www.oncotarget.com Oncotarget, 2022, Vol. 13, pp: 1246-1257

Research Perspective

Treasures from trash in cancer research

Fabiano Cordeiro Moreira[1,*], Dionison Pereira Sarquis[1,*], Jorge Estefano Santana de Souza[2], Daniel de Souza Avelar[1], Taíssa Maria Thomaz Araújo[1], André Salim Khayat[1], Sidney Emanuel Batista dos Santos[1,3] and Paulo Pimentel de Assumpção[1]

[1]Núcleo de Pesquisas em Oncologia/Universidade Federal do Pará, Belém, Pará, Brazil

[2]Instituto Metrópole Digital/Universidade Federal do Rio Grande do Norte, Natal, Brazil

[3]Instituto de Ciências Biológicas/Universidade Federal do Pará, Belém, Pará, Brazil

[*]Co-first authors

Correspondence to: *Paulo Pimentel de Assumpção*, ***email***: *assumpcaopp@gmail.com*

Keywords: cancer metagenomics; cancer sncRNA expression; RNA-Seq variant calling

Received: *February 23, 2022* ***Accepted***: *October 26, 2022* ***Published***: *November 17, 2022*

ABSTRACT

Introduction: Cancer research has significantly improved in recent years, primarily due to next-generation sequencing (NGS) technology. Consequently, an enormous amount of genomic and transcriptomic data has been generated. In most cases, the data needed for research goals are used, and unwanted reads are discarded. However, these eliminated data contain relevant information. Aiming to test this hypothesis, genomic and transcriptomic data were acquired from public datasets.

Materials and Methods: Metagenomic tools were used to explore genomic cancer data; additional annotations were used to explore differentially expressed ncRNAs from miRNA experiments, and variants in adjacent to tumor samples from RNA-seq experiments were also investigated.

Results: In all analyses, new data were obtained: from DNA-seq data, microbiome taxonomies were characterized with a similar performance of dedicated metagenomic research; from miRNA-seq data, additional differentially expressed sncRNAs were found; and in tumor and adjacent to tumor tissue data, somatic variants were found.

Conclusions: These findings indicate that unexplored data from NGS experiments could help elucidate carcinogenesis and discover putative biomarkers with clinical applications. Further investigations should be considered for experimental design, providing opportunities to optimize data, saving time and resources while granting access to multiple genomic perspectives from the same sample and experimental run.

INTRODUCTION

Advances in molecular biology and bioinformatics allow for unprecedented data generation, thereby promoting a greater understanding of many physiological and pathological phenomena in human organisms [1]. Nevertheless, the amount of data produced in each of these experiments usually surpasses the main focus of the proposed investigations [2, 3]. Currently, the strategies for finding molecular markers in cancer research have accumulated an enormous amount of unintended information. Most of these supposed useless data are often treated as trash and remain unexplored. However, in some cases, treasures hidden in these data are discarded [2, 4]. Here, we demonstrate potential strategies to benefit from nontargeted information resulting from high-throughput cancer investigations.

Human cancer genome and metagenomics

The last three decades have been extremely prolific regarding the generation of cancer genome data [5, 6]. Thousands of cancer genomes have been sequenced and analyzed around the world, and millions of dollars have

been invested in such a "gold rush", thus leading to an immense improvement in our understanding of disease [7, 8]. After sequencing, bioinformatics teams deeply explored the data according to the investigation goals. This difficult task of processing human genome data remains ongoing, especially in case of additional questions emerging after initial investigations. Nevertheless, a fraction of data is uncharted.

Recently, a new gold rush was launched: metagenomics. Again, large investments have been made toward efforts to discover the role of the microbiome in many cancer types [9–11]. Moreover, data generation is followed by bioinformatics, which has the task of interpreting the data and generating new hypotheses. New horizons in cancer knowledge are arising and on their way to be implemented in clinical practice.

These two gold rushes have several common features: most experiments collect fresh samples from human tumors and paired noncancer tissues; sequencing procedures result in a large and complex amount of data; bioinformatics teams select the desired information; and nonessential information ("trash") is discarded, aiming at focusing on the research goals and reducing confounding data [12–15]. Next-generation sequencing (NGS) data are available in public data banks, such as GEO (http://www.ncbi.nlm.nih.gov/geo) [16], and they usually provide access to raw genomic data, thereby increasing the credibility, transparency, and reproducibility of the results, while also allowing for additional investigations.

The steps of both cancer genomics and metagenomics experiments are similar. Both investigations are usually based on the acquisition of fresh samples from tumors followed by DNA extraction, library preparation and sequencing. In fact, human and nonhuman DNA are available at this point of the experiment. Nevertheless, according to the investigator, focus is driven to either human DNA for genomics or nonhuman DNA for metagenomics. The unexplored data from each case might harbor some of the objectives of other studies, paving the way for an integrative exploration of both human and nonhuman information.

Therefore, NGS genomic data have sufficient information for taxonomic investigation with similar precision of dedicated metagenomic experiments. In an attempt to test this hypothesis, we performed additional analyses using publicly available gastric, prostate and bladder cancer genomic data to explore metagenomic information.

Exploring additional small noncoding RNAs from miRNA sequencing

Total RNA-seq analyses are NGS experiments that have the most significant opportunity to explore new data. The majority of these experiments focus only on specific types of transcripts, such as mRNA, lncRNA, or miRNA. However, this strategy allows for exploration of several RNA molecules simultaneously. Additionally, from total RNA-seq data, it is possible to detect new genomic variants, providing potentially insights to such experiments. Furthermore, nonhuman microbiome expression data (metatranscriptome) are captured from total RNA-seq. Since both human and nonhuman data are usually sequenced, the identification of microbiome gene expression and several downstream investigations are possible, such as taxonomic profiles and host-microbiome interactions [17].

Most small noncoding RNA (sncRNA) investigations in cancer research address miRNA expression as potential cancer biomarkers and targets for therapy [18–20]. Usually, miRNA sequencing pipelines identify and select small RNA fragments and subsequently quantify each known miRNA [21, 22]. However, among these small fragments, many sequences are not representative of miRNAs and are excluded from analyses as contaminants. Some of these excluded sequences correspond to other classes of noncoding RNAs that increase the cancer process, as is the case for piwi-interacting RNAs (piRNAs) [23, 24]. An integrative strategy should include other noncoding RNAs (ncRNAs), thus harnessing the full potential of all samples and laboratory work, while opening new possibilities for the discovery of regulatory networks.

We hypothesize that evaluating the discarded sequences from miRNA sequencing data enables the identification of sncRNAs with potential value as cancer biomarkers or treatment approaches. To test this hypothesis, gastric, bladder and prostate cancer miRNA-seq data were explored.

Hidden markers in adjacent to tumor samples

Most gene expression investigations in solid tumors rely on comparisons between tumors and adjacent to tumor samples, considering adjacent samples as normal controls. Nevertheless, such specimens harbor molecular alterations that are insufficient to cause cancer but differ from normal tissues collected from noncancer patients [21, 25, 26]. The focus of such experiments is to identify differentially expressed genes between cancer and adjacent to cancer samples because these genes shed light on the molecular events involved in the carcinogenic process [27, 28]. However, initial molecular events are likely present in both adjacent to tumor and tumor samples; thus, searching for differences between them may not reveal these important carcinogenic molecular events [29].

In some cases, however, the analyzed transcripts contain neglected information, such as concomitant expression of oncogenes and expression of mutated genes from both adjacent to tumor and cancer tissues. Once more, these relevant data are likely to be discarded because

the goal is generally to identify differential expression and not concurrent variant patterns [27].

