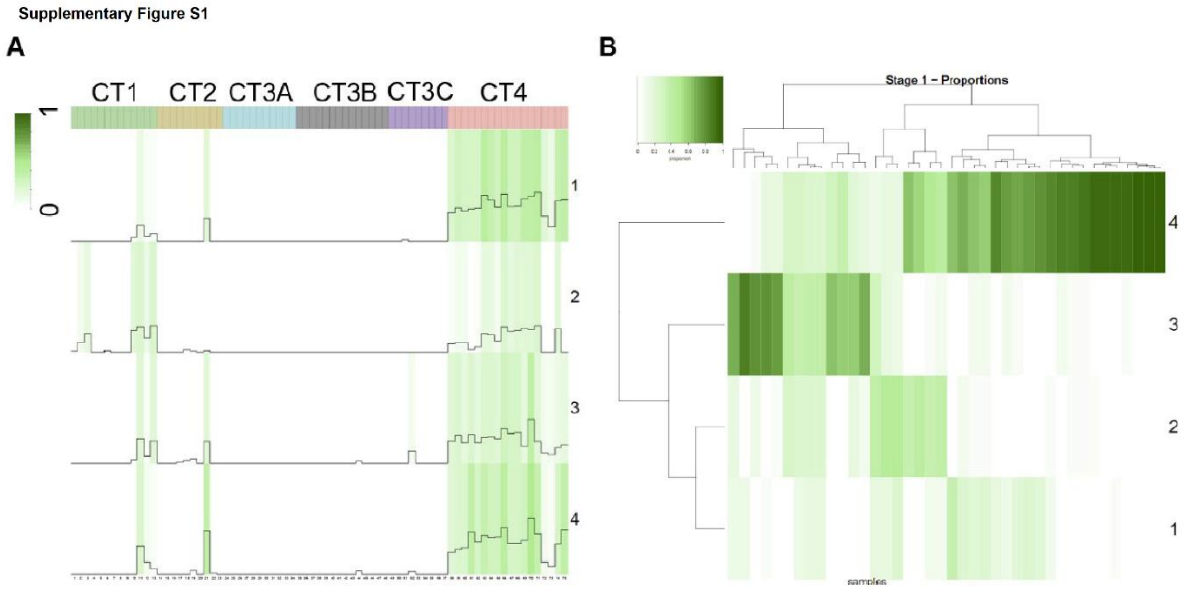
```
meme exRNA_smRC_peaks.fasta -brief 100000 -rna -oc exRNA_smRC_output -nostatus -evt .001
-mod zoops -nmotifs 10 -minw 3 maxw 6 -objfun classic -markov_order 0 1> stdout 2> stderr
```

The final results of the MEME computation are summarized here. Briefly, two 6 nucleotide motifs were found significantly over-enriched in two distinct groups of exRNA smRC peak consensus sequences, each representing approximately 11% of the total number of exRNA smRC peak consensus sequences interrogated. The motifs YCCACC (617 smRC peaks, RBP binding prediction: PCBP1, G3BP1, HNPRL, YBX1, ELAV1, E-value ~ 1e-46) and KKGAAR (626 smRC peaks, RBP binding prediction: ESRP2, HNRPRF, HNRPH1-3, SRSF1, E-value ~ 1e-8) were submitted for RBP motif homology assessment using ATtRACT[52] (https://attract.cnic.es/searchmotif) and RBPDB[53] (http://rbpdb.ccbr.utoronto.ca/).

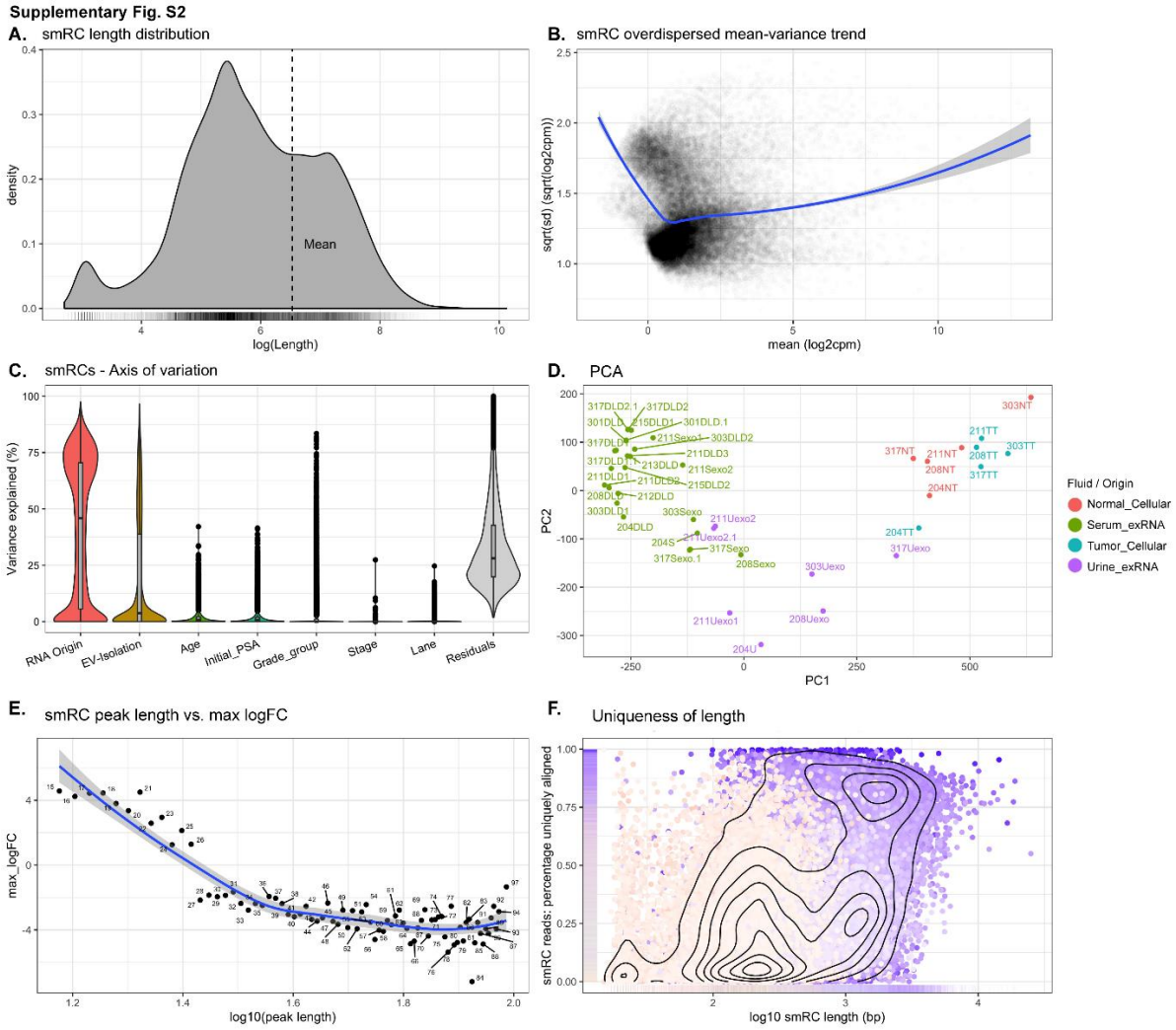
Examining the expression profile of exRNA smRCs enriched in either of these motifs across the prostate cancer 'smRC characterization' cohort reveals over-expression in exRNA compared to cellular smRCs (true by definition), for example in **Supplementary Fig. S8**, but interesting sub-patterns emerge. These include a bimodal downregulated expression of motif-enriched cellular smRCs, suggesting an enriched subset that might imply a role in exRNA packing

