







Article

# Gastric Normal Adjacent Mucosa Versus Healthy and Cancer Tissues: Distinctive Transcriptomic Profiles and Biological Features

Sabino Russi <sup>1</sup>, Giovanni Calice <sup>1</sup>, Vitalba Ruggieri <sup>1</sup>, Simona Laurino <sup>1</sup>,  
Francesco La Rocca <sup>1</sup>, Elena Amendola <sup>2</sup>, Cinzia Lapadula <sup>3</sup>, Debora Compare <sup>4</sup>,  
Gerardo Nardone <sup>4</sup>, Pellegrino Musto <sup>5</sup>, Mario De Felice <sup>6</sup>, Geppino Falco <sup>2,7,\*</sup>  
and Pietro Zoppoli <sup>1,\*</sup>

- <sup>1</sup> Laboratory of Preclinical and Translational Research, IRCCS–Referral Cancer Center of Basilicata (CROB), 85028 Rionero in Vulture (PZ), Italy
  - <sup>2</sup> Department of Biology, University of Naples Federico II, 80138 Naples, Italy
  - <sup>3</sup> Pathology Unit, IRCCS, Referral Cancer Center of Basilicata (CROB), 85028 Rionero in Vulture, Italy
  - <sup>4</sup> Department of Clinical Medicine and Surgery, University of Naples Federico II, 80131 Naples, Italy
  - <sup>5</sup> Unit of Hematology and Stem Cell Transplantation, IRCCS–Referral Cancer Center of Basilicata (CROB), 85028 Rionero in Vulture, Italy
  - <sup>6</sup> Istituto per l’Endocrinologia e l’Oncologia Sperimentale “Gaetano Salvatore” (IEOS), Consiglio Nazionale delle Ricerche (CNR), 80131 Naples, Italy
  - <sup>7</sup> Biogem, Istituto di Biologia e Genetica Molecolare, Via Camporeale, 83031 Ariano Irpino (AV), Italy
- \* Correspondence: geppino.falco@unina.it (G.F.); pietro.zoppoli@crob.it (P.Z.)

Received: 6 July 2019; Accepted: 22 August 2019; Published: 26 August 2019



**Abstract:** Gastric cancer (GC) is a leading cause of cancer-related deaths in the world. Molecular heterogeneity is a major determinant for the clinical outcomes and an exhaustive tumor classification is currently missing. Histologically normal tissue adjacent to the tumor (NAT) is commonly used as a control in cancer studies, nevertheless a recently published paper described the unique characteristics of the NAT in several tumor types. Little is known about the global gene expression profile of gastric NAT (gNAT) which could be an effective tool for a more realistic definition of GC molecular signature. Here, we integrated data of 512 samples from the Genotype-Tissue Expression project (GETx) and The Cancer Genome Atlas (TCGA) to analyze the transcriptome of healthy gastric tissues, gNAT, and GC samples. We validated TCGA-GETx data mining through inHouse gNAT and GC expression dataset. Differential gene expression together with pathway enrichment analyses, indeed, led to different results when using the gNAT or the healthy tissue as control. Based on our analyses, gNAT showed a peculiar gene signature and biological features, like the estrogen receptor pathways activation, suggesting a molecular behavior partially different from both healthy and GC tissues. Therefore, using gNAT as healthy control tissue in the characterization of tumor associated biological processes and pathways could lead to suboptimal results.

**Keywords:** gastric cancer; normal tissue adjacent to the tumor; gene expression profile

## 1. Introduction

Gastric cancer (GC) was the third leading cause of cancer mortality in 2018, responsible for 783,000 deaths (<http://www.who.int/news-room/fact-sheets/detail/cancer>) and of a poor 5-year survival in case of an advanced stage diagnosis or metastatic disease [1,2]. GCs are mostly adenocarcinomas, subdivided [3] into intestinal and diffuse types according to the Lauren classification and into papillary, tubular, mucinous (colloid) and poorly cohesive carcinomas according to the World Health

Organization [4]. In cancer studies, the histological normalcy also implies molecular normalcy. However, this assumption could not be applied to histologically normal adjacent tumor (NAT). It is well known that many molecular differences (versus normal tissues) characterize NAT such as allelic imbalance, telomere length [5], as well as transcriptomic and epigenetic aberrations [6]. Overall, the NAT tissue can be considered an intermediate, morphologically normal but molecularly altered pre-neoplastic state and these changes are evident up to 1 cm from the margins of the tumor [7]. About breast NAT, recent studies reported that the tumor microenvironment is essential for recurrence prediction and surgical strategies setting [8] and that, interestingly, NAT tissue is enriched for stromal [9] and wound response pathways [10]. It has also been highlighted that breast NAT tissue undergoes wound healing-like processes, extracellular matrix remodeling and an epithelial-to-mesenchymal transition (EMT) [11]. Transcriptomic analysis performed on prostate [12], liver [13], and colon [14] cancers also identified unique gene expression profiles for NAT, resulting from a crosstalk between tumor and adjacent tissue, principally mediated by cytokines and other tumor-secreted factors. Thus, comparing tumor and NAT tissues, usually considered as healthy control samples, many potential cancer biomarkers could be missed and/or wrongly pointed out, as it has been recently showed by Aran et al. [7]. Accordingly, based on the above-mentioned studies, we focused on the molecular characteristics of gastric NAT (gNAT), comparing its transcriptomic profile with tumor and non-diseased tissues, hereinafter defined as “healthy normal” samples. To reach our goal, we applied a system biology approach looking for global features like pathways and tissue composition. In particular, we integrated the transcriptomic data from the Genotype-Tissue Expression (GTEx) [15] and The Cancer Genome Atlas (TCGA) [16] projects. Then, we performed a comprehensive analysis of transcriptomic profiles from healthy gastric tissue, gNAT, and gastric tumor, including dimensionality reduction, differential gene expression, gene set enrichment and tissue composition analyses to provide a coherent definition of gNAT molecular phenotype. Remarkably, our analyses highlighted a possible bias depending on suboptimal sampling of normal gastric mucosa. Indeed, a subgroup of normal gastric samples in GTEx was characterized by muscular phenotype according to both gene set enrichment and anatomo-pathological description. Such subgroup shares many similarities with the tumor samples and its inclusion in any further analyses could impinge a clear assessment of the tumor phenotype. Furthermore, we showed that the gNAT tissue is distinct from both healthy and tumor tissues and represents an intermediate state, possibly resulting from NAT-tumor crosstalk.