Thus, adjacent tumor tissue, which is often used only as a gene expression control, potentially has somatic alterations common to cancer that may hold an essential role in understanding the first steps of carcinogenesis. Aiming to prove this concept, we explored sequencing data from paired gastric tumors and adjacent to tumor tissues.

MATERIALS AND METHODS

Data acquisition

All data were downloaded from SRA databank (https://www.ncbi.nlm.nih.gov/sra) [30]. For metagenome from genome data analysis, we employed three different cancer types: bladder cancer (PRJNA185252, 44 samples) [31], gastric cancer (PRJNA173904, 19 samples) [32], and prostate cancer (PRJNA412953, 15 samples; PRJEB6530, 20 samples) [33, 34]. For identification of additional sncRNAs from miRNA sequencing, we explored the same cancer types distributed as follows: bladder cancer (five cancer and five noncancer samples) [35]; peripheral blood of prostate cancer (32 cancer and 13 noncancer patients) [36]; and gastric cancer (eight cancer and eight noncancer samples) [20]. For variant calling from RNA-seq, we obtained data from 80 samples, which were generated from 20 gastric cancer patients and distributed as follows: 20 exome tumor samples, 20 exome blood samples, and 20 RNA-seq paired tumor and adjacent to tumor samples (Supplementary Table 1) [37].

Quality control

All downloaded samples were analyzed using FastQC (version 0.11.2) [38]. Trimming and filtering were performed using Trimmomatic (version 0.36) [39]. The parameters for each analysis were chosen based on the visual evaluation performed with FastQC, quality of data, data origin (sncRNA, RNA, or DNA), and type of sequencing (paired-end or single-end). The parameters and quality values (QVs) for each analysis are described in Supplementary Table 2.

Read alignment

To analyze metagenomic data from genomic sequencing, we used Centrifuge Aligner software (version 1.0.4-beta) [40], mapping reads to bacteria, archaea, viruses, and human sequences.

STAR aligner (version 2.7.0) [41] was used to map reads and identify additional sncRNAs and perform variant calling from RNA-seq data analysis. The genome version used in both analyses was HG19 (version 37.7; http://www.ensembl.org/info/data/ftp/). Since some sncRNAs may be repeated in several sites in human genome, the aligner parameters were adjusted, allowing at least 100 repetitions in genome. To improve identification of other sncRNAs, such as piRNAs in case multimapping occurred, the best alignment score was selected.

Exome samples were also mapped to HG19 human reference genome (version 37.7; http://www.ensembl.org/info/data/ftp/) with Burrow Wheeler Aligner (BWA; version 0.7.15) [42] using default parameters.

sncRNA expression data

For each sample, three annotations were performed for transcript quantification using htseq-count software (version 0.6) [43]. First, mirBase annotation [44] was used to quantify microRNA expression. Reads identified as miRNAs were quantified and filtered out from the .sam file. Next, piRbase annotation [45] was used to perform piRNA expression quantification, and reads identified as piRNAs were also quantified and filtered out from the .sam file. The remaining .sam file was then used for the identification and quantification of other transcripts with ENSEMBL annotation (https://www.ensembl.org). Since there are several overlapping sequences in piRbase, we used BEDtools (version 2.17) [46] to merge overlapping sequences into unique sequences, thus avoiding ambiguous recognition.

Taxonomic identification

Taxonomic identification consisted of identifying reads that did not align to the human genome but displayed a minimum alignment score of 60%. The results were reclassified with Recentrifuge (version 1.0.3-beta) [47]. This tool increases identification precision by using two approaches. One that considers a minimum score (here considered as 50) and a robust algorithm to remove contaminants. We employed genus-level taxonomic classification for comparison to literature data.

Variant calling from RNA-seq and DNA-seq data

Duplicated reads were removed using Picard Tools (version 2.18; http://broadinstitute.github.io/picard). The Genome Analysis Toolkit (GATK version 4.1.2) [48] was used for local realignment and recalibration.

All variants were called using BCFtools (version 1.8) [49] on exonic regions. Low coverage variants were filtered out (less than five variant reads or variant read depth less than 20% of total depth). We called somatic variant filtering out blood variants from tumor and adjacent to tumor samples. BCFtools was applied to compare and identify common variants between tumor and adjacent to tumor samples, and Ensembl Variant Effect Predictor (VEP; https://www.ensembl.org/Homo_sapiens/Tools/VEP) [50] was used for common variant annotation.

Statistical and graphical analyses

To compare taxonomic profiles, we filtered all data for the top 40 most relatively abundant genera (representing >90% of reads in all cancer types; Supplementary Table 3). We also converted the data to presence/absence to avoid sequencing biases and compared taxonomic profiles at genus level. The hypergeometric distribution was used to test significant overlap among microorganisms identified by the proposed whole-genome sequencing captured data (WGScd) and other literature data.

P values were adjusted for multiple testing using Benjamini-Hochberg false discovery rate (FDR) adjustments [51]. Alpha diversity was calculated using Simpson's diversity index with Vegan library [52], which was implemented in R, and significant differences between the WGScd and literature data were identified using the Kruskal–Wallis test followed by Dunn's post hoc test. The Vegan library was also employed to calculate and plot rarefaction curves.

To identify additional differentially expressed (DE) sncRNAs not addressed by the original authors, we used DESeq2 package [53], which was implemented in R, to analyze gastric, bladder, and prostate cancer experiments. SncRNAs satisfying the following criteria were tagged as differentially expressed: |log2(fold-change) | >1 and p value < 0.05. All graphics were created in the R statistical platform using the Venn [54] and ggplot2 [55] packages. The R codes for all statistical analyses and plots are provided in the Supplementary Material.

RESULTS

Metagenomic analysis from genomic data

Three different cancer types from bladder, prostate, and gastric tumors were explored to obtain additional findings. This methodology is hereafter referred to as "Whole Genome Sequencing captured data" (WGScd) to differentiate it from other metagenomic analyses. The results were compared to metagenomic literature data.

We downloaded data from 44 bladder cancer samples (PRJNA185252) and searched for nonhuman sequences aiming to obtain taxonomic information from the bladder cancer microbiome. After quality filtering and human sequence removal, an average of 180 thousand reads per sample were obtained. The most abundant bacterial genera found in bladder cancer tissues are demonstrated in Figure 1A and Supplementary Table 4. Since there is scarce information about bladder cancer metagenomics, urine metagenomic data were also included (Figure 1B) [56–60].

Despite finding some common genera, the hypergeometric enrichment test did not indicate any significant overlap between the analyzed data (p-adj > 0.05; Supplementary Table 5). However, two genera were present in all six studies: Finegoldia and Streptococcus. Data from Bučević Popović et al. (2018) were the only data that presented detailed taxonomic information from every studied case, and majority of the genera were also found by the WGScd approach (Figure 1B).

The metagenomic data obtained from genomic experiments from prostate and gastric cancers are shown in Figure 2. Rarefaction curves indicate that in most samples, the employed strategy generated sufficient data to represent the bacterial diversity of each sample (Supplementary Figure 1).

To compare the WGScd metagenomic findings with data from exclusively metagenomics studies [61–65], we first filtered out low abundance taxa of all samples (read counts < 10) and compared alpha diversity indices among samples with those found in other studies with same cancer types. The results showed that WGScd metagenomic analysis had alpha diversity values similar to those obtained from metagenomic sequencing (Kruskal–Wallis and Dunn's post hoc test; adjusted p value > 0.05; Supplementary Table 6; Figure 3A and 3B).