within cells, and an overall upregulation in nanoDLD isolation compared to UC within serum (and also overall compared to urine UC).

SUPPLEMENTARY FIGURES

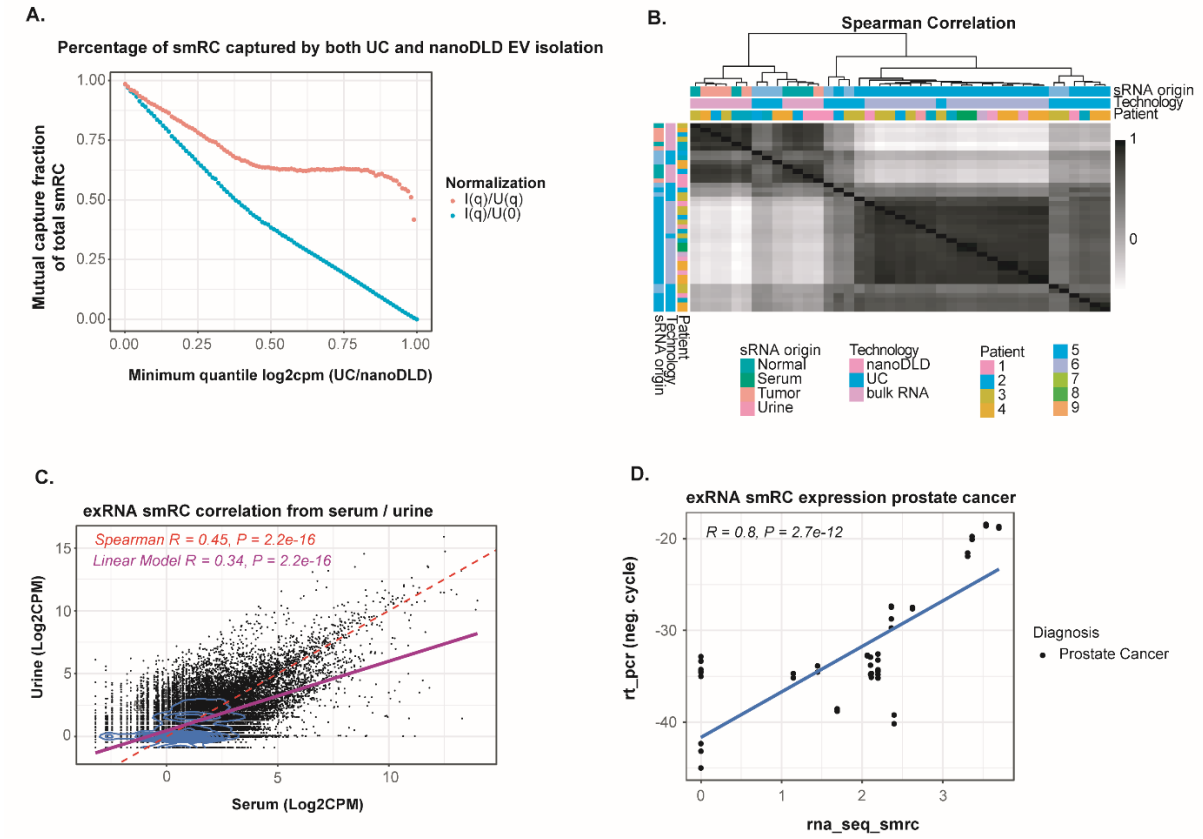


Supplementary Fig. S1. Annotated exRNA expression in cargo profiles. *(**A**) Heatmap of correlations of estimated constituent cargo profiles with our exRNA expression profiles. exRNA expression in units of normalized expression is correlated with key RNA species distinguishing the 6 cargo types (CTs, columns) previously identified[4]. CT4 is heavily enriched, i.e. ncRNA profiles 58-75 are heavily enriched, indicating highly EV specific origin of exRNA. (**B**) Heatmap of per-sample proportions of estimated constituent cargo profiles (columns) among the 6 cargo types (CTs).*



