medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Inflammatory breast cancer biomarker identification by simultaneous TGIRT-seq profiling of coding and non-coding RNAs in tumors and blood

Dennis Wylie[1*], Xiaoping Wang[2,3*], Jun Yao[1*], Hengyi Xu[1*], Elizabeth A. Ferrick-Kiddie[1], Toshiaki Iwase[2,3,5], Savitri Krishnamurthy[3,4], Naoto T. Ueno[2,3,5§], Alan M. Lambowitz[1§]

Short title: Cancer biomarker identification by TGIRT-seq

[1]Departments of Molecular Biosciences and Oncology, University of Texas at Austin, Austin, TX 78712

[2]Department of Breast Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

[3]Morgan Welch Inflammatory Breast Cancer Research Program and Clinic, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

[4]Department of Hematopathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

[5]University of Hawai‘i Cancer Center, Honolulu, HI 96813

*Co-first authors

[§]Co-corresponding authors

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

Inflammatory breast cancer (IBC) is the most aggressive and lethal breast cancer subtype, but lags in biomarker identification. Here, we used an improved Thermostable Group II Intron Reverse Transcriptase RNA sequencing (TGIRT-seq) method to simultaneously profile coding and non-coding RNAs from tumors, PBMCs, and plasma of IBC and non-IBC patients and healthy donors. Besides RNAs from known IBC-relevant genes, we identified hundreds of other overexpressed coding and non-coding RNAs ($p \leq 0.001$) in IBC tumors and PBMCs, including higher proportions with elevated intron-exon depth ratios (IDRs), likely reflecting enhanced transcription resulting in accumulation of intronic RNAs. As a consequence, differentially represented protein-coding gene RNAs in IBC plasma were largely intron RNA fragments, whereas those in healthy donor and non-IBC plasma were largely fragmented mRNAs. Potential IBC biomarkers in plasma included T-cell receptor pre-mRNA fragments traced to IBC tumors and PBMCs; intron RNA fragments correlated with high IDR genes; and LINE-1 and other retroelement RNAs that we found globally up-regulated in IBC and preferentially enriched in plasma. Our findings provide new insights into IBC and demonstrate advantages of broadly analyzing transcriptomes for biomarker identification. The RNA-seq and data analysis methods developed for this study may be broadly applicable to other diseases.

Key words: cancer, diagnostics, high-throughput RNA sequencing, hypertranscription, inflammatory response, intron RNA, LINE-1, plasma RNAs, retroelements, reverse transcriptase, RNA splicing, T cell receptor

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

High-throughput RNA sequencing (RNA-seq) is widely used to identify RNA biomarkers for diseases and environmental stimuli (1). Commonly, an initial comprehensive RNA-seq of one or more RNA biotypes is done to identify candidate biomarkers, which can then be validated by targeted assays in larger numbers of samples. Most desired are biomarkers that can be detected by minimally invasive blood-based assays and reflect molecular aberrations and immune responses throughout the body. For this purpose, relatively labile blood-based RNA biomarkers are potentially advantageous by enabling accurate diagnosis and real-time monitoring of disease progression and response to treatment. Thus far, most studies of cellular and blood-based RNA biomarkers have focused on polyadenylated mRNAs or mature miRNAs, but this approach misses numerous long non-coding RNAs (lncRNAs) and small non-coding RNAs (sncRNAs), as well intron and repeat RNAs that can be detected in blood but have not been widely explored for biomarker potential (2, 3).

Inflammatory breast cancer (IBC) is the most lethal and aggressive type of breast cancer (4). It comprises only 2-4% of breast cancer cases in the United States, but causes 8-10% of breast cancer deaths (5). Due to its high recurrence rate and propensity to metastasize, IBC patients have a shorter median overall survival time than non-inflammatory breast cancer (non-IBC) patients (6–8). Triple-negative IBC patients [TN-IBC, negative for estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor-2 (HER2)] have the worst prognosis among the receptor-based molecular subtypes (9). Although a tri-modality treatment (10), consisting of anthracycline-based neoadjuvant chemotherapy, modified radical mastectomy, and postoperative radiation therapy, has substantially improved patient survival outcomes, the median survival of patients with stage III IBC is only 4.75 years (11). FDA-approved targeted therapies specific for IBC are unavailable to date. Precise diagnosis based on molecular biology is the key to developing better treatment for IBC patients.

A number of molecular changes associated with the aggressiveness of IBC have been identifie previously, including RNA upregulation of NF-κB gene pathways (12–14); higher RNA expression levels of *COX-2 (12, 15)*, inflammatory cytokines *TNFα*, *IL-6*, and *IL-8 (16, 17)*, and the angiogenic factor *VEGF (18, 19)*; protein-level activation of JAK/STAT pathways (20, 21); and elevated protein levels of EGFR (22), E-cadherin (23), Rho-GTPase (24), and translation initiation factor eIF4G1 (25). A 79 gene signature for IBC was identified by integrated analysis of 3 distinct Affymetrix gene expression datasets (26). Additionally, a hybrid epithelial/mesenchymal

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

(E/M) or partial epithelial-to-mesenchymal transition (EMT) phenotype, including both epithelial and mesenchymal features, is often associated with metastatic IBC tumors (27). A particular challenge for IBC studies is that invasive tumors are often distributed as small foci without forming a discrete tumor mass. Thus, it is difficult to obtain high-quality core biopsies with high tumor cellularity, and the rarity of this disease limits patient sample size for optimal analyses. The historical challenges of identifying IBC-specific biomarkers for diagnosis and novel therapeutic targets using tissue-based analyses suggest that blood-based biomarker discovery is critically important for improving the clinical management and survival of IBC patients.

RNA-seq methods used for biomarker identification typically employ engineered retroviral reverse transcriptases (RTs) to convert RNAs into more readily sequenced cDNAs (1, 28). However, retroviral RTs are degenerate enzymes that have inherently low fidelity and processivity, reflecting loss of key regions used by ancestral RTs to bind template RNAs and constrain substrate binding pockets for higher fidelity (29, 30). A series of recent studies have shown that group II intron and other non-LTR-retroelement RTs, evolutionary ancestors of retroviral RTs, can out-perform retroviral RTs for RNA-seq and genome engineering applications (29, 31–36). Previously, we identified thermostable group II intron-encoded RTs (TGIRTs) from bacterial thermophiles that have higher fidelity, processivity, and strand displacement activity than retroviral RTs, as well as a proficient template-switching activity, which enables efficient seamless RNA-seq adapter addition with small amounts of RNA (31). These properties make it possible for the enzyme to give full-length end-to-end sequence reads of highly modified tRNAs and other structured sncRNAs and enabled us to develop TGIRT-seq, a fully validated RNA-seq method that enables comprehensive profiling of coding and non-coding RNAs from small amounts of starting material, including from plasma and extracellular vesicles (2, 3, 37–39).

Here, we used TGIRT-seq to identify potential IBC biomarkers by comprehensively profiling RNAs isolated from Formalin-Fixed Paraffin-Embedded (FFPE) tumor slices, PBMCs, and plasma from IBC and non-IBC patients and healthy donors. This technology identified numerous overexpressed coding and non-coding RNAs in IBC samples of all 3 tissue types, including plasma RNAs that could be traced to tumors and PBMCs as potential proxies for tumor biopsy in diagnosis and real-time monitoring of disease progression and response to treatment.

Results

RNA samples and TGIRT-seq library preparation

The samples collected at MD Anderson for this study included FFPE tumor slices, PBMCs, and

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

plasma from 10 IBC and 4 non-IBC patients plus PBMCs and plasma from 2 additional non-IBC patients and 13 healthy female donors. To supplement these samples, we purchased RNAs isolated from frozen matched tumor and healthy breast tissues from 4 other non-IBC patients (OriGene). Patient information, including hormone receptor (HR) and HER2 status, is summarized in Table S1.

To analyze the samples, we developed a new variation of the TGIRT-seq library preparation method in which RNAs extracted from cellular RNA samples were rRNA-depleted and separated into long and short fractions (200-nt cutoff; see Methods). The long RNAs were then chemically fragmented to enable optimal quantitation of mRNAs and lncRNAs (37, 40) and recombined with untreated small RNAs for TGIRT-seq library preparation. Plasma RNAs (1-2 ng of RNA from 200 μl of plasma) were used for TGIRT-seq library preparation directly without rRNA-depletion or size selection. The libraries were sequenced on an Illumina NextSeq to obtain 12.3 to 81.6 million reads per sample. Mapping statistics are summarized in the Supplemental file.

Principal component analysis

We used principal component analysis (PCA) to assess the degree to which disease state, tissue type, and mode of sample preservation contributed to TGIRT-seq-detected differences between the different samples (Fig. 1A). For all samples considered together, the first principal component (PC1) separated the plasma samples (right) from the tumor and PBMC samples (left), while the second principal component (PC2) split the PBMC samples (+, top left) from frozen tissue and FFPE tumor samples (● and ○, respectively; bottom left). These findings indicate that sample type was the major source of RNA sequencing differences, as expected.

Within each sample type grouping, the PCA analysis separated the IBC samples (red) from healthy and non-IBC samples (black and blue, respectively), cleanly for the PBMC samples and with some overlap for the breast tissue and plasma samples (Fig. 1A, left), indicating that disease state was a major source of RNA sequencing differences within each sample type. PC1 separated the non-IBC FFPE and frozen samples (blue open and closed circles, respectively), possibly reflecting disease-related differences in the relatively small number of samples and/or differences in the mode of tissue preservation. Perhaps for the latter reasons, 7 of the 10 IBC FFPE tumor samples (red) overlapped 3 of the frozen healthy breast tissue samples (black). The remaining 3 IBC tumor samples (FFPE-I04, FFPE-I07, and FFPE-I09), were outliers with lower gene detection rates than the other FFPE samples (Fig. 1A, left; Supplemental file).

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Notably, PCAs based on single categories of cellular RNA biotypes (protein-coding gene, lncRNAs, sncRNAs, and RepBase-annotated Repeat RNAs) also largely separated IBC samples (red) from the heathy and non-IBC samples (black and blue, respectively), the major exception being sncRNAs in FFPE IBC and non-IBC tumor slices (Fig. 1A, right), reflecting findings below that only a small number of sncRNAs showed significantly different expression levels between these classes. Collectively, these PCAs indicate that different categories of RNA biotypes, even Repeat RNAs, which are not commonly thought of as a rich source of RNA biomarkers, included sufficient differentially expressed (DE) RNAs to distinguish disease states albeit with lower resolution than simultaneous profiling of multiple RNA biotypes.

RNA biotypes detected by TGIRT-seq in different samples

Stacked bar graphs depicting the proportions of TGIRT-seq reads corresponding to different RNA biotypes (excluding rRNAs) for each of the 80 samples are shown in Fig. 1B. Reads mapping to protein-coding genes comprised the highest percentages of total reads in both the breast tissue and PBMC samples (31-80%), but lower percentages in plasma samples (3.0-40.6%), where cellular and mitochondrial (MT) tRNAs, sncRNAs, and Repeat RNAs, comprised higher and more variable percentages (Fig. 1B, top panel). 7SL, snoRNAs, and snRNAs comprised the highest percentage of sncRNA reads in the cellular RNA samples, with snoRNAs comprising higher percentages of the sncRNA reads in PBMC samples compared to the other sample types (Fig. 1B, middle panel). Y RNAs comprised the highest percentage of sncRNA reads in most of the plasma samples (Fig. 1B, middle panel). Significant differences in percentages in Fig. 1B were confirmed by paired Wilcoxon sign ranked tests (Supplemental file).

The protein-coding gene reads in breast tissue and PBMC samples mapped predominantly (>88%) to the annotated sense strand as expected, while the protein-coding gene reads in plasma included higher proportions mapping to the antisense strand and introns (Fig. S1), in agreement with previous TGIRT-seq analyses of plasma RNAs from healthy individuals (2, 3). Notably, the proportions of reads mapping to intron and antisense RNAs were higher in most of the IBC plasma samples than in the healthy and non-IBC plasma samples (Fig. S1), presaging findings below identifying intron and antisense RNA as potential IBC biomarkers in plasma.

Based on the percentage of total reads, cellular and MT tRNAs appeared to be upregulated relative to other RNA biotypes in IBC compared to non-IBC FFPE tumor samples (Mann-Whitney U (MWU) test p=0.0040), and even more dramatically upregulated in IBC

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

PBMCs (MWU *p*=1.0e-7, IBC vs combined healthy and non-IBC), likely reflecting PBMC activation, whereas tRNAs appeared to be downregulated in non-IBC PBMCs compared to both healthy and IBC PBMCs (MWU *p*=5.0e-5 and 8.0e-5, respectively), possibly reflecting immune suppression of PBMC responses (41). The read distribution for the plasma RNA samples was variable irrespective of disease state, with tRNAs being the most variably abundant RNA biotype (Fig. 1B, top panel). This variability may reflect different degrees of cell breakage, despite precautions in the method used for plasma preparation and/or variable degradation of more labile RNA species in clinical plasma samples, which were prepared at different times and used for multiple purposes. Although these caveats should be kept in mind, they did not prevent us from detecting reproducible IBC-specific RNAs in plasma, as described below.

RepBase-annotated repeat RNAs (denoted Repeat-element RNAs) constituted only low proportions of total reads in the cellular tumor and PBMC samples (1.6-7.0% in FFPE tumors, 0.42-1.5% in PBMCs), but much higher proportions of total reads in most of the plasma samples (up to 31%; paired Wilcoxon sign rank test *p*=6.1e-4 vs FFPE tumors and *p*=3.7e-9 vs PBMC), indicating preferential export and/or higher stability of these RNAs in plasma (Fig. 1B, top panel). Among the Repeat-element RNAs, simple repeats (mini- and micro-satellites) were more abundant in plasma than in the cellular sample types (Fig. 1B, bottom panel). Notably, retroelement RNAs, including Long Interspersed Nuclear Elements (LINEs), abundant but normally repressed retrotransposons in the human genome; Short Interspersed Nuclear Elements (SINEs), non-coding RNAs mobilized by LINE RTs; and endogenous retrovirus and other LTR-containing retroelement RNAs, comprised higher proportions of the Repeat-element RNAs in most of the IBC plasma samples than in most of the healthy or non-IBC plasma (Fig. 1B, bottom panel; MWU *p*=1.4e-5 for LINEs 8.7e-5 for SINEs, and 4.9e-4 for LTR-containing retroelements), reflecting wide-spread over-expression of these elements in IBC (see below).

TGIRT-seq differential gene expression (DE) and intron-depth ratio (IDR) analyses

Previous RNA-seq and microarray studies of protein-coding gene transcripts for both IBC and non-IBC tumor samples have focused on mature polyadenylated mRNAs or size-selected mature miRNAs (see Introduction). TGIRT-seq adds multiple dimensions to such analyses by combining better quantitation and 5' to 3' coverage of protein-coding genes, lncRNAs, and intron RNAs, with the ability to obtain full-length end-to-end reads of tRNAs and other structured sncRNAs (2, 3, 31, 37, 40). We performed genewise DE analysis by using the DESeq2 R package (42) with size

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

factors calculated based on all protein-coding gene reads in the sample in order to minimize the impact of variations in tRNA and other sncRNA composition of different sample types on differential expression results. All genes mentioned in the text are listed in the Supplemental file along with their $\log_2$-fold-changes (LFCs), *p*-values, false discovery rate (FDR) adjusted *p*-values (written adj. *p* below), and descriptions of cancer relevance from the literature if available. Unless noted otherwise, the phrase "significantly DE" indicates that the genes in question met the stringent criteria of adj. $p \leq 0.001$ and $|LFC| \geq 2$ according to DESeq2 analysis. Hallmark gene set and GO analysis of protein-coding gene transcripts was done by using the GO-MWU method to compare DE statistics for genes included or not included in a given gene set via Mann-Whitney U test (Fig. 2A). Pathways of interest from this analysis are discussed below.

IGV alignments showed that protein-coding gene transcripts detected by TGIRT-seq in cellular RNA samples could be divided into two categories: Type 1 corresponding to mature mRNAs with reads mapping selectively to exons, and Type 2 with reads distributed across exons and introns, likely derived from unspliced pre-mRNAs, if as in many cases, the reads spanned multiple introns of the same gene (example genes shown in Fig. 2B). Taking advantage of the ability of TGIRT-seq to quantitate intron RNAs, we developed a bioinformatic parameter termed Intron-exon Depth Ratio (IDR) defined as the average read depth over intronic segments compared to exonic segments, with a threshold IDR of 0.5 chosen to distinguish the Type 1 ($IDR<0.5$) from Type 2 ($IDR \geq 0.5$) gene expression patterns in this study (Fig. 2C). Protein-coding and lncRNA genes with $IDR \geq 0.5$ in IBC samples are indicated with an asterisk in the Figures and descriptions below. Although the percentages of protein-coding genes exceeding the IDR threshold were roughly equal in the IBC and non-IBC FFPE tumor sample sets (18.4 and 19.5%, respectively; Table S2A), 5 of 28 (17.9%) significantly overexpressed protein-coding genes in IBC vs non-IBC tumors (adj. $p \leq 0.001$, $LFC \geq 2$) had $IDR \geq 0.5$ in IBC, while no such overexpressed protein-coding genes were found in the reciprocal comparison of non-IBC vs IBC (Table S2B). Similarly, 120 of 194 (61.9%) significantly overexpressed protein-coding genes in IBC vs non-IBC PBMCs had $IDR \geq 0.5$ in IBC compared to 0 of 3 significantly overexpressed genes with $IDR \geq 0.5$ in non-IBC vs IBC PBMCs, with similar trends holding for much larger numbers of overexpressed genes with adj. $p \leq 0.05$ and $LFC \geq 1$ (Table S2A-C). The higher percentage of overexpressed protein-coding genes with high IDRs in IBC tumors and PBMCs suggests an imbalance between the rates of transcription and RNA splicing and/or intron RNA turnover leading to greater accumulation of unspliced pre-mRNAs and/or excised intron RNAs in IBC.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

DE analysis of mature miRNAs and longer transcripts of miRNA loci was done by first mapping reads to the miRBase miRNA hairpin reference followed by remapping to the miRBase miRNA mature reference. DE expressed miRNAs identified by TGIRT-seq included not only mature miRNAs but also longer transcripts, including previously annotated full-length pre-miRNAs (Fig. 2D). In addition to DESeq2, DE RNAs in plasma were also identified by peak calling to identify stable RNA fragments that are resistant to digestion by plasma RNases (3).

Genes were singled out for discussion and inclusion in heatmaps for different sample types according to several criteria: (1) their individual DE statistics; (2) their membership in important biological pathways or gene sets; (3) previous findings of relevance to cancer from the literature; and (4) detection in at least half of the discussed sample group.

Analysis of differential gene expression in tumors

We focus first on RNAs detected in tumors. Given the potential issues with differences in tissue preservation methods found in the PCA analysis (Fig. 1A), our primary comparison was IBC vs non-IBC FFPE tumor slices, which is a widely used comparison in the literature (12, 16, 26, 43–45). We then describe trends in other sample comparisons, including between frozen non-IBC tumors and matched frozen healthy breast tissues and between frozen and FFPE non-IBC tumors. For the IBC versus non-IBC FFPE tumor comparison, we identified 43 significantly DE genes (28 protein coding + 15 other) with adj. $p \leq 0.001$ and $|LFC| \geq 2$, with that number increasing to 637 (222 protein coding + 415 other) with adj. $p \leq 0.05$ and $|LFC| \geq 1$.

Gene set analysis performed on the BROAD Institute collection of hallmark gene sets (Fig. 2A left panel) showed upregulation of genes in the TNFα signaling via NFκB (abbreviated TNFα-via-NFκB) and epithelial-mesenchymal transition (EMT) pathways in IBC compared to non-IBC FFPE samples. An NFκB-focused pathway was identified previously as being more highly activated in IBC than non-IBC tumors (46), as was a custom-defined pathway focused on TNFα (26). Recently, the hallmark TNFα-via-NFκB pathway was found to be associated with genome-wide hypertranscription in aggressive tumors of many different cancer types (47, 48), a phenomenon potentially relevant to the prevalence of putative unspliced pre-mRNAs among DE genes in IBC tumors. Genes in BROAD hallmark gene sets E2F targets, G2M checkpoints, and MYC targets V1 were depressed in IBC FFPE compared to non-IBC FFPE tumor samples (Fig. 2A, left panel).

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

GO-MWU analysis performed on the GO collection of biological process gene sets (Fig. 2A right panel) identified the GO terms system process (GO:0003008) and biological adhesion (GO:0022610) as being enriched in genes with elevated expression in IBC relative to non-IBC FFPE tumor samples, and DNA metabolic process (GO:0006259) as being enriched in genes with depressed expression in IBC compared to non-IBC FFPE tumor samples. The gene set analyzed for the GO term biological adhesion shares 59 genes in common with the 194 genes analyzed from the BROAD hallmark EMT gene set enriched in IBC tumors.

Genes in the EMT pathway that were significantly overexpressed in IBC vs non-IBC FFPE tumors included *JUN*, *CTGF* (aka. *CCN2*), *CYR61* (aka. *CCN1*), *FOXC2*, *GEM*, and *TNFAIP3* (heat map Fig. 3, volcano plots Fig. S2, IGV alignments Fig. S3A). Several of these genes—*JUN*, *CYR61*, *GEM*, and *TNFAIP3*—are also members of the TNFα-via-NFκB pathway. Other genes in the TNFα-via-NFκB set that were significantly overexpressed in IBC included *DUSP1**, *EGR1**, *F3*, *FOS*, *PER1*, *NR4A1*, and *KLF9*. *FOS* and *JUN* encode components of Activator Protein 1 (AP-1), a modulator of inflammatory responses in cancers and immune disorders (49).

Three of the 10 most differentially IBC-elevated (by *p*-value) hallmark TNFα-via-NFκB signaling genes — *EGR1**, *JUN* (also a member of the hallmark EMT gene set), and *DUSP1**— were also the 3 most discriminatory genes between IBC and non-IBC samples in a study using an RT-qPCR panel to compare expression of 538 genes in 36 IBC and 22 non-IBC frozen tumor samples (43). *EGR1* is a transcriptional regulator that is essential for T and B cell development and mediates an EGFR pathway-regulated immunosuppressive tumor microenvironment in IBC, while *DUSP1* promotes tumorigenesis in a number of cancers and regulates immune escape in cancer cells (Supplemental file). Bieche et al. suggested that activation of AP-1, *EGR1**, and *DUSP1** in IBC may be evidence of intratumoral hypoxia in IBC (43), and we found support for this conjecture in the enrichment of genes with elevated expression in the BROAD hallmark hypoxia gene set in IBC-vs-non-IBC FFPE tumors (Fig. 2A, left panel).

Other significantly DE EMT and/or TNFα-via-NFκB genes included *PER1* consistent with research indicating a link between the disruption of circadian rhythms and the development of breast cancer; *F3* (coagulation factor III, aka tissue factor (TF)), whose expression increases in inflammatory and malignant conditions; *CTGF/CCN2* and *CYR61/CCN1*, encoding matricellular proteins involved in the regulation of many biological processes implicated in multiple types of cancer; *NR4A*, a nuclear transcription factor that is induced by inflammatory cytokines and may potentiate breast cancer cell migration and metastasis; *FOXC2*, a transcription factor involved in

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

metastasis in breast cancer; and *TNFAIP3*, which was part of a 6 gene signature for discrimination of IBC from non-IBC by RT-qPCR in a study focusing on NFκB-related genes ((14) and Supplemental file). Although not specifically included in the hallmark EMT gene set, *PIGR* (polymeric immunoglobulin receptor), an inflammatory mediator whose over-expression elicits an EMT in hepatic cell lines, also showed significantly elevated expression in IBC relative to non-IBC tumors (Fig. 3, Supplemental file).

Although protein-level activation of the JAK-STAT pathway in IBC patient samples has been reported (20, 21), we did not find much evidence for RNA-level enrichment of the hallmark IL6/JAK/STAT3 signaling gene set in genes with differential expression between IBC and non-IBC FFPE tumor samples (Fig. 2A). The only significantly overexpressed gene ($p \leq 0.001$ and LFC≥2) in this gene set in IBC tumors was *JUN* (Fig. 3), which is also a member of the TNFα-via-NFκB and EMT gene sets (see above). One JAK-STAT gene, *IL4R*, had adj. $p=0.016$ and LFC=1.2, while all other members of this gene set, as well as other JAK and STAT genes not included in this gene set had either adj. $p>0.05$ or LFC<1 for the FFPE IBC vs non-IBC comparison.

Another top DE gene, *RGS1** (regulator of G protein signaling 1) exhibited both significantly and consistently higher expression levels in IBC than in non-IBC FFPE tumor samples (Fig. 3 and Fig. S2). *RGS1* was identified as a potential biomarker for the progression of tumors in melanoma, with a recent study showing that *RGS1* expression in cytotoxic T lymphocytes (CTLs) reduces the trafficking of CTLs to tumors and is associated with shorter survival of patients with breast cancer (Supplemental file).

IBC tumors have also been reported to be enriched in a cancer stem cell (CSC) population defined by protein expression levels as ALDH+ (50) or CD44+CD24-/low (51). RNA expression levels of *ALDH1A1* were not significantly elevated in our IBC compared to non-IBC tumor samples (adj. $p=0.18$, though unadjusted $p=0.0074$), while *CD24* RNA expression levels were moderately decreased (adj. $p=0.021$; Fig. 3).

EGFR, whose overexpression independently predicts high recurrence rate and shorter survival for IBC patients, and *RARRES1* (encoding the protein TIG1), whose high expression is correlated with shorter survival times of IBC patients (22, 52), were overexpressed to different degrees in subsets of IBC patient samples, but not significantly when comparing the entire IBC sample set (Fig. 3, adj. $p=0.39$ and 0.36, respectively). Among other genes or proteins previously noted as elevated in IBC tumors (see Introduction), we found little evidence for differences in RNA

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

expression levels for *COX-2* (aka *PTGS2*), *IL6**, *IL8* (aka *CXCL8*), *CDH1* (which encodes E-cadherin), *EIF4G1*, or the various *VEGF* genes (adj. $p>0.05$ for all, Supplemental file). One Rho GTPase gene, *RHOB*, had adj. $p=0.022$ and LFC=1.7).

Other significantly overexpressed RNAs in IBC vs non-IBC FFPE tumor samples included: 7 lncRNAs, 5 corresponding to annotated mature lncRNA (AC006924.1, AC036176.3, AC044784.1, LINC02562, and SEPT7P6) and 2 having high read coverage across introns (LINC02015*, which predicts survival in glioblastoma multiform patients, and antisense RNA NFIA-AS2*, which correlates with poor prognosis in glioma; Fig. 3, Fig. S3B and Supplemental file); SNORD44, which has an antitumor effect when overexpressed in colorectal cancer cells; 5 simple repeat or satellite sequences (ATGAA)n, (ATTCC)n, (CTCT)n, (TGAGA)n, and (TCCATTCCAT)n; and LINE-1 element L1M3b, which has been reported to modulate chromatin assembly (Fig. 3, Fig. S2 and Supplemental file). Two other snoRNAs, SNORD123, which was among the key discriminative sncRNAs used in predicting eight different cancer types, and SNORA21, were just below the stringent fold-change or FDR cutoffs (SNORD123: adj. $p=0.003$, LFC = +6.8; SNORA21: adj. $p=4.3e-6$, LFC = -1.8). Despite the elevated expression levels of tRNAs in IBC vs non-IBC FFPE tumors (Fig. 1B, top panel), we found no significant differences in the relative abundance of different full-length tRNA isodecoder species, 5'- or 3' halves, or 3' tRFs between these samples at adj. $p\leq0.001$, LFC$\geq$2 (Fig. S2).

We identified one mature miRNA (mir-125b-5p) as significantly upregulated in IBC versus non-IBC FFPE tumor slices, and 5 others (let-7b-5p, mir-125a-5p, mir15b-5p, and mir-199a-5p, and mir-24-3p) as upregulated but with slightly higher adj. $p\leq0.01$ (Fig. 3 and Fig. S2). Although several of these miRNAs were found previously to have cancer relevance (Supplemental file), none had been identified previously as upregulated in IBC tumors. Among the 12 miRNAs identified previously as upregulated in IBC (53–56), we found 10 to be nominally elevated (LFCs 0.05 to 3.32) either as mature miRNAs or as fragments derived from pre-miRNA hairpins but adj. $p>0.05$, while miR-520a-5p appeared to downregulated (adj. $p= 0.75$, LFC=-0.95), and we did not detect miR-720 in our datasets (Supplemental file).

Comparisons between frozen healthy breast tissue and non-IBC tumor samples for the same set of IBC-focused genes showed mixed results, but with a trend toward lower expression of many of these genes in the frozen non-IBC tumor samples (Fig. 3). These findings suggest that the overexpression of some genes in IBC vs non-IBC FFPE tumors could reflect their down-regulation in non-IBC tumors. Exceptions included *TNFAIP3**, *GEM*, *F3*, and *SMIM3*, which had

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

similar expression levels between the paired frozen healthy breast tissue and non-IBC tumors (fold changes 0.84-1.16; LFCs = -0.25 to 0.22), and *RGS1**, which was among the 10 most significantly elevated genes in the FFPE IBC vs non-IBC comparison (LFC=3.93, adj. p=1.02e-9) and was only moderately elevated in frozen non-IBC tumors vs paired healthy breast tissue (LFC=1.02, adj. p = 0.40), indicating up-regulation in IBC tumors.

Analysis of differential gene expression in PBMCs

Our analysis of PBMCs identified 577 IBC-vs-non-IBC DE genes (212 protein coding and 365 other) at adj. $p \leq 0.001$ and $|LFC| \geq 2$, increasing to 3,380 (1,741 protein coding and 1,639 other) at adj. $p \leq 0.05$ and $|LFC| \geq 1$. Heat maps, Volcano plots, and IGV plots for notable genes in PBMC samples are shown in Fig. 4, Fig. S4, and Fig. S5, respectively.

The most DE-enriched BROAD hallmark gene set that showed elevated expression in IBC compared to both healthy and non-IBC PBMCs consisted of 144 genes downregulated in response to UV radiation (Fig. 2A, left panel). Among these genes, *NEK7* (NIMA related kinase 7), a key player in cell cycle progression required for cytokinesis, had the lowest adj. p-values for the IBC versus non-IBC and healthy comparisons (4.8e-13 and 9.7e-19 relative to non-IBC and healthy, respectively), and *SDC2** (Syndecan 2), which encodes a transmembrane (type I) heparin sulfate proteoglycan, a class of molecules involved at multiple steps in the progression of different cancers, had the largest effect size (LFCs 3.3 and 3.6 relative to non-IBC and healthy, respectively; Fig. 4A and Supplemental file).