## 2. Results

### 2.1. Integrative Analysis of TCGA and GTEx RNA-Seq Data

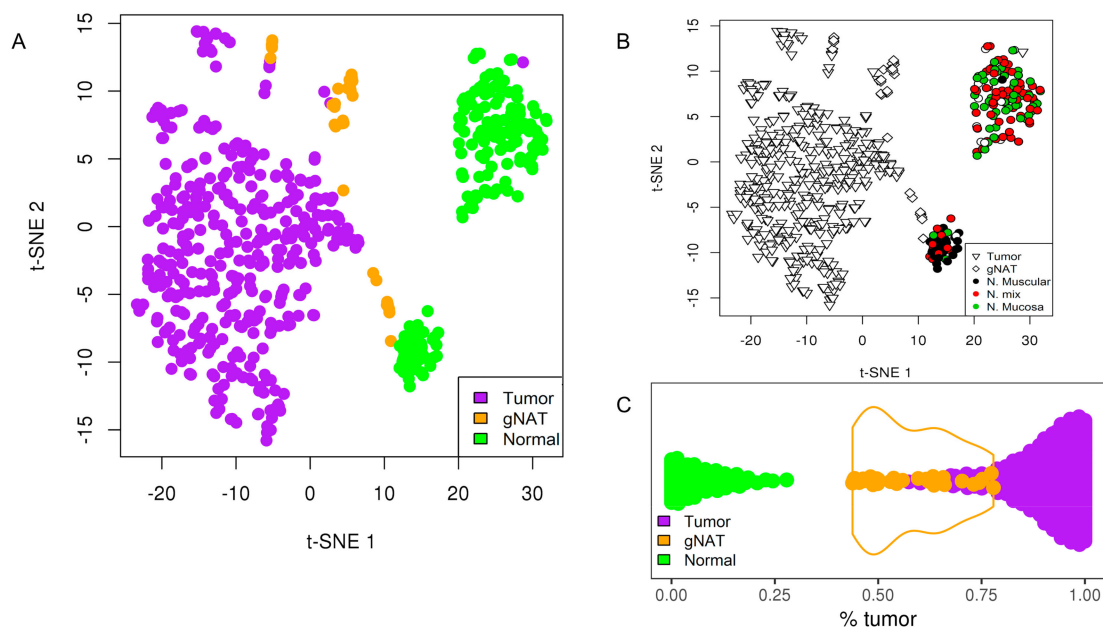
To compare samples’ transcriptomic profiles of TCGA and GTEx projects, we adopted an identical analysis approach. In particular, we obtained raw RNAseq reads of both GTEx and TCGA samples using the same pipeline [17] and compiled a dataset comprising of 162 healthy normal samples, 32 gNAT samples and 380 primary gastric tumor samples (Table 1). On average, cancer patients were older than healthy donors of 18.3 years, while women were 7% more abundant in the healthy group. We assessed differential batch effects and we verified datasets comparability by evaluating the expression and variation of housekeeping genes [18]. We correlated their median expression levels across all samples and we found a high degree of agreement between the two datasets (Pearson  $R = 0.95$ ,  $p$ -value  $\ll 1 \times 10^{-6}$ ) (Figure S1).

**Table 1.** Distribution of samples and demographic characteristics of patients and healthy donors included in the study. GTEx: Genotype-Tissue Expression project; TCGA: The Cancer Genome Atlas; gNAT: gastric normal tissue adjacent to the tumor.

Tissue Type	Number of Samples	Sex (% of Female)	Age (Mean $\pm$ SD)
GTEx healthy	162	42.6	47.4 $\pm$ 12.4
TCGA gNAT	32	34.4	66.4 $\pm$ 9.1
TCGA tumor	380	35	65.7 $\pm$ 10.6

## 2.2. Evaluation of the Samples Molecular Variability

In this study, gNAT and healthy tissues data were obtained from TCGA and GTEx datasets, respectively. A major limitation inherent to the integration of multiple independently collected datasets is disparity among sample sets and the different sequencing protocols. By standardizing analysis pipelines according to Q. Wang et al. and removing possible technical distortions such as differences in sample preparation and batch effects through EDASeq and RUVSeq [19,20] packages (Figure S3), data were successfully merged enabling the analyses of RNA-Sequencing data from different sources [21,22]. We investigated the most important sources of variability through a dimensionality reduction process that, reducing the variables (genes) under consideration, enabled us to distinguish three biological groups, with gNAT samples graphically localized between tumor and healthy samples (Figure 1A). Interestingly, the healthy samples were further divided in two clusters, both separated from the tumor and the gNAT groups, one resulting to be more similar to the tumor along the first component. According to the anatomopathological classification of the samples, it resulted to be composed by muscular tissue more than by gastric mucosa (Figure 1B), thus we named the two normal clusters as “Normal Muscular” and “Normal Mucosa”.