After correcting for multiple testing, the hypergeometric enrichment test did not indicate any significant overlap among cancer metagenomic results (p-adj > 0.05; Supplementary Table 7). In both cancer types, we found several genera present in WGScd analyses and all metagenomic experiments: four in gastric cancer (Helicobacter, Neisseria, Prevotella, and Streptococcus) and ten in prostate cancer (Escherichia, Pseudomonas, Ralstonia, Acinetobacter, Corynebacterium, Rhodococcus, Staphylococcus, Sphingomonas, Streptococcus, and Acidovorax) (Figure 3C and 3D).

sncRNAs analysis

For bladder cancer, we analyzed an average of 16 million (Mi) known transcript reads per sample, including ~81% miRNA reads, ~9% piRNA, and ~10% other transcripts. There was an average of 0.8 Mi reads per sample in gastric cancer, of which ~43% were miRNA reads, ~22% were piRNA reads, 13% were small nucleolar RNA (snoRNA) reads, and ~22% were other transcripts. For prostate cancer peripheral blood, we quantified 1.3 Mi reads per sample, from which ~21% were piRNA reads, ~11% were snoRNA reads, ~3% were snRNAs, ~2% were miRNA reads, and ~63% were other transcripts (Figure 4A).

Comparing gastric cancer sncRNA expression with that of noncancer gastric samples, we identified 57 DE piRNAs, of which 46 were upregulated and 11 were downregulated (Figure 4B). Regarding sncRNA expression in prostate cancer peripheral blood samples

with that of noncancer patients, we were able to identify two upregulated piRNAs and one upregulated snoRNA (Figure 4C). Comparing bladder cancer sncRNA expression with noncancer bladder samples, we identified 102 DE piRNAs, of which 29 were upregulated and 73 were downregulated (Figure 4D). Supplementary Table 8 contains the list of DE sncRNAs.

Analyzing common variations in tumor and adjacent-to-tumor samples

After filtering, 7,443 somatic variants in tumor samples and 7,469 variants in adjacent samples were identified. Comparing tumors with adjacent to tumor samples, we found 1,635 common variants in 1,084

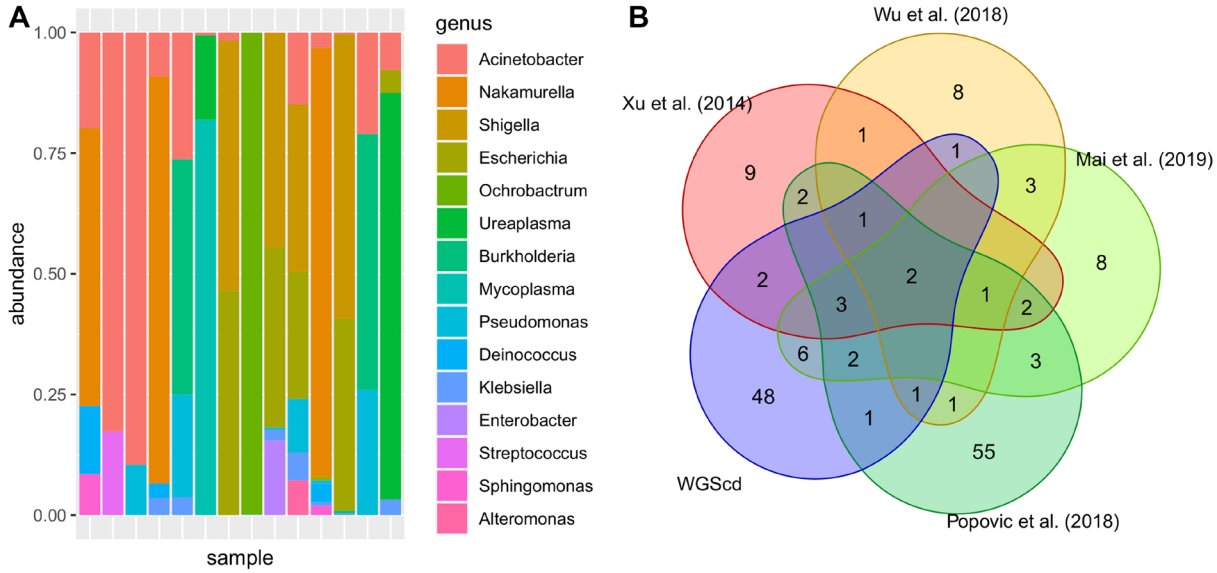

Figure 1: Most abundant bacteria taxa found in bladder analysis. In (**A**), relative genus abundance among samples. In (**B**), presence/absence Venn diagram: Bladder cancer tissue metagenomic profile obtained from Whole Genomic Sequencing captured data (WGScd) compared with literature research of urine bladder cancer metagenomic profile obtained from sequencing rRNA 16s amplicon. Taxon data were converted to genus since not all works present results in species resolution.

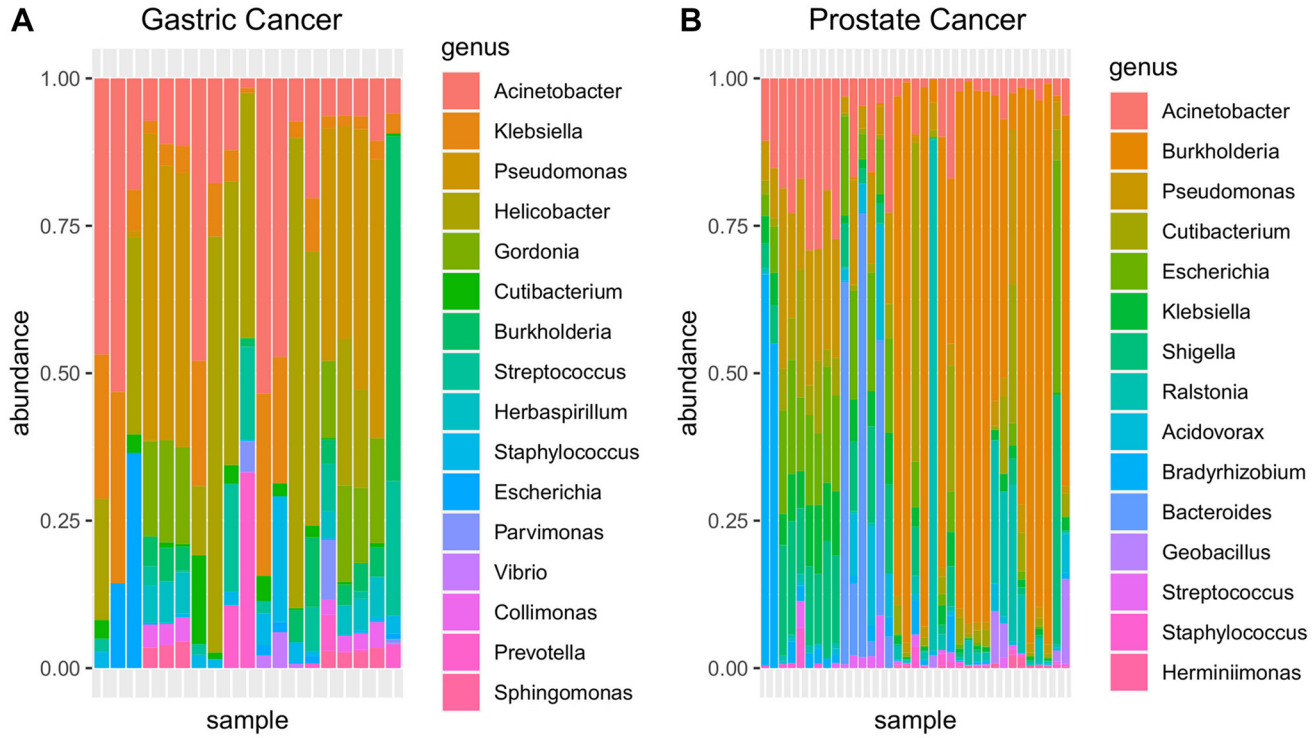

Figure 2: Metagenomic relative abundance in the genus rank from genomic sequencing of (**A**) gastric and (**B**) prostate cancers (WGScd).

genomic positions, most of which were single nucleotide variants (1,021; 94.9%). Small deletions (30; 2.8%) and small insertions (25; 2.3%) were also found. Most variants had been previously reported (1,036; 95.6%), and 48 (4.4%) were unreported mutations.