The Broad hallmark TNFα-via-NFκB gene set was the most strongly enriched in upregulated genes in IBC compared to non-IBC PBMCs, yet showed little signs of such enrichment in the IBC-vs-healthy comparison (Fig. 2A). However, several key genes from this gene set were upregulated in IBC PBMCs compared to both non-IBC and healthy PBMCs (Fig. 4A, Fig. S4A, B, and Supplemental file). These included *NFKB1** itself (adj. p=0.0011/5.4e-8 for IBS vs non-IBC/IBC vs healthy); *SERPINB2*, a regulator of adaptive immune response (adj. p=0.00018/0.0025); *OLR1**, which leads to malignant transformation through activation of inflammation (adj. p=1.4e-6/0.028); *NAMPT**, which is overexpressed in PBMCs from patients with adult T-cell leukemia (both adj. $p \leq 0.001$); and *ATP2B1**, a plasma membrane calcium ATPase involved in the calcium signaling required for granule secretion in granulocytes (both adj. $p \leq 0.001$) (Supplemental file).

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Protein-coding gene expression was also used to estimate PBMC cell-type abundances via the CIBERSORTx algorithm (57) (Fig. 5). This analysis indicated elevation of both eosinophils and neutrophils, granulocytic white blood cells that are key players in inflammation (58), in IBC patient PBMCs (MWU p=0.0028 and 0.0010, respectively, for eosinophil and neutrophil abundances in IBC PBMC samples compared to the combined healthy and non-IBC sample sets). There was also significant evidence (p=0.010) for a reduction in resting mast cells, another class of granulocytes, in IBC PBMCs relative to the combined non-IBC patient and healthy donor PBMC sample sets.

Among the genes composing the CIBERSORTx marker panel with expression specificity for eosinophils or neutrophils, 3 exhibited higher expression levels in IBC PBMCs than either non-IBC patient or healthy donor PBMC samples: *SAMSN1*, most strongly expressed in eosinophils (57), with LFCs 2.4 and 1.5 for IBC versus non-IBC and healthy, respectively, adj. p≤0.001 for both comparisons; *VNN3*, most strongly expressed in neutrophils (57), with LFCs 2.3 and 1.2 for IBC vs non-IBC and healthy, respectively, adj. p≤0.001 for IBC vs non-IBC and 0.0029 for IBC vs healthy); and *MGAM*, also most strongly expressed in neutrophils (57), with LFCs 2.1 and 1.6 for IBC vs non-IBC and healthy, respectively, adj. p≤0.001 for IBC vs healthy, adj. p≤0.01 for IBC vs non-IBC (Fig. 4A). *BRAF*, a well-known cancer marker, which is expressed at highest levels in B cells but also at higher levels in eosinophils and neutrophils than other cell types (57), also showed signs of elevated expression in IBC PBMCs (LFCs 1.3 and 1.6 relative to non-IBC and healthy, respectively, adj. p≤0.001 for both comparisons).

As was the case for the FFPE IBC vs non-IBC comparison, we saw little evidence of enrichment the hallmark IL6/JAK/STAT3 signaling gene set among the differentially expressed genes in IBC PBMCs (Fig. 2A). None of the genes in this gene set were differentially expressed at adj. p≤0.001 and LFC≥2, but 3 (*MAP3K8*, *IL1R2*, and *CXCL1*) had adj. p≤0.01 and LFC≥1, and 2 more (*IL1R1* and *IL1B*) had adj. p≤0.05 and LFC≥1. Though not part of the hallmark IL6/JAK/STAT3 signaling gene set, one STAT gene, *STAT4*, showed an increase (LFC=1.6, adj. p=2.2e-10) in IBC relative to non-IBC samples (Supplemental file).

Forty lncRNAs were significantly upregulated in IBC PBMCs compared to healthy donor and non-IBC patient PBMCs (Fig. 4A). Ten of these significantly DE lncRNAs had a previously described relationship to cancer, including AL354793.1, identified as a biomarker for triple-negative breast cancer; LINC01192, associated with low survival of triple-negative breast cancer; and antisense NR2F2-AS1, shown to promote cancer cell proliferation in a number of different

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

cancers (Supplemental file). Notably, GCSHP3, a pseudogene that is downregulated in gastric cancer (Supplemental file), was expressed at high levels in the two HR- IBC patient samples and at moderate levels in a low borderline HR+ IBC patient sample (patient 10), but strongly downregulated in the remaining six HR+ IBC patient PBMC samples, mirroring its expression pattern in gastric cancer (Fig. 4A). Two additional lncRNAs NEAT1 and MALAT1 (NEAT2), which function in DNA repair and play roles in regulating the growth and progression of breast and other cancers (Supplemental file), were upregulated in IBC compared to healthy donor PBMC samples (LFC=1.7 and 0.57, adj. p=1.3e-6 and 0.00044, respectively), but less so or not at all in IBC compared to non-IBC patient PBMCs samples (LFC=1.0 and 0.068, adj. p=0.074 and 0.89, respectively; Fig. 4A and Supplemental file).

The proportions of cellular and mt tRNA reads were significantly higher in PMBCs from IBC patients than in healthy donors and non-IBC patients (Fig. 1B, top panel; MWU p=1.0e-7 for both cellular and mt tRNAs, IBC vs combined healthy and non-IBC samples), likely reflecting activation of immune and inflammatory responses in IBC PBMCs. This putative activation was accompanied by significant increases in the relative abundance of many tRNA isodecoder families, as well as some tRNA halves and 3'-tRFs fragments, potentially relevant to cell regulatory processes, with the 5' half of TRE-CTE notable for being significantly downregulated in IBC patient vs both healthy donor and non-IBC patient PBMCs (Fig, 4A, Volcano plots Fig. S4A and B).

Twenty-seven snoRNAs were significantly up-regulated in IBC patient compared to healthy donor and/or non-IBC patient PBMCs (Fig. 4B, Fig. S4A and B). These included SNORA11 and SNORA65, which were reported to be signatures of ameloblastoma; SNORA13, which has been related to chemoresistance of osteosarcomas; and SNORD72, which is overexpressed in hepatocellular carcinoma (Supplemental file). Four snoRNAs were significantly downregulated in both IBC and non-IBC patient PBMCs relative to healthy donor PBMCs (SNORA16A, SNORD28, SNORD41, and SNORD12C; Fig. 4B, Fig. S4A and B). SNORA105A was notable for being downregulated in IBC compared to both healthy and non-IBC PBMCs (Fig. 4B, Fig. S4A and B).

Forty mature miRNAs identified by mapping to the miRBase mature miRNA reference set and 22 miRNA transcripts identified by mapping to the miRBase hairpin reference set were significantly upregulated in IBC PBMC samples compared to both healthy and non-IBC PBMC samples (Fig. 4B, Fig. S4A and B). Significantly upregulated mature miRNAs identified by

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

mapping to the miRBase mature miRNA reference set included miR-21, whose expression levels correlate with the histological grade of IBC (59), and miR-223 and miR-142 both identified as hematopoietic miRNAs capable of increasing the rate at which progenitor cells differentiate into T cells (60) with miR-223 identified as a potential diagnostic marker and therapeutic target for inflammatory disorders (61) (Fig. 4B, Fig. S4A and B, and Supplemental file). IGV alignments showed that all the upregulated miRNA transcripts identified by mapping to the hairpin reference also included mature miRNAs that were 1-2 nt shorter or longer than annotated miRNA sequence (denoted with a superscript "m+pre" in that section of Fig. 4B).

Twelve upregulated miRNA transcripts that mapped to the miRBase hairpin reference also contained reads corresponding to longer RNAs, including both full-length pre-miRNAs and RNAs that were likely derived from pre-miRNA but shorter than the annotated hairpin sequence (both indicated by superscript "pre" in Fig. 4B, examples shown in IGV alignments in Fig. S5E). Both the pre-miRNA hairpin and mature form of miR-let-7a1 and miR-103A2, which promotes proliferation and migration of esophageal squamous cells (Supplemental file), were overexpressed in IBC PBMCs compared to both healthy donor and non-IBC PBMCs (Fig. 4B and Supplemental file). miR-650 pre-miRNA hairpin fragments were significantly elevated in both healthy and non-IBC compared to IBC PBMCs (Fig. 4B and Fig. S5E).

Finally, 16 Repeat-element RNAs were significantly upregulated in IBC compared to healthy and non-IBC PBMC, including 12 simple repeats or satellite sequences, 3 LTR elements (LTR89, MLT1-int, and PRIMA41-int) and one LINE-1 element L1PB (Fig. 4B, Fig. S4A and B).

Identification of potential IBC biomarkers in plasma

Previous TGIRT-seq analyses found that human plasma from healthy individuals contains a mixture of largely fragmented mRNAs and lncRNAs, likely reflecting digestion of accessible non-protein bound regions by plasma RNase, together with full-length mature tRNAs and other structured sncRNAs, mature miRNAs, pre-miRNAs, other transcripts of miRNA loci, and a variety of Repeat-element RNAs (2, 3). TGIRT-seq of the clinical plasma samples in this study gave datasets with RNA sequences mapping to up to 22,000 different genes encompassing all of the above categories (Supplemental file). In addition to significantly over-represented RNAs identified by DESeq2, we also used TGIRT-seq peak calling (3) to detect significantly over-represented RNA fragments (read coverage ≥10 reads at the peak maximum, q-value≤0.01). Significantly over-represented RNAs identified by DESeq2 and detected in at least half of the IBC

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

samples included 30 protein-coding gene RNAs, 6 lncRNAs, 1 sncRNA (RNY5), 39 Repeat-element RNAs, and those identified by peak calling and also detected in at least half of the IBC samples included 2 protein-coding gene RNAs, 2 lncRNAs (one intronic), and 64 Repeat-element RNAs (heat maps Fig. 6, Volcano plots Fig. S6, IGV alignments Fig. S7). The number of significantly over-represented RNAs identified in IBC plasma by DESeq2—in all comparisons, IBC vs non-IBC, IBC vs healthy, and IBC vs combined non-IBC and healthy sample sets—increased from 76 to 546 for adj. $p \leq 0.05$ and LFC≥1 (Supplemental file). A number of mature miRNAs (let-7a-5p, let-7c-5p, let-7f2-5p, miR-15a-5p, miR-1246-5p, miR-19b-3p and miR-223-3p) and partially or aberrantly processed miRNA transcripts (mir-619, mir-1273a, mir-1285-1, mir-3929, and mir-5096) detected in 5 or more IBC samples were over-represented (LFC≥2) in IBC compared to healthy and non-IBC samples but fell short of statistical significance, likely due to low read counts and variable quality of the plasma samples (Fig. 6A).

Protein-coding gene RNAs that were significantly over-represented in IBC plasma included 4 corresponding to short, discrete, largely exonic RNA fragments, possibly stabilized in plasma by bound proteins (Fig. 6A). For one gene, *DIRAS3*, a tumor suppressor gene whose normally silenced maternal allele can be re-activated to induce autophagy in breast cancer cells, the over-represented species were 72- and 45-nt RNA fragments derived from the 3' end of the coding sequence and 3' UTR, respectively (Fig. 6A and C). The other 3 over-represented exon-containing RNAs were 23- or 26-nt RNA fragments of 3 different T-cell receptor beta joining region one segments (*TRBJ1-1*, *-2*, and *-6*; Fig. 6A, B, D, and E). The *TRBJ1-6* RNA was detected in all 10 of the IBC samples (highly abundant in 9, moderately abundant in 1 and not detected in any of the healthy or non-IBC samples) (Fig. 6A). The *TRBJ1-1* RNA was abundant in 4 of the IBC samples and not detected in any of the healthy or non-IBC samples, while the *TRBJ1-2* RNA was found in all 10 of the IBC samples (highly abundant in 7 and moderately abundant in 3) and in lower abundance in one each of the healthy and non-IBC samples (Fig. 6B).

Surprisingly, the remaining 26 protein-coding gene RNAs that were significantly and consistently over-represented in IBC patient plasma corresponded to 50- to 80-nt intron RNA fragments (denoted with superscript "i" in Fig. 6A; IGV examples in Fig. 6F, G and Fig. S7A, top line). Intron RNA fragments were also the overrepresented RNA species for all 6 lncRNA genes that were significantly over-represented in IBC plasma (Fig. 6A, IGV plots in Fig. S7B). Some of the over-represented intron RNAs were highly specific for IBC plasma (*e.g.*, *PHF24* and *PLVAP*, AC117373.1), while others were sparsely present in healthy and/or non-IBC plasma (Fig. 6A).

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

For RNY5, the over-represented RNAs in IBC plasma were 5' fragments of 28-33 nt (Fig. 6H), similar to those that were found to be enriched in cancer cell extracellular vesicles and cause rapid death of human primary cells but not cancer cells, suggesting a mechanism by which cancer cells might establish a favorable environment for proliferation (62). No tRNAs, tRNA halves or 3' tRFs were found to be significantly over-represented in IBC plasma at adj. $p \leq 0.001$ and $|LFC| \geq 2$ compared to non-IBC and healthy plasma (Fig. S6A-C). A 165-nt RNA identified as over-represented in IBC plasma by peak calling mapped to an intergenic region of chromosome 7 (denoted "Intergenic" in Fig. 6B).

The 39 Repeat-element RNAs identified as significantly over-represented in IBC plasma samples by DESeq2 included 18 LINEs, 8 SINEs, 8 LTR-containing retroelements, 2 DNA transposons, and 3 satellite sequences (Fig. 6A), while the 64 Repeat-element RNA fragments identified by peak calling in IBC ranged from 151 to 478 nt and included 43 mapping to LINEs, 5 to LTR-containing retroelements, 2 to SINEs, and the remainder to low complexity sequences, which include simple repeats and satellite sequences (Fig. 6B). Notably, all the LINE-1 RNAs identified as over-represented in IBC plasma by DESeq2 corresponded to the LINE-1 sense strand, while all those identified as over-represented by peak calling corresponded to the LINE-1 antisense strand (Supplemental file).

Reciprocally, we identified numerous genes whose transcripts were significantly over-represented and present in at least half of the healthy and non-IBC plasma samples, but sparsely represented in IBC plasma samples (heat maps Fig. 7A and B, Volcano plots Fig. S6A-C, IGV examples in Fig. 7D and Fig. S7). These included 125 protein-coding gene RNAs, 1 lncRNA (*SNHG6*, suggested to promote cancer development by miRNA sponging), 8 sncRNAs (7 snoRNA and snRNA U4ATAC), the 5' half of AlaCGC and 3'-tRFs of AsnGTT and ArgTCG, 5 mature miRNAs (mir-16-1-5p, mir-30d-5p, mir-199a-5p, mir-101-3p, and mir-142-3p), 4 simple repeats (CCTCTTn, TCCTCCn, TTCTCTCn, and CTTCTCTTn), and 2 LTR retroelement RNAs (HERV1-LTRc and LTR12D) (Fig. 7A and B). Three other protein-coding gene RNAs and 1 lncRNA were found to be over-represented in non-IBC plasma compared to both healthy and IBC plasma (Fig. 7C).

In contrast to IBC plasma, where the over-represented RNAs for protein-coding genes were largely intronic, with 3 exceptions (*PTK2B*, *SNCA*, and *ZMPSTE24*), the over-represented RNAs for protein-coding genes in healthy and non-IBC plasma mapped largely to exons, indicating that they were fragmented mRNAs (examples in Fig. 7A and D and Fig. S7A, bottom two lines). This

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

stark difference between IBC and non-IBC plasma samples is illustrated by the finding that although introns comprised substantial proportions of the bases in reads mapping to protein-coding genes in all plasma samples regardless of disease state (healthy, 65%; non-IBC, 40%; and IBC, 87%), they comprised >91% of the bases in reads from over-represented protein-coding genes in IBC plasma compared to only 10 and 6% for over-represented protein-coding genes in healthy and non-IBC plasma samples, respectively (Fig. 7E).

Origin of *TRBJ1* RNAs in IBC patient plasma

The cellular origin of the 3 *TRBJ1* RNA fragments identified above as potential plasma RNA biomarkers for IBC was revealed by IGV alignments (Fig. 8A), which among other differences showed peaks corresponding to the same *TRBJ1-2* and *-6* and to a lesser extent *TRBJ1-1* and *-5* RNAs rising above a background of reads spanning these regions in IBC tumor and PBMCs. Despite the presence of RNAs mapping across the *TRBJ1* locus in the cellular samples, the reads detected in plasma were short with relatively discrete lengths, suggesting they corresponded to RNA fragments protected from plasma RNase by bound proteins (3).

Closer scrutiny revealed that all of the exonic *TRBJ1* reads from plasma samples, regardless of disease state, extended a short distance into a downstream intron, while those in the cellular samples were a mixture of exon-intron and spliced mRNA reads in which the different *TRBJ1* segments were spliced to the downstream *TRBC1* constant region (Fig. 8B and C). The spliced *TRBJ1-TRBC1* mRNA read are expected for *TRBJ1-n* segments that have been productively rearranged by VDJ recombination in T-cells. Among the cellular RNA samples, the preponderance of *TRBJ1-2* and *-6* reads that extend into the intron was highest in IBC FFPE tumors (Fig. 8B), where such *TRBJ1* reads may have originated either from unspliced pre-mRNA in tumor infiltrating T cells or from aberrant transcription of the *TRBJ1-n* locus in non-T-cells, the latter possibly reflecting wide-spread enhanced transcription in IBC tumors (Fig. 8B and C). Regardless of their origin, the *TRBJ1-2* and *TRBRJ1-6* exon-intron RNAs, which were prevalent in IBC patient plasma samples but sparse or undetectable in non-IBC patient or healthy donor plasma samples, are strong candidates for plasma RNA biomarkers for IBC.

Origin of intron RNAs fragments in IBC patient plasma

We wondered if the prevalence of significantly over-represented intron RNAs in IBC plasma samples might be related to the accumulation of unspliced pre-mRNAs and/or excised intron

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

RNAs from genes with high IDRs in IBC tumors and PBMCs (Table S2). To investigate this further, we compared IDRs for protein-coding genes between different sample types and found overall higher IDRs in PBMCs from IBC patients vs those from non-IBC patients or healthy donors with this difference magnified in the IBC plasma (Fig. 9A). Although the average abundances of individual protein-coding gene RNAs in both non-IBC and IBC plasma samples correlated with those in FFPE tumor and PBMC samples across all genes with different IDRs (Fig. 9B, different colored trend lines and dots), this correlation was higher for genes with high IDRs (≥1) in FFPE tumors and particularly in PBMCs (Fig. 9B, red dots and trend lines). The elevation of average plasma RNA levels for higher-IDR genes was most pronounced for those genes with relatively low expression levels in FFPE tumor or PBMC samples (Fig. 9B, x-axis $\log_2$-expression coordinates in the 2.5-7.5 range), with RNAs from a clear subset of high IDR genes within this range of expression levels present at much higher levels in plasma than RNAs from lower IDR genes (Fig. 9B, red points above trend lines at left side of each panel). To quantify these phenomena, we calculated Spearman correlations (r_s) between IDR and the difference in plasma vs FFPE tumor or PBMC RNA levels across all genes. We found that this correlation was significantly positive (Spearman correlation test p-value<1e-10) for all 4 comparisons in the 4 panels of Fig. 9B, but with higher correlations for IBC tumors and PBMCs (r_s =0.12 and 0.15, respectively) than for non-IBC FFPE tumors and PBMCs (r_s=0.05 for both). These findings suggest that both IDR and disease-specific differences contributed to the prevalence of intron RNA fragments in IBC plasma.

A substantial proportion of the 26 protein-coding genes identified by DESeq2 as having over-represented intron RNA fragments in IBC plasma (Fig. 6A) had higher and in some cases dramatically higher IDR values in IBC tumor and/or PBMC samples than in non-IBC or healthy donor samples (Fig. 9C). For these 26 genes we identified a total of 2,481 intron fragments (where overlapping fragments were considered one fragment), including 40 intron RNA fragments that were over-represented in IBC plasma and found in at least half of the IBC plasma samples, making them candidates for IBC biomarkers (2 examples shown in Fig. 9D; others in Fig. S8). Collectively, these finding indicate that many of the intron RNAs found in IBC plasma samples were derived from high IDR genes in IBC tumors and/or PBMCs, coupled with preferential export and/or stabilization of the intron RNA fragments in IBC plasma.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Overexpression of LINE and other retroelements elements in IBC

Finally, findings above identified LINE and other Repeat-element RNA fragments as being among the most prevalent differentially represented RNAs in IBC plasma samples compared to those from non-IBC patients or healthy donors (Fig. 6). To investigate if this might reflect global activation of different Repeat-element families in IBC, we aggregated reads mapping to multiple loci of each Repeat-element family and used scatter plots to compare the aggregated counts for those families in IBC samples to those in healthy donor and non-IBC patient samples. The scatter plots indicated overexpression of multiple Repeat-element families in IBC compared to healthy and non-IBC breast tissue and PBMC samples, including known active human LINE-1s (L1Hs), as well as and primate-specific LINE-1s (L1PA2-17), LINE-2s, SINEs, endogenous retroviruses and other LTR-containing retroelements, and to a lesser extent DNA transposons and other Repeat-element RNAs (Fig. 10A and Fig. S9). The higher abundance of LINE and other Repeat-element RNAs in IBC tumors and PBMCs was magnified in plasma, suggesting preferential enrichment of these RNAs in plasma relative to other RNA biotypes (Fig. 10A and Fig. S9). By contrast, pairwise comparisons of non-IBC and healthy donor samples showed no similar enrichment of LINE and other repeat elements RNAs in non-IBC tumors, PBMCs, or plasma (Fig. 10A and Fig. S9).

For further analysis, we analyzed reads mapping to L1Hs, an active human-specific LINE-1 element (63). Such reads, which may include those from other LINE-1 elements with similar sequences, were distributed throughout the length of L1Hs with somewhat higher density near the 3' end of L1Hs RNA, possibly reflecting contributions from truncated LINEs, and included both sense and antisense strand LINE RNA reads, with the latter predominating, particularly in plasma (Fig. 10B and C). Full-length LINE-1 antisense RNAs could result from activation of an as yet undetected LINE-1 promoter, but more likely from read through transcription from downstream genes or transcription of genes in which LINE-1s had inserted within introns in the antisense orientation. The ratios of LINE-1 elements inserted in the antisense and sense orientations in human cell lines was found to 1.13 and 1.8 in different studies (64, 65). Notably, the ratios of antisense to sense strand reads mapping to L1Hs in plasma samples were substantially higher than these values or the ratios of antisense to sense strand reads mapping to L1Hs in the cellular RNA samples (Fig. 10C), indicating preferential enrichment of LINE-1 antisense RNAs in plasma.

To explore the origin of L1Hs sense and antisense strand reads, we examined the status of the *TP53* gene, a known LINE-1 repressor in the IBC patient samples, and found that the abundance

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

of L1Hs sense-strand reads in IBC tumors and PBMCs (Fig. 10C, top panel) was correlated with putative *TP53* somatic mutations and/or lower levels of *TP53* RNA in IBC tumors (Fig. 10C bottom panel; see Fig. S10 for mapping of putative *TP53* mutations) (66, 67). By contrast, the abundance of L1Hs antisense reads in IBC tumors and PBMCs did not show a similar correlation, suggesting that they arose via transcription of host or downstream genes in IBC, potentially enhancing their utility as blood-based biomarkers (Fig. 10C). In non-IBC tumors and PBMCs, the levels of L1Hs sense-strand reads remained equal to or lower than those in the corresponding frozen healthy breast tissue or FFPE IBC tumor samples, despite the presence of putative *TP53* somatic mutations in non-IBC tumor samples (Fig. 10C), likely reflecting diverse, layered mechanisms that control the overall abundance of LINE-1 RNAs in human cells (63, 68). The presence of both sense and antisense LINE-1 reads in IBC samples is consistent with the presence of LINE-1 dsRNAs, which have been suggested to induce an inflammatory response in multiple human cell types via RIG-I and MDA5 sensing and activation of interferon signaling (69), potentially highly relevant to IBC.

Discussion

Here, we used TGIRT-seq to profile protein-coding and non-coding RNAs in tumor slices, PBMCs and plasma from IBC patients compared to non-IBC patients and healthy donors, to our knowledge the first such comprehensive RNA analysis for any cancer or disease. We identified hundreds of significantly DE genes at adj. $p \leq 0.001$ and $|LFC| \geq 2$ in IBC tumors and PBMCs, increasing to thousands with adj. $p \leq 0.05$ and $|LFC| \geq 1$. The DE protein-coding genes identified as having significantly elevated RNA expression levels in IBC tumors included key members of the BROAD TNFα-via-NFκB and EMT gene sets, which were also found to be DE in IBC in previous studies (12–14, 27). By contrast, except for *JUN*, an inflammatory response modulator that is also a member of the TNFα-via-NFκB and EMT gene sets (49), we found little evidence for elevated RNA expression of genes of the Broad IL6/JAK/STAT pathway, which appears to be activated largely at the protein level in IBC (17, 43).

Overall, we found that PBMCs from IBC patients were more highly activated than those from non-IBC patients or most healthy donors, as judged by elevated levels of cellular and Mt tRNA, which could support increased rates of protein synthesis, as well as elevated levels of snRNAs, which could support increased rates of RNA splicing (Fig. 1B). Major pathways identified as being activated in IBC PBMCs were Broad TNFα-via-NFκB and UV response down, including

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

*ATP2B1**, a member of both pathways involved in granule secretion in granuloctyes (Supplemental file). The CIBERSORTx algorithm, which estimates the abundances of different cell types from gene expression data, indicated elevated abundance of granulocytic eosinophils and neutrophils in PBMCs from IBC patients compared to those from non-IBC patients and healthy donors. In addition to being associated with innate immune responses to parasitic infections, eosinophils are also known to play a role in an anti-tumor response (70). Despite the likely presence of tumor infiltrating PBMCs, we found only limited overall correlation of LFCs for significantly DE genes encoding different RNA biotypes between IBC PBMCs and tumors (Figure S11).

Taking advantage of the ability of TGIRT-seq to quantitate levels of intron RNAs, we developed a bioinformatic parameter termed Intron-exon Depth Ratio (IDR), which distinguished protein-coding genes with two different expression patterns, one with lower IDRs corresponding to mRNAs with reads mapping largely to exons, and the other with higher IDRs reflecting higher levels of reads arising from different introns of the gene (Fig. 1B). Analysis of this parameter showed that 18% and 62% of significantly overexpressed protein-coding genes (adj. $p \leq 0.001$ and $|LFC| \geq 2$) in IBC tumor and PBMC samples, respectively, had IDRs ≥ 0.5 in IBC, while no such significantly overexpressed protein-coding gene was found in non-IBC tumors or PBMCs, with similar patterns holding for much larger numbers of DE genes with lower significance and fold-change cutoffs (adj. $p \leq 0.05$ and $|LFC| \geq 1$; Table S2). The higher percentage of DE high IDR genes in IBC suggests an imbalance between the rates of transcription and RNA splicing and/or intron RNA turnover. In general, this imbalance could reflect higher rates of transcription and/or lower rates of RNA splicing or intron RNA turnover, all of which could vary in different ways for each gene or intron. In IBC tumors, this imbalance might reflect activation of genes involved in oncogenesis and cellular counter measures. In IBC PBMC samples, this imbalance likely reflects activation of immune responses resulting in higher rates of transcription of overexpressed genes, so that RNA splicing and/or intron RNA turnover become rate limiting.

The wide-spread imbalance in the rates of transcription and RNA splicing in IBC tumors and PBMCs may be related to hypertranscription, a phenomenon recently reported for a number of aggressive cancers (47, 48). If so, TGIRT-seq might provide a general method for identifying different degrees of hypertranscription and distinguishing it from more restricted elevated transcription of specific gene sets in cancers. Although further studies will be needed to sort this out, a practical consequence of IDR analysis was the identification of numerous protein-coding

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

and lncRNA genes with high IDRs in IBC and the identification of unspliced pre-mRNAs and intron RNA fragments from these genes as potential biomarkers in IBC plasma samples. RNA-seq methods that focus on poly(A)-containing mRNAs would not only miss such features but might also identify genes with high IDRs as under-expressed due to lower levels of spliced mRNAs.

Consistent with previous TGIRT-seq analysis of human plasma (2, 3), we found that intron RNAs contributed substantial proportion of protein-coding gene bases in plasma than in cellular RNA samples irrespective of disease state (Fig. 7E). Surprisingly, however, while most of the abundant differentially represented protein-coding gene RNAs in plasma samples from IBC patients were pre-mRNA or intron RNA fragments, most of those in plasma samples from healthy and non-IBC patients were fragmented mRNAs with reads mapping to multiple exons (Fig. 7E, examples shown in Fig. 6F and G and Fig. 7D). Further analysis revealed that this difference was related not only to the higher number of differentially expressed genes with high IDRs in IBC tumors and PBMCs, but also to disease-specific differences that favor accumulation of unspliced pre-mRNA and intron RNA fragments in IBC plasma (Fig. 9). The latter could reflect preferential breakage of diseased cells, closer proximity to blood or lymph for dispersed IBC tumors, or wide-spread enhanced transcription that saturates the splicing apparatus and triggers mechanisms for cellular export of intron RNAs in IBC. Importantly, a significant proportion of intron RNAs identified in IBC plasma samples corresponded to those present in higher abundance in IBC tumors and/or PBMCs, supporting their utility as blood-based markers with direct links to disease progression (Fig. 9). It will be of interest to investigate whether higher levels of intron RNAs and their preferential enrichment in plasma are also characteristic of other aggressive cancers.

Wide-spread enhanced transcription due to DNA hypomethylation and changes in chromatin structure likely also contribute to overexpression of LINE and other retroelements in IBC tumors and PBMCs and the enrichment of these RNAs in IBC patient plasma (Figs. 2B, 6A, and 10). LINE and other retroelements are activated in many cancers (71, 72), but to our knowledge such activation has not been reported for PBMCs. The presence of both sense and antisense reads that extend across the length of the active hominoid LINE-1 (L1Hs) in IBC tumors and PBMCs is consistent with findings suggesting that LINE-1 elements can form dsRNAs that induce an inflammatory response in human cell types via RIG-I and MDA5 sensing and activation of interferon signaling (69), potentially highly relevant to IBC. The enrichment of LINE RNAs in plasma led to the identification of multiple sense and antisense LINE RNAs as potential plasma biomarkers for IBC (Fig. 6). Antisense LINE RNAs, whose over-expression in IBC may reflect

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

enhanced transcription of downstream genes or host genes with LINEs inserted within introns in the antisense orientation, have the potential to be particularly useful as biomarkers, provided they have a unique identifiable sequence that can be mapped to a specific genomic locus.