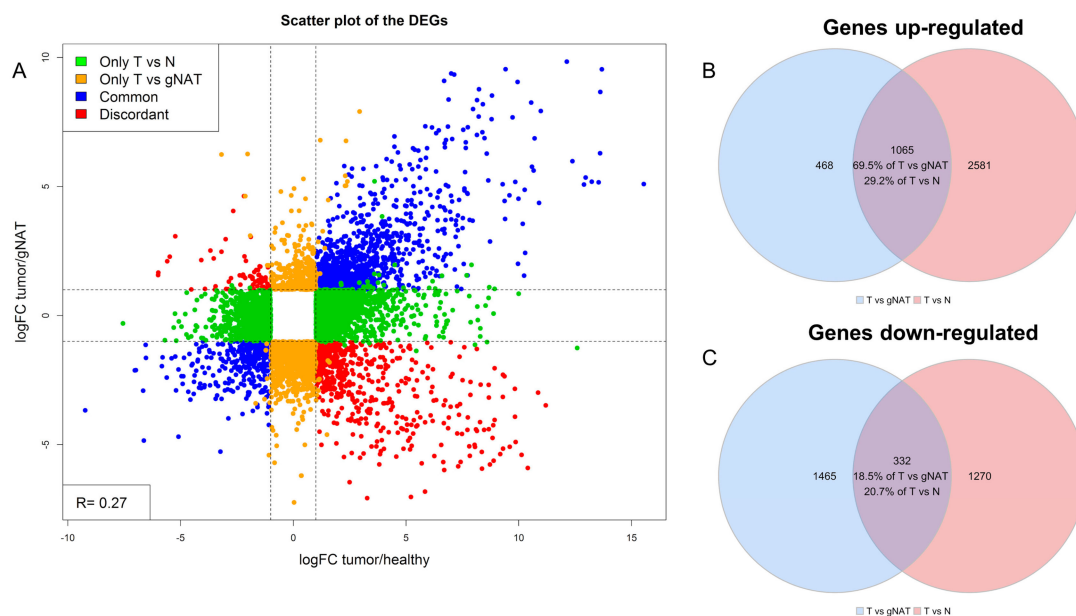
**Figure 1.** Scatter plot of the samples after dimensionality reduction procedure. (A) Gastric cancer (GC) in purple, gNAT in orange and normal tissue in green; (B) GC and gNAT with no fill while normal tissue in black red or green according to the anatomopathological classification as muscular, mixed or mucosa. (C) Deconvolution highlighting the tumor/normal fraction for each sample. GC in purple, gNAT in orange and normal tissue (only mucosa) in green. In orange violin plot overlay of the gNAT distribution.

Furthermore, we comparatively evaluated the two clusters through differential expression and gene enrichment analyses. Among the differentially expressed genes (DEGs), we found 2711 genes up regulated in mucosa cluster enriching gastric acid secretion and 2568 up regulated genes in muscular cluster enriching vascular smooth muscle contraction (Figures S4 and Figure S5). Moreover, the heatmap of the top 900 DEGs (adjusted  $p$ -value  $\ll 1 \times 10^{-6}$  and  $\text{abs}(\log_2\text{FC}) > 3$ ) was reported in Figure S6. However, since the muscular samples are not representative of the normal gastric tissue, we considered not appropriate to include them in further analyses. As above mentioned, gNAT samples appeared distinguished from the other tissues' samples. As shown in Figure 1C, by using a deconvolution pipeline [23] able to calculate the “normal:tumor” fraction for all samples, we found that gNAT samples were positioned between tumor and healthy tissue samples. Strikingly, gNAT group was similar to

the tail of the tumor sample distribution, possibly suggesting a microscopic contamination of gNAT samples with tumor and vice versa. However, even considering such overlap, gNAT appeared to be a distinct tissue type. To validate our findings, we searched public data repositories for independent studies that collected samples from all three tissue types included in our analyses. Our search yielded a microarray cohort (E\_MTAB\_1338) with sufficient sample sizes of GC, gNAT, and healthy tissues. The same methodology described above applied on this dataset resulted in a consistent distribution of samples (Figure S7). Consistently with the TCGA-GTEX samples pattern (Figure 1A), even in this small dataset, we found two healthy tissue clusters and one of the subgroups characterized by the activation of the gastric acid secretion (adjusted  $p$ -value < 0.01), as pointed in Figure S8.

### 2.3. Evaluation of the Adequacy of the gNAT as Control in Cancer Research

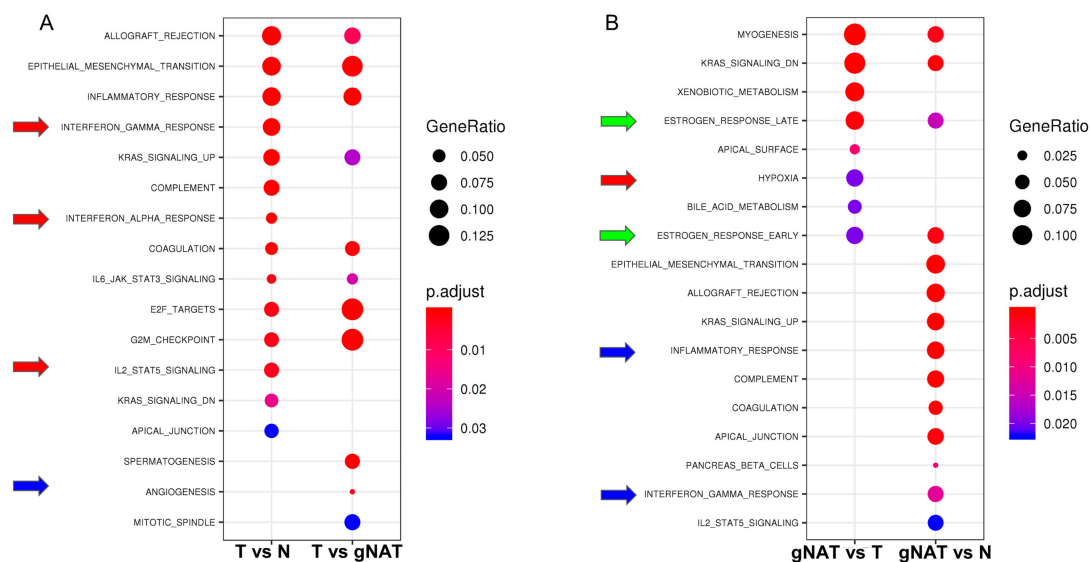
To assess the impact of different/sub-optimal control samples on differential expression analyses results, we performed further evaluations. Firstly, we compared results obtained by considering gNAT and healthy normal samples as controls for tumor samples. Comparison between tumor and gNAT resulted into 3330 DEGs while 5248 DEGs were obtained by comparing tumor with normal samples. Interestingly, the overall Pearson correlation between fold-changes was poor ( $R = 0.27$ ), suggesting that gNAT and healthy normal samples have significantly different molecular signatures. As depicted in Figure 2A, among the DEGs resulted from the comparative analysis, 1397 genes showed the same behavior, 634 DEGs were discordant, while 3217 and 1299 were significant only in one of the two comparisons. Overlap between up and down regulated genes deriving from the comparisons was depicted in the Venn diagrams in Figure 2B,C.