This analysis was able to identify 23 high-impact common variants (Supplementary Table 9; Figure 5), including 11 frameshifts, two start-loss variants, one stop-loss variant, and one stop-gained variant. From total, 144 common variations were previously reported for gastric cancer on the COSMIC cancer dataset (https://cancer.sanger.ac.uk/cosmic) [66].

DISCUSSION

Obtaining financial support for cancer research remains a challenge. Additionally, recruiting patients for such investigations also represents a critical step. Therefore, optimizing cancer research by reducing the number of expensive rounds of sequencing experiments and exploring the produced data by additional innovative approaches represents an option to overcome the limitations of collecting human biological samples and the scarce availability of resources to cover the costs of high-throughput experiments. Thus, three different approaches are proposed to better

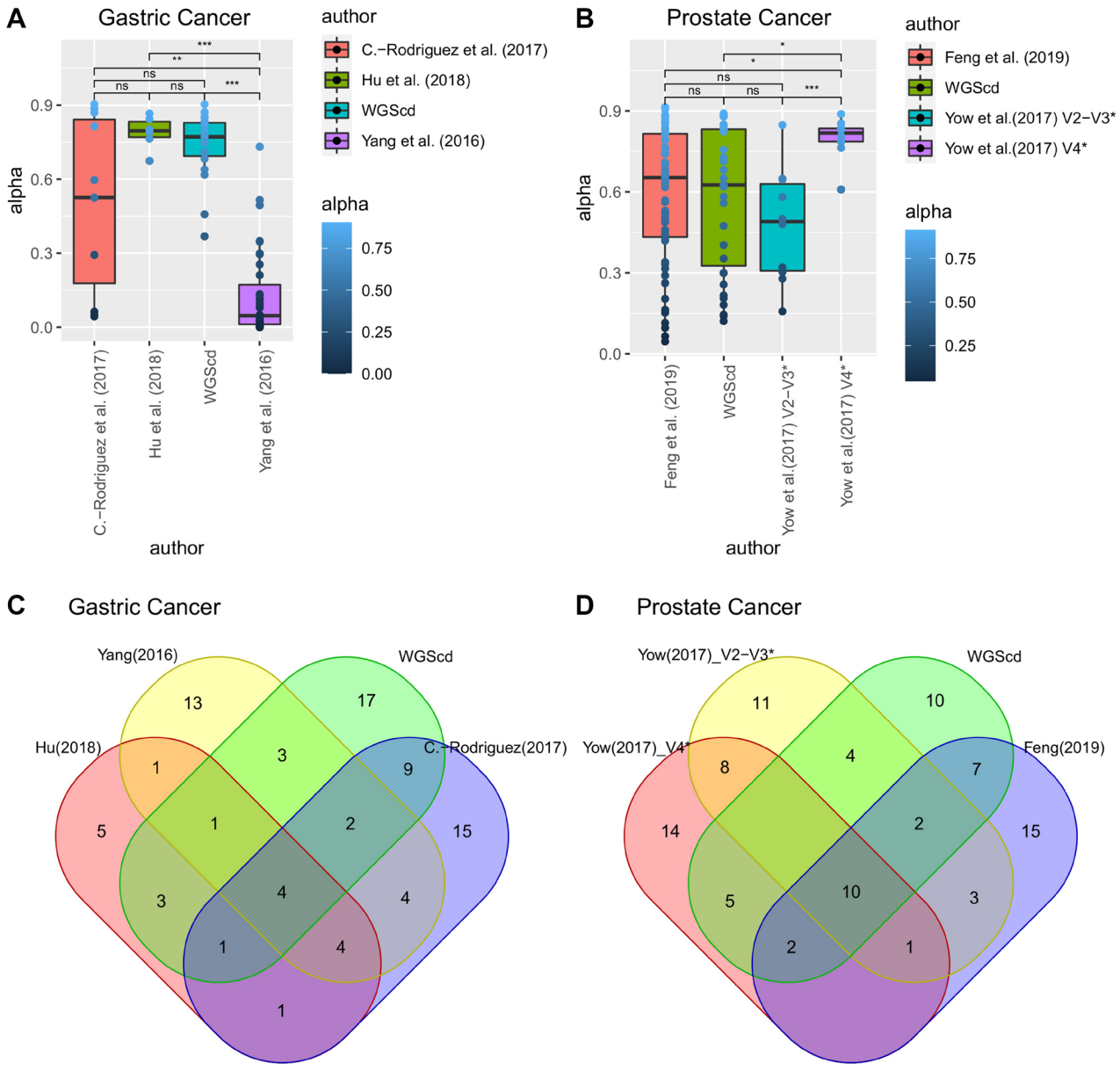

Figure 3: In (**A**) and (**B**), alpha diversity box plot: comparison with literature data indicates that metagenomic analyses from WGScd seem to be as capable of representing community diversity as are dedicated metagenomic analyses. In (**C**) and (**D**), presence/absence Venn diagram: Gastric and prostate cancers microbiome profile obtained from WGScd compared with dedicated metagenomic data of gastric and prostate cancers. Taxon data were converted to genera since not all works present results in species resolution. *Yow et al. (2017) performed two different analyses: v2/v3 rRNA 16s regions and v4 rRNA 16s region.