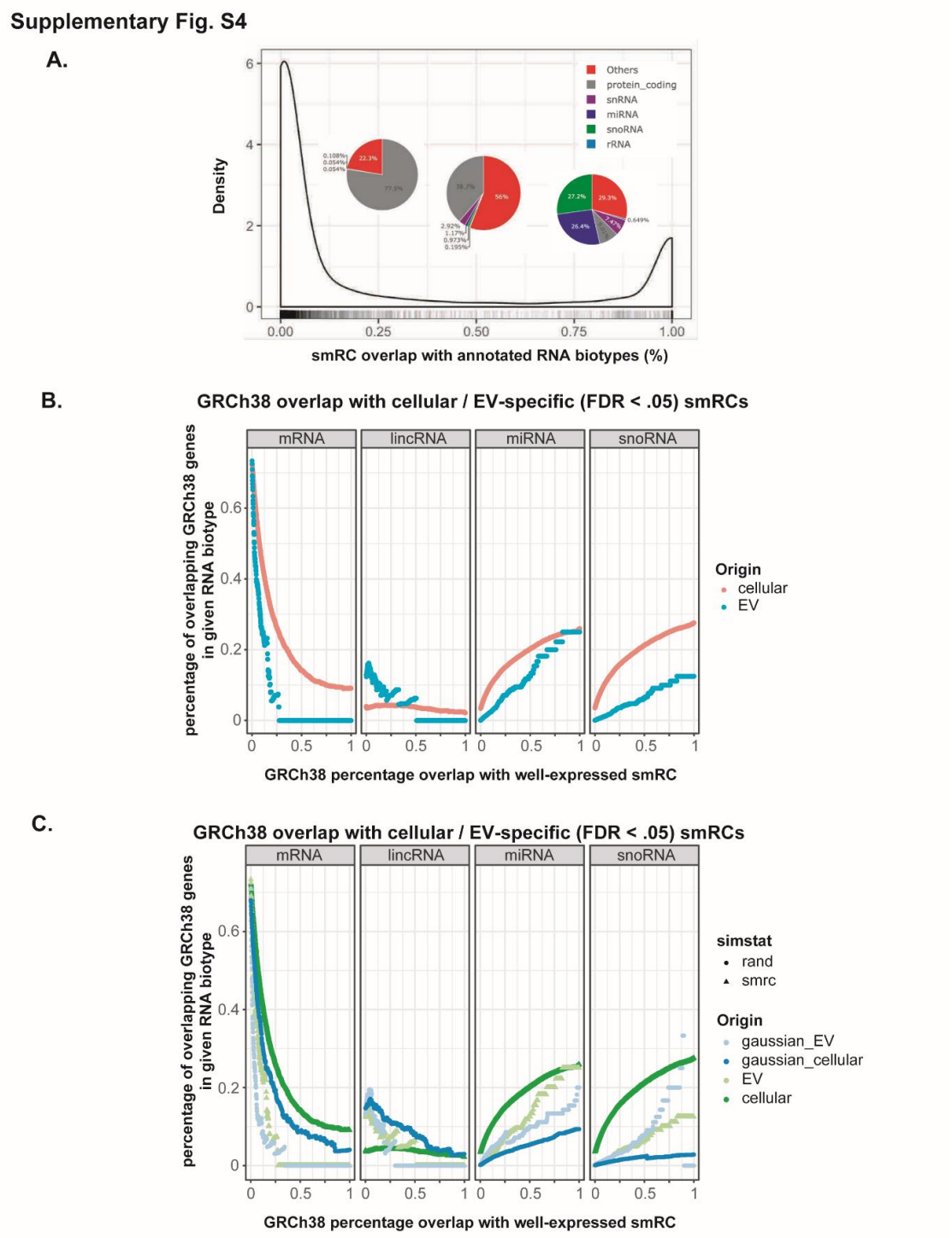
Supplementary Fig. S2. smRC properties of prostate cancer 'smRC characterization cohort'. *(**A**) Density plot of smRC length. (**B**) Mean-variance profile across all samples. (**C**) smRC axis of variation. The relative contribution of each axis of expression variation is displayed across the training prostate cancer dataset in order of magnitude. RNA origin (i.e., EV-derived or cellular) contributed the most to the observed variance. (**D**) Principal component analysis (PCA). (**E**) Maximum value logFC among all significant smRCs as a function of the length of the smRC peak consensus sequence. (**F**) Mapping uniqueness of smRC as a function of the length.*