A major goal of this study was to identify potential plasma RNA biomarkers for IBC that could be used for noninvasive liquid biopsies to differentiate IBC from non-IBC as well as monitor diseases progression and response to treatment. Among the many DE RNAs in IBC, prime candidates for such blood-based IBC biomarkers included 3 different *TRBJ1* exon-intron RNA fragments, possibly related to aberrant enhanced transcription and/or induction of a specific immune response in IBC. These 3 *TRBJ1* RNAs could be traced to tumors and PBMCs in IBC patients, respectively, supporting their utility as blood-based RNA biomarkers with direct links to disease progression. The *TRBJ1-6* pre-mRNA fragments, which we detected in all 10 IBC plasma samples but none of the healthy or non-IBC plasma samples, were likewise not detected by TGIRT-seq of plasma prepared from a healthy individual at different times (PRJNA279866), from different lots of commercial plasma pooled from multiple healthy individuals (PRJNA640428), nor from COVID-19 patient PBMCs or plasma (15 and 20 patients, respectively; unpublished data), advancing them as a prime candidate for an IBC-specific plasma RNA biomarker.

Our study has the following limitations. Although TGIRT-seq identified numerous potential RNA biomarkers for IBC in tumor, PBMC, and plasma samples, validating the clinical utility of these biomarkers for diagnostic, predictive, and/or prognostic purposes will require high-throughput targeted assays of larger numbers of samples. Our study found that intron and retroelement RNAs are enriched in plasma, but mechanisms by which this occurs require further investigation. Finally, although we identified somatic *TP53* mutations and/or decreased *TP53* RNA levels in IBC tumors as factors that contribute to activation of LINE elements in IBC, other factors that modulate this activation, suppress LINE activation despite *TP53* mutations in non-IBC tumors, and activate LINE elements in IBC PBMCs require further investigation.

Methods

Patients and samples

Ten IBC patient FFPE tumor tissue and paired PBMC and plasma samples, 4 non-IBC patient FFPE tumor tissue and paired PBMC and plasma samples, 2 additional non-IBC patient PBMC and plasma samples, and 13 healthy donor PBMC and plasma samples were obtained from the Morgan Welch Inflammatory Breast Cancer Research Program and Clinic at MD Anderson Cancer Center (MDACC). Each patient and healthy donor gave written informed consent. This study was approved by the MDACC Institutional Review Board. IBC was diagnosed by strictly following the criteria of the American Joint Committee on Cancer (73). All IBC patient tissues

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

were reviewed by a pathologist (Savitri K.) to ensure accurate diagnosis and tumor cellularity >60%. Frozen matched tumor and normal breast tissue RNAs from 4 deidentified non-IBC patients were purchased from OriGene (CR525106 and CR560146, CR520829 and CR561033). Individual patient information, including HR and HER2 status for all MD Anderson patients and provided by the supplier for the frozen commercial samples, is summarized in Table S1.

RNA preparation

Total cellular RNA was extracted from FFPE tumor slices and PBMCs at MD Anderson by using a miRVana miRNA isolation kit (ThermoFisher), shipped to UT Austin on dry ice, and stored frozen at -80 ºC until ready for use. For TGIRT-seq library preparation, the thawed tumor and PBMC RNAs (500 ng for most samples but as little as 200 ng for a few tumor samples) were incubated with Baseline-ZERO DNase (Lucigen; 2 units, 30 min at 37 ˚C) to digest DNA followed by rRNA depletion using an Illumina RiboZero Gold (Human/Mouse/Rat) kit. The rRNA-depleted RNAs were then cleaned up with a Zymo RNA clean and concentrator kit by using the manufacturer's two-fraction protocol to separate long and short RNAs. The long RNAs (>200 nt) were chemically fragmented to 70-90 nt length by using an NEBNext Magnesium RNA Fragmentation Module (New England Biolabs; 94˚C for 5 min) and cleaned up with a Zymo RNA clean and concentrator kit using a modified 8X ethanol protocol (v/v ratio 1:2:8 for RNA sample: kit RNA Binding Buffer: 100% ethanol) to minimize loss of very small RNAs (74, 75). The chemically fragmented long RNAs were then combined with the non-chemically fragmented short RNAs (<200 nt), and the reconstituted total RNAs were treated with T4 polynucleotide kinase (Lucigen; 50 U for 30 min at 37 ˚C) to remove 3' phosphates and 2',3'-cyclic phosphates, which impede TGIRT template switching, followed by a final clean-up with a Zymo RNA clean and concentrator kit using the modified 8X ethanol protocol above.

Plasma was prepared at MD Anderson from the same blood samples as PBMCs by sedimentation into a Ficoll cushion being careful to avoid cell breakage (2, 3) and shipped to UT Austin on dry ice. Plasma RNAs were isolated using a miRNeasy Serum/Plasma Advanced kit (Qiagen) with a modified lysis buffer step in which the plasma samples were incubated with proteinase K (20 mg/mL stock, 1% v/v) in the presence of SDS (0.5% stock, 1% v/v) and linear acrylamide carrier (5 mg/mL stock, 1-2 µl per sample) for 15 min at 50 ˚C in order to improve the recovery of protein-bound RNAs. After neutralization, centrifugation, and column clean up as specified in the kit protocol, the isolated RNAs were treated with T4 polynucleotide kinase (Lucigen; 50 units, 30 min at 37 ˚C) to remove 3' phosphates and with DNase I (Zymo: 30 U, 30 min at 37 ˚C) to degrade residual DNA, followed by a final clean up with a Zymo RNA clean and concentrator kit using the modified 8X ethanol protocol described above.

The Purchased frozen matched non-IBC tumor and healthy tissue RNAs (500 ng) from 4 non-IBC patients were treated with Baseline-ZERO DNase for 30 min at 37 ˚C to remove residual DNA and then processed by using the Zymo kit two-fraction protocol to obtain total RNA samples with chemically fragmented long RNAs and intact short RNAs, as described above for the FFPE tumor and PBMC samples.

For all samples, RNA concentration and quality were checked by mRNA assay with a Pico Kit on an Agilent 2100 Bioanalyzer. The recovered RNAs (2-50 ng for cellular RNAs and 0.5-2 ng for plasma RNAs) were used for TGIRT-seq library preparation.

Construction and sequencing of TGIRT-seq libraries

TGIRT-seq libraries were constructed as described (74, 75) by using TGIRT-template switching from a synthetic RNA template/DNA primer duplex to the 3' end of the target RNA for 3' RNA-seq adapter addition and a single-stranded DNA ligation to the 3' end of the cDNA using

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Thermostable 5' AppDNA/RNA Ligase (New England Biolabs) for 5' RNA-seq adapter addition. The TGIRT-template switching reaction was done with 1 μM TGIRT-III RT (InGex, LLC) for 15 min at 60 ^{o}C). The resulting cDNAs were amplified by PCR with primers that add capture sites and indices for Illumina sequence (denaturation 98°C for 5 sec, followed by 12 cycles of 98°C for 5 sec, 65°C for 10 sec, and 72°C for 10 sec) The PCR products were purified by using Agencourt AMPure XP beads (Beckman Coulter), and the libraries were sequenced on an Illumina NextSeq 500 to obtain 2 x 75-nt paired-end reads.

Read mapping

Illumina TruSeq adapters and PCR primer sequences were trimmed from the reads with Cutadapt v2.8 (76) (sequencing quality score cut-off at 20; p-value <0.01) and reads <15-nt after trimming were discarded. To minimize mismapping, we used a sequential mapping strategy. First, reads were mapped to the human mitochondrial genome (Ensembl GRCh38 Release 93) and the *Escherichia coli* genome (GeneBank: NC_000913) using HISAT2 v2.1.0 (77) with customized settings (-k 10 --rfg 1,3 --rdg 1,3 --mp 4,2 --no-mixed --no-discordant -- no-spliced-alignment) to identify reads derived from mitochondrial and *E. coli* RNAs (denoted Pass 1). Unmapped reads from Pass1 were then mapped to a collection of customized references sequences for human sncRNAs (miRNA, tRNA, Y RNA, Vault RNA, 7SL RNA, 7SK RNA) and rRNAs (2.2-kb 5S rRNA repeats from the 5S rRNA cluster on chromosome 1 (1q42, Gene- Bank: X12811) and 43-kb 45S rRNA containing 5.8S, 18S and 28S rRNAs from clusters on chromosomes 13,14,15, 21, and 22 (GeneBank: U13369), using HISAT2 with the following settings -k 20 --rdg 1,3 --rfg 1,3 --mp 2,1 --no-mixed --no-discordant --no-spliced-alignment --norc (denoted Pass 2). Unmapped reads from Pass 2 were then mapped to the human genome reference sequence (Ensembl GRCh38 Release 93) using HISAT2 with settings optimized for nonspliced mapping (-k 10 --rdg 1,3 --rfg 1,3 --mp 4,2 --no-mixed --no-discordant --no-spliced- alignment) (denoted Pass 3) followed by splice aware mapping (-k 10 --rdg 1,3 --rfg 1,3 --mp 4,2 --no-mixed --no-discordant --dta) (denoted Pass 4). For reads that mapped to multiple genomic loci with the same mapping score in passes 3 to 4, the alignment with the shortest distance between the two paired ends (*i.e.*, the shortest read span) was selected. In the case of ties (*i.e.*, reads with the same mapping score and read span), reads mapping to a chromosome were selected over reads mapping to scaffold sequences, and in other cases, the read was assigned randomly to one of the tied choices. The filtered multiply mapped reads were then combined with the uniquely mapped reads from Passes 3-4 by using SAMtools v1.10 (78).

To generate counts for different genomic features, the mapped reads were intersected with Ensembl GRCh38 Release 93 gene annotations plus the RNY5 gene and its 10 pseudogenes, which were not annotated in this release. Annotations for repeat elements were from RepBase (https://www.girinst.org/repbase/), and annotations for hairpin miRNA and mature miRNA were from miRBase (https://www.mirbase.org/). Coverage of each feature was calculated by BEDTools v2.29.2 (https://bedtools.readthedocs.io/en/latest/content/bedtools-suite.html). To avoid miscounting reads with embedded sncRNAs that were not filtered out in Pass 2 (*e.g.*, snoRNAs), reads were first intersected with sncRNA annotations, and the remaining reads were then intersected with the annotations for protein-coding gene RNAs, lincRNAs, antisense RNAs, and other lncRNAs to get the read count for each annotated feature. Reads coverages were calculated from merged BAM files and converted to BigWig format using UCSC commandline utility tools (https://genome.ucsc.edu/index.html), and then visualized by using Integrative Genomics Viewer v2.6.2 (IGV) (https://software.broadinstitute.org/software/igv/).

Normalization and differential expression analysis

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Read counts from tumor tissue samples were normalized by using DESeq2 (42) with sample size factors estimated based on counts from nuclear protein-coding genes. Genewise differential expression analyses were also performed using DESeq2, while all other analyses were applied to log-transformed DESeq2-normalized expression values (with $\log_2$ transformation applied after adding a pseudocount of 1 to avoid the singular behavior of $\log_2 0$). Candidate genes were checked visually by IGV to see if differential expression reflected read coverage across the gene or was limited to specific regions and to filter out protein-coding, lncRNA, and miRNA genes whose differential expression was due to reads mapping to simple repeats or other low complexity sequences as defined by RepBase (https://www.girinst.org/repbase/). For DESeq2 analysis of Repeat-element RNAs, reads mapping to the same repeat element were aggregated by the name of the element and normalized by protein-coding gene reads.

Identification of gene sets enriched for differential expression

Elevated or depressed DESeq2 test statistics among genes composing either the BROAD Institute's list of hallmark gene sets (79) or Gene Ontology (GO) biological process gene sets (80) was assessed by using GO-MWU (Gene Ontology Mann-Whitney U testing, https://github.com/z0on/GO_MWU) (81), which analyzed each GO category G within specified size limits (minimum of 10 observed genes, maximum of 10% of all observed genes) and tests whether the scores (DESeq2 test statistics) associated with genes within G were significantly more likely to be found near the top or bottom of the global ranked list of genes than would randomly shuffled genes. GO categories with highly overlapping sets of associated genes were merged (retaining the name of the largest merged category) by complete linkage hierarchical clustering (based on gene set dissimilarities defined by the ratio of the number of overlapping genes to the number of genes in the smaller gene set) and then applying a dendrogram cut height of 0.25 (thereby merging together a cluster of gene sets only when the most dissimilar pair of them overlaps in ≥75% of genes contained in the smaller of the pair).

Peak calling

For each sample, a BAM file with read pairs that mapped to the human-genome was processed as described above. Reads that overlapped sncRNAs (Ensembl annotation, www.ensembl.org), rRNA loci, Mt DNA (to eliminate mis-mapping of Mt RNAs to nuclear Mt DNA segments (NUMTs), and human genome blacklist regions (https://sites.google.com/site/anshulkundaje/projects/blacklists) were removed using BEDTools. BAM files from libraries within the same group (Healthy, non-IBC, and IBC) were combined and then split into separate BAM files containing only reads mapping to the forward or reverse strand. BAM files for each strand were then used as input for the MACS2 callpeak algorithm (https://pypi.org/project/MACS2/) using the combined BAM file for both strands from healthy or non-IBC plasma datasets (n=15 and 6, respectively) as the base-line control, via: macs2 callpeak --treatment {combined BAM from IBC dataset, forward or reverse strand} --control {combined BAM from Healthy or non-IBC dataset, forward or reverse strand } --outdir {outdir} --name {comparison_strand} --nomodel --format BAMPE --qvalue 0.01 --keep-dup all --gsize hs --min-length 150.

The called peaks from the forward and reverse strands were then combined and filtered to obtain candidate peaks with read coverage ≥10 reads at the peak maximum, a false discovery rate ≤ 0.01 (*q*-value assigned by MACS2), and a requirement that the peak be detected in at least 5 of the 10 IBC plasma samples to avoid batch effects.

The annotation of peaks was done with BEDTools (https://bedtools.readthedocs.io/en/latest/content/bedtools-suite.html) intersect command against

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

genomic features from Ensembl genes (https://useast.ensembl.org/index.html) and RepBase-annotated repeats (https://www.girinst.org/repbase/). An overlap score, defined as the number of bases overlapping the genomic feature divided by the peak width, was computed for each overlapping record between a peak and a genomic feature. A maximal overlap score of 1 indicates the peak can be explained by the full-length genomic feature. A best feature annotation for each peak was selected by the highest overlap score. In cases of identical overlap scores, a feature was selected with the priority: repeat regions (RepBase), protein-coding genes, lncRNAs.

Intron-exon Depth Ratio (IDR) analysis

IDRs were considered both as a single quantity θ representing all genes and separately as an individual quantity θ_g for each gene, through a simple model $y = \theta x / [(\theta-1)x+1]$, where y is the estimated fraction of reads for a protein-coding gene in FFPE IBC tumor samples derived from introns and x is the fraction of the genomic annotation of the gene's length corresponding to introns. This model arises from the assumption that the per-base rate of intronic region inclusion in sequencing reads is a fixed quantity θ times the per-base exonic region inclusion rate. The intronic read fractions y_g for gene g are estimated as 1-(# reads mapped to g using transcriptomic reference/# reads mapped to g using genomic reference), capped at a maximum value of 1 and floored at a minimum value of 0. Mapping to the transcriptomic reference was done using HISAT2 (http://daehwankimlab.github.io/hisat2/) and Ensembl GRCh38 Release 93 reference, while genomic mapping was performed as described in read mapping section above. Genomic and transcriptomic mapping results included slightly different gene sets; we assigned IDR values only to genes included in both results sets.

All-genes θ (with no subscript) values were estimated by least-squares using data for all genes; in the FFPE IBC sample set, for example, this yields a value of ~0.19, indicating that read depth for intronic regions tends to be about 0.19 times the read depth for the exonic regions on average across all genes. Individual genes g may have much higher or lower IDR values θ_g, where θ_g is the value of $\theta(g)$ which solves $y_g = \theta x_g / [(\theta-1)x_g+1]$ for the values y_g and x_g associated specifically with gene g only. IDR values θ_g for each gene were calculated separately for each combination of diagnostic group (healthy, non-IBC, and IBC) and sample type (frozen tissue, FFPE, and PBMC).

Data deposition

The TGIRT-seq datasets in this manuscript as listed in the Supplemental File have been deposited in the National Center for Biotechnology Information Sequence Read Archive (accession number: PRJNA954747).

Conflict-of-interest statement: AML is an inventor on patents owned by the University of Texas at Austin for TGIRT enzymes and other stabilized reverse transcriptase fusion proteins and methods for non-retroviral reverse transcriptase template switching. AML, JY, and HX are inventors on a patent application filed by the University of Texas for the use of intron RNAs as biomarkers. DW, XW, JY, HX, EAF-K, NTU, and AML are inventors on a patent application filed jointly by UT Austin and MD Anderson entitled "Methods and Compositions for Diagnosing, Treating and/or Preventing Inflammatory Breast Cancer" based on potential RNA biomarkers identified in this study.

Acknowledgements:

We thank the patients and healthy donors for contributing samples and the staff of the Morgan Welch Inflammatory Breast Cancer Research Program and Clinic for collecting those samples. We thank Lauren Ehrlich (University of Texas at Austin), Tak Mak (University of Toronto), John

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Moran (University of Michigan), and Ryan Nottingham (University of Texas at Austin) for comments on the manuscript and Shanshan Wang (University of Texas at Austin) for contributing to TGIRT-seq library preparation. The Texas Advanced Computing Center (TACC) and the Center for Biomedical Research Support (CBRS) at the University of Texas at Austin provided high-performance computing resources. High-throughput sequencing was done by the Genomic Sequencing and Analysis Facility at the University of Texas at Austin.

Funding: National Institutes of Health grant R35 GM136216 (AML), Welch Foundation grant F-1607 (AML), Breast Cancer Research Foundation grant BCRF-20-164 (NTU), National Institutes of Health grant 1R01CA258523-01A1 (NTU), the Morgan Welch Inflammatory Breast Cancer Research Program (NTU), and the State of Texas Rare and Aggressive Breast Cancer Research Program Grant (NTU).

References

1. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nat Rev Genet* 2019;20(11):631–656.
2. Qin Y, et al. High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. *RNA* 2016;22(1):111–128.
3. Yao J, et al. Identification of protein-protected mRNA fragments and structured excised intron RNAs in human plasma by TGIRT-seq peak calling. *Elife* 2020;9:e60743.
4. Cristofanilli M, et al. Update on the management of inflammatory breast cancer. *The Oncologist* 2003;8(2):141–148.
5. Hance KW, et al. Trends in inflammatory breast carcinoma incidence and survival: the surveillance, epidemiology, and end results program at the National Cancer Institute. *Journal of the National Cancer Institute* 2005;97(13):966–975.
6. Jaiyesimi IA, Buzdar AU, Hortobagyi G. Inflammatory breast cancer: a review. *Journal of Clinical Oncology* 1992;10(6):1014–1024.
7. Yamauchi H, Ueno NT. Targeted therapy in inflammatory breast cancer. *Cancer* 2010;116(S11):2758–2759.
8. Yamauchi H, et al. Inflammatory breast cancer: what we know and what we need to learn. *The Oncologist* 2012;17(7):891.
9. Masuda H, et al. Long-term treatment efficacy in primary inflammatory breast cancer by hormonal receptor-and HER2-defined subtypes. *Annals of Oncology* 2014;25(2):384–391.
10. Rueth NM, et al. Underuse of trimodality treatment affects survival for patients with inflammatory breast cancer: an analysis of treatment and survival trends from the National Cancer Database. *Journal of Clinical Oncology* 2014;32(19):2018.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

11. Schlichting JA, et al. Inflammatory and non-inflammatory breast cancer survival by socioeconomic position in the Surveillance, Epidemiology, and End Results database, 1990-2008. *Breast Cancer Res Treat* 2012;134(3):1257–1268.
12. Van Laere S, et al. Distinct molecular signature of inflammatory breast cancer by cDNA microarray analysis. *Breast Cancer Research and Treatment* 2005;93(3):237–246.
13. Van Laere SJ, et al. Nuclear 11 factor-kappaB signature of inflammatory breast cancer by cDNA microarray validated by 12 quantitative real-time reverse transcription-PCR, immunohistochemistry, and nuclear factor-13 kappaB DNA-binding. *Clinical Cancer Research* 2006;12:3249–56.
14. Lerebours F, et al. NF-kappa B genes have a major role in inflammatory breast cancer. *BMC Cancer* 2008;8(1):41.
15. Wang X, et al. EGFR signaling promotes inflammation and cancer stem-like activity in inflammatory breast cancer. *Oncotarget* 2017;8(40):67904.
16. Van Laere S, et al. Distinct molecular phenotype of inflammatory breast cancer compared to non-inflammatory breast cancer using Affymetrix-based genome-wide gene-expression analysis. *British Journal of Cancer* 2007;97(8):1165–1174.
17. Jhaveri K, et al. Hyperactivated mTOR and JAK2/STAT3 pathways: molecular drivers and potential therapeutic targets of inflammatory and invasive ductal breast cancers after neoadjuvant chemotherapy. *Clinical Breast Cancer* 2016;16(2):113–122.
18. Shirakawa K, et al. Tumor-infiltrating endothelial cells and endothelial precursor cells in inflammatory breast cancer. *International Journal of Cancer* 2002;99(3):344–351.
19. Van der Auwera I, et al. Increased angiogenesis and lymphangiogenesis in inflammatory versus noninflammatory breast cancer by real-time reverse transcriptase-PCR gene expression quantification. *Clinical Cancer Research* 2004;10(23):7965–7971.
20. Marotta LLC, et al. The JAK2/STAT3 signaling pathway is required for growth of CD44+ CD24–stem cell–like breast cancer cells in human tumors. *The Journal of Clinical Investigation* 2011;121(7):2723–2735.
21. Overmoyer BA, et al. JAK2/STAT3 activity in inflammatory breast cancer supports the investigation of JAK2 therapeutic targeting. *Cancer Reseach* 2012;72(24 Suppl).
22. Cabioglu N, et al. Expression of growth factor and chemokine receptors: new insights in the biology of inflammatory breast cancer. *Annals of Oncology* 2007;18(6):1021–1029.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

23. Alpaugh ML, et al. Relationship of sialyl-Lewisx/a underexpression and E-cadherin overexpression in the lymphovascular embolus of inflammatory breast carcinoma. *The American Journal of Pathology* 2002;161(2):619–628.
24. van Golen KL, et al. A novel putative low-affinity insulin-like growth factor-binding protein, LIBC (lost in inflammatory breast cancer), and RhoC GTPase correlate with the inflammatory breast cancer phenotype. *Clinical Cancer Research* 1999;5(9):2511–2519.
25. Silvera D, et al. Essential role for eIF4GI overexpression in the pathogenesis of inflammatory breast cancer. *Nature Cell Biology* 2009;11(7):903–908.
26. Van Laere SJ, et al. Uncovering the molecular secrets of inflammatory breast cancer biology: an integrated analysis of three distinct affymetrix gene expression datasets. *Clinical Cancer Research* 2013;19(17):4685–4696.
27. Kai K, et al. CSF-1/CSF-1R axis is associated with epithelial/mesenchymal hybrid phenotype in epithelial-like inflammatory breast cancer. *Sci Rep* 2018;8(1):9427.
28. Martín-Alonso S, Frutos-Beltrán E, Menéndez-Arias L. Reverse Transcriptase: From Transcriptomics to Genome Editing. *Trends Biotechnol* 2021;39(2):194–210.
29. Stamos JL, Lentzsch AM, Lambowitz AM. Structure of a Thermostable Group II Intron Reverse Transcriptase with Template-Primer and Its Functional and Evolutionary Implications. *Mol Cell* 2017;68(5):926-939.e4.
30. Belfort M, Lambowitz AM. Group II Intron RNPs and Reverse Transcriptases: From Retroelements to Research Tools. *Cold Spring Harb Perspect Biol* 2019;11(4):a032375.
31. Mohr S, et al. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* 2013;19(7):958–970.
32. Liu B, et al. A split prime editor with untethered reverse transcriptase and circular RNA template. *Nat Biotechnol* 2022;40(9):1388–1393.
33. Grünewald J, et al. Engineered CRISPR prime editors with compact, untethered reverse transcriptases. *Nat Biotechnol* [published online ahead of print: September 26, 2022]; doi:10.1038/s41587-022-01473-1
34. Guo L-T, et al. Sequencing and Structure Probing of Long RNAs Using MarathonRT: A Next-Generation Reverse Transcriptase. *J Mol Biol* 2020;432(10):3338–3352.
35. Upton HE, et al. Low-bias ncRNA libraries using ordered two-template relay: Serial template jumping by a modified retroelement reverse transcriptase. *Proc Natl Acad Sci U S A* 2021;118(42):e2107900118.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

36. Maguire S, Guan S. Rolling circle reverse transcription enables high fidelity nanopore sequencing of small RNA. *PLoS One* 2022;17(10):e0275471.
37. Nottingham RM, et al. RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. *Rna* 2016;22(4):597–613.
38. Shurtleff MJ, et al. Broad role for YBX1 in defining the small noncoding RNA composition of exosomes. *Proc Natl Acad Sci U S A* 2017;114(43):E8987–E8995.
39. Faucher-Giguère L, et al. High-grade ovarian cancer associated H/ACA snoRNAs promote cancer cell proliferation and survival. *NAR Cancer* 2022;4(1):zcab050.
40. Boivin V, et al. Simultaneous sequencing of coding and noncoding RNA reveals a human transcriptome dominated by a small number of highly expressed noncoding genes. *RNA* 2018;24(7):950–965.
41. Deshmukh SK, et al. Inflammation, immunosuppressive microenvironment and breast cancer: opportunities for cancer prevention and therapy. *Ann Transl Med* 2019;7(20):593.
42. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 2014;15(12):550.
43. Bièche I, et al. Molecular profiling of inflammatory breast cancer: identification of a poor-prognosis gene expression signature. *Clin Cancer Res* 2004;10(20):6789–6795.
44. Bekhouche I, et al. High-resolution comparative genomic hybridization of inflammatory breast cancer and identification of candidate genes. *PLoS One* 2011;6(2):e16950.
45. Bertucci F, et al. Gene expression profiles of inflammatory breast cancer: correlation with response to neoadjuvant chemotherapy and metastasis-free survival. *Ann Oncol* 2014;25(2):358–365.
46. Wang X, et al. Targeting Signaling Pathways in Inflammatory Breast Cancer. *Cancers (Basel)* 2020;12(9):E2479.
47. Cao S, et al. Estimation of tumor cell total mRNA expression in 15 cancer types predicts disease progression. *Nat Biotechnol* 2022;40(11):1624–1633.
48. Zatzman M, et al. Widespread hypertranscription in aggressive human cancers. *Sci Adv* 2022;8(47):eabn0238.
49. Trop-Steinberg S, Azar Y. AP-1 Expression and its Clinical Relevance in Immune Disorders and Cancer. *Am J Med Sci* 2017;353(5):474–483.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

50. Charafe-Jauffret E, et al. Aldehyde dehydrogenase 1-positive cancer stem cells mediate metastasis and poor clinical outcome in inflammatory breast cancer. *Clin Cancer Res* 2010;16(1):45–55.
51. Fillmore CM, Kuperwasser C. Human breast cancer cell lines contain stem-like cells that self-renew, give rise to phenotypically diverse progeny and survive chemotherapy. *Breast Cancer Res* 2008;10(2):R25.
52. Wang X, et al. TIG1 promotes the development and progression of inflammatory breast cancer through activation of Axl kinase. *Cancer Research* 2013;73(21):6516–6525.
53. Qi Y, et al. Expression signatures and roles of microRNAs in inflammatory breast cancer. *Cancer Cell International* 2019;19(1):23.
54. Debeb BG, et al. miR-141-mediated regulation of brain metastasis from breast cancer. *JNCI: Journal of the National Cancer Institute* 2016;108(8).
55. Van der Auwera I, et al. Integrated miRNA and mRNA expression profiling of the inflammatory breast cancer subtype. *British journal of cancer* 2010;103(4):532–541.
56. Anfossi S, et al. High serum miR-19a levels are associated with inflammatory breast cancer and are predictive of favorable clinical outcome in patients with metastatic HER2+ inflammatory breast cancer. *PloS One* 2014;9(1):e83113.
57. Newman AM, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* 2019;37(7):773–782.
58. Palmblad J. The role of granulocytes in inflammation. *Scand J Rheumatol* 1984;13(2):163–172.
59. Lerebours F, et al. miRNA expression profiling of inflammatory breast cancer identifies a 5-miRNA signature predictive of breast tumor aggressiveness. *International Journal of Cancer* 2013;133(7):1614–1623.
60. Sun W, et al. miR-223 and miR-142 attenuate hematopoietic cell proliferation, and miR-223 positively regulates miR-142 through LMO2 isoforms and CEBP-$$. *Cell research* 2010;20(10):1158–1169.
61. Aziz F. The emerging role of miR-223 as novel potential diagnostic and therapeutic target for inflammatory disorders. *Cellular Immunology* 2016;303:1–6.
62. Chakrabortty SK, et al. Extracellular vesicle-mediated transfer of processed and functional RNY5 RNA. *RNA* 2015;21(11):1966–1979.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