**Figure 2.** Tumor (T) vs. gNAT and T vs. normal differentially expressed genes. (A) Scatter plot of the  $\log_2FC$ . DEGs in green exclusively in T vs. N, in orange exclusively in T vs. gNAT, in blue in common while in red discordant. (B) Overlap of the up regulated genes. (C) Overlap of the down-regulated genes.

To investigate whether using different control samples could have an impact on the biological features associated to GC, we performed a hypergeometric test using the 50 hallmark and the 5917 Gene Ontology (GO) gene sets from mSigDB [24]. Altogether, 8 out of 17 significant (adjusted  $p$ -value < 0.05) hallmarks were in common, 6 were specific of the comparison of tumor samples with normal and 3 of tumor vs. gNAT (Figure S9A). In particular, using the gNAT as control masked some hallmarks such as interferon response or the IL2/STAT5 signaling also pinpointing misleading results as angiogenesis

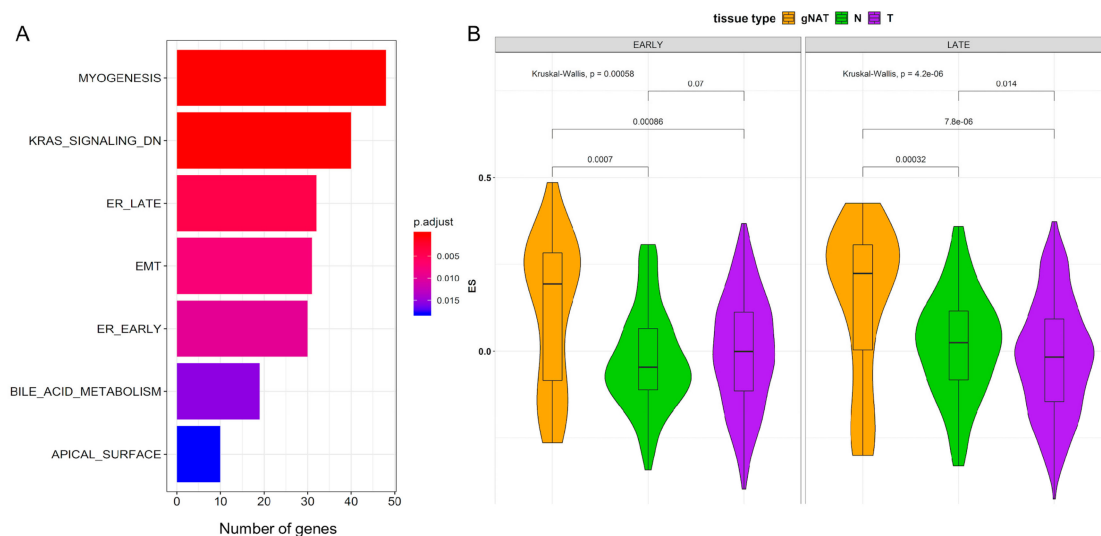
(Figure 3A). When we compared the GO enrichments, only 220 out of 1214 (adjusted  $p$ -value  $< 0.05$ ) were in common, 863 were specific of the comparison with normal and 131 of the comparison with gNAT (Figure S9B). In particular, the gNAT masked some GO as innate and adaptive immune responses (Figure S10). Overall, we concluded that considering gNAT as a control for GC samples did not allow detection of the majority of DEGs in tumors neither accurately identified all the biological features of the GC.



**Figure 3.** Gene set analysis (GSA) of the hallmark gene sets using the differentially expressed genes (DEGs) between tumor or gNAT vs. each of the other two tissues, respectively. Adjusted  $p$ -value in red-blue color scale. Gene ratio in dot size scale. Red, blue and green arrows highlight interesting exclusive and common gene sets, respectively. (A) Tumor centered analysis, (B) gNAT centered analysis.

#### 2.4. Molecular and Biological Characterization of the gNAT

To further explore the divergence between gNAT and tumor samples, we performed a differential expression analysis. By comparing gNAT with tumor samples, we identified 1797 upregulated and 1533 downregulated DEGs (Tables S1 and S2). To gain further insight into the global patterns that differentiate the gNAT from tumor specimens, we performed a gene set analysis (GSA) using the hallmark and the GO gene sets [24]. GSA on the hallmark and the GO gene sets showed 16% of the all hallmarks and 6% of the all GO, respectively, as significantly different between gNAT and tumor (adjusted  $p$ -value  $< 0.05$ ). In particular, hypoxia together with early and late estrogen receptor (ER) pathways hallmarks specifically resulted from this comparison (Figure 3B). Remarkably, many categories have been identified through GO enrichment (Figure S10B). We also assessed DEGs between gNAT and healthy tissue, finding 5415 DEGs (3711 upregulated and 1704 down regulated) (Tables S1 and S2). Altogether, 14 hallmarks and 911 GO were significantly perturbed (adjusted  $p$ -value  $< 0.05$ ). Hallmark enrichment analysis highlighted, among the others, early and late ER pathways as well as inflammatory and interferon gamma responses (Figure 3B). Remarkably, among the GO interesting categories, inflammatory and innate immune responses appeared to be exclusive of gNAT vs. healthy tissue comparison while extracellular matrix emerged from both comparisons (Figure S10B). Finally, 1323 genes have been found significantly up regulated in gNATs when compared with both tumor and normal gastric tissue samples. GSA performed on the significantly up regulated genes resulted in 7 activated hallmark categories (Figure 4A), possibly highlighting the molecular features of the gNAT.



**Figure 4.** (A) GSA of the hallmark gene sets using the genes up regulated in gNAT vs. both the other 2 tissues. Adjusted  $p$ -value in red-blue color scale, (B) Violin plot of the hallmark early and late estrogen receptor (ER) pathways activity in gNAT, normal and tumor, respectively in orange, green and purple. ES obtained by gene set enrichment analysis (GSEA).