Supplementary Fig. S3



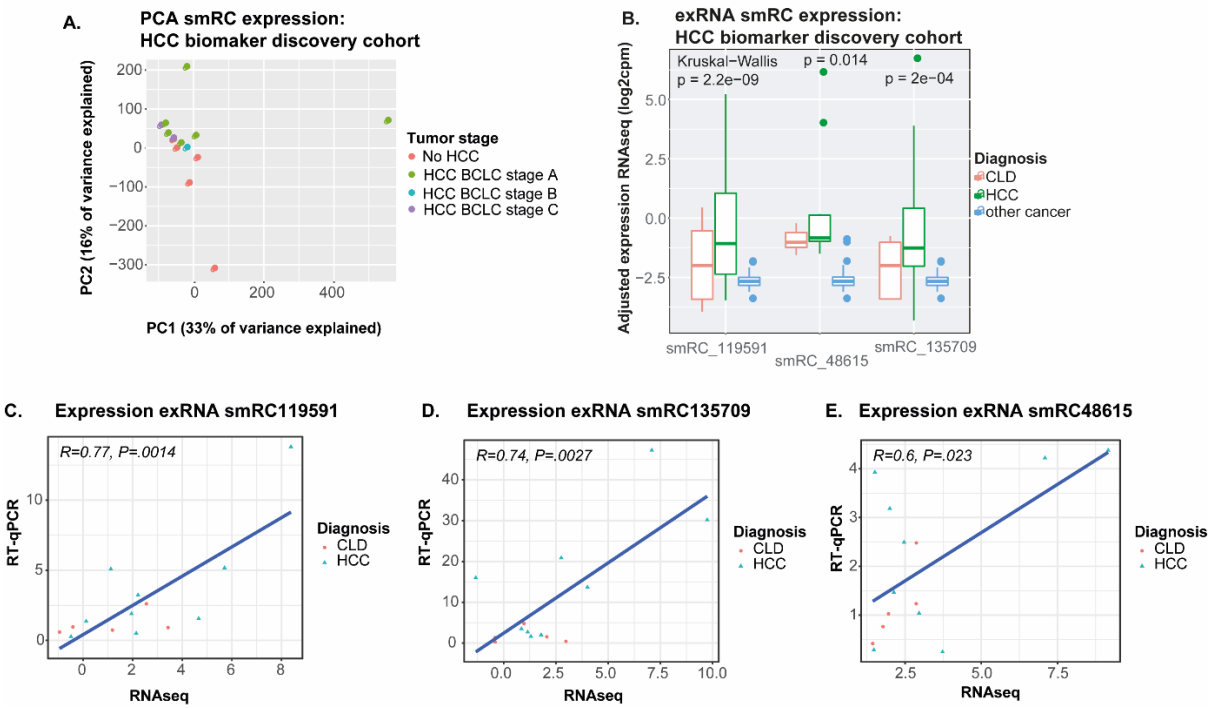
Supplementary Fig. S3. smRC correlation properties across different biofluids and technologies. *(**A**) Percentage of smRC captured by both UC and nanoDLD EV isolation. (**B**) Correlation plot across prostate cancer samples. (**C**) Correlation plot for EV-derived smRC expression across different biofluids (i.e., serum versus urine) using UC. (**D**) Correlation of single smRC expression between RNAseq and RT-PCR in the prostate cancer cohort.*

Supplementary Fig. S4



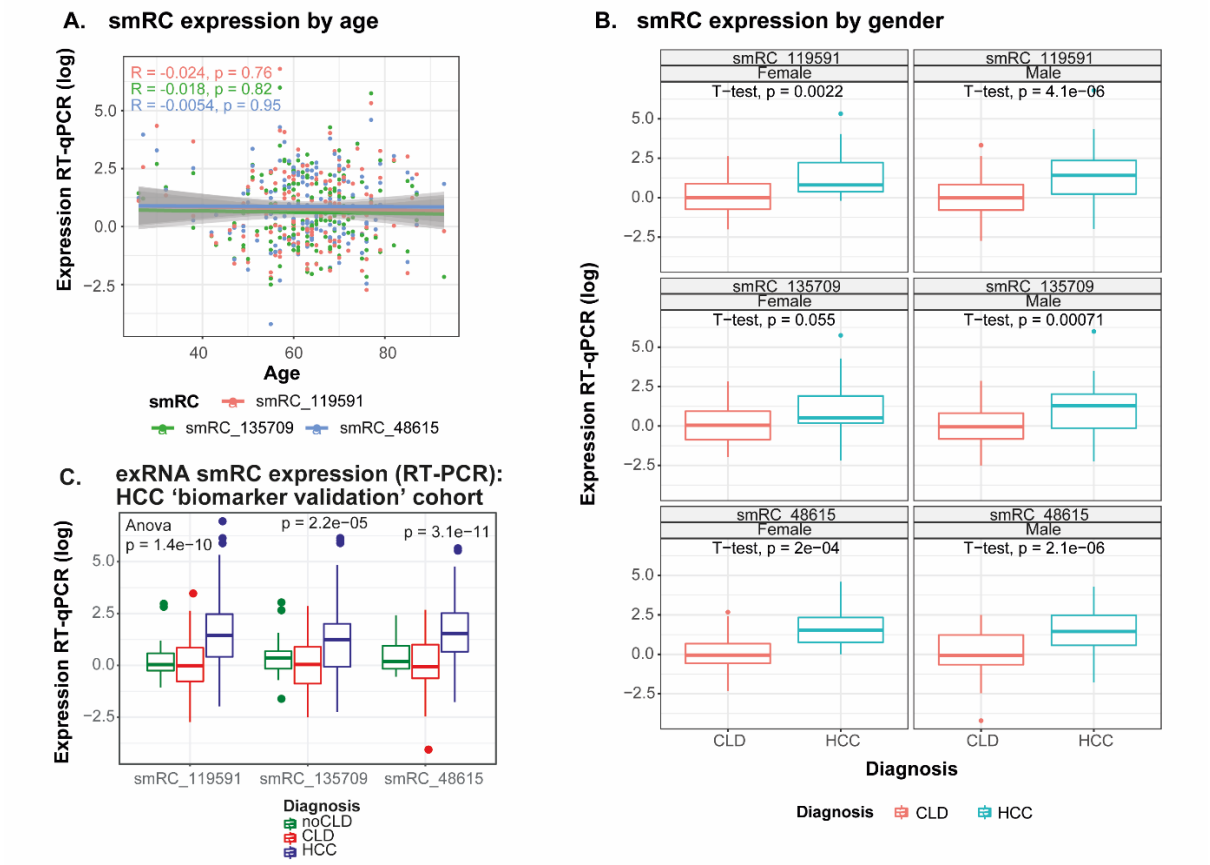
Supplementary Fig. S4. Level of smRC overlap with annotated hg38 biotypes. *(**A**) Distribution of percentage overlap of smRCs onto all known hg38 RNA biotypes. Low overlap (<<1) indicates smRC does not contain whole RNA bioptype, high or total overlap (~1) indicates RNA biotype contained within smRC. (**B**) Plot of given RNA biotype abundance percentage (among all RNA biotypes in hg38 annotation) versus smRC overlap percentage as above. Abundance percentage quantifies the frequency of a given RNA biotype among all others. (**C**) Same as (B), only with curves derived from a random genomic distribution matching number and size of smRCs.*