63. Protasova MS, Andreeva TV, Rogaev EI. Factors Regulating the Activity of LINE1 Retrotransposons. *Genes (Basel)* 2021;12(10):1562.
64. Flasch DA, et al. Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication. *Cell* 2019;177(4):837-851.e28.
65. Smit AF. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* 1999;9(6):657–663.
66. Tiwari B, et al. p53 directly represses human LINE1 transposons. *Genes Dev* 2020;34(21–22):1439–1451.
67. McKerrow W, et al. LINE-1 expression in cancer correlates with p53 mutation, copy number alteration, and S phase checkpoint. *Proc Natl Acad Sci U S A* 2022;119(8):e2115999119.
68. Tunbak H, et al. The HUSH complex is a gatekeeper of type I interferon through epigenetic regulation of LINE-1s. *Nat Commun* 2020;11(1):5387.
69. Zhao K, et al. LINE1 contributes to autoimmunity through both RIG-I- and MDA5-mediated RNA sensing pathways. *J Autoimmun* 2018;90:105–115.
70. Gatault S, et al. Involvement of eosinophils in the anti-tumor response. *Cancer Immunol Immunother* 2012;61(9):1527–1534.
71. Rodić N, et al. Long interspersed element-1 protein expression is a hallmark of many human cancers. *Am J Pathol* 2014;184(5):1280–1286.
72. Ardeljan D, et al. The Human Long Interspersed Element-1 Retrotransposon: An Emerging Biomarker of Neoplasia. *Clin Chem* 2017;63(4):816–822.
73. Singletary SE, et al. Revision of the American Joint Committee on Cancer staging system for breast cancer. *J Clin Oncol* 2002;20(17):3628–3636.
74. Xu H, et al. Improved TGIRT-seq methods for comprehensive transcriptome profiling with decreased adapter dimer formation and bias correction. *Scientific reports* 2019;9(1):1–17.
75. Xu H, Nottingham RM, Lambowitz AM. TGIRT-seq Protocol for the Comprehensive Profiling of Coding and Non-coding RNA Biotypes in Cellular, Extracellular Vesicle, and Plasma RNAs. *Bio-Protocol* 2021;11(23). doi:10.21769/BioProtoc.4239
76. Kechin A, et al. cutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing. *J Comput Biol* 2017;24(11):1138–1143.
77. Kim D, et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019;37(8):907–915.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

78. Li H, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25(16):2078–2079.
79. Liberzon A, et al. The molecular signatures database hallmark gene set collection. *Cell Systems* 2015;1(6):417–425.
80. Consortium GO. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Research* 2019;47(D1):D330–D338.
81. Wright RM, et al. Gene expression associated with white syndromes in a reef building coral, Acropora hyacinthus. *BMC Genomics* 2015;16(1):371.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

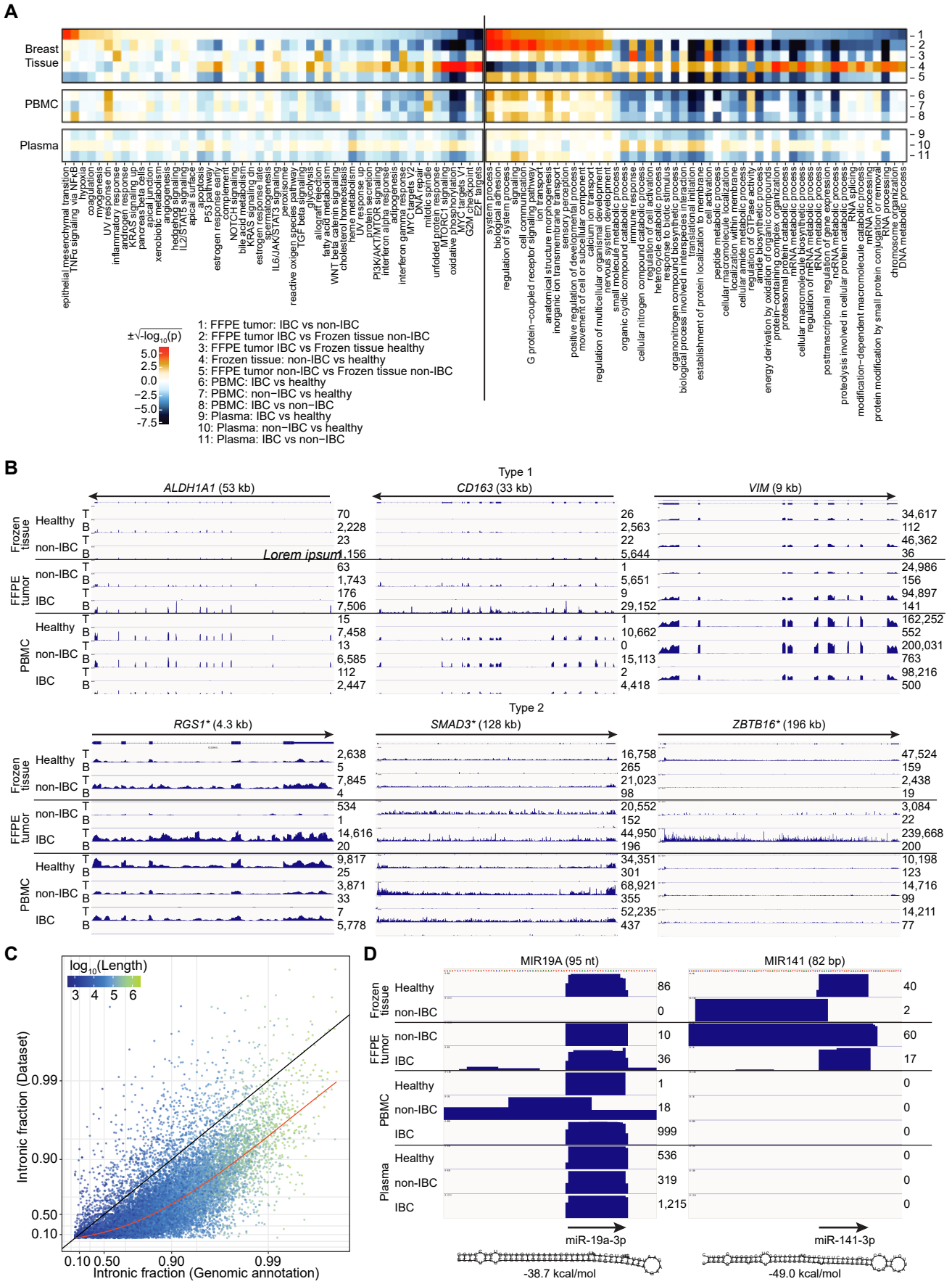
Figure 1

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. 1. Principal component analysis and distribution of reads mapping to different RNA biotypes in different sample sets. (**A**) Principal components analysis (PCA) applied to all genes in all sample sets (left panel) and PCA run separately on genes subsetted by biotype and samples subsetted by sample type (right panel). (**B**) Stacked bar graphs showing the percentage of reads in each sample assigned to different RNA biotypes (Ensembl GRCh38 Release 93) (top panel: all biotype categories excluding rRNA; middle panel: sncRNA categories; bottom panel: Repeat element categories). Sample tissue type (Frozen, FFPE, PBMC, or Plasma) and disease status (Healthy, non-IBC, or IBC) are indicated above, and sample names corresponding to healthy donor or patient ID numbers are indicated below (H, healthy; B, non-IBC; I, IBC). Paired Wilcoxon sign rank test results for differences in biotype read distributions between sample types can be found in the Supplemental file.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 2



medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. 2. Pathway analysis and categories of protein-coding genes and miRNA transcripts identified by TGIRT-seq. (**A**) Statistical significance of enrichment assessed by GO-MWU of BROAD Institute 50 Hallmark or Gene Ontology (GO) gene sets (left and right panel, respectively) for differentially elevated genes (shades of red or yellow) or depressed genes (shades of blue) for the pairwise comparisons numbered on the right and defined in the key below. The color of each tile in the heatmap indicates the square-root of $-\log_{10}(p)$, where p is the MWU enrichment p-value, multiplied by a sign factor of 1 for enrichment in elevated genes and -1 for enrichment in depressed genes. The columns of the heatmap are labeled by the names of the gene sets tested and are clustered to position gene sets with similar MWU significance together. (**B**) Integrated Genome Viewer (IGV) plots for representative protein-coding genes for which reads mapped largely to exons from mature mRNAs (***Type 1, top row***) or across the entire gene region including intronic sequences (***Type 2, second row***). T and B, top and bottom DNA strands, respectively. The number of reads for RNAs in each IGV plot is indicated to the right. (**C**) Intron-exon depth ratio (IDR) analysis. Each point indicates the estimated fraction of reads for a protein-coding gene in FFPE IBC tumor samples derived from introns (y-axis, $-\log(1-y)$ scaling) compared to the fraction of the genomic annotation of the gene's length corresponding to introns (x-axis, $-\log(1-x)$ scaling). Points are colored according to the $\log_{10}$-transformed length of the genomic interval encoding the gene (exons and introns included): longer genomic intervals tend to be associated with relatively high genomic intron fractions, as is evident from color gradient from left to right. The black line indicates equality $y=x$, corresponding to the set of points representing genes with IDR=1; points above or below this line represent genes with IDR greater or less than 1, respectively. The red trend line plots the solution to the IDR equation (see Methods) when applied to data for all genes taken together, yielding a typical IDR value of ~0.19. (**D**) Examples of IGV plots for miRNA genes for which the observed sense-strand reads mapped to either the mature miRNA sequence or longer transcripts within the pre-miRNA sequence, including full-length pre-miRNAs. IGV plots are labeled as in panel B. RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi) predicted secondary structures of the pre-miRNA hairpins and their calculated minimum free energy values are shown below the corresponding IGV plots. The number of reads is indicated to the right.
samples. The color scale is shown below the heat map of panel A. Lanes are numbered according to the healthy donor or patient ID number (Table S1). miRNAs mapping to the hairpin miRNA reference were categorized as: m, mature miRNAs, if reads were 1 or 2 nucleotides shorter or longer than the annotated mature miRNA sequence; pre, pre-miRNAs if reads were derived from pre-miRNA only; or m+pre, if reads were derived from both mature and pre-miRNAs.espectively.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 3

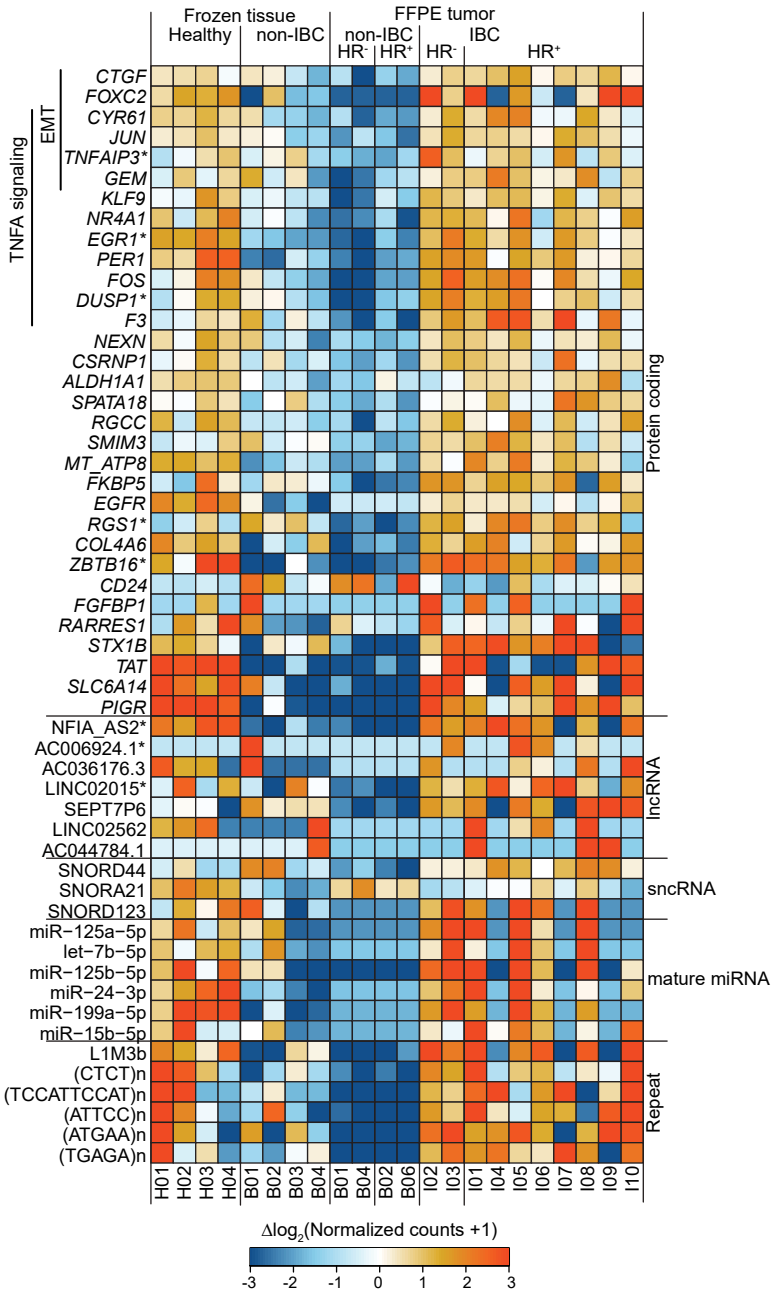
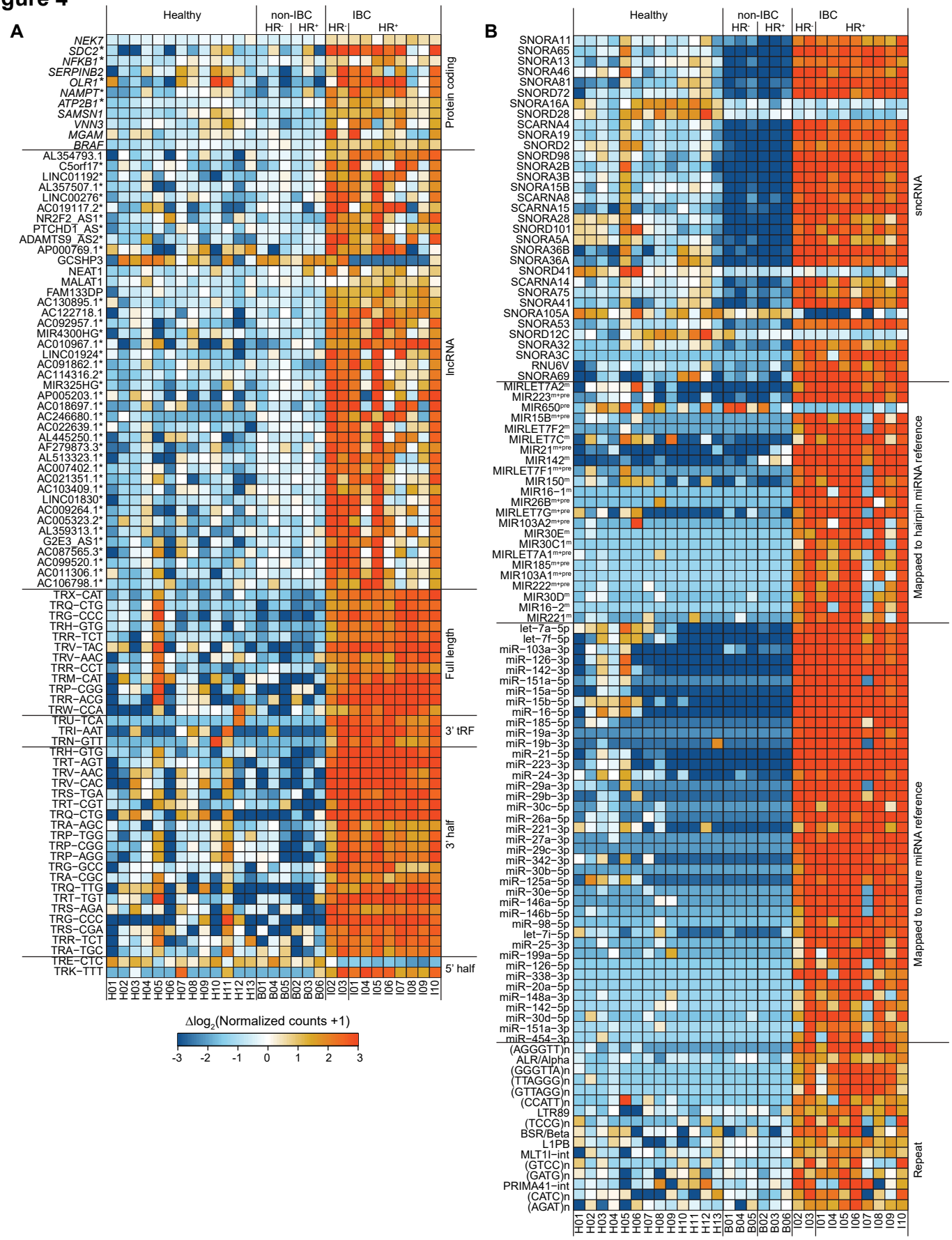


Fig. 3. Heat map of expression levels of significantly differentially expressed and other notable genes in healthy or tumor breast tissue samples. The significantly overexpressed genes ($p \leq 0.001$, LFC$\geq$2) shown in the heat map are limited to those detected in half or more of the IBC samples. The color of the tile for gene *g* in sample *s* encodes the difference in the (log-transformed DESeq2-normalized) expression level for sample *s* from overall mean expression for gene *g* for all samples of same type (frozen tissue or FFPE). The color scale is shown below the heat map. Lanes are numbered according to the patient ID number (Table S1).

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 4



medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. 4. Heat map of expression levels of significantly differentially expressed genes and other notable genes in PBMCs. Expression profiles for (**A**) select protein-coding genes, lncRNAs, and tRNAs including tRNA halves and 3'-tRFs, and (**B**) sncRNAs, miRNAs, or Repeat-element RNAs in the PMBC sample set. The significantly overexpressed genes shown in the heat map ($p \leq 0.001$, LFC$\geq$2) are limited to those detected in half or more of the IBC samples. Tiles are colored by expression difference relative to genewise mean taken across all PBMC samples. The color scale is shown below the heat map of panel A. Lanes are numbered according to the healthy donor or patient ID number (Table S1). miRNAs mapping to the hairpin miRNA reference were categorized as: m, mature miRNAs, if reads were 1 or 2 nucleotides shorter or longer than the annotated mature miRNA sequence; pre, pre-miRNAs if reads were derived from pre-miRNA only; or m+pre, if reads were derived from both mature and pre-miRNAs.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 5

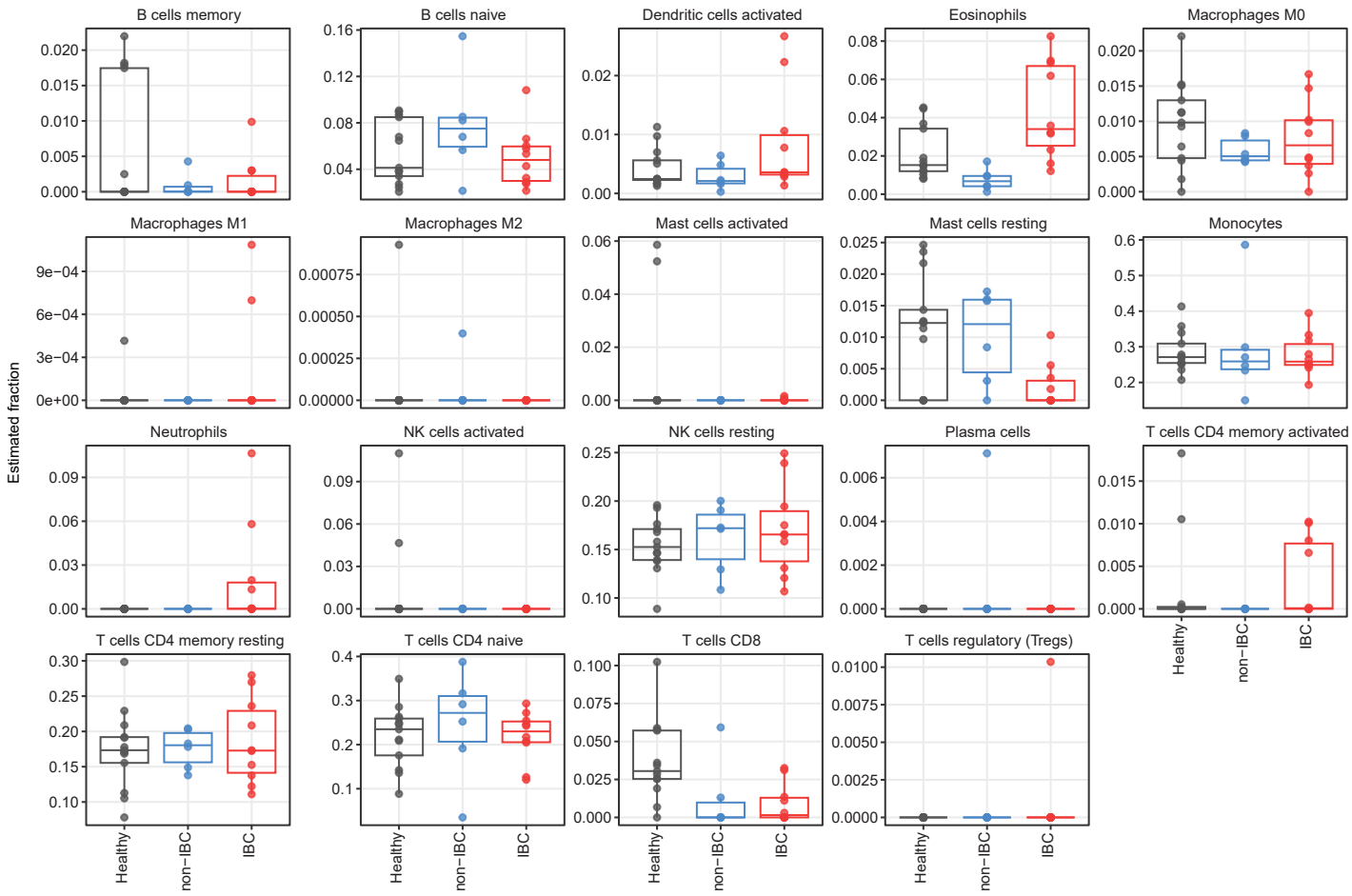
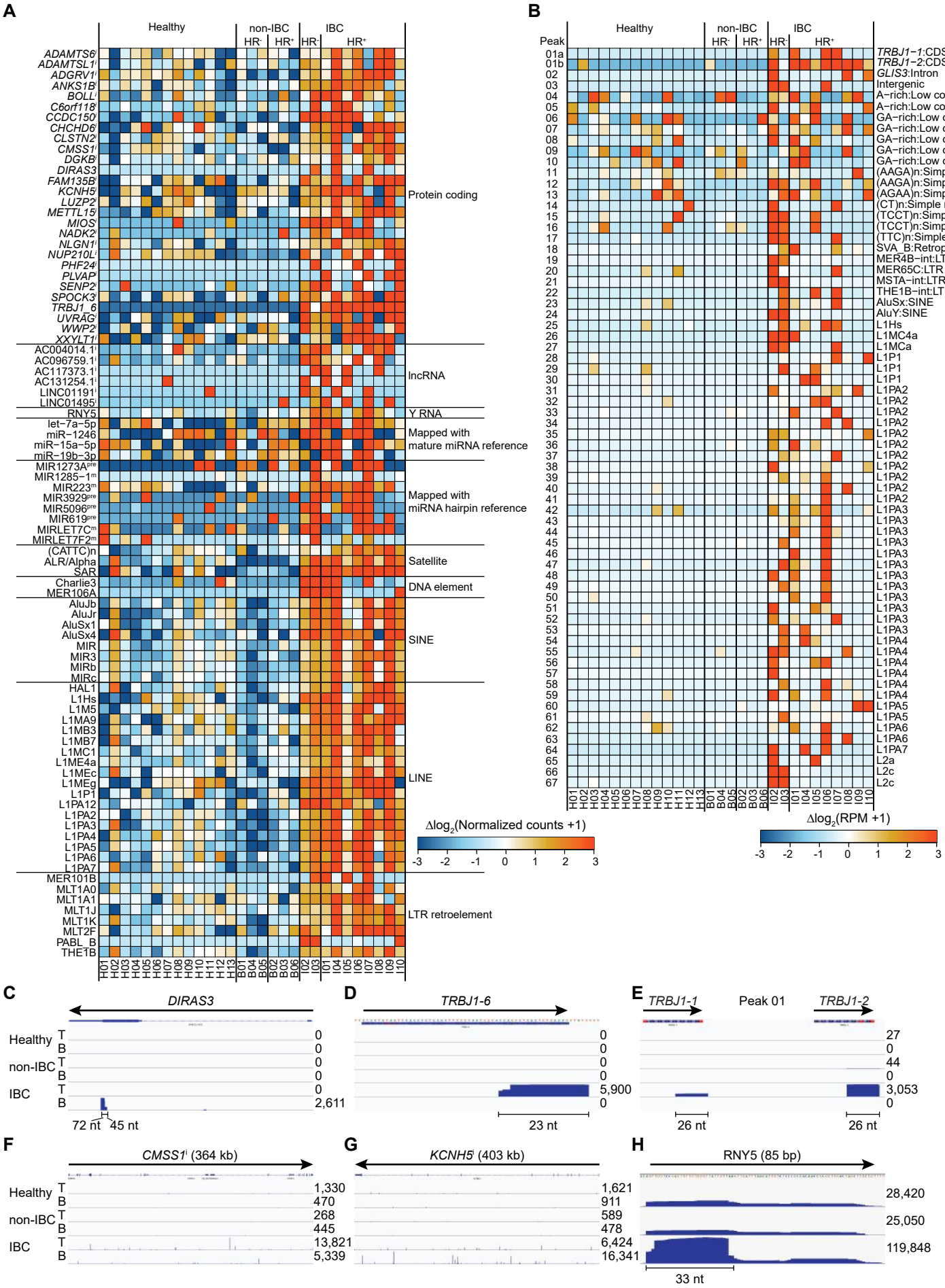


Fig. 5. CIBERSORTx-estimated fractions of PBMC cell types for all samples grouped by diagnosis. Points represent individual sample cell type fraction estimates (*y*-axis) separated by patient diagnosis (Healthy, non-IBC, or IBC). Midpoints of boxes indicate median values for each sample group, with bottom and top of boxes indicating 25th and 75th percentile estimates for within-group distributions of cell type fractions.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 6

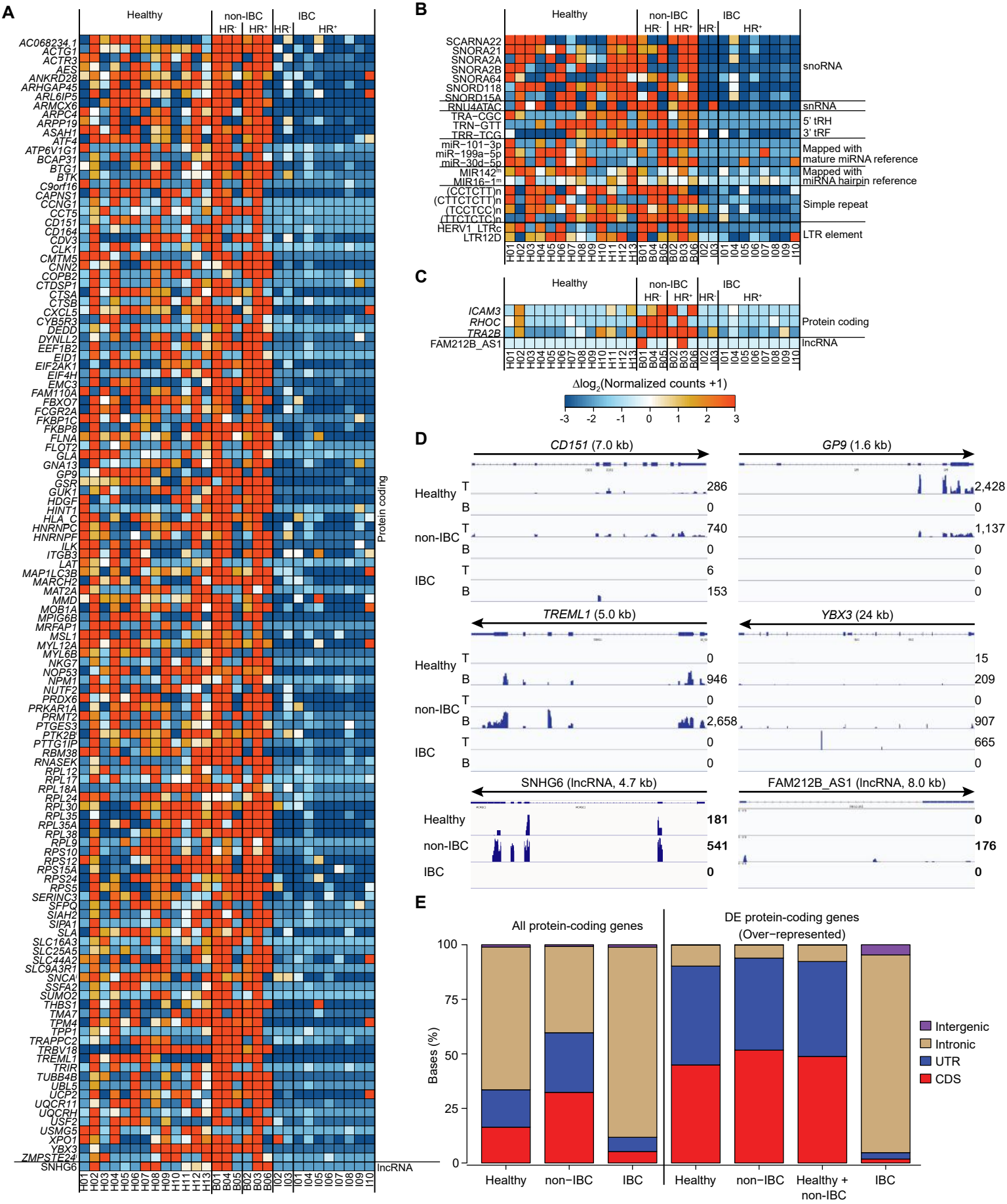


medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. 6. Heatmap of the relative abundance and IGV plots for select genes in the plasma sample set. (**A**) Genes and features of interest, including those identified by DESeq2 analysis as significantly overrepresented ($p \leq 0.001$, LFC$\geq$2) in the IBC plasma datasets. Tiles are colored by expression differences relative to the corresponding genewise means taken across all plasma samples. Color scale at bottom right of panel. miRNAs mapping to the hairpin miRNA reference were categorized as: m, mature miRNAs, if reads were 1 or 2 nucleotides shorter or longer than the annotated mature miRNA sequence; pre, pre-miRNAs if reads were derived from pre-miRNA only; or m+pre, if reads were derived from both mature and pre-miRNAs. (**B**) Heatmap for called peaks in the plasma sample set using MACS2. Tiles are colored by log-transformed difference in RPM-normalized counts from the genewise mean across all plasma samples. Color scale at bottom right. Lanes in heat maps are numbered according to the healthy donor or patient ID number (Table S1). (**C-H**) IGV plots for the indicated genes and peaks. T and B, top and bottom DNA strands, respectively. The number of reads for RNAs in each IGV plot is indicated to the right.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 7