Interestingly, early and late estrogen responses together with extracellular matrix seemed to exclusively characterize gNAT (Figure 3B). Inflammatory response, although characteristic of gNAT in its comparison with normal tissue, has not been found enriched in the comparison with tumor. Consistently, a lower expression of estrogen receptors in GC compared to gNAT was reported in literature, confirming our findings [25,26]. Moreover, gene set enrichment analysis (GSEA) on the hallmark categories also confirmed early and late responses to the estrogen as active biological processes (Table S3). In order to shed light on those hallmarks, a single sample gene set enrichment analysis (ssGSEA) was performed on TCGA dataset, as shown in Figure 4B. To further characterize the role of ER pathways in gNATs and GCs we performed univariate analysis of the most interesting clinical variables (grade, TNM, anatomic positions, etc.) (Table S4). Interestingly, we found that both ER pathways were significantly ( $p$ -value < 0.001) associated with tumor histological grade (Figure S11). In particular, lower enrichment scores (ES) were observed in G3 grade while gNATs showed higher, although not significant (probably due to the few numbers of gNATs), ESs levels. Very interestingly, the therapy success (Table S4) and the anatomic localization of the samples (Figure S12A) significantly correlated with the ES of both ER pathways in gNATs. One criticism exploring the ER response can be related to the samples gender but there is no association between ER activity and gender (Figure S12B). In Figure S13 is shown the association of tissue type and grade with late ER pathway activity according to gender (A) or Japanese Gastric Cancer Association (JCGA) anatomical site (B), respectively. Interestingly there was a strong and significant difference in the distribution of the activity of the ER late pathway between tumor and gNAT in patients with G3 grade proximal/middle localized tumor.

### 2.5. Hypothesis Validation Through inHouse GC RNAseq Dataset Generation

In order to validate our approach, we compared, after performing a quality control (Figure S14), the global gene expression profiles of 9 GC and 9 gNAT samples through RNAseq analyses (inHouse dataset). There were 2563 DEGs ( $\text{abs}(\log_2\text{FC}) > 1$  and adjusted  $p$ -value < 0.05), 1625 were found up regulated in tumor vs. gNAT comparison and 938 up regulated in gNAT vs. tumor. Figure 5C shows the hallmark and GO significantly activated categories in both inHouse tumor vs. gNAT and gNAT vs. tumor comparisons. Considering the reduced size of the inHouse samples, compared to the TCGA dataset, we set the threshold at  $p$ -value < 0.05. As reported in Table S5 and Figure 5, good

overlaps between the tumor vs. gNAT upregulated genes in TCGA and inHouse datasets as well as between (A) the hallmark categories (4 out of 6 and 8) and (B) the GO categories (158 out of 484 and 340) were observed.



**Figure 5.** Distribution and description of enriched hallmark and GO gene sets in gNAT vs. Tumor between inHouse and TCGA dataset. (A) depicts the hallmark results, (B) depicts the GO results, (C) depicts GSA results. The *p*-value in red-blue color scale. Gene ratio in dot size scale.

In order to confirm the reliability of the inHouse RNAseq, we processed our 18 samples by digital PCR, analyzing the expression of 6 genes: 1 housekeeping, 1 non-expressed gene and 4 gNAT highly expressed genes, the latter belonging to the hallmark categories enriched in both TCGA and inHouse datasets (myogenesis, late ER and KRAS signaling down). The selection of a good internal standard genes is well documented problem and the suitable genes vary according with tissue specificity and treatments [27]. Here, we selected DEED as housekeeping being its expression among the less variable in TCGA-GTEx dataset. Correlation between digital PCR and RNAseq expression levels was remarkably high ( $R > 0.97$ ) and the “non-expressed” gene was not detected by digital PCR, as well (Table S6).

### 3. Discussion

Our study provides insight into differences among gNAT, GC, and healthy tissue samples as concerns the gene expression profiles. Importantly, we found that gNAT samples are characterized by a peculiar biological behavior different from both healthy and GC tissues. Hence, considering gNAT specimens as control of tumor samples in GC, molecular characterization studies could be affected, by some extent, by false positive and negative as also recently reported in literature for other types of cancer [7]. Using a dimensional reduction procedure, we realized that some of healthy tissue samples were more similar to the muscle tissue than to the gastric mucosa. This finding could be explained by the quality of samples and the collection procedures (few amounts of specimens and/or quick degradation of gastric mucosa). In order to ensure the suitability of normal samples for an appropriate comparison, we decided to discard the muscular-like ones from further analyses. Subsequently, we highlighted the existence of qualitative and quantitative differences in GC DEGs detected by using gNAT or healthy tissue as control. Correlation between the fold change profiles was low and most genes were discordant or at least non-concordant. Interestingly, some gene sets emerged from comparisons of tumor vs. healthy tissue samples. In particular, interferon response and IL2/STAT5 signaling pathways (Figure 3A)