Supplementary Fig. S5



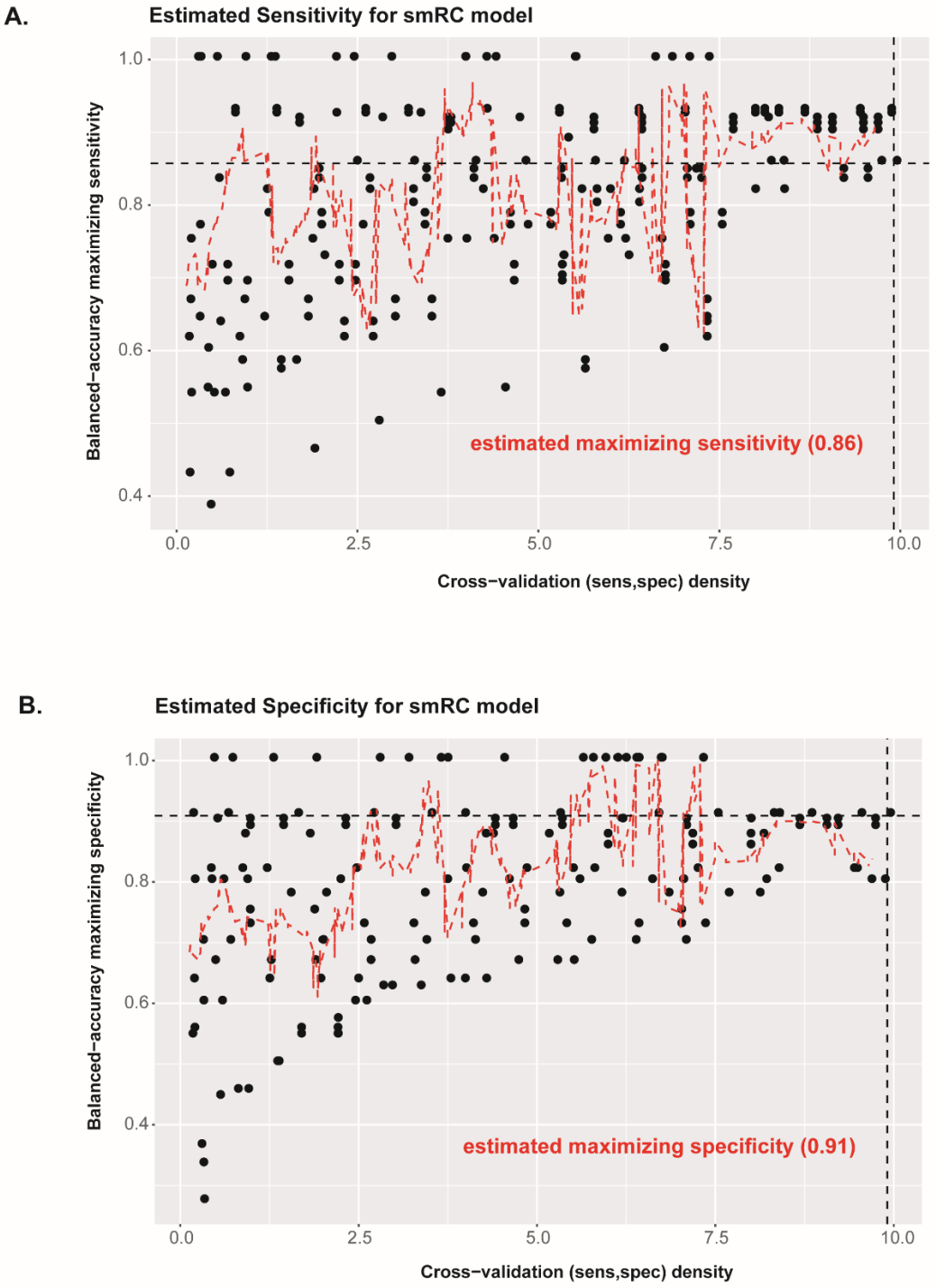
Supplementary Fig. S5. smRC expression in 'HCC biomarker discovery' cohort. *(**A**) Principal component analysis (PCA) for HCC biomarker discovery cohort. (**B**) Expression for each smRC between chronic liver disease controls (CLD, n=5), HCC patients (n=10), and patients with other non-HCC malignancies (n=142). (**C-E**) Correlation of 3-smRC-signature expression between RNAseq and RT-PCR in the HCC discovery cohort.*