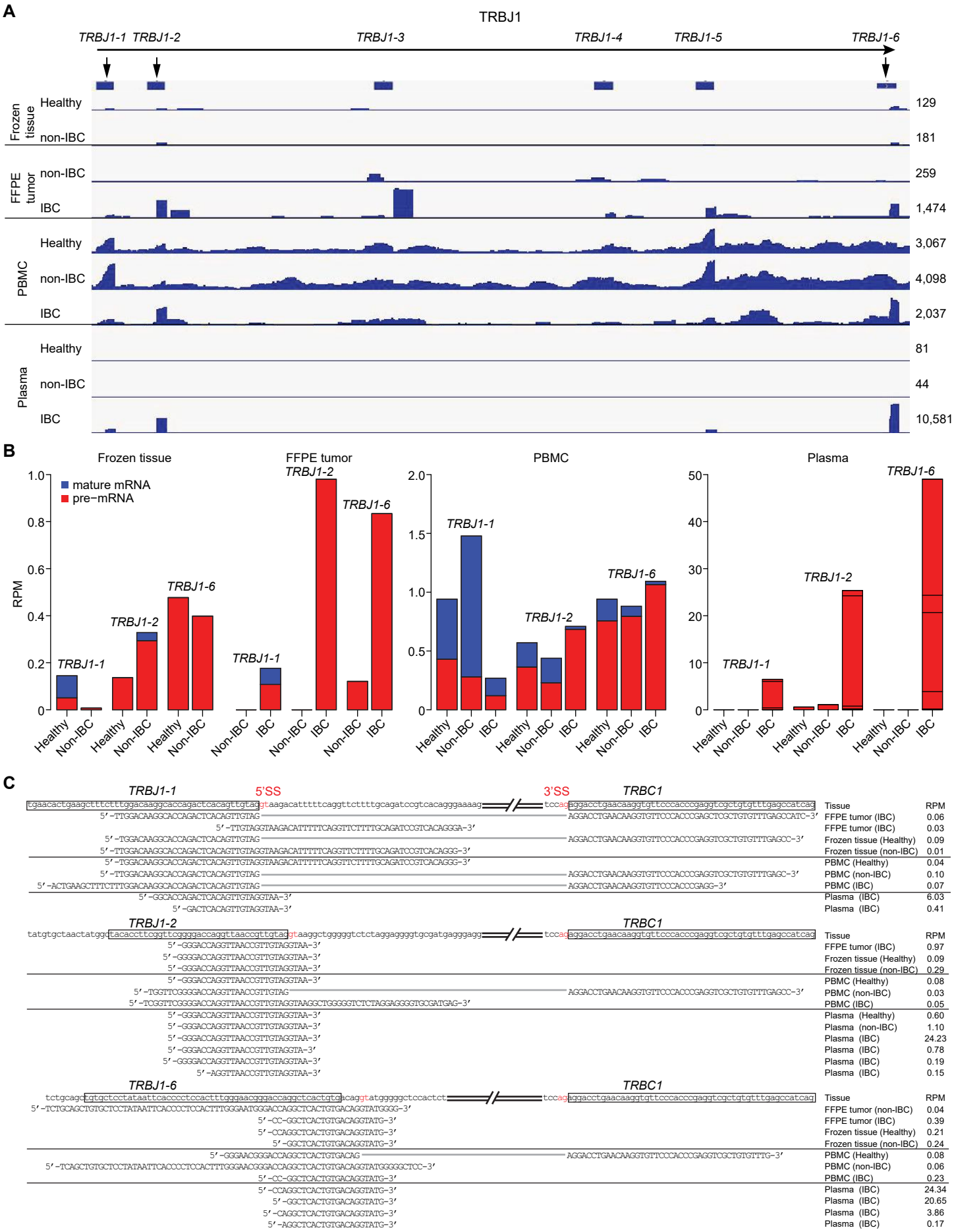


medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. 7. Heatmap of the relative abundance for genes and features that were significantly over-represented in healthy and/or non-IBC plasma samples relative to the IBC plasma samples. (**A-C**) heatmaps for the indicated sample and RNA comparisons. Tiles are colored by expression differences relative to the corresponding genewise means taken across all plasma samples. Lanes in heat maps are numbered according to the healthy donor or patient ID number (Table S1). miRNAs mapping to the hairpin miRNA reference were categorized as: m, mature miRNAs, if reads were 1 or 2 nucleotides shorter or longer than the annotated mature miRNA sequence. (**D**) IGV plots of overrepresented genes in healthy or non-IBC plasma compared to IBC plasma. T and B, top and bottom DNA strands, respectively. The number of reads for RNAs in each IGV plot is indicated to the right. (**E**) Stacked bar graphs showing the percentage of bases in reads mapped to coding sequences (CDS), untranslated regions (UTRs), introns, or intergenic regions for all protein-coding genes detected in the indicated plasma samples (left 3 bars) or for significantly over-represented protein-coding genes ($p \leq 0.001$, LFC$\geq$2) detected in $\geq$ 50% of the indicated plasma samples (right 4 bars).

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 8

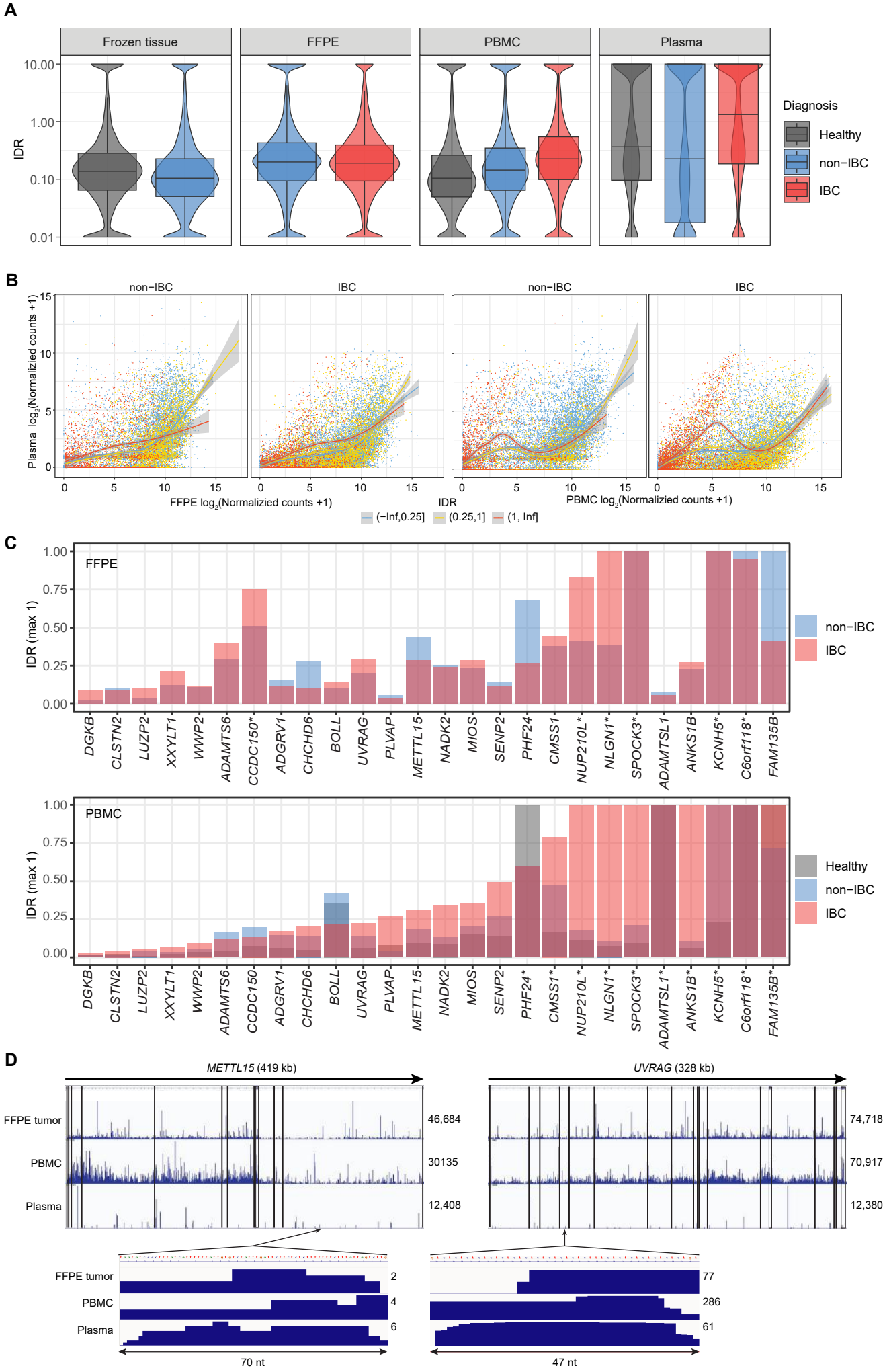


medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. 8. Cellular origin of *TRBJ1* RNAs that were over-represented in IBC plasma samples. (**A**) IGV plots for RNAs mapping to genomic locus encoding *TRBJ1-1* through *TRBJ1-6* in combined datasets for different samples. The number of reads for RNAs in each IGV plot is indicated to the right. (**B**) Stacked bar graph showing RPM for spliced mRNA (blue) and unspliced pre-mRNA (red) for *TRBJ1-1*, *TRBJ1-2*, and *TRBJ1-6* in different sample types. Sequences corresponding to the splice site for *TRBJ1-1*, *TRBJ1-2*, or *TRBJ1-6* to *TRBC1* were extracted and categorized by sample class, sample type, and RNA type (unspliced pre-mRNA or spliced mRNA). Horizontal thin black lines in the bar graphs for plasma delineate unspliced pre-mRNA reads with different sequences. (**C**). Sequence alignments for the reads encompassing the splice-site (SS) region of *TRBJ1-1*, *TRBJ1-2*, or *TRBJ1-6* to *TRBC1* (5'SS, 3'SS, and intron 5' GT and 3' AG in red). The displayed reads are representative abundant RNA fragments from each category for samples in the same disease category and tissue type. All unique RNA fragments detected in plasma datasets with more than 5 reads are shown. Reads containing spliced exons are connected by gray lines.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 9

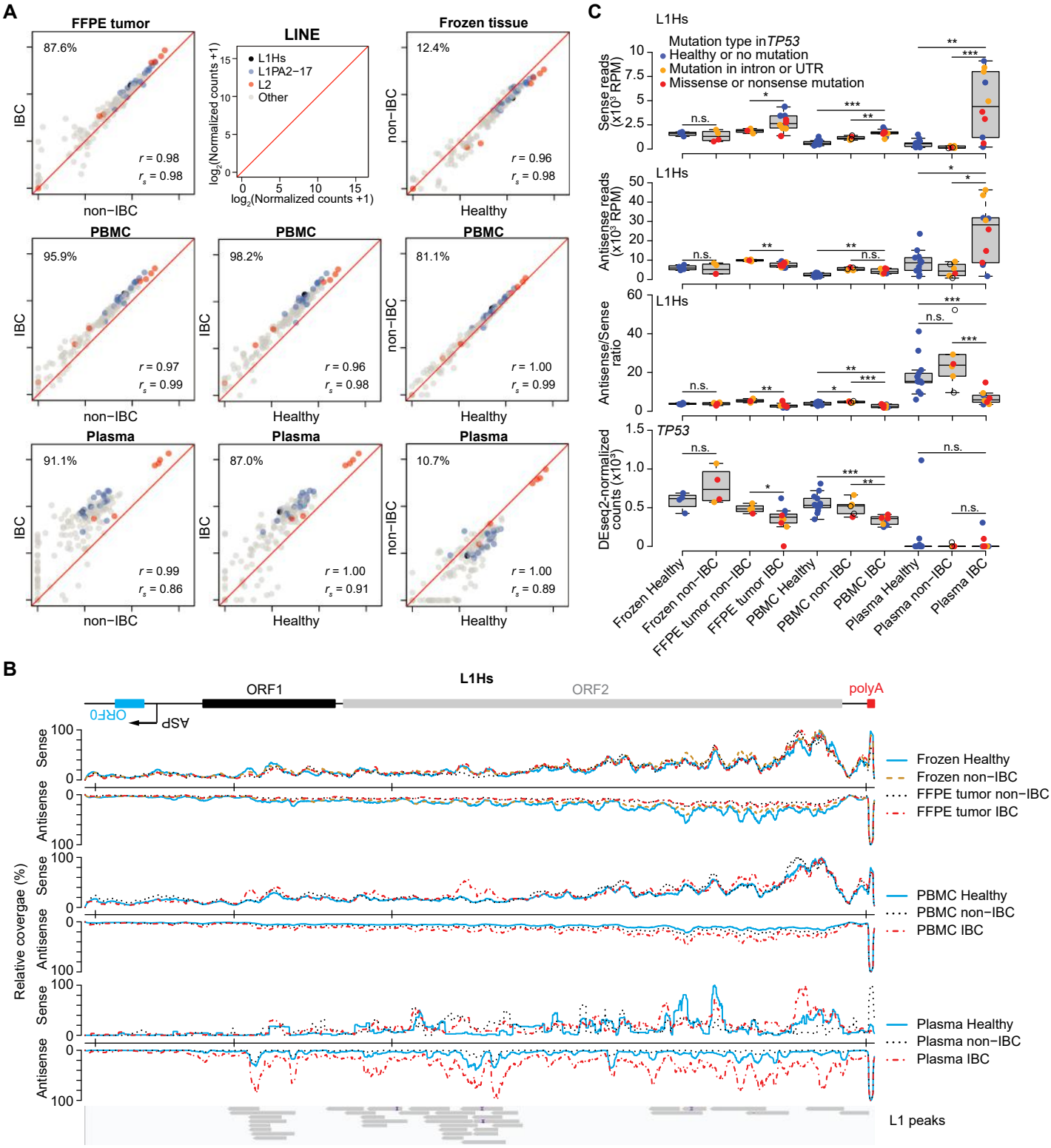


medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. 9. Cellular origin of intron RNAs that are over-represented in IBC plasma samples. (**A**) Violin plots with overlaid boxplots (bottom/top of boxes indicate $25^{th}/75^{th}$ percentiles, respectively) summarizing the distribution of $\log_{10}$-transformed IDR values for samples grouped by diagnosis and tissue. (**B**) Average plasma expression level of each gene plotted against the average FFPE tumor (left two panels) or PBMC (right two panels) expression level of that gene in either non-IBC samples (1st and 3rd panels) or IBC samples (2nd and 4th panels). Each point represents a single gene, colored according to binned IDR values in FFPE tumor or PBMC samples with color codes shown below the plots. Generalized additive model trendlines are fit for each group of IDR-binned genes, with 95% confidence intervals based on standard error estimates indicated by the ribbons around the trendlines. The trendlines show clear separation particularly for the region around 5 on the *x*-axis in each panel, with high-IDR genes (red points and trendline) tending to show higher average plasma expression than lower-IDR genes with similar FFPE tumor or PBMC expression levels. (**C**) Bar plot of IDR values for FFPE (top) and PBMC (bottom) sample sets, with diagnostic group indicated by bar color, for intron RNA fragments found in plasma. Asterisks denote genes with IDR≥0.5 in IBC tumors and/or PBMCs, separately calculated for each sample type. (**D**) IGV plots showing representative intron fragments (sense strand only) detected in IBC FFPE tumors, PBMC, and plasma for 2 genes whose intron RNAs were significantly over-represented in IBC plasma. Vertical black lines and boxes indicate UTRs and exons of the gene. Expanded regions at the bottom show examples of tumor and PBMC read coverage for intron RNA fragments found in plasma. The number of reads for RNAs in each IGV plot is indicated to the right.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 10



medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. 10. Activation of LINE-1 RNAs in IBC tumor and PBMCs and strandedness of LINE-1 RNA reads in plasma. (**A**) Scatter plots comparing the relative abundance of different families of LINE RNAs detected in breast tissue, PBMC or plasma samples. DESeq2 normalized counts of LINE elements were averaged by tissue type and disease status and $\log_2$ transformed for pairwise comparisons. Different subfamilies of the LINE element are color coded as shown in the top center panel of the Figure. The numbers at the top left of each scatterplot indicate the percentage of LINE RNAs above the diagonal, a measure of over-expression of LINE elements in each comparison. Pearson's (r) and Spearman's (r_s) correlation coefficients are shown at the bottom right. (**B**) Coverage plots of reads mapped to the L1Hs consensus sequence in combined dataset from different tissues. Coverage depths were normalized to the maximum coverage of the sense or antisense reads in each sample type. (**C**) Box plots comparing the abundance (RPM) of reads mapped to the L1Hs consensus sequence, and the ratio of reads mapped to L1Hs antisense and sense strand in different tissue types (top 3 plots) correlated with a box plot of DESeq2 normalized counts for *TP53* in different tissue types (bottom plot). Wilcoxon tests were used to compare different groups of samples *: $p \leq 0.05$; **: $p \leq 0.01$; ***: $p \leq 0.001$; n.s. not significant. Samples are colored by *TP53* mutation status in breast tissues (Frozen healthy, Frozen non-IBC tumors, FFPE non-IBC, and FFPE IBC tumors). Somatic *TP53* mutations were detected by comparing each of the cancer samples with their putatively healthy counterparts (non-IBC frozen tumor tissue vs matched healthy frozen tissue, FFPE non-IBC or IBC vs matched PBMCs; see Fig. S10). Blue: no mutation; orange: mutation in intron or UTR, red: missense or nonsense mutations in the coding sequence. Patient samples in PBMC and plasma box plots were colored based on the *TP53* genotype of the corresponding tumor dataset; the two non-IBC PBMC and plasma samples that lacked corresponding tumor samples are shown as open circles. Abbreviation: ASP, LINE-1 antisense promoter.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table S1. Donor information

Donor ID* (TGIRT-seq)	Tissue type (Dataset ID**)				Age	Race	ER +/-	PR +/-	HER2 +/-	HR	IBC
	Frozen tissue	FFPE	PBMC	Plasma							
H01			PBMCH01	PlasmaH01							
H02			PBMCH02	PlasmaH02							
H03			PBMCH03	PlasmaH03							
H04			PBMCH04	PlasmaH04							
H05			PBMCH05	PlasmaH05							
H06			PBMCH06	PlasmaH06							
H07			PBMCH07	PlasmaH07							
H08			PBMCH08	PlasmaH08							
H09			PBMCH09	PlasmaH09							
H10			PBMCH10	PlasmaH10							
H11			PBMCH11	PlasmaH11							
H12			PBMCH12	PlasmaH12							
H13			PBMCH13	PlasmaH13							
B01		FFPEB01	PBMCB01	PlasmaB01	51-55	White	Negative	Negative	Negative	Negative	N
B02		FFPEB02	PBMCB02	PlasmaB02	61-65	White	Positive	Positive	Negative	Positive	N
B03			PBMCB03	PlasmaB03	41-45	White	Positive	Positive	Negative	Positive	N
B04		FFPEB04	PBMCB04	PlasmaB04	66-70	Black	Negative	Negative	Negative	Negative	N
B05			PBMCB05	PlasmaB05	46-50	Black	Negative	Negative	Negative	Negative	N
B06		FFPEB06	PBMCB06	PlasmaB06	46-50	Black	Positive	Positive	Positive	Positive	N
I01		FFPEI01	PBMCI01	PlasmaI01	71-75	White	Positive	Negative	Negative	Positive	Y
I02		FFPEI02	PBMCI02	PlasmaI02	66-70	White	Negative	Negative	Negative	Negative	Y
I03		FFPEI03	PBMCI03	PlasmaI03	41-45	White	Negative	Negative	Negative	Negative	Y
I04		FFPEI04	PBMCI04	PlasmaI04	56-60	White	Positive	Positive	Negative	Positive	Y
I05		FFPEI05	PBMCI05	PlasmaI05	21-25	White	Positive	Negative	Negative	Positive	Y
I06		FFPEI06	PBMCI06	PlasmaI06	51-55	White	Positive	Positive	Negative	Positive	Y
I07		FFPEI07	PBMCI07	PlasmaI07	41-45	White	Positive	Positive	Negative	Positive	Y
I08		FFPEI08	PBMCI08	PlasmaI08	51-55	White	Positive	Positive	Negative	Positive	Y
I09		FFPEI09	PBMCI09	PlasmaI09	46-50	White	Positive	Negative	Negative	Positive	Y
I10		FFPEI10	PBMCI10	PlasmaI10	56-60	White	Positive#	Negative	Negative	Positive	Y
H01	FTH01				46-50	White					
H02	FTH02				61-65	Unknown					
H03	FTH03				56-60	White					
H04	FTH04				71-75	White					
B01	FTB01				46-50	White	Negative	NA	Negative	NA	N
B02	FTB02				61-65	Unknown	Positive	Positive	Negative	Positive	N
B03	FTB03				56-60	White	Positive	Positive	Negative	Positive	N
B04	FTB04				71-75	White	Negative	NA	Negative	NA	N

*: Donor ID: H, healthy; B, non-IBC; I, IBC

**: Dataset ID: PBMC, peripheral blood mononuclear cells; FFPE, formalin-fixed paraffin-embedded breast tissue; FT, forzen breast tissue; H, healthy; B, non-IBC; I, IBC

#: ER 1%, classified as low positive in a grading system which 1-10% is low positive and 10-100% is positive. All other ER positive FFPE tumors were between 30 and 100

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table S2. Intron Depth Ratio Summary Statistics

(A)

Group	Overall IDR Percentile* 25th	50th (Median)	75th	% IDR≥0.5
FT Healthy	0.06	0.14	0.28	13.9
FT non-IBC	0.05	0.10	0.23	12.3
FFPE non-IBC	0.09	0.20	0.43	19.5
FFPE IBC	0.09	0.19	0.39	18.4
PBMC Healthy	0.05	0.10	0.26	14.6
PBMC non-IBC	0.06	0.14	0.35	17.3
PBMC IBC	0.10	0.23	0.55	24.6
Plasma Healthy	0.10	0.37	>10	36.5
Plasma non-IBC	0.02	0.23	>10	30.0
Plasma IBC	0.18	1.34	>10	48.7

(B) Overexpressed (OE) Protein-coding Genes, *p*≤0.001, LFC≥2.

Comparison	OE Genes: #**	OE Genes w/ IDR≥0.5: # (%)
Frozen non-IBC vs. Healthy	20	2 (10.0)
Frozen Healthy vs. non-IBC	148	17 (11.5)
FFPE IBC vs. non-IBC	28	5 (17.9)
FFPE non-IBC vs. IBC	0	-
PBMC IBC vs. Healthy	498	297 (59.6)
PBMC Healthy vs. IBC	32	5 (15.6)
PBMC IBC vs. non-IBC	194	120 (61.9)
PBMC non-IBC vs. IBC	3	0 (0)
PBMC non-IBC vs. Healthy	0	-
PBMC Healthy vs. non-IBC	3	3 (100)
Plasma IBC vs. Healthy	65	37 (56.9)
Plasma Healthy vs. IBC	126	29 (23.0)
Plasma IBC vs. non-IBC	86	55 (64.0)
Plasma non-IBC vs. IBC	115	14 (12.2)
Plasma non-IBC vs. Healthy	41	15 (36.6)
Plasma Healthy vs. non-IBC	46	21 (45.7)

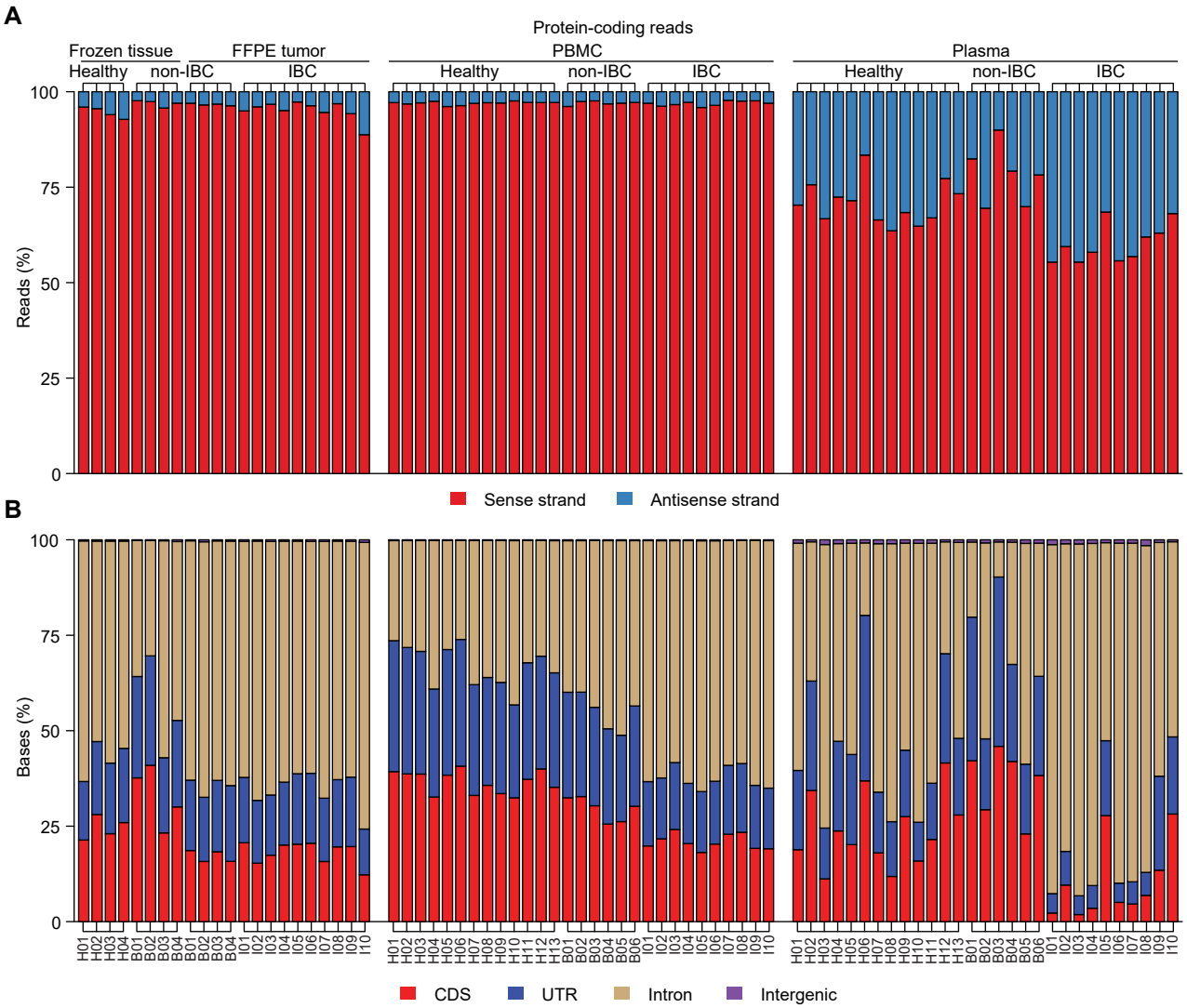
(C) OE Protein-coding Genes, *p*≤0.05, LFC≥1.

Comparison	OE Genes: #**	OE Genes w/ IDR≥0.5: # (%)
Frozen non-IBC vs. Healthy	272	10 (3.7)
Frozen Healthy vs. non-IBC	552	49 (8.7)
FFPE IBC vs. non-IBC	204	55 (27.0)
FFPE non-IBC vs. IBC	10	0 (0)
PBMC IBC vs. Healthy	2346	975 (41.6)
PBMC Healthy vs. IBC	1262	163 (12.9)
PBMC IBC vs. non-IBC	1268	578 (45.6)
PBMC non-IBC vs. IBC	368	43 (11.7)
PBMC non-IBC vs. Healthy	29	8 (27.6)
PBMC Healthy vs. non-IBC	153	17 (11.1)
Plasma IBC vs. Healthy	280	192 (68.6)
Plasma Healthy vs. IBC	406	111 (27.3)
Plasma IBC vs. non-IBC	286	212 (74.1)
Plasma non-IBC vs. IBC	416	43 (10.3)
Plasma non-IBC vs. Healthy	49	16 (32.7)
Plasma Healthy vs. non-IBC	80	37 (46.2)

*IDR Percentiles are values for which 25, 50, or 75%, respectively, of protein-coding genes have an IDR ≤ indicated number.
**Overexpressed gene numbers here include only those genes for which IDR could be assigned by comparing genomic and transcriptomic mapping results.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure S1

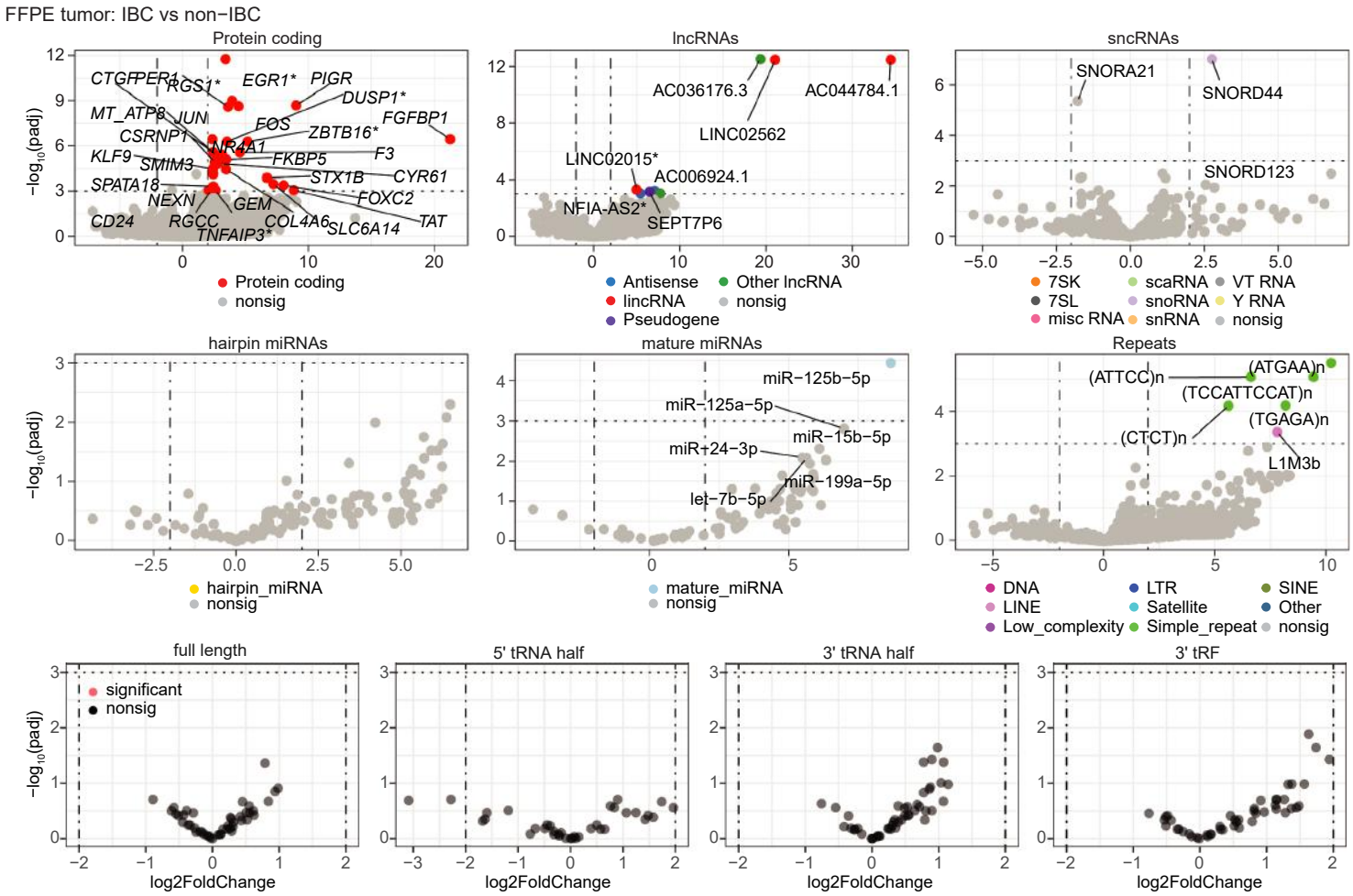


medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. S1. Analysis of reads mapping to protein-coding genes. (**A**) Percentage of reads mapped to the sense or antisense stand of protein-coding genes. (**B**) Percentage of bases of protein-coding gene reads mapped to coding sequences (CDS), untranslated regions (UTR), introns, or intergenic regions. Sample tissue type (Frozen, FFPE, PBMC, or Plasma) and disease status (Healthy, non-IBC, or IBC) are indicated above, and sample names corresponding to healthy donor or patient ID numbers are indicated below (H, healthy; B, non-IBC; I, IBC).