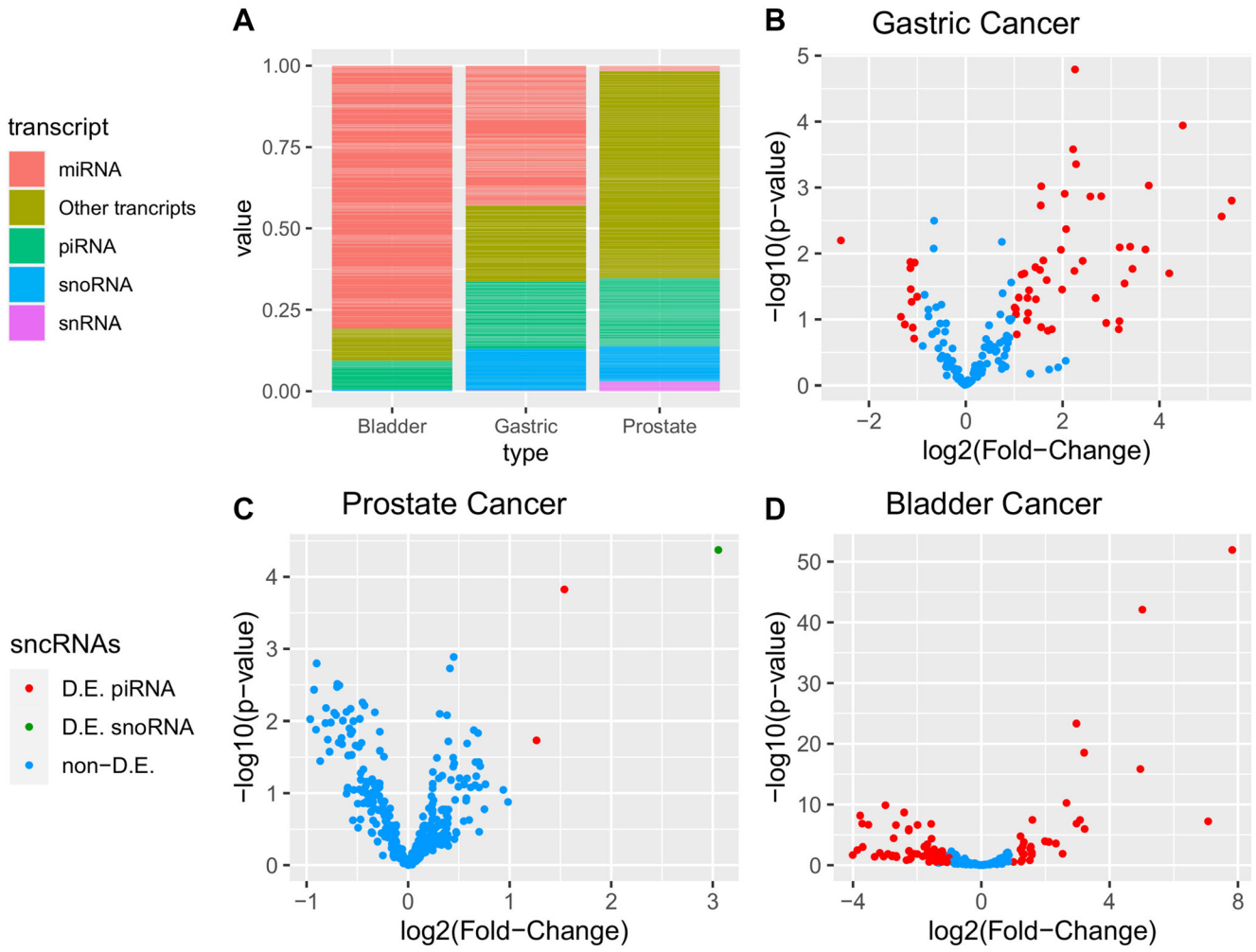

Figure 4: Additional sncRNAs differential expression analysis of gastric, prostate, and bladder cancers obtained from miRNAs expression analyses data. In (**A**), sncRNAs relative abundance of each sequencing. In (**B**), (**C**) and (**D**) volcano plot identifying differentially expressed (DE) sncRNAs (adjusted p-value < 0.05; |log2(fold-change)| > 1).

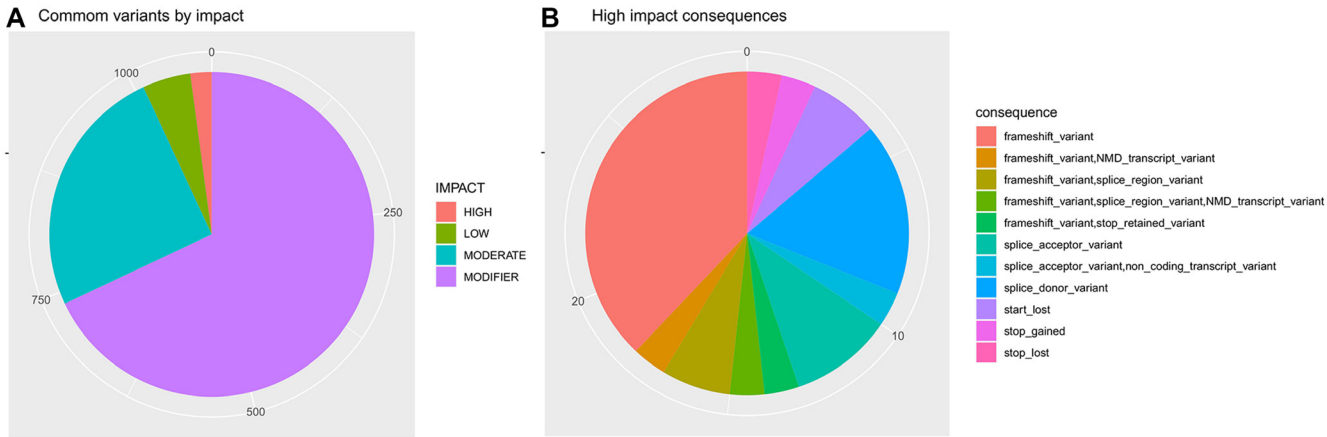

Figure 5: Somatic variants identified in both tumor and adjacent tissue. In (**A**) common variants by impact. In (**B**) potential consequences from high impact common variants. Variants impact and consequences were predicted by the ENSEMBL VEP (https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html).

explore NGS data: (i) obtaining metagenomic data from genomic sequencing; (ii) capturing additional sncRNAs from miRNA sequencing; and (iii) analyzing common variants in tumor and adjacent to tumor samples.

Metagenomics has become a new player in cancer research that has reached clinical practice [67]. Although it has been explored in many types of tumors, the role of the microbiome in bladder cancer remains obscure. Bladder epithelium and urine have been considered sterile in healthy individuals; however, new evidence recently demonstrated that the urinary tract also harbors a specific microbiome [56] that may participate in many diseases, including cancer.

Our results showed that metagenomic analyses using genomic data are viable. However, several discussions may emerge from this approach regarding data reliability, such as whether contamination may mask results and whether combined acquisition of human and nonhuman sequences produces the same results as those from solely human or solely nonhuman experiments.

Regarding contamination, both metagenomic and WGScd analyses share similar vulnerabilities. Tissue collection, laboratory handling, and even sterile reagents are not free of contaminants, since sequencing does not require exclusively viable microorganisms, and fragments of inert DNA may be sequenced and included in downstream investigations. Finding such contaminants remains a challenge, even for metagenomics experiments. Filtering contaminants should not be disregarded, both during classical metagenomics experiments and WGScd, and a critical assessment of results is essential for any potential clinical application. Nevertheless, as demonstrated, a significant part of the bacterial presence in NGS data [68–71] is obtained from sequenced tissue and potentially represents tissue microbiome.

A comparison of WGScd results with metagenomics experimental results strongly suggested similarities between the two methods. According to these preliminary results, it seems viable to capture metagenomics information and save time, lab work, and financial resources by looking at already produced data from genomic sequences.

Another question arises from the applicability of WGScd for other tumor sites. To shed light on this question, additional analyses were carried out, and gastric and prostate cancer data were also tested. Compared with other reports, our results seem to be as reliable as those from metagenomic investigations.

An additional application of the proposed strategy is the investigation of rare tumors and sites with complex accessibility, such as the brain. These data could be reinvestigated from a metagenomic standpoint to provide insights on potential interactions that could be addressed both for carcinogenic understanding and clinical applications.

Experiments should be designed to search for both types of data and perform integrative investigation of genomics and metagenomics from the same samples. In addition to conserving time, saving human and financial resources, and reducing the number of recruited patients, as discussed, this integrative analysis provides an additional advantage of joining genomics and metagenomics from the same clinical situation instead of analyzing each from a different set of patients.

Another approach for better exploring NGS data is simultaneous analysis of diverse sncRNAs. A significant number of miRNA NGS experiments are currently reported due to their potential role as biomarkers. They are essential for cellular and tissue homeostasis and are involved in posttranscriptional gene regulation [18]. Sequencing is relatively inexpensive, and the results for cancer research are relevant. However, other promising sncRNAs are involved in critical biological processes, such as piRNAs [72], and could be analyzed from the same raw dataset.

By exploring this possibility in three cancer types, a large number of additional ncRNAs were identified and quantified, confirming our hypothesis. This relevant information adds strength and value to such analyses, introducing new players and enabling an integrative interpretation of the role of these sncRNAs in cancers and other biological processes.