Supplementary Fig. S6



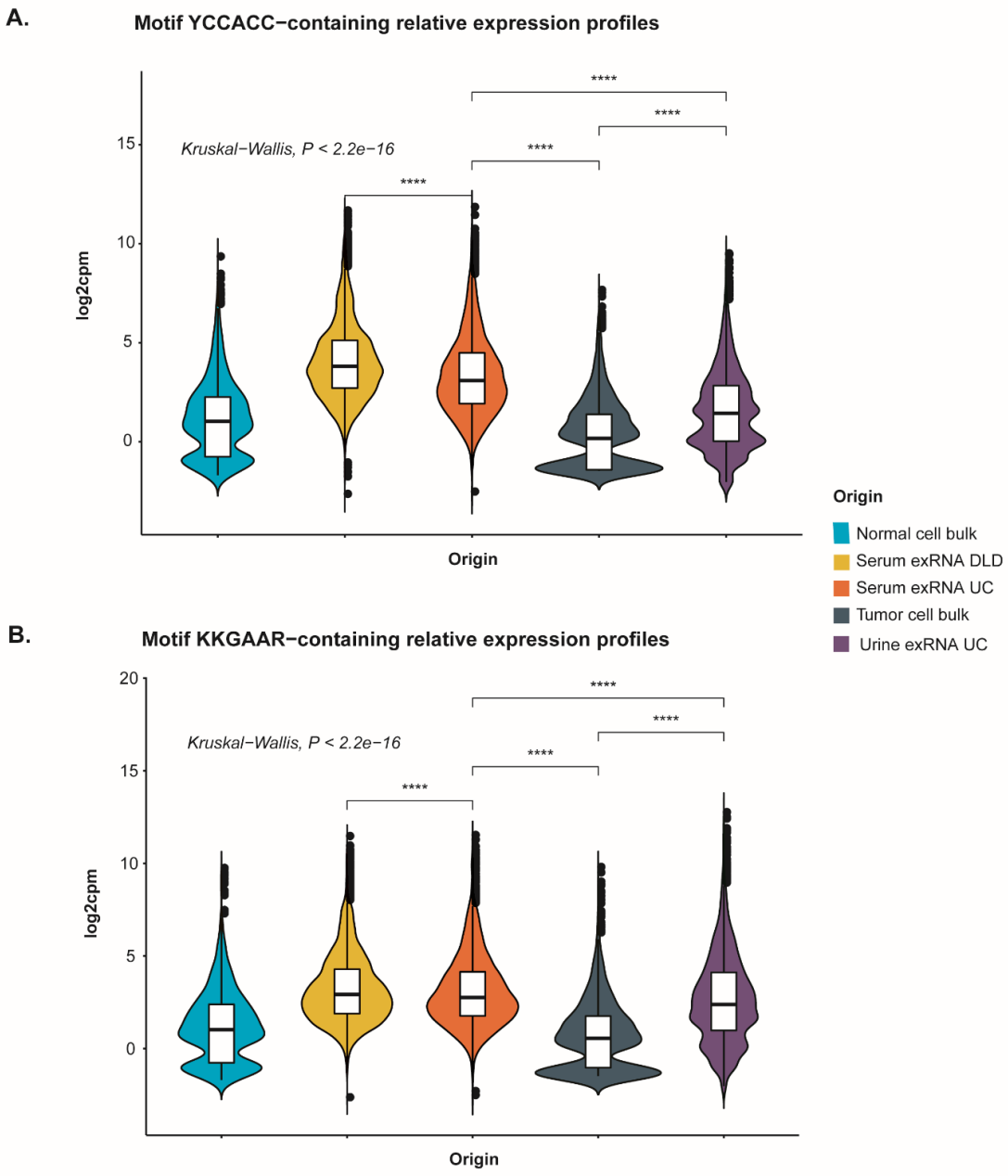
Supplementary Fig. S6. smRC expression in 'HCC biomarker validation' cohort and correlation with clinical variables. *Correlation of smRC expression with age (**A**) and gender (**B**). (**C**) Expression for each smRC between HCC patients (n=105), chronic liver disease controls (CLD, n=85), and patients without chronic liver disease (noCLD, n=19) (RT-qPCR data).*

Supplementary Fig. S7



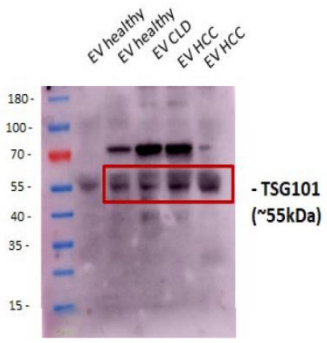
Supplementary Fig. S7. Sensitivity and specificity of 3-smRC model. *Balanced accuracy−maximizing sensitivity* ***(A)*** *and specificity* ***(B)****, respectively, versus kernel density estimation of all [sens, spec] simulation pairs (with n = 30 moving average) for the smRC model to discriminate early stage HCC from controls at high risk.*

Supplementary Fig. S8



Supplementary Fig. S8. Motif containing smRC expression in prostate cancer 'smRC characterization' cohort across RNA origin. *Expression profile of exRNA smRCs enriched in either of the motifs (**A**, YCCACC, **B**, KKGAAR) reveals over-expression in exRNA compared to cellular smRCs (true by definition) with a bimodal downregulated expression of motif-enriched cellular smRCs.*

Supplementary Fig. S9



Supplementary Fig. S9. Complete image of Western Blotting analysis targeting TSG101. *The section included in the manuscript is highlighted in red.*

SUPPLEMENTARY TABLES

Supplementary Table S1. Clinical characteristics of discovery cohort for HCC patients and controls.

	Early Stage HCC (n=10)	CLD, risk for HCC (n=5)	P-Value
Age (Years)	67	63	0.94
Sex (Male)	7 (70%)	3 (60%)	1
Cirrhosis (Yes)	8 (80%)	4 (80%)	1
Etiology			
HCV	4 (40%)	2 (40%)	1
HBV	3 (30%)	1 (20%)	1
NASH	3 (30%)	2 (20%)	1
Tumor stage (BCLC)			
Early Stage (Stage A)	6 (60%)	n.a.	n.a.
Intermediate Stage (BCLC B)	2 (20%)	n.a.	n.a.
Advanced Stage (BCLC C)	2 (20%)	n.a.	n.a.
Largest nodule (cm)	3.5	n.a.	n.a.
AFP (ng/mL*)	20.6	4.6	0.46

Continuous variables are displayed as median. *Upper limit of normal 9ng/mL. AFP, alpha fetoprotein, BCLC, Barcelona Clinic for Liver Cancer, HBV/HCV, chronic hepatitis B/C, NASH, non-alcoholic steatohepatitis