as well as innate and adaptive immune responses (Figure S10A) were missed when comparing tumor with the gNAT specimens. Similarly, angiogenesis (Figure 3A) was detected only in tumor vs. gNAT comparison. Accordingly, we assumed that gNAT is not an appropriate control in gene expression studies aimed to characterize GC molecular signatures and biological features. To better pinpoint gNAT specific characteristics, we examined, firstly separately then together, the gene expression differences among gNAT, healthy gastric mucosa and tumor samples. Comparing gNAT vs. healthy tissue samples (Figure 3B), the enrichment analysis of the hallmark categories highlighted activation of early and late estrogen as well as inflammatory and interferon gamma responses. Remarkably, following GO enrichment (Figure S10B), extracellular matrix, inflammatory and innate immune responses were highlighted. On the other side, considering the differences between gNAT and tumor tissue specimens (Figure 3B), the enrichment analysis identified the activation of hallmark categories like hypoxia together with early and late ER pathways. Interestingly, GO enrichment (Figure S10B) highlighted results similar to those obtained by the comparison of gNAT vs. normal. Moreover, we showed that many gNAT gene profiles were distinct from both healthy and tumor tissues, among which we identified a set of genes specifically overexpressed in gNAT, demonstrating their association with hallmark and GO categories. Therefore, GC gene expression studies not including healthy tissues as control could lead to misidentify gNAT specific genes as selectively under-expressed in the tumor, despite their normal expression levels. Among the results of our analyses, detection of early and late ER pathways activation appeared a quite intriguing finding. In particular, early and late estrogen response gene sets seem to be significantly enriched in the gNAT samples from proximal/middle sites when compared to those from distant sites. No differences on anatomical origin were instead detected among tumor samples. Since in GTEx dataset stomach anatomical sites annotations are missing, it would be useful to consider such information in future investigations. The role of ERs in gastric cancer, possible mechanisms underlying it and clinical relevance of deregulated ERs in GC patients have been widely investigated. Interestingly, ER- $\alpha$  and ER- $\beta$  have been found to be down regulated in tumor tissues when compared to adjacent mucosa samples [26]. Intriguingly, our results suggested an activation of the ERs pathways in proximal/middle adjacent mucosa compared with tumor tissue samples showing, instead, ER normal expression levels. This finding could be explained by the ability of tumor mass to act as an endocrine organ, controlling metabolism and homeostasis of both neighboring and distant tissues [28]. Noteworthy, several studies demonstrated the presence of high expression and activity levels of estrogen production enzymes that could drive the secretion of biologically active estrogens in GC [29,30]. As for any computationally based study, independent confirmation is necessary to draw conclusions. For this reason, we analyzed 9 gastric tumor and 9 gNAT samples producing a RNAseq dataset (inHouse dataset) on which we performed differential expression analysis. The DEGs, hallmark and GO categories obtained from InHouse analyses were consistent with those resulting from TCGA.

## 4. Materials and Methods

### 4.1. Data Collection and Processing

The Cancer Genome Atlas (TCGA) (<https://www.cancer.gov/tcga>) and The Genotype-Tissue Expression (GTEx) Project (<https://www.gtexportal.org/home/>) repositories collected high throughput and clinical data of many tumors and normal tissues from many organs. Although unifying cancer and normal RNA sequencing data from different sources represent a bioinformatics challenge, TCGA and GTEx RNAseq data were successfully merged by Q. Wang et al. enabling cross-study analysis of RNA-Sequencing data [21,22]. From Memorial Sloan Kettering repository (<https://github.com/mskcc>) and from UCSC Xena browser (<https://xenabrowser.net/datapages/>) we retrieved dataset (TCGA STAD and GTEx) and clinical data, respectively. In particular, after crossing RNAseq dataset and clinical data, we obtained the data of 380 gastric cancer RNAseq, 32 adjacent normal tissue and 162 normal samples. It is important to pinpoint that all the samples we call “normal” were collected by GTEx from non-diseased tissue sites. Gene annotation retrieved from TCGA using TCGAbiolinks [31]



package. All the information about TCGA samples' clinical data, pathology reports and tissue are easily retrievable by the cBioPortal (<https://www.cbioportal.org/>) website while the "GTEx Tissue Harvesting Work Instruction" provides all the available information about the GTEx tissues.

To reduce the GC-content bias and standardize the distribution of the counts in each sample we applied EDAsq [32] normalization pipeline adjusting within-lane gene specific effects (GC-content) and between-lane distributional differences (global-scaling using the upper-quartile). Batch effects and differences in sample preparation can have substantial ramifications on the outcomes. Thus, we used stringent removal of unwanted variation, in particular we employed the RUVg method from the RUVSeq package [19], which performs factor analysis on residuals using a negative gene set that has constant covariates. The negative set we used was a list of housekeeping genes [18], and the factors of unwanted variation were added in the design matrix for the regression-like model used by edgeR [33] package to perform differential expression analysis. From ArrayExpress repository [34] we retrieved the 108 samples E-MTAB-1338 [35] dataset containing 71 GC, 21 gNAT, and 16 normal mucosa tissue. Such data were normalized using quantile normalization from beadarray [36] package. Ten GC and 10 gNAT samples obtained from IRCCS-CROB bio-bank (Ethical committee approval N: 20180042426) populated the inHouse dataset (submitted to ArrayExpress repository with ID: E-MTAB-8135).

#### 4.2. inHouse Gene Expression Profiles

For library preparation, a barcoded cDNA library first generated with SuperScript®VILO™ cDNA Synthesis kit (Life Technologies Corporation, Carlsbad, CA, USA) from 10 ng of total RNA. Then cDNA amplified using Ion AmpliSeq™ Transcriptome Human Gene Expression Kit (Life Technologies Corporation) to accurately maintain expression levels of all targeted genes. The average size of each amplicon is ~150 bp. Amplified cDNA Libraries evaluated for quality and quantified using Agilent Bioanalyzer High sensitivity chip. Libraries were then diluted to 100pM and pooled equally, with eight individual samples per pool. Pooled libraries amplified and enriched by using IonChef Instrument (Life Technologies Corporation) according to manufacturer instructions. Templated libraries sequenced on S5™ sequencing system using Ion 540 Kit-Chef kit and chip, obtaining the normalized counts as result of the Target Amplicon-seq pipeline. Moreover, we performed a multidimensional scaling plot of the samples and according to Figure S14 we discarded the "2" (T and S) paired samples as the 2S sample resulted different from the other gNATs.

#### 4.3. Dimensionality Reduction

Dimensionality reduction performed using the Rtsne [37] package on the log<sub>2</sub> CPM values (RNA-seq), or log<sub>2</sub> expression values (microarray). The deconvolution procedure was performed using the DeconRNAsq [23] package and the result of this procedure is a proportion of the "tumor contribution" to the sample.

#### 4.4. Differential Expression Analysis and Venn

Differential expression analysis performed using edgeR. For both RNAseq data, only genes with at least 10 reads in half of the smallest group were included for the analysis, also to handle the pronounced differences in library sizes between TCGA and GTEx (Figure S2). A gene was considered as differentially expressed (DEG) if (1) corrected (FDR)  $p$ -value < 0.05 and (2) > 2-fold ( $\log_2FC > 1$ ) expression change. All the comparisons depicted in the Venn diagrams refer to the 9954 genes (universe) overlapping between inHouse and TCGA dataset.