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure S2



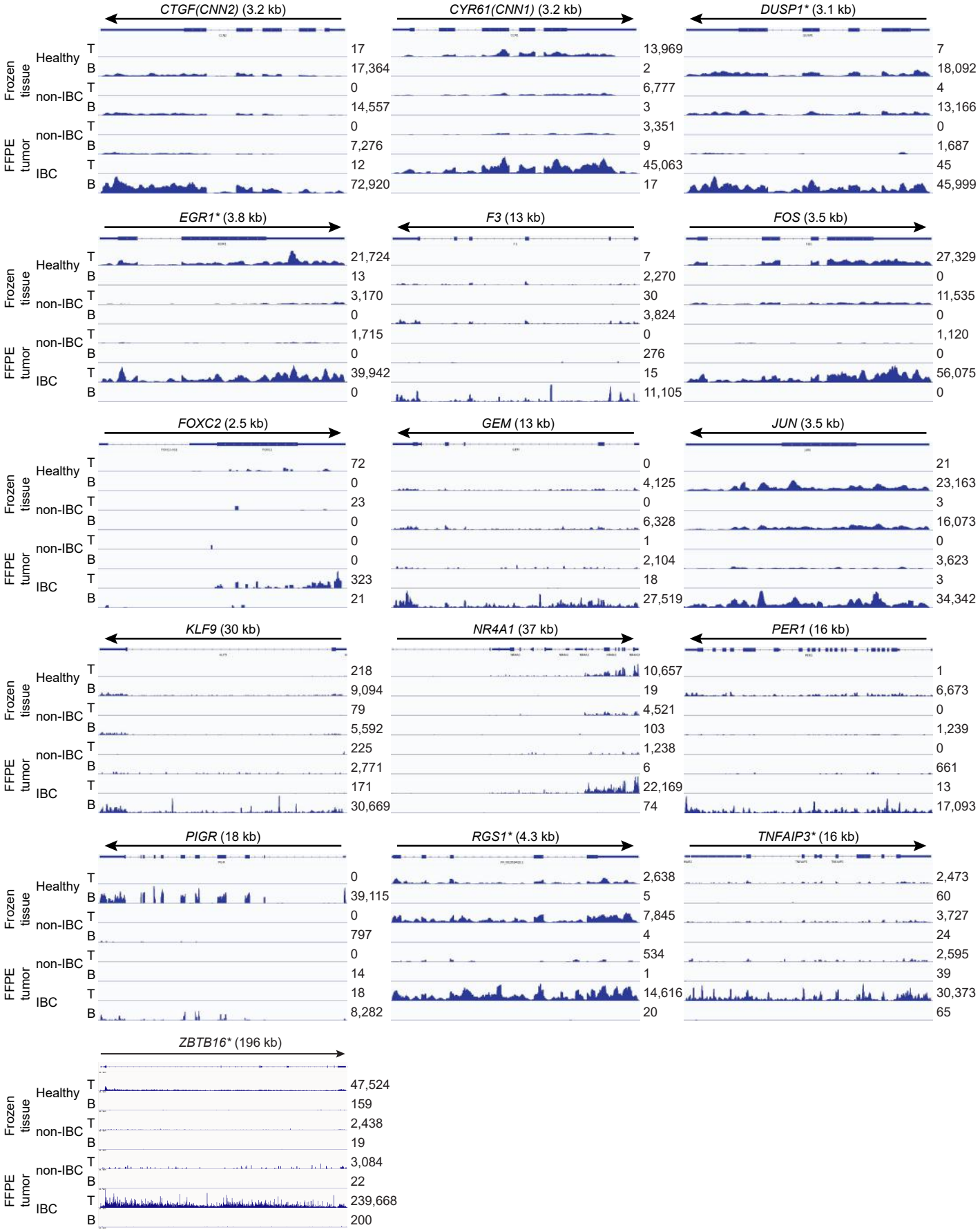
medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. S2. Volcano plots for DESeq2 comparisons of FFPE breast tissue samples. The *y*-axis indicates the log-transformed adj. *p*-value while the *x*-axis indicates the effect size (DESeq2-estimated LFC). Each point represents one gene with biotypes color coded for those genes with adj. $p \leq 0.001$ and $|LFC| \geq 2$. Significantly overrepresented and other notable genes that passed filtering steps and were detected in at least 50% of the samples of the group in which they were overrepresented are labeled by their gene symbol.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

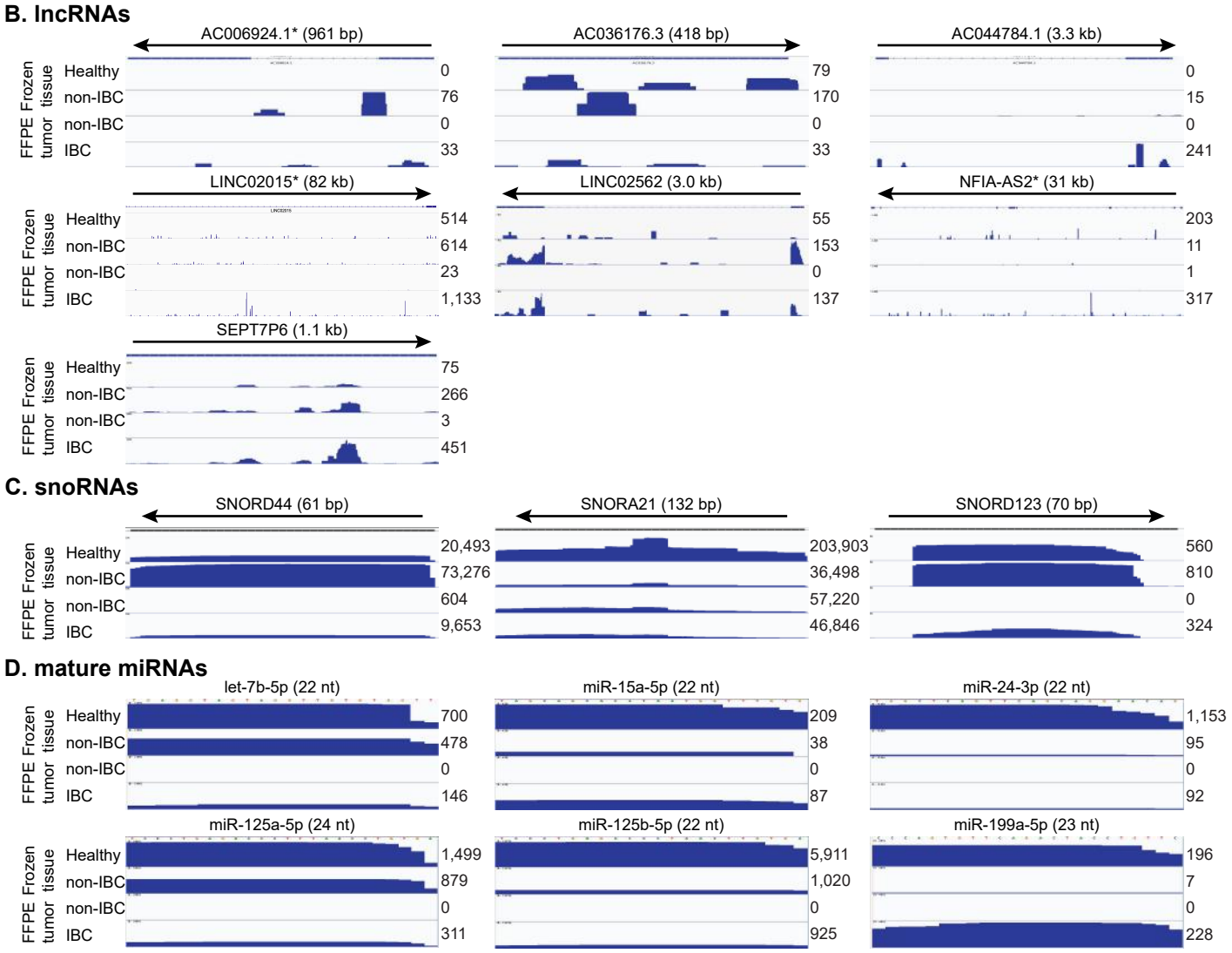
Figure S3

A. Protein-coding genes



medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure S3



medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. S3. IGV plots of differentially expressed or notable genes in breast tissue samples. (**A**) protein-coding genes, (**B**) lncRNAs, (**C**) snoRNAs, and (**D**) mature miRNAs. T and B, top strand and bottom DNA strands, respectively. The number of reads for RNAs in each IGV plot is indicated to the right. The primary comparison in Results is for FFPE IBC versus FFPE non-IBC tumors. Read counts between FFPE and Frozen tissue samples are not directly comparable.

Figure S4

A PBMC: IBC vs Healthy

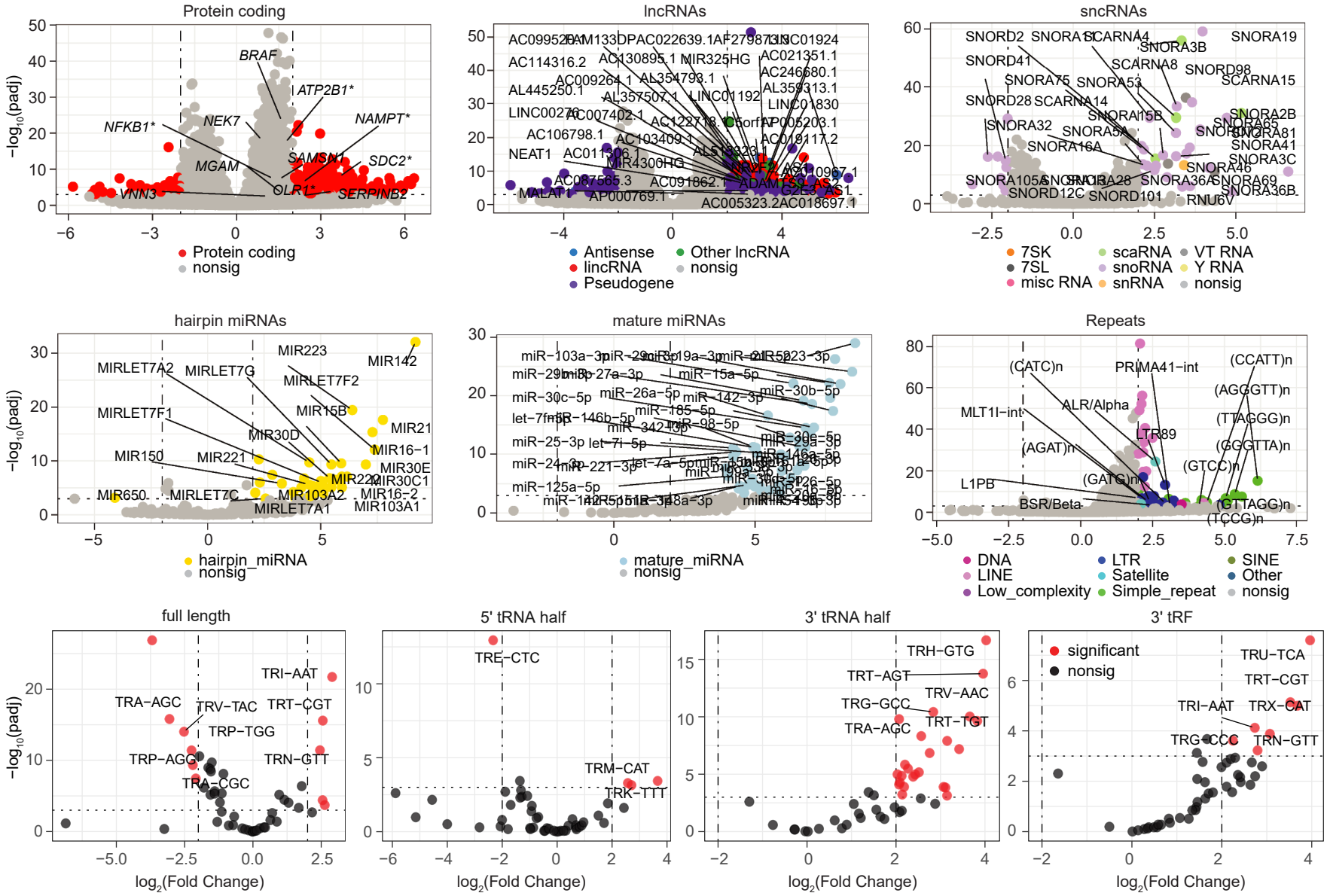
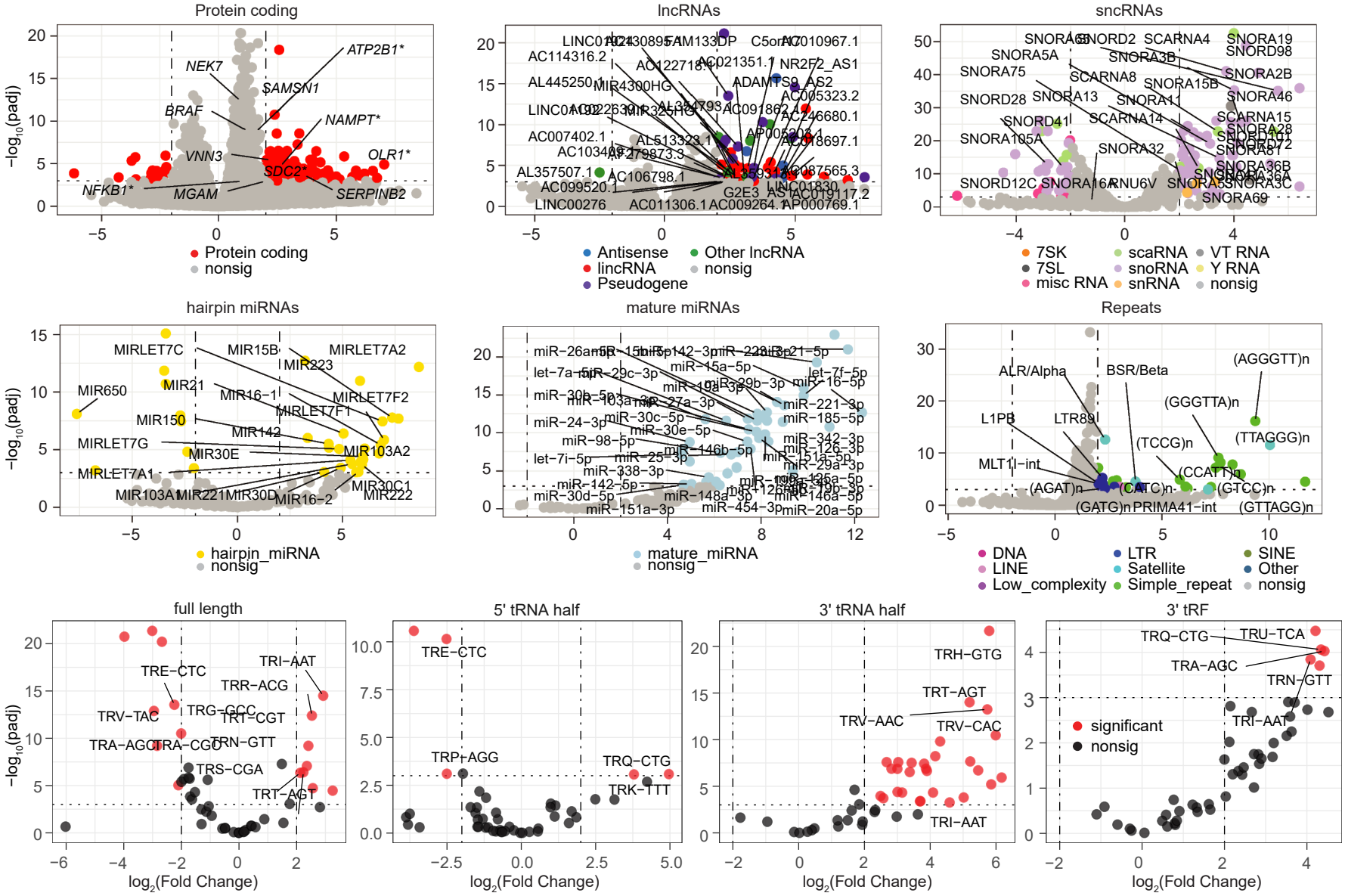


Figure S4

B PBMC: IBC vs non−IBC



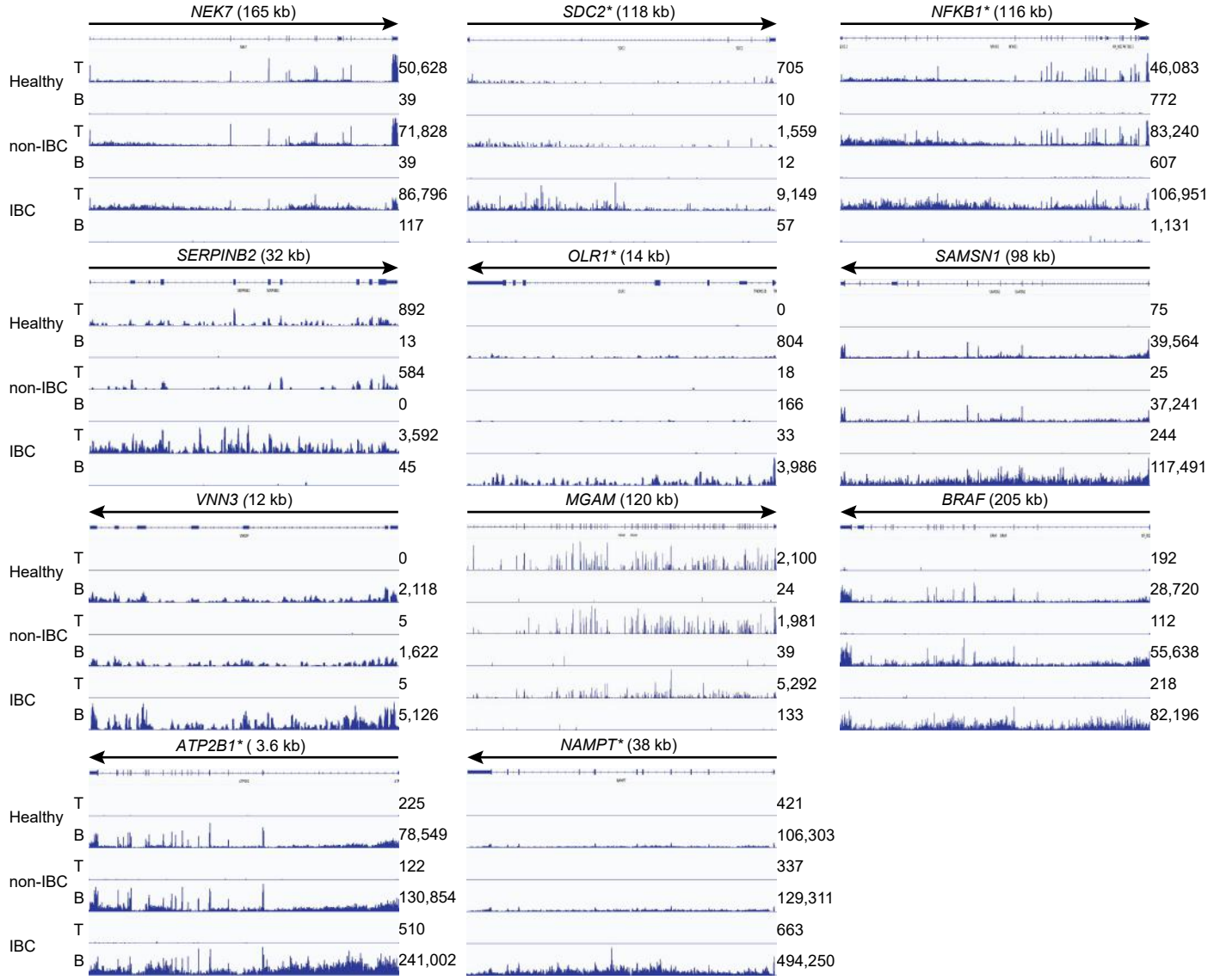
medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. S4. Volcano plots for DESeq2 comparisons of PBMC samples. (**A**) IBC patient vs Healthy donor. (**B**) IBC vs non-IBC patients. The *y*-axis indicates log-transformed adj. *p* while x-axis indicates LFC. Each point represents one gene with biotype color coded for those genes with adj. $p \leq 0.001$ and $|LFC| \geq 2$. Significantly overrepresented and other notable genes that passed filtering steps and were detected in at least 50% of the samples of the group in which they were overrepresented are labeled by their gene symbol. Protein-coding genes with IDR$\geq$0.5 are marked with an asterisk as done elsewhere in the text and Figures; lncRNA genes with IDR$\geq$0.5 were not marked with an asterisk due to crowding of gene names (see Fig. 4 and Supplemental file for IDR status).

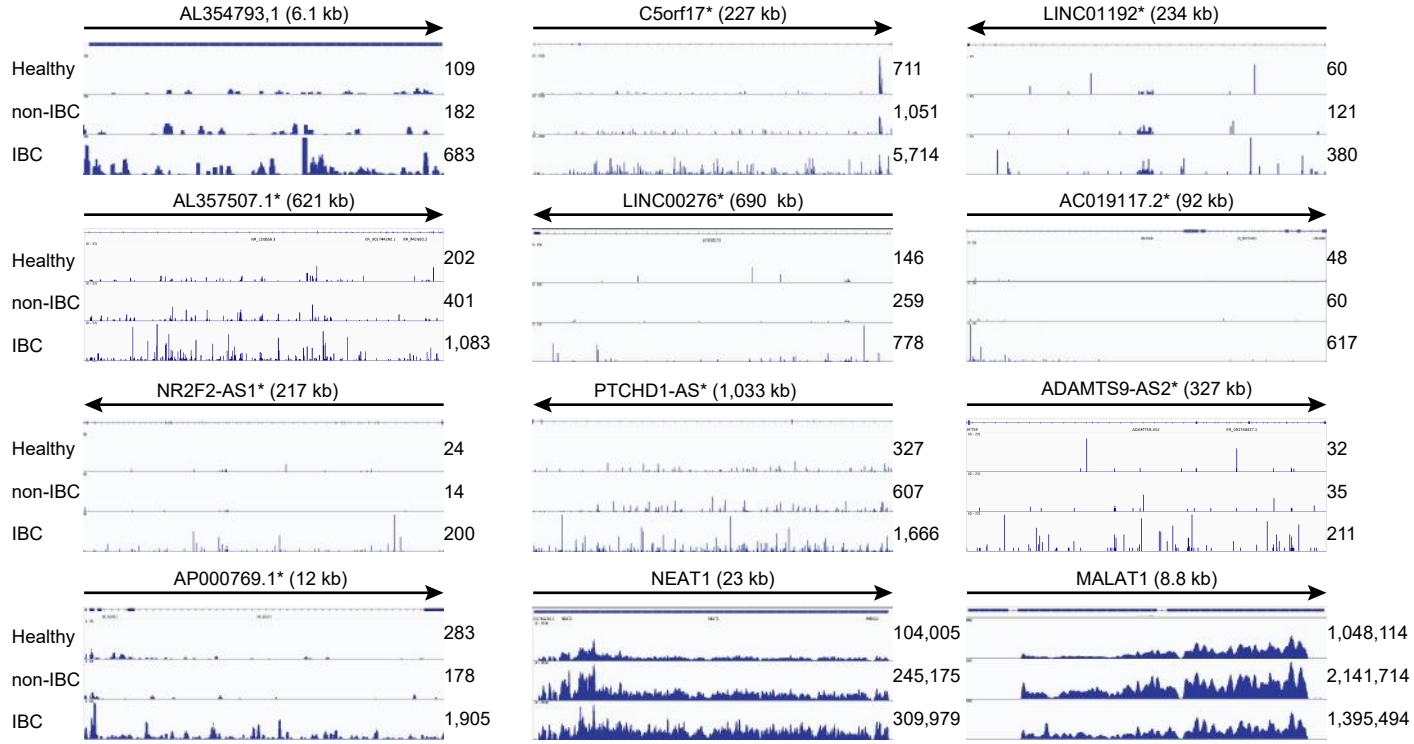
medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure S5

A. Protein-coding genes



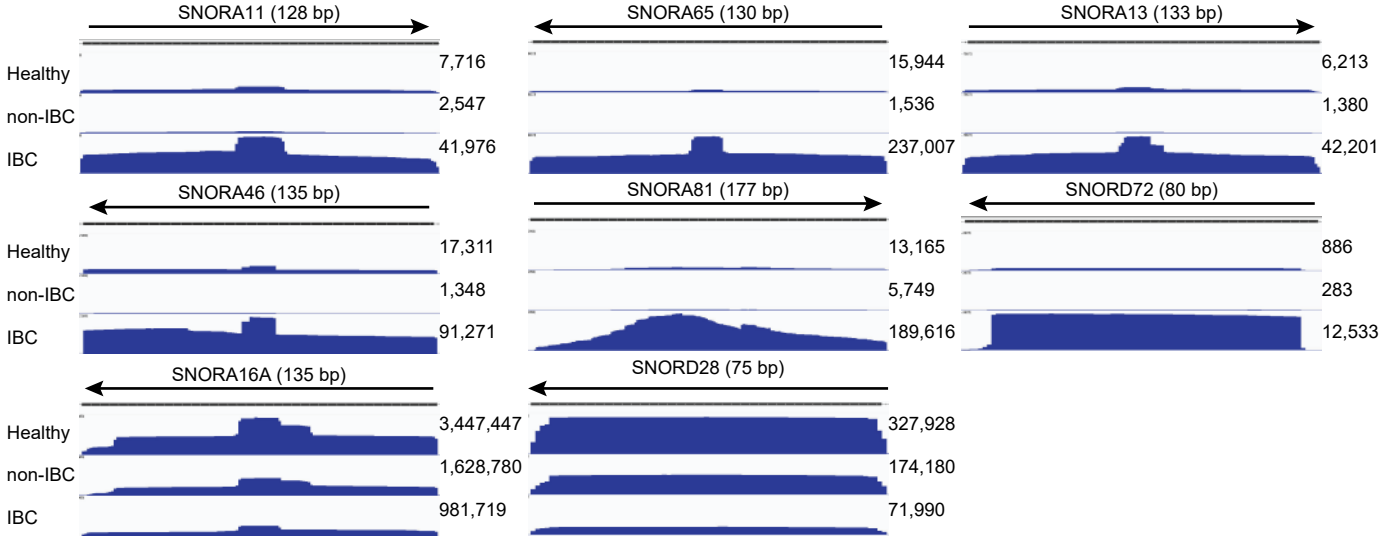
B. lncRNAs



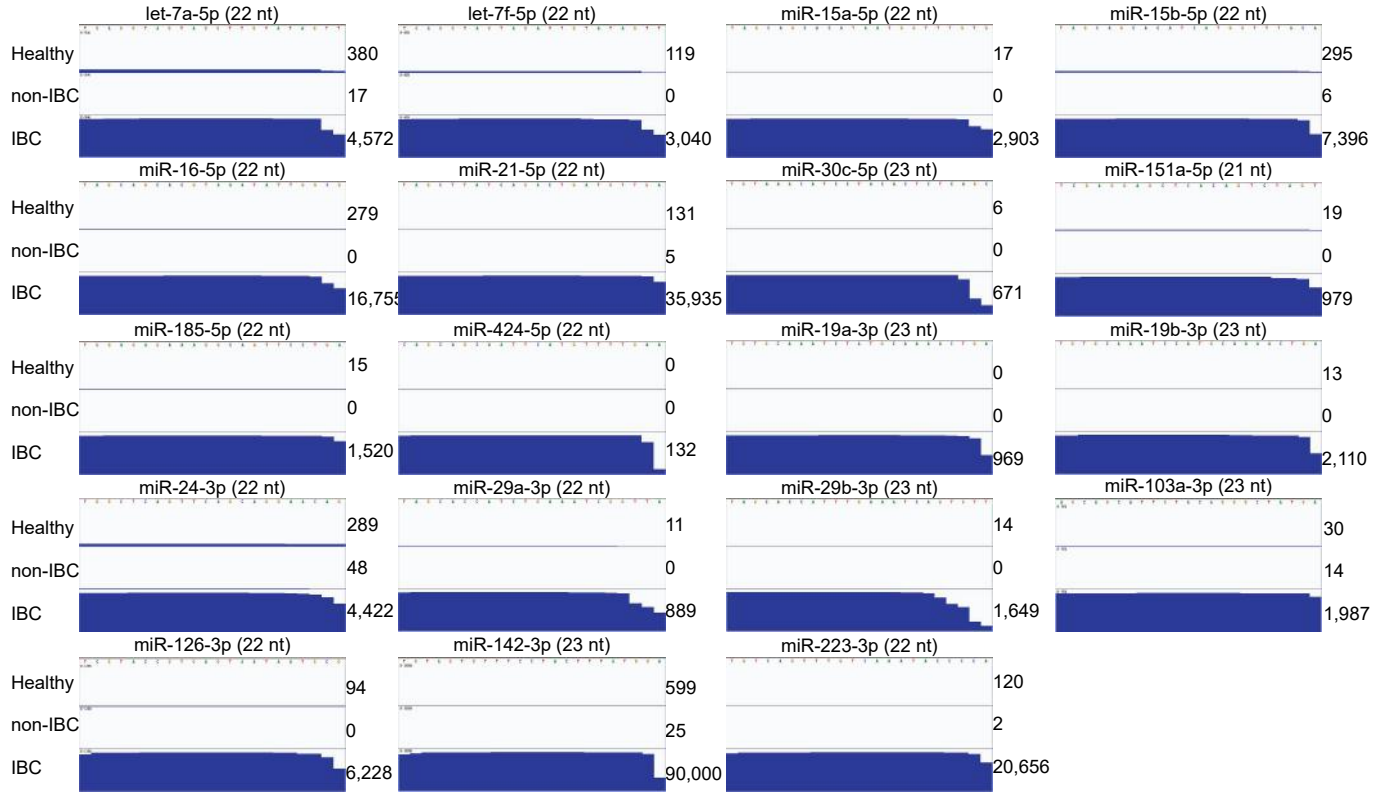
medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure S5

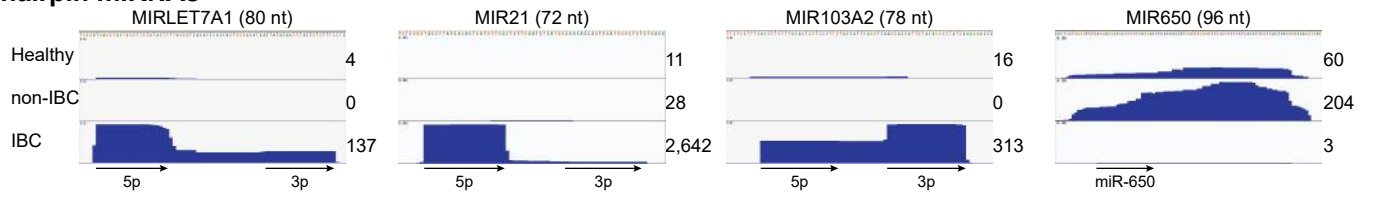
C. sncRNAs



D. mature miRNAs



E. hairpin miRNAs



medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. S5. IGV plots of differentially expressed or notable genes in PBMC samples. (**A**) protein-coding genes, (**B**) lncRNAs, (**C**) sncRNAs, (**D**) mature miRNAs, and (**E**) hairpin miRNAs. T and B, top strand and bottom DNA strands, respectively. The number of reads for RNAs in each IGV plot is indicated to the right.