Supplementary Table S2

	Bootstrap Validation of Penalized AFP + smRC model						Bootstrap Validation of Penalized AFP + smRC model					
Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n	Original Sample	Training Sample	Test	Optimism Index	Corrected	n
Dxy	0.810	0.850	0.780	0.060	0.750	1000	0.910	0.930	0.880	0.050	0.860	1000
R2	0.590	0.630	0.570	0.060	0.530	1000	0.740	0.770	0.720	0.050	0.680	1000
Intercept	0.000	0.000	0.010	−0.01	0.010	1000	0.000	0.000	−0.02	0.020	−0.02	1000
Slope	1.000	1.000	0.910	0.090	0.910	1000	1.000	1.000	0.920	0.080	0.920	1000
Emax	0.000	0.000	0.020	0.020	0.020	1000	0.000	0.000	0.020	0.020	0.020	1000
D	0.600	0.630	0.550	0.080	0.520	1000	0.830	0.850	0.760	0.090	0.740	1000
U	−0.01	−0.01	0.000	−0.01	0.000	1000	−0.01	−0.01	0.000	−0.02	0.000	1000
Q	0.610	0.640	0.540	0.100	0.510	1000	0.840	0.860	0.760	0.110	0.730	1000
B	0.130	0.110	0.140	−0.02	0.150	1000	0.080	0.070	0.100	−0.03	0.110	1000
g	2.760	3.100	2.780	0.320	2.440	1000	4.910	5.480	4.990	0.500	4.420	1000
gp	0.390	0.400	0.390	0.010	0.380	1000	0.440	0.440	0.440	0.000	0.430	1000
AUC	0.910	0.920	0.891	0.030	0.874	1000	0.955	0.967	0.942	0.024	0.930	1000

Dxy: Somers' rank correlation between the observed HCC status and predicted HCC probabilities; Emax: maximum absolute calibration error on probability scale; B: Brier score; U: unreliability index; D: discrimination; Q: quality (Q = D – U); g: Gini's mean difference of log-odds between HCC and CLD; gp: Gini's mean difference in probability scale; AUC; Area Under the Receiver Operating Curve

Supplementary Table S3.

RNA biotype	Biofluid	statistic	P-value	Alternative hypothesis
mRNA	cell	0.527	0	CDF x above CDF y
mRNA	exRNA	0.106	1.32e-05	CDF x above CDF y
lincRNA	cell	0.00	1	CDF x above CDF y
lincRNA	exRNA	0.250	0	CDF x above CDF y
miRNA	cell	0.913	0	CDF x above CDF y
miRNA	exRNA	0.393	0	CDF x above CDF y
snoRNA	cell	1.000	0	CDF x above CDF y
snoRNA	exRNA	0.158	0	CDF x above CDF y
mRNA	cell	0.008	1	CDF x below CDF y
mRNA	exRNA	0.068	.0098135	CDF x below CDF y
lincRNA	cell	0.689	0	CDF x below CDF y
lincRNA	exRNA	0.058	0.03455966	CDF x below CDF y
miRNA	cell	0	1	CDF x below CDF y
miRNA	exRNA	0.068	0.0098135	CDF x below CDF y
snoRNA	cell	0	1	CDF x below CDF y
snoRNA	exRNA	0.245	0	CDF x below CDF y

Supplementary Table S4. RT-qPCR assay sequences for orthogonal smRC validation in prostate cancer dataset

smRC	Genomic location (hg38)	Target sequence for RT-qPCR assay
prostate_1	chr8:21329709-21329879	CUAGGCCAGUGGUCUUUAUGU
prostate_2	chr2:148881489-148881928	AUAGGUUUGGUCCUAGCCUUUCUAUUAGCUCUUAGUAAGAUUACACAUGCAAGCAUCCCCAUUCCAGUGAGUUCACCCUCUAAAUCACC
prostate_3	chr2:222918489-222918711	GGGGGAAGGAGGAGAAAAUUCACAUGUAAACUUGUUC

Supplementary Table S5. RT-qPCR assay sequences of 3-smRC signature and genomic location

smRC	Genomic location (hg38)	Region length (bp)	Major peak length (bp)	Average Expression (log2cpm)	logFC	adjusted p-value	Peak consensus sequence	Target sequence for RT-qPCR assay	Included in 3-smRC signature	AUC
smRC_119591	chr8:137627017-137627182	166	15	2.260	3.258	0.01583	CCUCUUCUUAACACC	UUGUCCUCUUCUUAACACC	Yes	0.75
smRC_125851	chr9:95513777-95515830	2054	15	1.081	3.554	0.00951	CCCCUUAUUUACCCC	UUUCCUCCCCUUAUUUACCCC	No	NA
smRC_135709	chr10:70817194-70818087	894	15	2.328	3.498	0.00418	CCUUCCCGUACUACC	CUCCCUUCCCGUACUACC	Yes	0.68
smRC_48615	chr3:103950043-103953627	3585	15	3.102	2.513	0.03956	CUCUUUACAGUGACC	UGUCUCUUUACAGUGACC	Yes	0.78

Supplementary Table S6

Read mapping status	PrCA smRC characterization cohort (reads)	HCC smRC biomarker discovery cohort (reads)
Total (post cutadapt)	494 828 430	409 592 240
Unmapped	213 988 996 (43.2%)	138 477 962 (33.8%)
Unmapped because m > 50	10 370 278 (2.1%)	11 205 827 (2.7%)
Unmapped because guidance failed	38 714 298 (7.8%)	7 066 592 (1.7%)
Uniquely mapped (U)	86 369 038 (17.5%)	203 895 810 (49.8%)
Multiply mapped (m < 3, R)	2 136 572 (0.4%)	2 072 756 (0.4%)
Multiply mapped with u-rescue (P)	143 249 248 (28.9%)	46 873 293 (11.4%)
Primary alignments (U + R + P)	**229 618 286 (46.4%)**	**250 769 103 (61.2%)**

References

4 Murillo OD, Thistlethwaite W, Rozowsky J, *et al.* exRNA Atlas Analysis Reveals Distinct Extracellular RNA Cargo Types and Their Carriers Present across Human Biofluids. *Cell* 2019;**177**:463–77.e15.

19 Théry C, Witwer KW, Aikawa E, *et al.* Minimal information for studies of extracellular vesicles 2018 (MISEV2018): a position statement of the International Society for Extracellular Vesicles and update of the MISEV2014 guidelines. *J Extracell Vesicles* 2018;**7**:1535750.