#### 4.5. Gene-Set Enrichment Analyses

MSigDB [24] Hallmark, GO and KEGG [38] gene sets overrepresentation for each different DEGs lists obtained applying ClusterProfiler [39] and pathview [40] packages. We considered statistically significant the gene sets resulting from the analysis of TCGA-GTEx data with a FDR adjusted  $p$ -value < 0.05 while, to compensate the relative lower number of samples in the inHouse dataset,

we accepted  $p$ -value  $< 0.05$  as threshold. ssGSEA implemented in the GSVA [41] package was used to score samples according to the normalized counts of the genes on MSigDB Hallmark and GO gene sets. The enrichment score (ES) reflects the degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes. Higher the ES, more activated the pathway/biological process.

Statistical analysis performed using the computing environment R [42].

#### 4.6. Digital PCR

From 1000 ng of total RNA reverse transcribed with SuperScript IV VILO Master Mix (Invitrogen), according to manufacturer's protocol, about 10 ng of cDNA were used. Droplet Digital PCR (ddPCR) was performed using the QX200™ Droplet Digital™ PCR system (Bio-Rad Laboratories, Hercules, CA, USA) including the droplet generator and the reader. Reactions were prepared in 20  $\mu$ L volumes with QX200™ ddPCR™ EvaGreen Supermix (Bio-Rad Laboratories) and the following primer sets, retrieved from hg19 AmpliSeq Transcriptome 21K\_v1 DataSheet, were used: NCAM1 (Fw-TGTGGACATCACCTGCTACTTC, Rev-ATGGGCTCCTTGGACTCATC); TFF2 (Fw-AGTGCTGCTTCTCCAACCTTCAT, Rev-TGATAAGGCGAAGTTTCTTCTTTGGT); LTF (Fw-CCTGTCAGCTGCATAAAGAGAGA, Rev-GTAGACTTCCGCCGCTACA); TFF1 (Fw-GCCCTCCCAGTGTGCAAATAA, Rev-GCCCTCCCAGTGTGCAAATAA); CYP1A1 (Fw-CATCCGGGACATCACAGACA, Rev-GAGATAGCAGTTGTGACTGTGTCAA). We used DEDD (Fw-TCAGATGTGTAGCAAGCGGC, Rev-CAGTATTCAGCCCGAACCCG) as housekeeping gene being one of the most stable in TCGA dataset (data not shown). After droplets generation in DG8™ Cartridges (Bio-Rad Laboratories), ddPCR™ 96-Well PCR Plates containing reactions in duplex were loaded onto Verity 96-well Thermal Cycler (Applied Biosystems, Foster City, CA, USA) and cycled as follows: 5 min at 95 °C, 40X (30 sec at 95 °C, 60 s at 60 °C), 5 min at 4 °C, 5 min at 90 °C, and held at 4 °C. QuantaSoft Software v1.7 (Bio-Rad Laboratories) was used to analyze the output of QX200™ Droplet Reader (Bio-Rad Laboratories).

## 5. Conclusions

Our results suggest that using gNAT tissue as control in gene expression analysis could mislead the identification of tumor important pathways. Although gNAT as the control tissue in gene expression analysis allows the identification of the major part of tumor biomarkers and pathways, our results suggest that healthy tissue as control improves the molecular and biological characterization of GC. Interestingly, gNAT tissue shows gene profiles and some biological pathways distinct from both healthy tissue and tumor, although the underlying mechanisms remain to be validated. As supported by recent reports, our findings warrant further investigation on the complex interplay between gNAT and GC tissue. Understanding molecular mechanisms orchestrating this crosstalk could indeed pave the way for identification of novel tumor biomarkers and druggable targets for treatment of GC.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2072-6694/11/9/1248/s1>, Figure S1: Scatter plot of the correlation of the median expression levels of housekeeping genes in normal and gNAT samples. Greater the standard deviation (SD), greater the radius of the points, Figure S2: Boxplot of the library reads in TCGA GC and gNAT and in GTEx normal samples; GC in purple, gNAT in orange and normal tissue in green, Figure S3: Boxplot of the relative log expression (RLE) pre and post (EDAseq) normalization. A shows the relative log expression (RLE) without normalization. There are differences in RLE between TCGA (GC (purple) and gNAT (orange)) samples and GTEx healthy samples (green) while in B, after normalization, no apparent differences are observed, Figure S4: KEGG enrichment of the vascular smooth muscle contraction pathway (hsa04270) using genes up regulated in normal muscular vs. normal mucosa, Figure S5: KEGG enrichment of the gastric acid secretion pathway (hsa04971) using genes up-regulated in normal mucosa vs. normal muscular, Figure S6: Heatmap of the top 900 DEGs between normal mucosa vs. normal muscular samples. log2FCs in red-green color scale. In legend, above the HM, normal mucosa in cyan while normal muscular in red, Figure S7: Scatter plot of the E\_MTAB\_1338 samples after dimensionality reduction procedure. (A) GC in purple, gNAT in orange and normal tissue in green, Figure S8: KEGG pathways enrichment in E\_MTAB\_1338 database using DEGs between the right-low cluster and the middle-high cluster of normal samples, respectively. A Gastric acid secretion pathway (hsa04971). B Vascular smooth muscle contraction pathway (hsa04270). Only the Gastric acid secretion pathway is significantly enriched, Figure S9: Overlap of the enriched hallmark and GO gene sets; A and B depict the results of the overlap between gNAT vs. normal and gNAT vs. GC; C and D depict the results of

the overlap between GC vs. gNAT vs. GC vs. normal, Figure S10: gene set analysis (GSA) of the GO categories using the DEGs between tumor or gNAT vs. each of the other 2 tissues, respectively. Adjusted p-value in red-blue color scale. Gene ratio in dot size scale. Red, blue and green arrows highlight interesting exclusive and common gene sets, respectively. (A) Tumor centered analysis, (B) gNAT centered analysis, Figure S11: Association of the histological grade with hallmark early and late ER pathways activity in gNAT and GC, Figure S12: Association of JGCA anatomical part (A) and gender (B) with hallmark early and late ER pathways activity in gNAT and GC, Figure S13: Association of therapy tissue type and grade with hallmark late ER pathway activity according to gender (A) or JCGA anatomical site (B), respectively, Figure S14: multidimensional scaling plot of the RNA samples in which distances correspond to leading log-fold-changes between each pair of RNA samples. GC in purple, gNAT in orange, labels represents the name of the samples, Table S1: Summary of the differentially expressed genes among adjacent, normal and tumor samples., Data S1: TCGA DEGs tumor vs. adjacent vs. normal, Data S2: Enriched hallmark in TCGA gNAT, Data S3: univariate analysis of the most interesting clinical variables (grade, TNM, anatomic positions, etc.) on the TCGA data, Data S4: inHouse DEGs in GC vs. gNAT, Data S5: digital PCR validations.