Figure S6

A Plasma: IBC vs Healthy + non-IBC

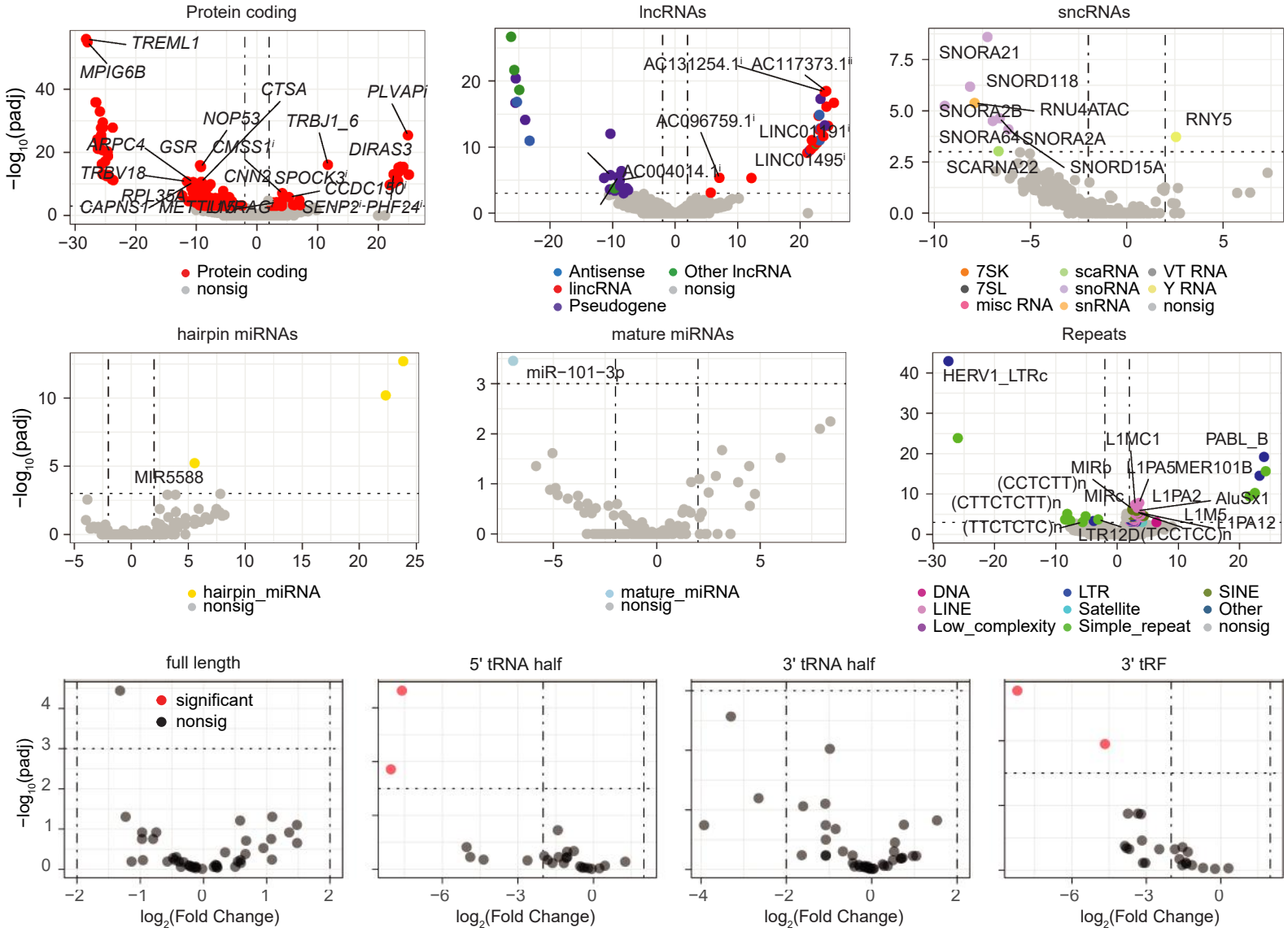


Figure S6

B Plasma: IBC vs Healthy

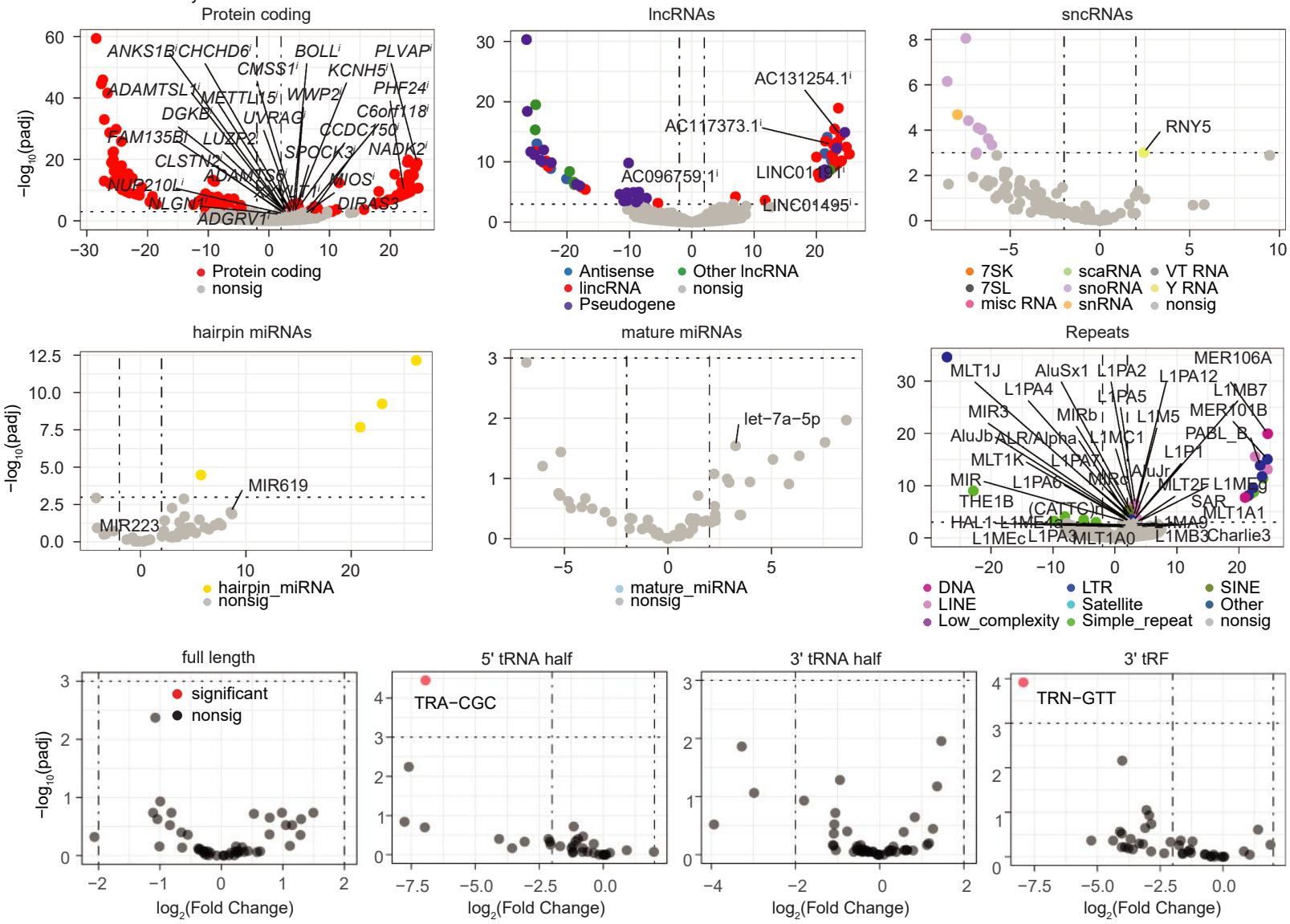
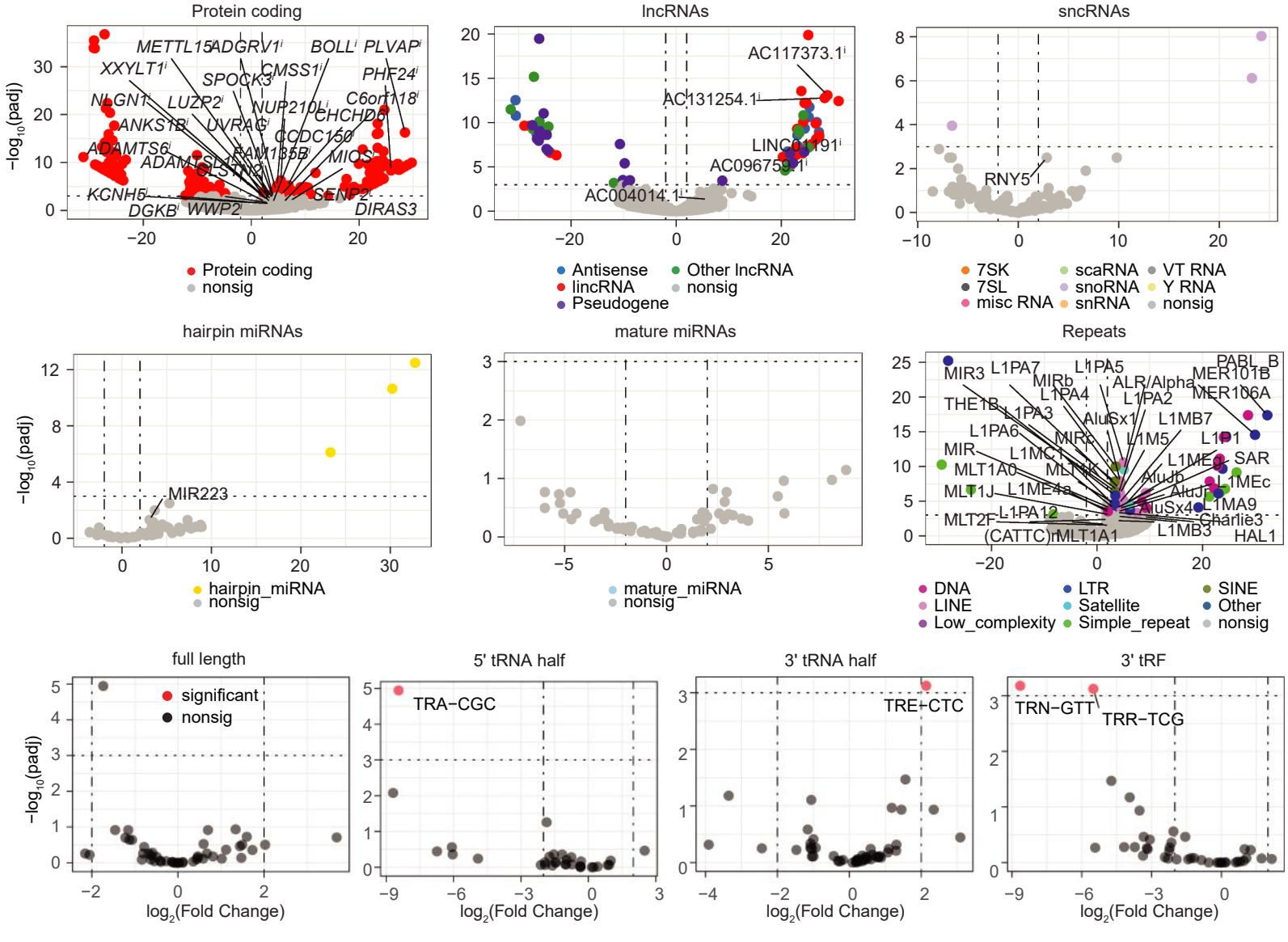


Figure S6

C Plasma: IBC vs non−IBC



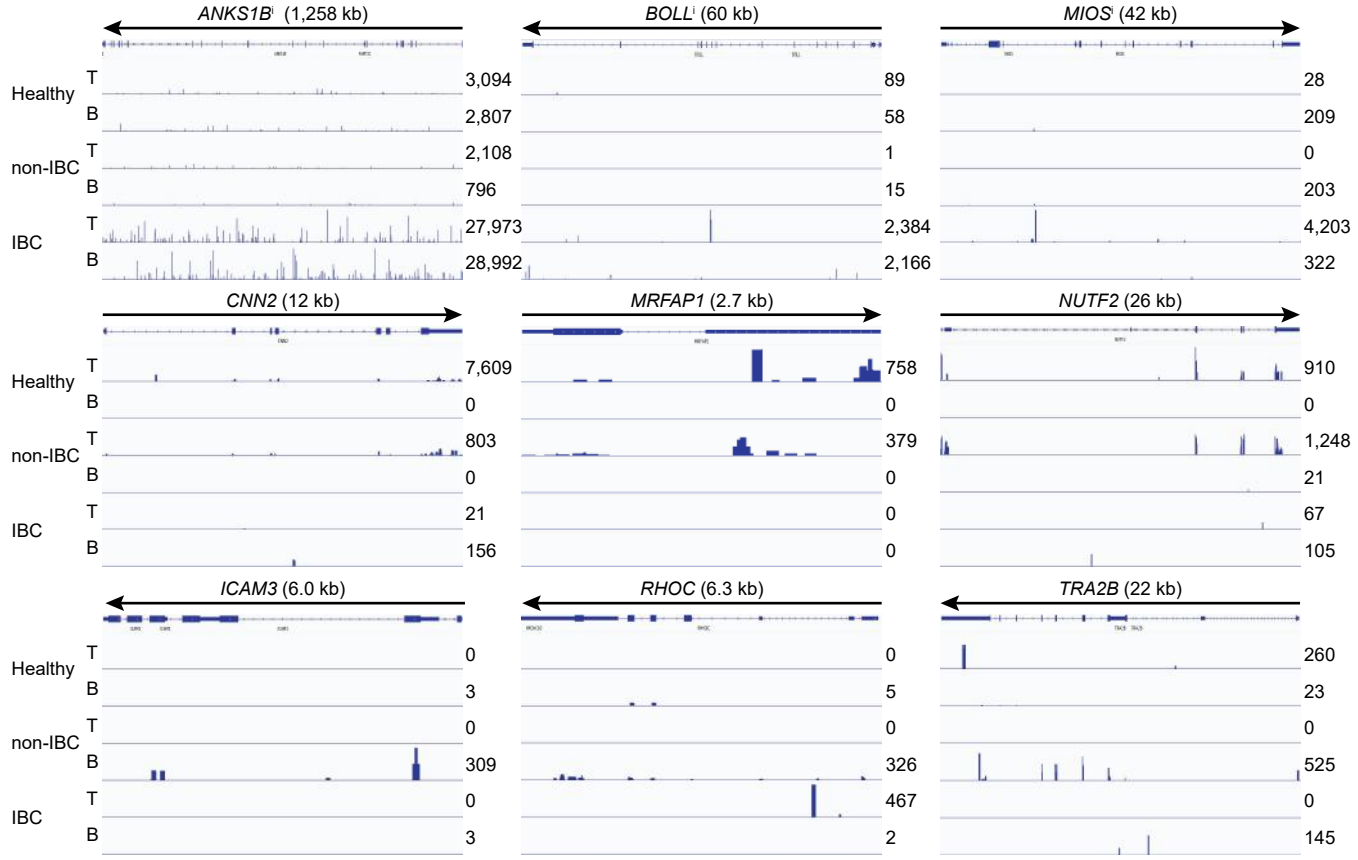
medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. S6. Volcano plots for DESeq2 two-group comparisons in plasma samples. The *y*-axis indicates log-transformed adj. *p* while x-axis indicates LFC. Each point represents one gene with biotypes color coded for those genes with adj. $p \leq 0.001$ and $|LFC| \geq 2$. Significantly overrepresented and other notable genes that passed filtering steps and were detected in at least 50% of the samples of the group in which they were overrepresented are labeled by their gene symbol.

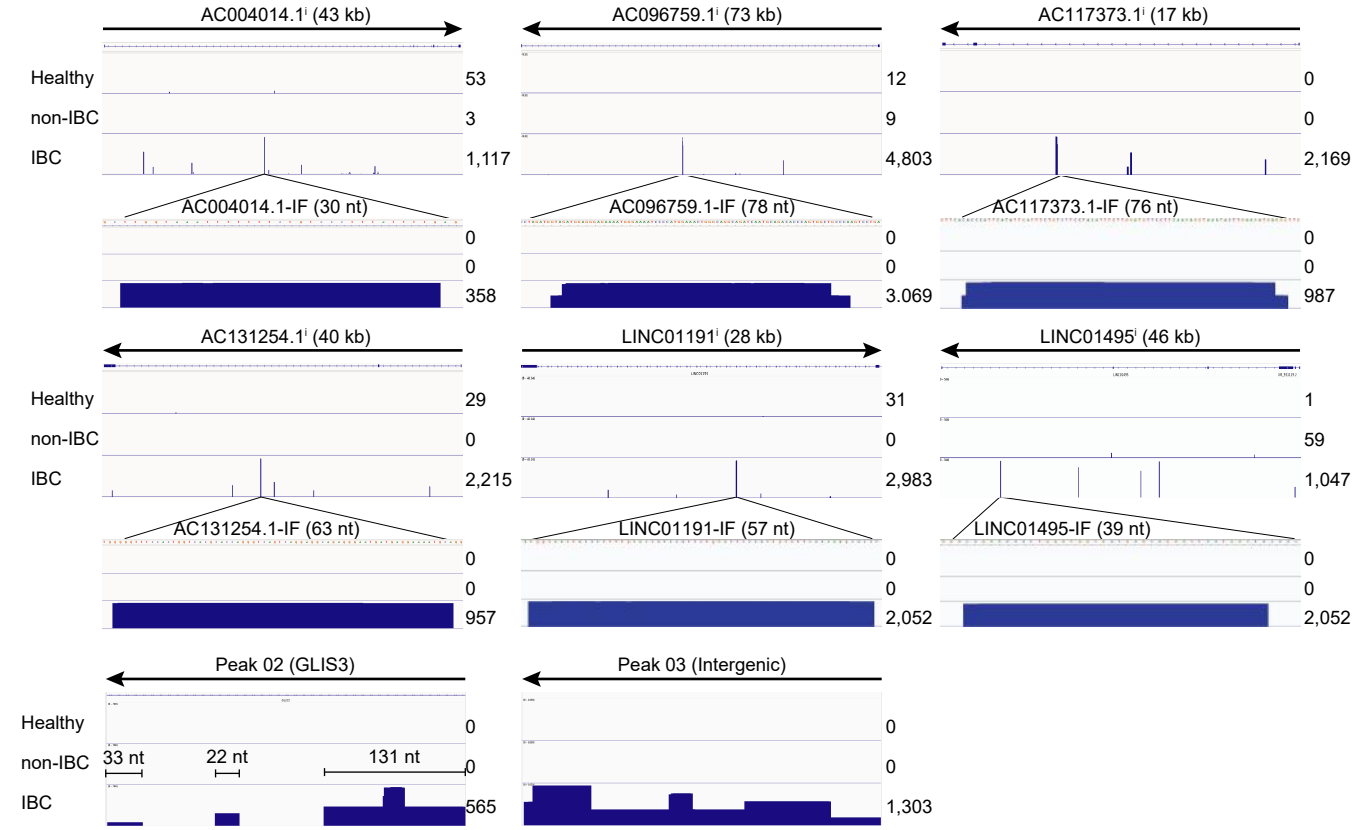
medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure S7

A. Protein-coding genes



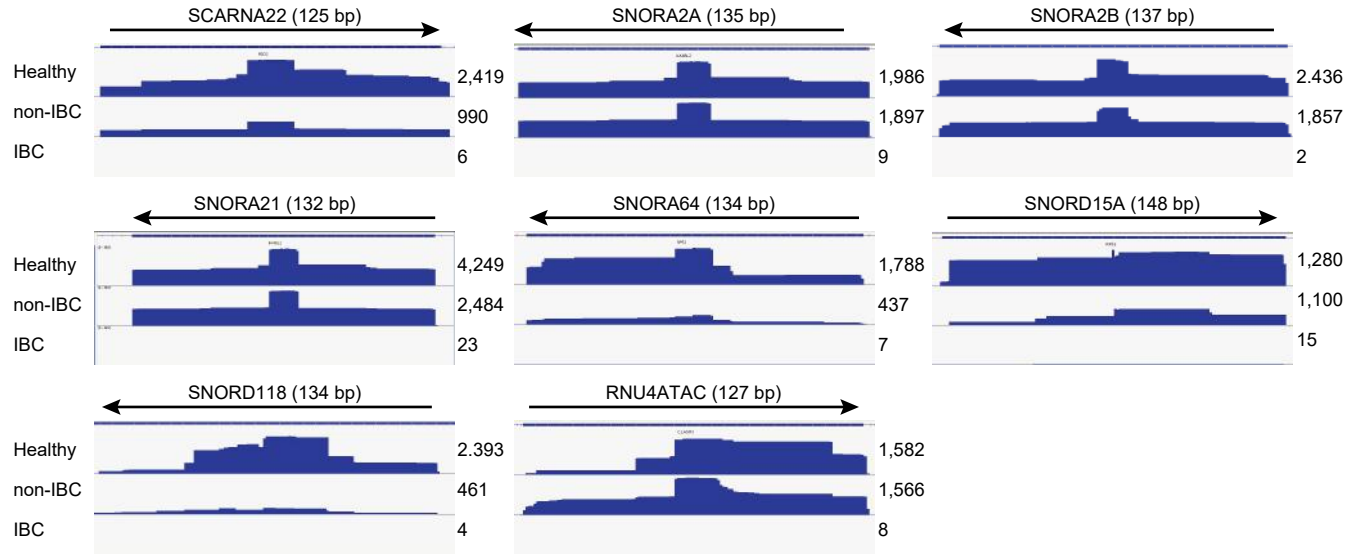
B. lncRNAs



medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure S7

C. sncRNAs



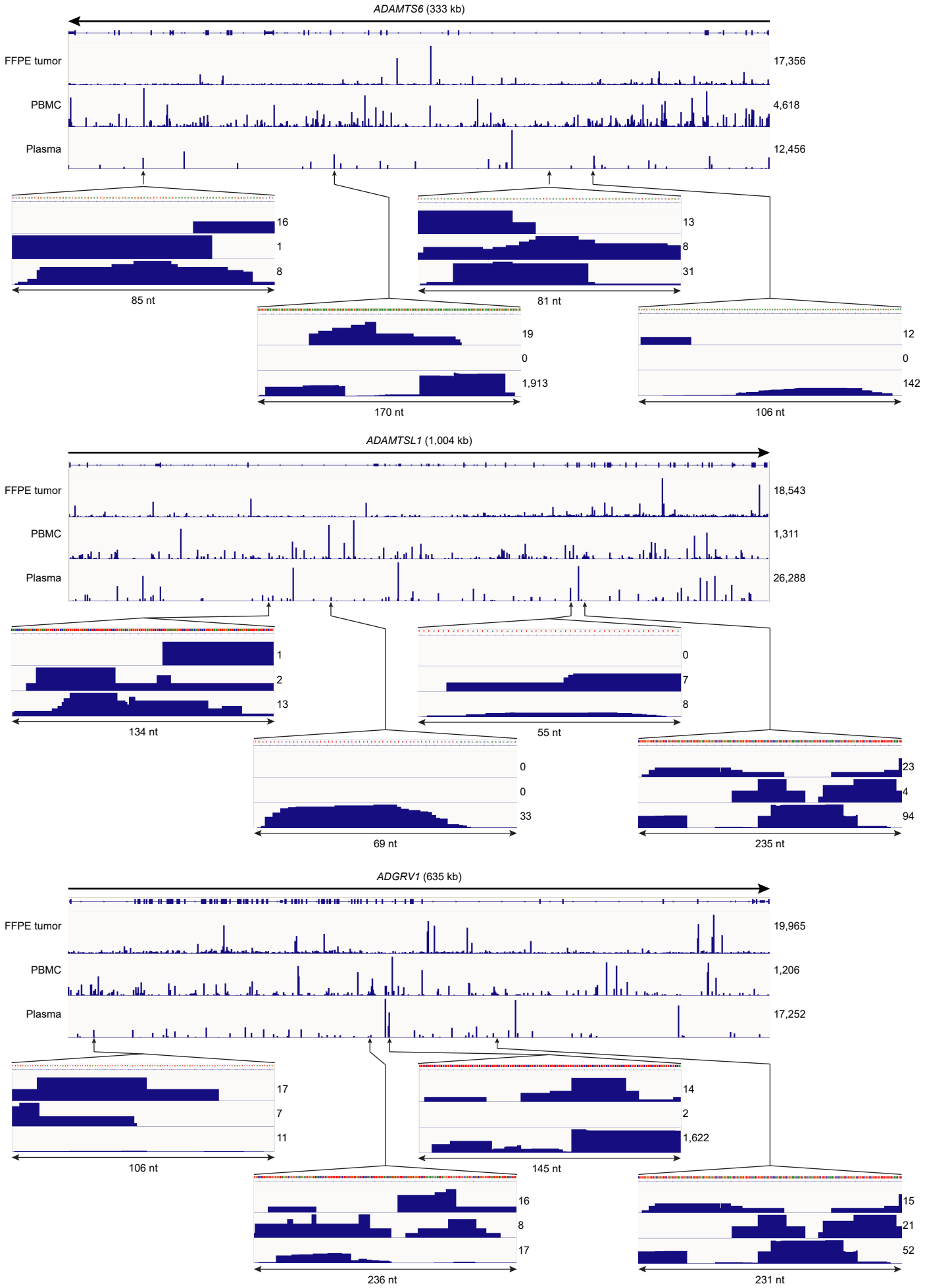
D. miRNAs

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. S7. IGV plots of differentially expressed or notable genes in plasma samples. (**A**) protein-coding genes, (**B**) lncRNAs, (**C**) sncRNAs, and (**D**) miRNAs. T and B, top and bottom DNA strands, respectively. The number of reads is indicated to the right.