However, it should be noted that library preparation has an essential role in ncRNA identification [73]. Some sncRNA sequencing requires size selection of larger RNA fragments when compared to exclusive miRNA sequencing. Library preparation from each analyzed dataset had different sizes, namely, 18–30 nucleotides (nts) for bladder tissue, 10–40 nts for prostate tissue, and 15–35 nts for gastric tissue. Larger size selections allowed for identification of a greater variety of sncRNAs, although with lower expression levels. Conversely, smaller size selection resulted in a lower variety of molecules but higher expression levels. This should be taken into consideration for experimental design.

Finally, RNA-seq data were analyzed to find common variants for both tumor and adjacent to tumor tissue in gastric cancer. Using RNA-seq data to identify genomic variants is challenging given the technical computational limitations due to intrinsic complexity of transcriptome, which increases the rate of false-positives compared to DNA-seq data [74, 75]. Alignment of RNA-seq data is more complex than that of DNA-seq data because in mRNA, introns are removed by splicing, which in turn could be identified as deletions. Similarly, RNA editing and polyadenylation processes introduce additional mismatches not found in usual DNA-seq alignment [75]. The false-positives introduced by RNA editing can be minimized since most RNA editing sites are already described [76] and variations in these genomic positions can be removed.

In this particular analysis, removing RNA editing variants was not considered since we aimed to find common variants from tissue adjacent RNA-seq data and

tumor DNA-seq data. Somatic variants in both tissues at an RNA edit position are more likely to be true genomic variants than false-positive findings.

Another putative limitation is calling only mutations in expressed transcript, leaving out both unexpressed genes and intronic/intergenic regions. However, using RNA-seq data to call variants allows for the identification of tissue-specific variant expressions, which are relevant for translational approaches.

Our results identified several common deleterious variants in both tissues. Although these findings may need further experimental investigation, based on the field cancerization hypothesis, this approach may shed light on early steps of carcinogenesis. The proof of concept is again strengthened, especially when regarding the strategy of obtaining as much data as possible from each experiment, allowing more comprehensive interpretations and optimizing resources.

Altogether, our results strengthen the hypothesis that abundant additional and potentially useful information can be extracted from NGS. Moreover, the integrated investigation of every available information should provide a broader and more robust interpretation of the molecular scenario from each experiment.

Data availability

All data generated or analysed during this study are included in this article. Further enquiries can be directed to the corresponding author.

Author contributions

FCM, DPS and DSA performed data collection. FCM, DPS and DSA contributed data or analysys tools. FCM, DPS and DSA performed analysis. FCM, DPS, JESS, DSA, TMTA, ASK, SEBS and PPA wrote and reviews the paper. All authors have reviewed and approved the manuscript in its final state.

ACKNOWLEDGMENTS

We acknowledge Universidade Federal do Pará (PROPESP and FADESP) for the technical support and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for fellowship support.

CONFLICTS OF INTEREST

Authors have no conflicts of interest to declare.

REFERENCES

1. Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, Ferrari R. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform. 2018; 19:286–302. https://doi.org/10.1093/bib/bbw114. [PubMed]

2. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. Journal of Big Data. 2019; 6:1–25. https://doi.org//10.1186/s40537-019-0217-0.

3. D'Argenio V. The High-Throughput Analyses Era: Are We Ready for the Data Struggle? High Throughput. 2018; 7:E8. https://doi.org/10.3390/ht7010008. [PubMed]

4. Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, Spreafico R, Hafler DA, McKinney EF. From Big Data to Precision Medicine. Front Med (Lausanne). 2019; 6:34. https://doi.org/10.3389/fmed.2019.00034. [PubMed]

5. Wheeler DA, Wang L. From human genome to cancer genome: the first decade. Genome Res. 2013; 23:1054–62. https://doi.org/10.1101/gr.157602.113. [PubMed]

6. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458:719–24. https://doi.org/10.1038/nature07943. [PubMed]

7. Porta-Pardo E, Valencia A, Godzik A. Understanding oncogenicity of cancer driver genes and mutations in the cancer genomics era. FEBS Lett. 2020; 594:4233–46. https://doi.org/10.1002/1873-3468.13781. [PubMed]

8. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015; 19:A68–77. https://doi.org/10.5114/wo.2014.47136. [PubMed]

9. Sinha R, Ahsan H, Blaser M, Caporaso JG, Carmical JR, Chan AT, Fodor A, Gail MH, Harris CC, Helzlsouer K, Huttenhower C, Knight R, Kong HH, et al. Next steps in studying the human microbiome and health in prospective studies, Bethesda, MD, May 16-17, 2017. Microbiome. 2018; 6:210. https://doi.org/10.1186/s40168-018-0596-z. [PubMed]

10. Goodman B, Gardner H. The microbiome and cancer. J Pathol. 2018; 244:667–76. https://doi.org/10.1002/path.5047. [PubMed]

11. Bhatt AP, Redinbo MR, Bultman SJ. The role of the microbiome in cancer development and therapy. CA Cancer J Clin. 2017; 67:326–44. https://doi.org/10.3322/caac.21398. [PubMed]

12. Nagashima T, Yamaguchi K, Urakami K, Shimoda Y, Ohnami S, Ohshima K, Tanabe T, Naruoka A, Kamada F, Serizawa M, Hatakeyama K, Matsumura K, Ohnami S, et al. Japanese version of The Cancer Genome Atlas, JCGA, established using fresh frozen tumors obtained from 5143 cancer patients. Cancer Sci. 2020; 111:687–99. https://doi.org/10.1111/cas.14290. [PubMed]

13. Walker SP, Tangney M, Claesson MJ. Sequence-Based Characterization of Intratumoral Bacteria-A Guide to Best Practice. Front Oncol. 2020; 10:179. https://doi.org/10.3389/fonc.2020.00179. [PubMed]

14. Smith A, Pierre JF, Makowski L, Tolley E, Lyn-Cook B, Lu L, Vidal G, Starlard-Davenport A. Distinct microbial communities that differ by race, stage, or breast-tumor subtype in breast tissues of non-Hispanic Black and non-Hispanic White women. Sci Rep. 2019; 9:11940. https://doi.org/10.1038/s41598-019-48348-1. [PubMed]

15. Feng Y, Jaratlerdsiri W, Patrick SM, Lyons RJ, Haynes AM, Collins CC, Stricker PD, Bornman MSR, Hayes VM. Metagenomic analysis reveals a rich bacterial content in high-risk prostate tumors from African men. Prostate. 2019; 79:1731–38. https://doi.org/10.1002/pros.23897. [PubMed]

16. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R. NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res. 2005; 33:D562–66. https://doi.org/10.1093/nar/gki022. [PubMed]

17. Sangiovanni M, Granata I, Thind AS, Guarracino MR. From trash to treasure: detecting unexpected contamination in unmapped NGS data. BMC Bioinformatics. 2019; 20:168. https://doi.org/10.1186/s12859-019-2684-x. [PubMed]

18. Pereira AL, Magalhães L, Moreira FC, Reis-das-Mercês L, Vidal AF, Ribeiro-Dos-Santos AM, Demachki S, Anaissi AKM, Burbano RMR, Albuquerque P, Dos Santos SEB, de Assumpção PP, Ribeiro-Dos-Santos ÂKC. Epigenetic Field Cancerization in Gastric Cancer: microRNAs as Promising Biomarkers. J Cancer. 2019; 10:1560–69. https://doi.org/10.7150/jca.27457. [PubMed]