40 Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 2011;**17**:10–2.

41 Mortazavi A, Williams BA, McCue K, *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;**5**:621–8.

42 Moxon S, Schwach F, Dalmay T, *et al.* A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics* 2008;**24**:2252–3.

43 Stocks MB, Moxon S, Mapleson D, *et al.* The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics* 2012;**28**:2059–61.

44 Johnson NR, Yeoh JM, Coruh C, *et al.* Improved Placement of Multi-mapping Small RNAs. *G3* 2016;**6**:2103–11.

45 Langmead B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* 2010;**Chapter 11**:Unit 11.7.

46 Axtell MJ. ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* 2013;**19**:740–51.

47 Rozowsky J, Kitchen RR, Park JJ, *et al.* exceRpt: A Comprehensive Analytic Platform for Extracellular RNA Profiling. *Cell Syst* 2019;**8**:352–7.e3.

48 Onuchic V, Hartmaier RJ, Boone DN, *et al.* Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types. *Cell Rep* 2016;**17**:2075–86.

49 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.

50 Law CW, Chen Y, Shi W, *et al.* voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;**15**:R29.

51 Bailey TL, Boden M, Buske FA, *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 2009;**37**:W202–8.

52 Giudice G, Sánchez-Cabo F, Torroja C, *et al.* ATtRACT-a database of RNA-binding proteins and associated motifs. *Database* 2016;**2016**. doi:10.1093/database/baw035

53 Cook KB, Kazan H, Zuberi K, *et al.* RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res* 2011;**39**:D301–8.

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	4
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	4
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	6-7
	4	Study objectives and hypotheses	6-7
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	8
Participants	**6**	Eligibility criteria	8
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	8
	8	Where and when potentially eligible participants were identified (setting, location and dates)	8
	9	Whether participants formed a consecutive, random or convenience series	8
Test methods	**10a**	Index test, in sufficient detail to allow replication	6, 9-10 & supplement
	10b	Reference standard, in sufficient detail to allow replication	n/a
	11	Rationale for choosing the reference standard (if alternatives exist)	6
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	9-10 & supplement
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	n/a
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	8-10 & supplement
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	8-10 & supplement
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	9-10 & supplement
	15	How indeterminate index test or reference standard results were handled	9-10 & supplement
	16	How missing data on the index test and reference standard were handled	9-10 & supplement
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	9-10 & supplement
	18	Intended sample size and how it was determined	9-10
RESULTS			
Participants	**19**	Flow of participants, using a diagram	12
	20	Baseline demographic and clinical characteristics of participants	25
	21a	Distribution of severity of disease in those with the target condition	15-16, 25
	21b	Distribution of alternative diagnoses in those without the target condition	15-16, 25

	22	Time interval and any clinical interventions between index test and reference standard	8
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	16-17, Fig 6
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	16-17, Fig 6
	25	Any adverse events from performing the index test or the reference standard	n/a
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	18-20
	27	Implications for practice, including the intended use and clinical role of the index test	18-20
OTHER INFORMATION			
	28	Registration number and name of registry	n/a
	29	Where the full study protocol can be accessed	n/a
	30	Sources of funding and other support; role of funders	21

STARD 2015

AIM

STARD stands for “Standards for Reporting Diagnostic accuracy studies”. This list of items was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Authors can use the list to write informative study reports. Editors and peer-reviewers can use it to evaluate whether the information has been included in manuscripts submitted for publication.

Explanation

A **diagnostic accuracy study** evaluates the ability of one or more medical tests to correctly classify study participants as having a **target condition.** This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future. A medical test can be an imaging procedure, a laboratory test, elements from history and physical examination, a combination of these, or any other method for collecting information about the current health status of a patient.

The test whose accuracy is evaluated is called **index test.** A study can evaluate the accuracy of one or more index tests. Evaluating the ability of a medical test to correctly classify patients is typically done by comparing the distribution of the index test results with those of the **reference standard**. The reference standard is the best available method for establishing the presence or absence of the target condition. An accuracy study can rely on one or more reference standards.

If test results are categorized as either positive or negative, the cross tabulation of the index test results against those of the reference standard can be used to estimate the **sensitivity** of the index test (the proportion of participants *with* the target condition who have a positive index test), and its **specificity** (the proportion *without* the target condition who have a negative index test). From this cross tabulation (sometimes referred to as the contingency or “2x2” table), several other accuracy statistics can be estimated, such as the positive and negative **predictive values** of the test. Confidence intervals around estimates of accuracy can then be calculated to quantify the statistical **precision** of the measurements.

If the index test results can take more than two values, categorization of test results as positive or negative requires a **test positivity cut-off**. When multiple such cut-offs can be defined, authors can report a receiver operating characteristic (ROC) curve which graphically represents the combination of sensitivity and specificity for each possible test positivity cut-off. The **area under the ROC curve** informs in a single numerical value about the overall diagnostic accuracy of the index test.

The **intended use** of a medical test can be diagnosis, screening, staging, monitoring, surveillance, prediction or prognosis. The **clinical role** of a test explains its position relative to existing tests in the clinical pathway. A replacement test, for example, replaces an existing test. A triage test is used before an existing test; an add-on test is used after an existing test.

Besides diagnostic accuracy, several other outcomes and statistics may be relevant in the evaluation of medical tests. Medical tests can also be used to classify patients for purposes other than diagnosis, such as staging or

prognosis. The STARD list was not explicitly developed for these other outcomes, statistics, and study types, although most STARD items would still apply.

DEVELOPMENT

This STARD list was released in 2015. The 30 items were identified by an international expert group of methodologists, researchers, and editors. The guiding principle in the development of STARD was to select items that, when reported, would help readers to judge the potential for bias in the study, to appraise the applicability of the study findings and the validity of conclusions and recommendations. The list represents an update of the first version, which was published in 2003.

More information can be found on http://www.equator-network.org/reporting-guidelines/stard.