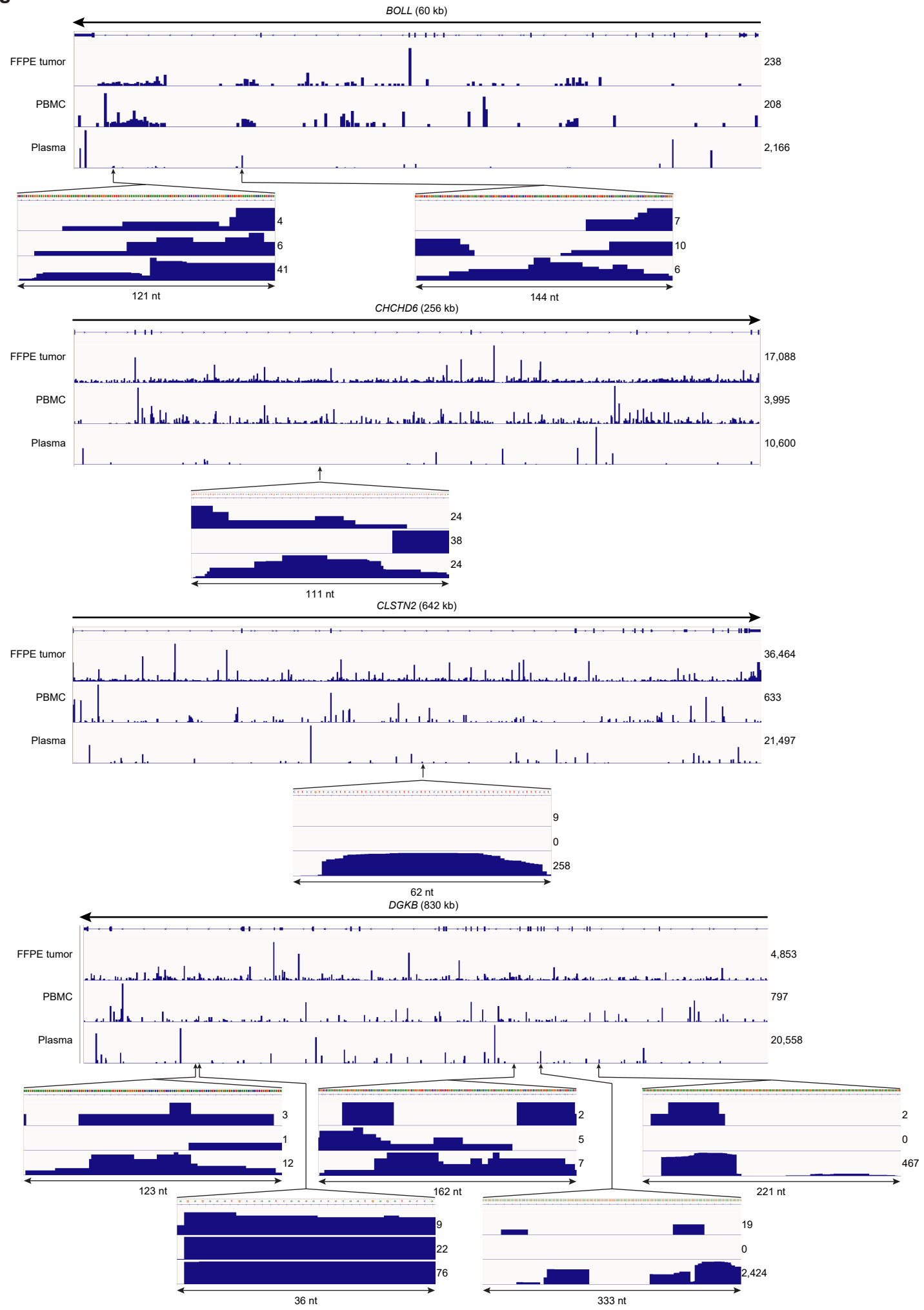
medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure S8



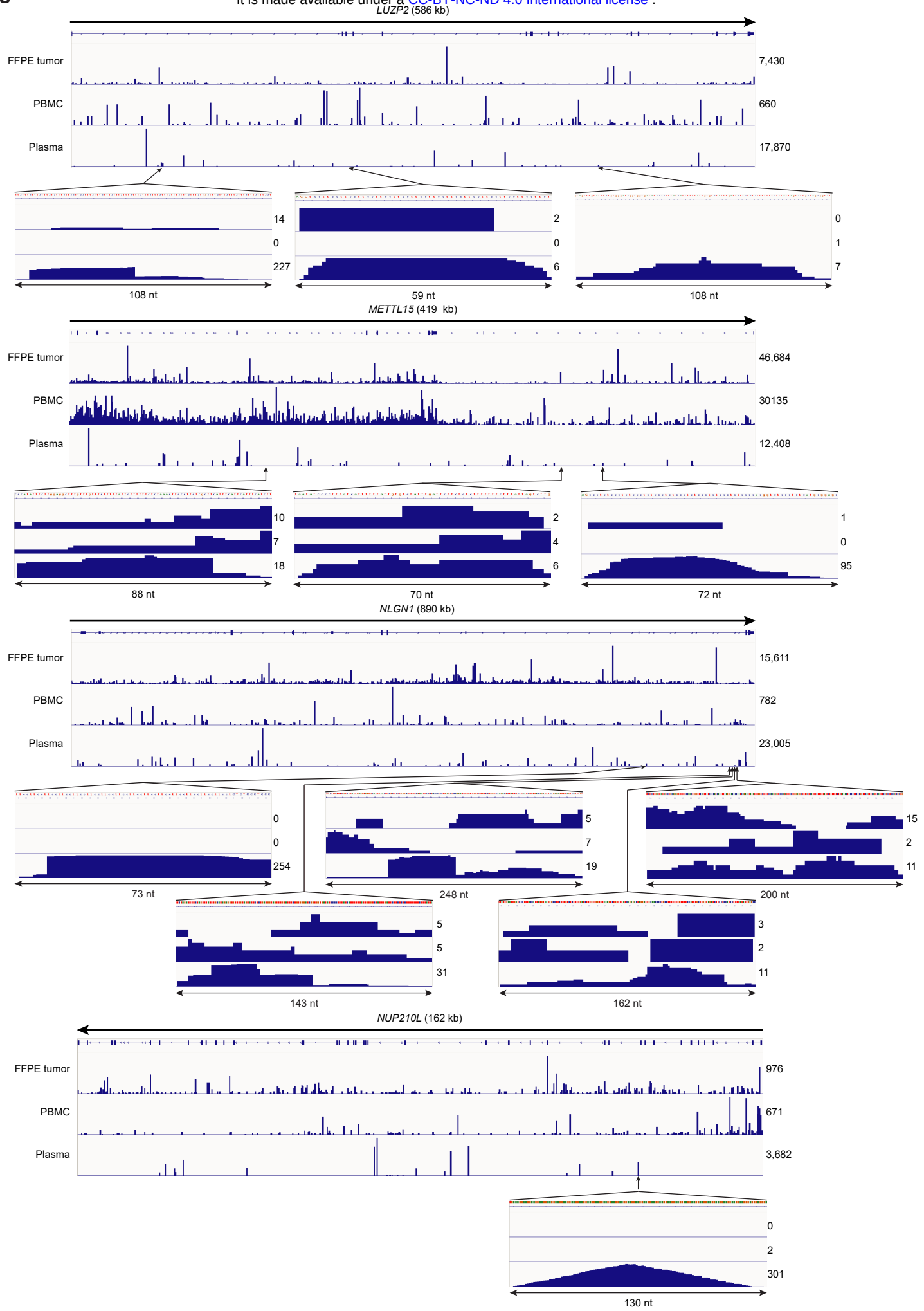
medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure S8



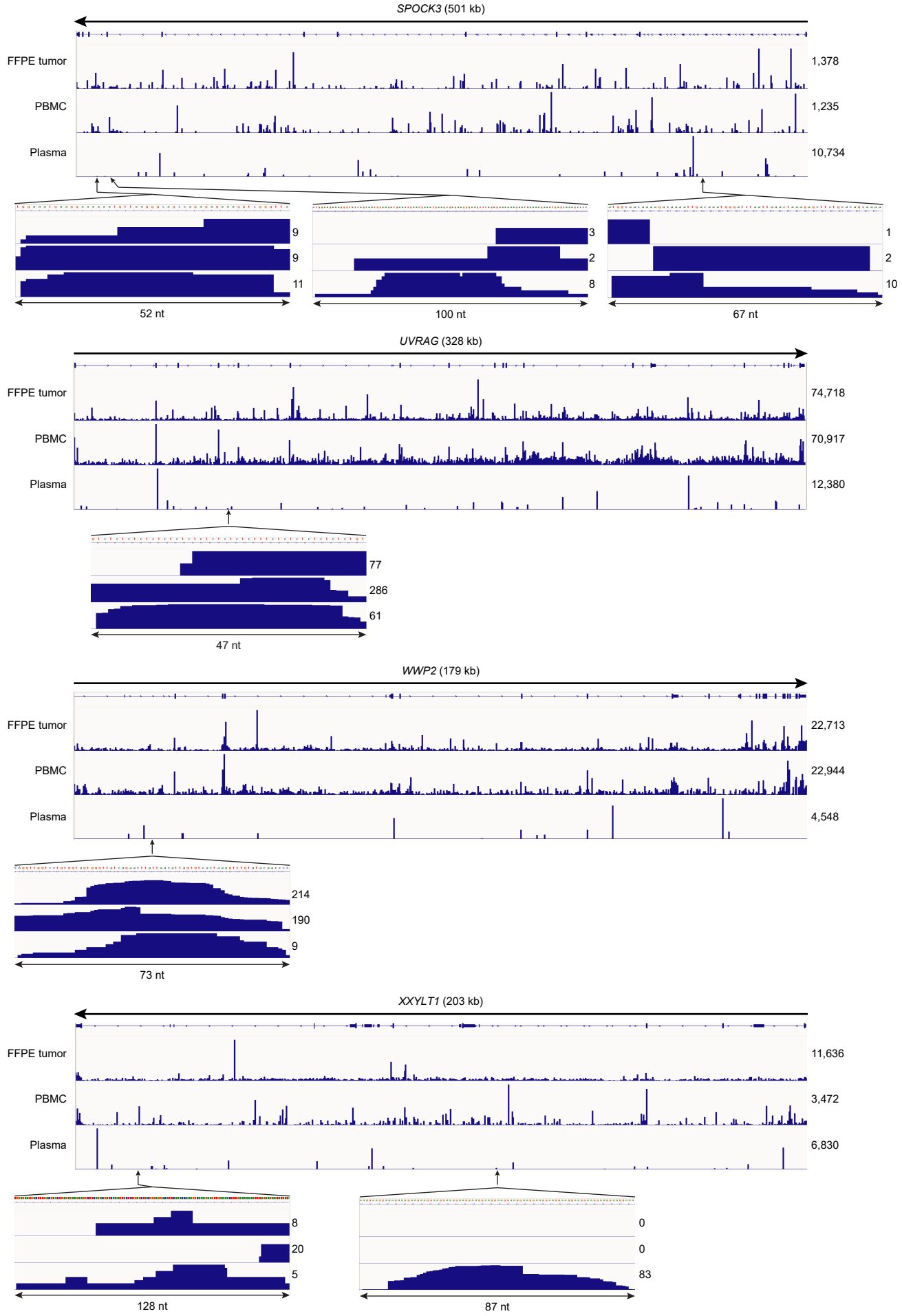
medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure S8



medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure S8



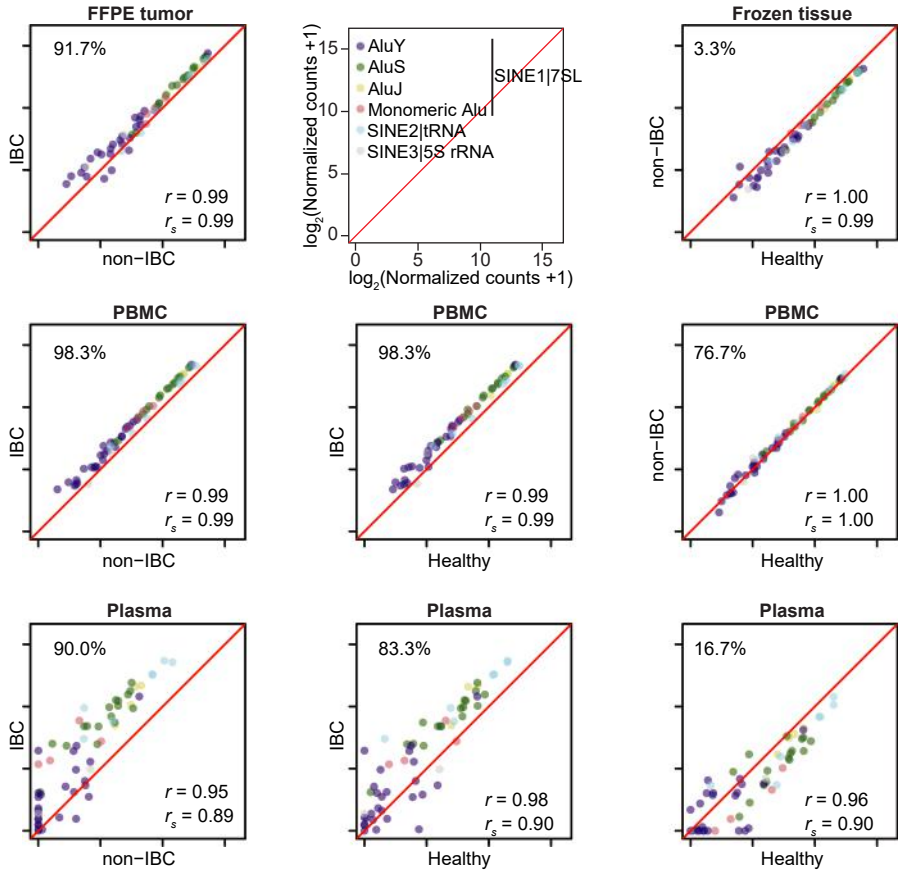
medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. S8. IGV plots for protein-coding genes whose intron RNAs were significantly over-represented in IBC plasma and for which intron RNA fragments could be traced to tumors or PBMCs. The IGV plots show coverage in combined datasets for IBC FFPE tumors, PBMC, and plasma with expanded regions at the bottom for intron RNA fragments found in 5 or more IBC plasma samples. The number of reads is indicated to the right.

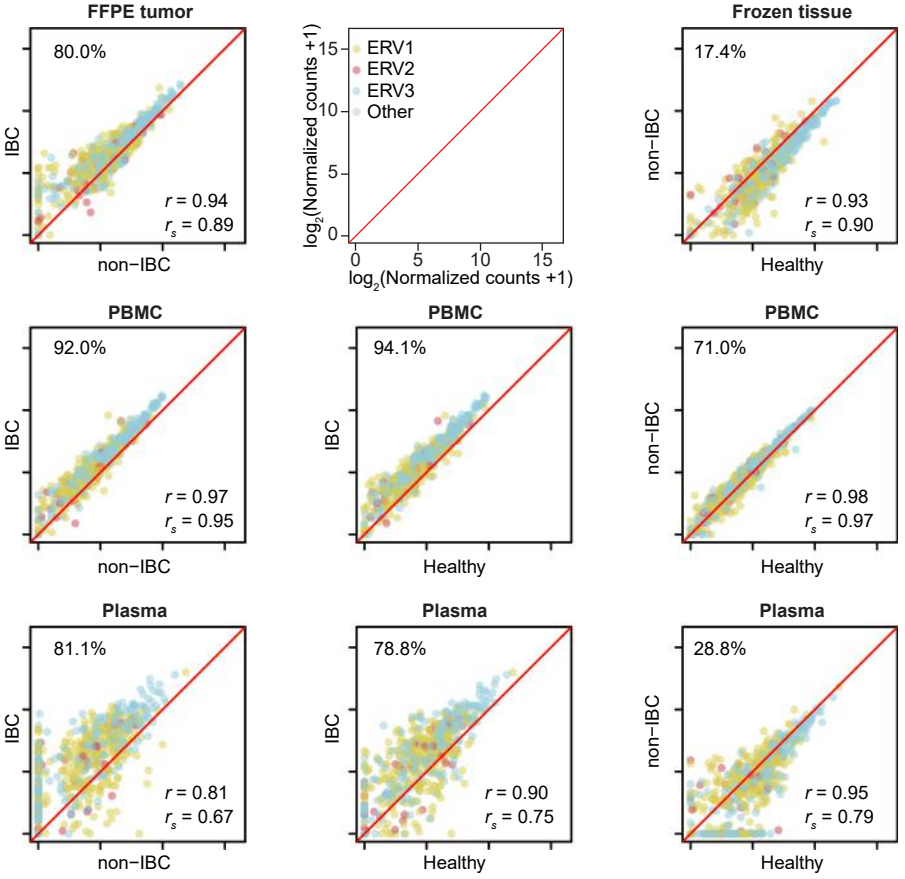
medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure S9

A. SINEs



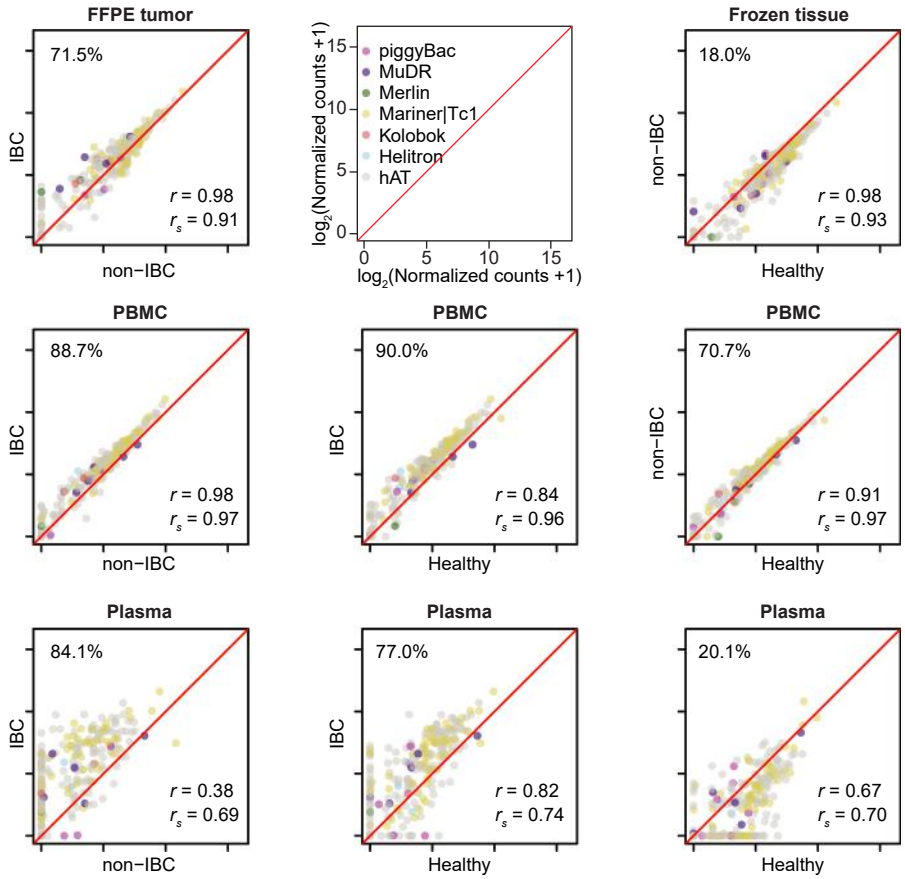
B. LTR-containing retroelements



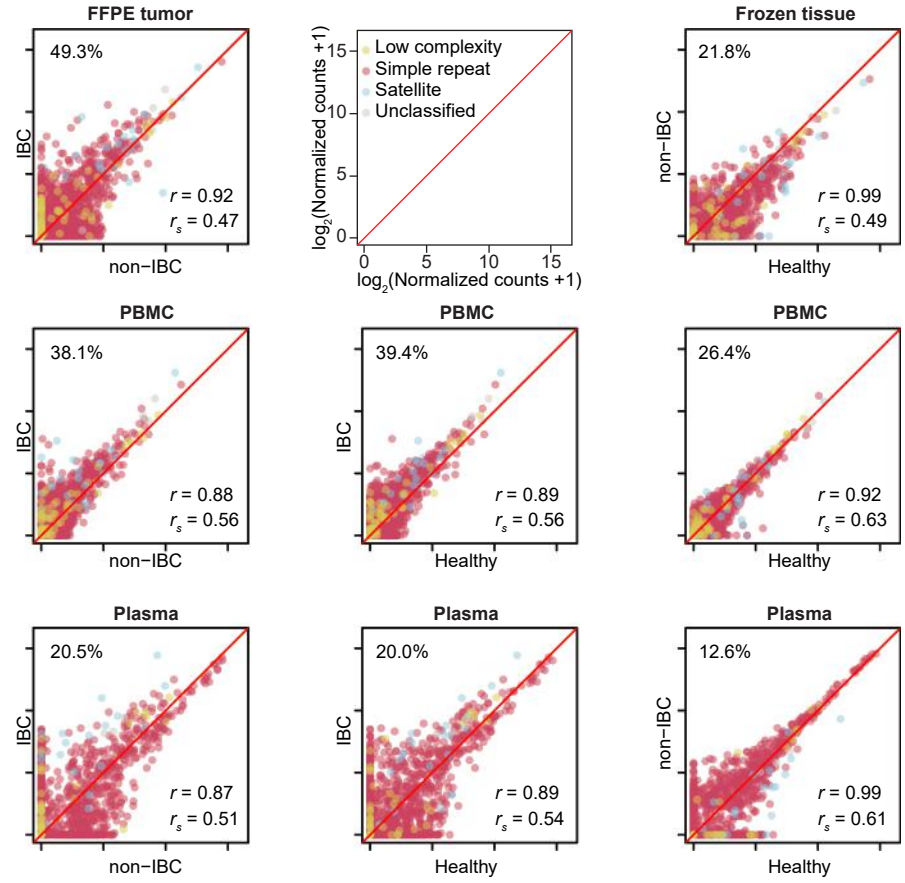
medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure S9

C. DNA elements



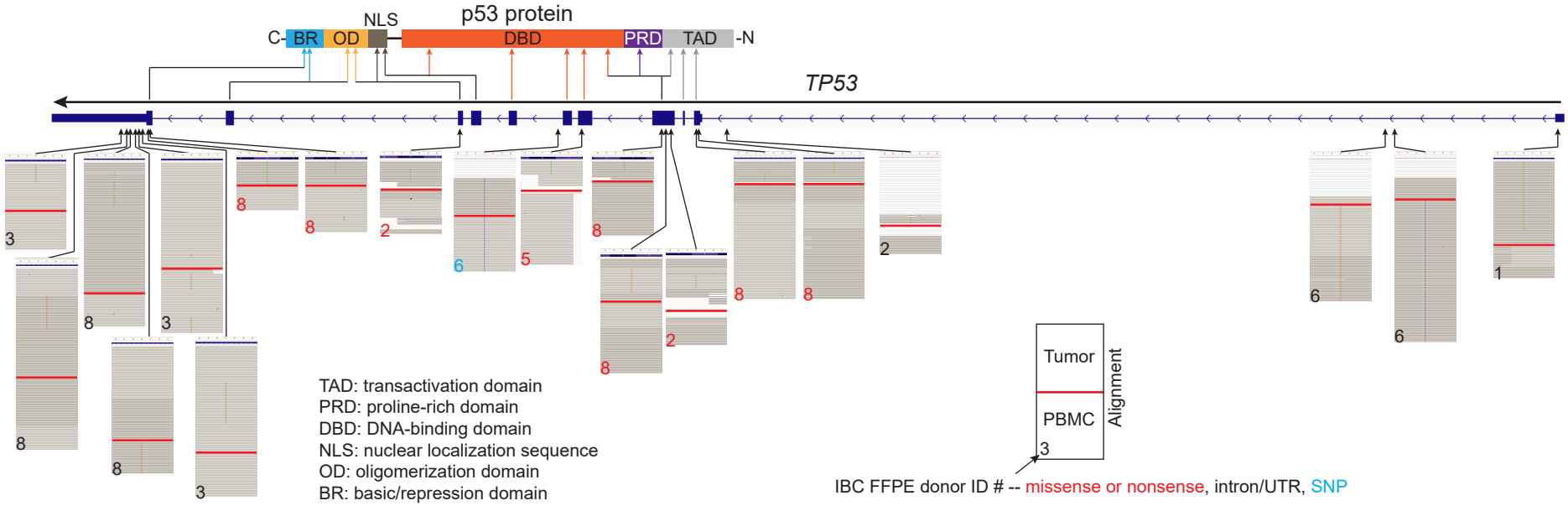
D. Other

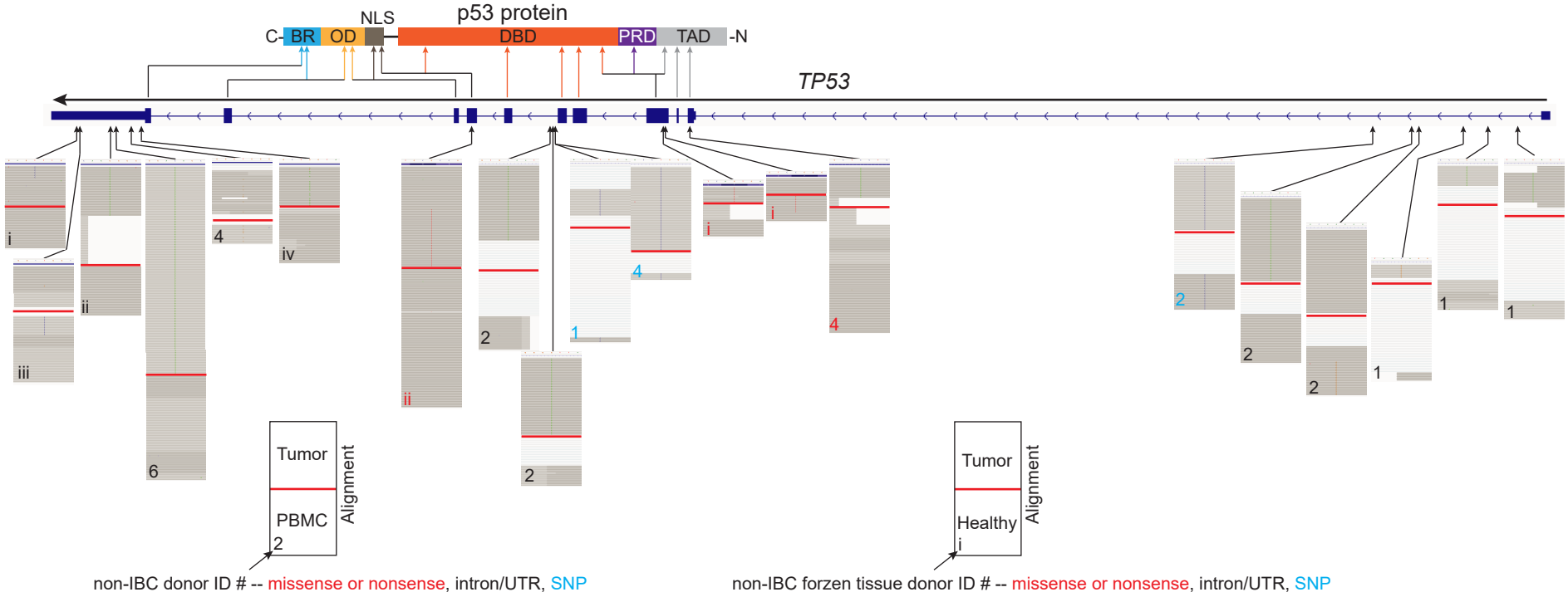


medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. S9. Scatter plots comparing the relative abundance of different Repeat-element RNAs detected in breast tissue, PBMC or Plasma. (**A**) SINE elements, (**B**) LTR-containing retroelements, (**C**) DNA elements, and (**D**) Other repeats. DESeq2-normalized counts of Repeat-elements were averaged by tissue type and disease status and $\log_2$ transformed for pairwise comparisons. The numbers at the top left of each scatterplot indicate the percentage of Repeat-element RNAs above the diagonal, a measure of over-expression of Repeat elements in each comparison. Pearson's (r) and Spearman's (r_s) correlation coefficients are shown at the bottom right. Subfamilies of the repeat elements are color coded as indicated in the top center box for each family of repeats.

Figure S10


p53 protein
C-
BR
OD
NLS
DBD
PRD
TAD
-N
TP53
TAD: transactivation domain
PRD: proline-rich domain
DBD: DNA-binding domain
NLS: nuclear localization sequence
OD: oligomerization domain
BR: basic/repression domain
Tumor
PBMC
Alignment
IBC FFPE donor ID # -- missense or nonsense, intron/UTR, SNP


p53 protein
C-
BR
OD
NLS
DBD
PRD
TAD
-N
TP53
Tumor
PBMC
Alignment
non-IBC donor ID # -- missense or nonsense, intron/UTR, SNP
Tumor
Healthy
Alignment
non-IBC forzen tissue donor ID # -- missense or nonsense, intron/UTR, SNP

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Fig. S10. Putative mutations in *TP53* gene in IBC and non-IBC tumor samples. Putative mutations in the *TP53* gene in FFPE and frozen tumor samples were identified by VarScan (http://dkoboldt.github.io/varscan/) using the matched PBMC or frozen healthy breast tissue sample of the same patient to distinguish somatic cell mutations from germline mutations or SNPs. The IGV screenshots show the location of each identified mutation in the *TP53* gene. Aligned reads from tumor and PBMC or healthy breast tissue of the same patient are divided by a horizontal red line. The number in the lower left corner of each IGV screenshot indicates the patient ID number color coded by the nature of the variation seen in the tumor. SNPs were identified as variations also seen in PBMCs or normal breast tissue from the same patient.

medRxiv preprint doi: https://doi.org/10.1101/2023.05.26.23290469; this version posted May 30, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure S11

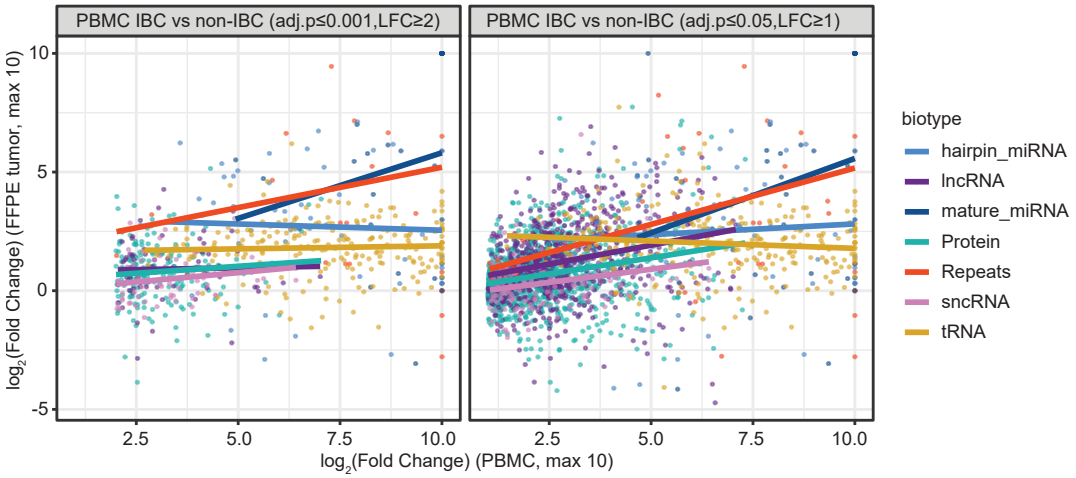


Fig. S11. Scatter plots comparing IBC vs non-IBC DE effect sizes (LFC**s) in FFPE tumors and PBMCs.** Each point represents an identified overexpressed gene in IBC vs non-IBC patient PBMCs with adj. $p \leq 0.001$, LFC$\geq$2 (left panel) or adj. $p \leq 0.05$, LFC$\geq$1 (right panel). The *y*-axis encodes DESeq2 LFC for the FFPE tumor IBC vs non-IBC comparison while the *x*-axis encodes the LFC for PBMC IBC vs non-IBC comparison. Color indicates gene biotype, as indicated by legend at right. Trendlines indicate the linear model fit summarizing relationship between two comparisons for all plotted genes, fit separately for each indicated gene biotype.