19. da Silva Oliveira KC, Thomaz Araújo TM, Albuquerque CI, Barata GA, Gigek CO, Leal MF, Wisnieski F, Rodrigues Mello Junior FA, Khayat AS, de Assumpção PP, Rodriguez Burbano RM, Smith MC, Calcagno DQ. Role of miRNAs and their potential to be useful as diagnostic and prognostic biomarkers in gastric cancer. World J Gastroenterol. 2016; 22:7951–62. https://doi.org/10.3748/wjg.v22.i35.7951. [PubMed]

20. Lan H, Lu H, Wang X, Jin H. MicroRNAs as potential biomarkers in cancer: opportunities and challenges. Biomed Res Int. 2015; 2015:125094. https://doi.org/10.1155/2015/125094. [PubMed]

21. Pereira A, Moreira F, Vinasco-Sandoval T, Cunha A, Vidal A, Ribeiro-Dos-Santos AM, Pinto P, Magalhães L, Assumpção M, Demachki S, Santos S, Assumpção P, Ribeiro-Dos-Santos Â. miRNome Reveals New Insights Into the Molecular Biology of Field Cancerization in Gastric Cancer. Front Genet. 2019; 10:592. https://doi.org/10.3389/fgene.2019.00592. [PubMed]

22. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P. Library construction for next-generation sequencing: overviews and challenges. Biotechniques. 2014; 56:61–64. https://doi.org/10.2144/000114133. [PubMed]

23. Barberán-Soler S, Vo JM, Hogans RE, Dallas A, Johnston BH, Kazakov SA. Decreasing miRNA sequencing bias using a single adapter and circularization approach. Genome Biol. 2018; 19:105. https://doi.org/10.1186/s13059-018-1488-z. [PubMed]

24. Coenen-Stass AML, Magen I, Brooks T, Ben-Dov IZ, Greensmith L, Hornstein E, Fratta P. Evaluation of methodologies for microRNA biomarker detection by next generation sequencing. RNA Biol. 2018; 15:1133–45. [PubMed]

25. Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, Goga A, Sirota M, Butte AJ. Comprehensive analysis of normal adjacent to tumor transcriptomes. Nat Commun. 2017; 8:1077. https://doi.org/10.1038/s41467-017-01027-z. [PubMed]

26. Assumpção MB, Moreira FC, Hamoy IG, Magalhães L, Vidal A, Pereira A, Burbano R, Khayat A, Silva A, Santos S, Demachki S, Ribeiro-Dos-Santos Â, Assumpção P. High-Throughput miRNA Sequencing Reveals a Field Effect in Gastric Cancer and Suggests an Epigenetic Network Mechanism. Bioinform Biol Insights. 2015; 9:111–17. https://doi.org/10.4137/BBI.S24066. [PubMed]

27. de Assumpção PP, Khayat AS, Thomaz Araújo TM, Barra WF, Ishak G, Cruz Ramos AMP, Dos Santos SEB, Dos Santos ÂKCR, Demachki S, de Assumpção PB, Calcagno DQ, Dos Santos NPC, de Assumpção MB, et al. Traps and trumps from adjacent-to-tumor samples in gastric cancer research. Chin J Cancer Res. 2018; 30:564–67. https://doi.org/10.21147/j.issn.1000-9604.2018.05.10. [PubMed]

28. Huang X, Stern DF, Zhao H. Transcriptional Profiles from Paired Normal Samples Offer Complementary Information on Cancer Patient Survival--Evidence from TCGA Pan-Cancer Data. Sci Rep. 2016; 6:20567. https://doi.org/10.1038/srep20567. [PubMed]

29. de Assumpção PP, Dos Santos SE, Dos Santos ÂK, Demachki S, Khayat AS, Ishak G, Calcagno DQ, Dos Santos NP, de Assumpção CB, de Assumpção MB, Sortica VA, Araújo TM, Moreira FC, et al. The adjacent to tumor sample trap. Gastric Cancer. 2016; 19:1024–25. https://doi.org/10.1007/s10120-015-0539-3. [PubMed]

30. Shumway M, Cochrane G, Sugawara H. Archiving next generation sequencing data. Nucleic Acids Res. 2010; 38:D870–71. https://doi.org/10.1093/nar/gkp1078. [PubMed]

31. Guo G, Sun X, Chen C, Wu S, Huang P, Li Z, Dean M, Huang Y, Jia W, Zhou Q, Tang A, Yang Z, Li X, et al. Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation. Nat Genet. 2013; 45:1459–63. https://doi.org/10.1038/ng.2798. [PubMed]

32. Lee YS, Cho YS, Lee GK, Lee S, Kim YW, Jho S, Kim HM, Hong SH, Hwang JA, Kim SY, Hong D, Choi IJ, Kim BC, et al. Genomic profile analysis of diffuse-type gastric cancers. Genome Biol. 2014; 15:R55. https://doi.org/10.1186/gb-2014-15-4-r55. [PubMed]

33. Jaratlerdsiri W, Chan EKF, Gong T, Petersen DC, Kalsbeek AMF, Venter PA, Stricker PD, Bornman MSR, Hayes VM. Whole-Genome Sequencing Reveals Elevated Tumor Mutational Burden and Initiating Driver Mutations in

African Men with Treatment-Naïve, High-Risk Prostate Cancer. Cancer Res. 2018; 78:6736–46. https://doi.org/10.1158/0008-5472.CAN-18-0254. [PubMed]

34. Wyatt AW, Mo F, Wang K, McConeghy B, Brahmbhatt S, Jong L, Mitchell DM, Johnston RL, Haegert A, Li E, Liew J, Yeung J, Shrestha R, et al. Heterogeneity in the inter-tumor transcriptome of high risk prostate cancer. Genome Biol. 2014; 15:426. https://doi.org/10.1186/s13059-014-0426-y. [PubMed]

35. Itesako T, Seki N, Yoshino H, Chiyomaru T, Yamasaki T, Hidaka H, Yonezawa T, Nohata N, Kinoshita T, Nakagawa M, Enokida H. The microRNA expression signature of bladder cancer by deep sequencing: the functional significance of the miR-195/497 cluster. PLoS One. 2014; 9:e84311. https://doi.org/10.1371/journal.pone.0084311. [PubMed]

36. Daniel R, Wu Q, Williams V, Clark G, Guruli G, Zehner Z. A Panel of MicroRNAs as Diagnostic Biomarkers for the Identification of Prostate Cancer. Int J Mol Sci. 2017; 18:1281. https://doi.org/10.3390/ijms18061281. [PubMed]

37. Mun DG, Bhin J, Kim S, Kim H, Jung JH, Jung Y, Jang YE, Park JM, Kim H, Jung Y, Lee H, Bae J, Back S, et al. Proteogenomic Characterization of Human Early-Onset Gastric Cancer. Cancer Cell. 2019; 35:111–24.e10. https://doi.org/10.1016/j.ccell.2018.12.003. [PubMed]

38. Andrews S. FastQC: a quality control tool for high throughput sequence data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

39. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30:2114–20. https://doi.org/10.1093/bioinformatics/btu170. [PubMed]

40. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016; 26:1721–29. https://doi.org/10.1101/gr.210641.116. [PubMed]

41. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29:15–21. https://doi.org/10.1093/bioinformatics/bts635. [PubMed]

42. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–60. https://doi.org/10.1093/bioinformatics/btp324. [PubMed]

43. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015; 31:166–69. https://doi.org/10.1093/bioinformatics/btu638. [PubMed]

44. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 2006; 34:D140–44. https://doi.org/10.1093/nar/gkj112. [PubMed]

45. Wang J, Zhang P, Lu Y, Li Y, Zheng Y, Kan Y, Chen R, He S. piRBase: a comprehensive database of piRNA sequences. Nucleic Acids Res. 2019; 47:D175–80. https://doi.org/10.1093/nar/gky1043. [PubMed]

46. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–2. https://doi.org/10.1093/bioinformatics/btq033. [PubMed]

47. Martí JM. Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. PLoS Comput Biol. 2019; 15:e1006967. https://doi.org/10.1371/journal.pcbi.1006967. [PubMed]

48. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013; 43:11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43. [PubMed]

49. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience. 2021; 10:giab008. https://doi.org/10.1093/gigascience/giab008. [PubMed]

50. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome Biol. 2016; 17:122. https://doi.org/10.1186/s13059-016-0974-4. [PubMed]

51. Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. J Educ Behav Stat. 2000; 25:60–83. https://doi.org/10.3102/10769986025001060.

52. Oksanen J, Blanchet F, Kindt R, Legendre P, Minchin R, O'Hara R. vegan: Community ecology package version 2.0–10. J Stat Softw. 2013; 48:103–32.

53. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15:550. https://doi.org/10.1186/s13059-014-0550-8. [PubMed]

54. Mamakani K, Myrvold W, Ruskey F. Generating all Simple Convexly-drawable Polar Symmetric 6-Venn Diagrams. In: Iliopoulos CS, Smyth WF, (eds). Combinatorial Algorithms. IWOCA 2011. Lecture Notes in Computer Science. Springer: Berlin, Heidelberg. 2011; 7056:275–86. https://doi.org/10.1007/978-3-642-25011-8_22.

55. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag: New York. 2016. Available from: https://ggplot2.tidyverse.org.

56. Mansour B, Monyók Á, Makra N, Gajdács M, Vadnay I, Ligeti B, Juhász J, Szabó D, Ostorházi E. Bladder cancer-related microbiota: examining differences in urine and tissue samples. Sci Rep. 2020; 10:11042. https://doi.org/10.1038/s41598-020-67443-2. [PubMed]

57. Bučević Popović V, Šitum M, Chow CT, Chan LS, Roje B, Terzić J. The urinary microbiome associated with bladder

cancer. Sci Rep. 2018; 8:12157. https://doi.org/10.1038/s41598-018-29054-w. [PubMed]

58. Wu P, Zhang G, Zhao J, Chen J, Chen Y, Huang W, Zhong J, Zeng J. Corrigendum: Profiling the Urinary Microbiota in Male Patients With Bladder Cancer in China. Front Cell Infect Microbiol. 2018; 8:429. https://doi.org/10.3389/fcimb.2018.00429. [PubMed]

59. Xu W, Yang L, Lee P, Huang WC, Nossa C, Ma Y, Deng FM, Zhou M, Melamed J, Pei Z. Mini-review: perspective of the microbiome in the pathogenesis of urothelial carcinoma. Am J Clin Exp Urol. 2014; 2:57–61. [PubMed]

60. Mai G, Chen L, Li R, Liu Q, Zhang H, Ma Y. Common Core Bacterial Biomarkers of Bladder Cancer Based on Multiple Datasets. Biomed Res Int. 2019; 2019:4824909. https://doi.org/10.1155/2019/4824909. [PubMed]

61. Hu YL, Pang W, Huang Y, Zhang Y, Zhang CJ. The Gastric Microbiome Is Perturbed in Advanced Gastric Adenocarcinoma Identified Through Shotgun Metagenomics. Front Cell Infect Microbiol. 2018; 8:433. https://doi.org/10.3389/fcimb.2018.00433. [PubMed]

62. Yang I, Woltemate S, Piazuelo MB, Bravo LE, Yepez MC, Romero-Gallo J, Delgado AG, Wilson KT, Peek RM, Correa P, Josenhans C, Fox JG, Suerbaum S. Different gastric microbiota compositions in two human populations with high and low gastric cancer risk in Colombia. Sci Rep. 2016; 6:18594. https://doi.org/10.1038/srep18594. [PubMed]

63. Castaño-Rodríguez N, Goh KL, Fock KM, Mitchell HM, Kaakoush NO. Dysbiosis of the microbiome in gastric carcinogenesis. Sci Rep. 2017; 7:15957. https://doi.org/10.1038/s41598-017-16289-2. [PubMed]

64. Yow MA, Tabrizi SN, Severi G, Bolton DM, Pedersen J, Giles GG, Southey MC, and Australian Prostate Cancer BioResource. Characterisation of microbial communities within aggressive prostate cancer tissues. Infect Agent Cancer. 2017; 12:4. https://doi.org/10.1186/s13027-016-0112-7. [PubMed]

65. Feng Y, Ramnarine VR, Bell R, Volik S, Davicioni E, Hayes VM, Ren S, Collins CC. Metagenomic and metatranscriptomic analysis of human prostate microbiota from patients with prostate cancer. BMC Genomics. 2019; 20:146. https://doi.org/10.1186/s12864-019-5457-z. [PubMed]

66. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR, Wooster R. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer. 2004; 91:355–58. https://doi.org/10.1038/sj.bjc.6601894. [PubMed]

67. Ng C, Li H, Wu WKK, Wong SH, Yu J. Genomics and metagenomics of colorectal cancer. J Gastrointest Oncol. 2019; 10:1164–70. https://doi.org/10.21037/jgo.2019.06.04. [PubMed]

68. Stinson LF, Keelan JA, Payne MS. Identification and removal of contaminating microbial DNA from PCR reagents: impact on low-biomass microbiome analyses. Lett Appl Microbiol. 2019; 68:2–8. https://doi.org/10.1111/lam.13091. [PubMed]

69. Robinson KM, Crabtree J, Mattick JS, Anderson KE, Dunning Hotopp JC. Distinguishing potential bacteria-tumor associations from contamination in a secondary data analysis of public cancer genome sequence data. Microbiome. 2017; 5:9. https://doi.org/10.1186/s40168-016-0224-8. [PubMed]

70. Laurence M, Hatzis C, Brash DE. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. PLoS One. 2014; 9:e97876. https://doi.org/10.1371/journal.pone.0097876. [PubMed]

71. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014; 12:87. https://doi.org/10.1186/s12915-014-0087-z. [PubMed]

72. Vinasco-Sandoval T, Moreira FC, Vidal AF, Pinto P, Ribeiro-Dos-Santos AM, Cruz RLS, Fonseca Cabral G, Anaissi AKM, Lopes KP, Ribeiro-Dos-Santos A, Demachki S, de Assumpção PP, Ribeiro-Dos-Santos Â, Santos S. Global Analyses of Expressed Piwi-Interacting RNAs in Gastric Cancer. Int J Mol Sci. 2020; 21:7656. https://doi.org/10.3390/ijms21207656. [PubMed]

73. Calcagno DQ, Mota ER, Moreira FC, de Sousa SBM, Burbano RR, Assumpção PP. Role of PIWI-Interacting RNA (piRNA) as Epigenetic Regulation. In: Patel V, Preedy V, (eds). Handbook of Nutrition, Diet, and Epigenetics. Springer, Cham. 2019; 187–209. https://doi.org/10.1007/978-3-319-55530-0_77.

74. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. Am J Hum Genet. 2013; 93:641–51. https://doi.org/10.1016/j.ajhg.2013.08.008. [PubMed]

75. Sheng Q, Zhao S, Li CI, Shyr Y, Guo Y. Practicability of detecting somatic point mutation from RNA high throughput sequencing data. Genomics. 2016; 107:163–69. https://doi.org/10.1016/j.ygeno.2016.03.006. [PubMed]

76. Lo Giudice C, Tangaro MA, Pesole G, Picardi E. Investigating RNA editing in deep transcriptome datasets with REDItools and REDIportal. Nat Protoc. 2020; 15:1098–131. https://doi.org/10.1038/s41596-019-0279-7. [PubMed]