**Author Contributions:** Conceptualization, P.Z., G.F. and S.R.; methodology, P.Z. and G.C.; validation, S.R., F.L.R., S.L. and E.A.; resources, C.L.P. and D.C.; data curation, G.C., C.L.P. and S.L.; writing—original draft preparation, S.R. and P.Z.; writing—review and editing, V.R., P.M., G.N., M.D.F. and G.F.; supervision, P.Z. and G.F.

**Funding:** This work was supported by current research funds, Italian Ministry of Health, to IRCCS-CROB, Rionero in Vulture, Potenza, Italy.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Van Cutsem, E.; Sagaert, X.; Topal, B.; Haustermans, K.; Prenen, H. Gastric cancer. *Lancet* **2016**, *388*, 2654–2664. [[CrossRef](#)]
2. Zong, L.; Abe, M.; Seto, Y.; Ji, J. The challenge of screening for early gastric cancer in China. *Lancet* **2016**, *388*, 2606. [[CrossRef](#)]
3. Lauren, P. The two histological main types of gastric carcinoma: Diffuse and so-called intestinal-type carcinoma. An attempt at a histo-clinical classification. *Acta Pathol. Microbiol. Scand.* **1965**, *64*, 31–49. [[CrossRef](#)] [[PubMed](#)]
4. Bosman, F.T.; Carneiro, F.; Hruban, R.H.; Theise, N.D. *WHO Classification of Tumours of the Digestive System*; IARC Press: Lyon, France, 2010.
5. Heaphy, C.M.; Bisoffi, M.; Fordyce, C.A.; Haaland, C.M.; Hines, W.C.; Joste, N.E.; Griffith, J.K. Telomere DNA content and allelic imbalance demonstrate field cancerization in histologically normal tissue adjacent to breast tumors. *Int. J. Cancer* **2006**, *119*, 108–116. [[CrossRef](#)]
6. Heaphy, C.M.; Griffith, J.K.; Bisoffi, M. Mammary field cancerization: Molecular evidence and clinical importance. *Breast Cancer Res. Treat.* **2009**, *118*, 229–239. [[CrossRef](#)] [[PubMed](#)]
7. Aran, D.; Camarda, R.; Odegaard, J.; Paik, H.; Oskotsky, B.; Krings, G.; Goga, A.; Sirota, M.; Butte, A.J. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat. Commun.* **2017**, *8*, 1077. [[CrossRef](#)] [[PubMed](#)]
8. Graham, K.; Ge, X.; de Las Morenas, A.; Tripathi, A.; Rosenberg, C.L. Gene expression profiles of estrogen receptor-positive and estrogen receptor-negative breast cancers are detectable in histologically normal breast epithelium. *Clin. Cancer Res.* **2011**, *17*, 236–246. [[CrossRef](#)]
9. Casbas-Hernandez, P.; Sun, X.; Roman-Perez, E.; D’Arcy, M.; Sandhu, R.; Hishida, A.; McNaughton, K.K.; Yang, X.R.; Makowski, L.; Sherman, M.E.; et al. Tumor intrinsic subtype is reflected in cancer-adjacent tissue. *Cancer Epidemiol. Biomark. Prev.* **2015**, *24*, 406–414. [[CrossRef](#)]
10. Troester, M.A.; Lee, M.H.; Carter, M.; Fan, C.; Cowan, D.W.; Perez, E.R.; Pirone, J.R.; Perou, C.M.; Jerry, D.J.; Schneider, S.S. Activation of host wound responses in breast cancer microenvironment. *Clin. Cancer Res.* **2009**, *15*, 7020–7028. [[CrossRef](#)]
11. Trujillo, K.A.; Heaphy, C.M.; Mai, M.; Vargas, K.M.; Jones, A.C.; Vo, P.; Butler, K.S.; Joste, N.E.; Bisoffi, M.; Griffith, J.K. Markers of fibrosis and epithelial to mesenchymal transition demonstrate field cancerization in histologically normal tissue adjacent to breast tumors. *Int. J. Cancer* **2011**, *129*, 1310–1321. [[CrossRef](#)]
12. Chandran, U.R.; Dhir, R.; Ma, C.; Michalopoulos, G.; Becich, M.; Gilbertson, J. Differences in gene expression in prostate cancer, normal appearing prostate tissue adjacent to cancer and prostate tissue from cancer free organ donors. *BMC Cancer* **2005**, *5*, 45. [[CrossRef](#)]

13. Tung, E.K.-K.; Mak, C.K.-M.; Fatima, S.; Lo, R.C.-L.; Zhao, H.; Zhang, C.; Dai, H.; Poon, R.T.-P.; Yuen, M.-F.; Lai, C.-L.; et al. Clinicopathological and prognostic significance of serum and tissue Dickkopf-1 levels in human hepatocellular carcinoma. *Liver Int.* **2011**, *31*, 1494–1504. [[CrossRef](#)]
14. Sanz-Pamplona, R.; Berenguer, A.; Cordero, D.; Molleví, D.G.; Crous-Bou, M.; Sole, X.; Paré-Brunet, L.; Guino, E.; Salazar, R.; Santos, C.; et al. Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer. *Mol. Cancer* **2014**, *13*, 46. [[CrossRef](#)]
15. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **2013**, *45*, 580–585. [[CrossRef](#)]
16. The Cancer Genome Atlas (TCGA). Research Network. Available online: <http://cancergenome.nih.gov> (accessed on 3 June 2019).
17. Rahman, M.; Jackson, L.K.; Johnson, W.E.; Li, D.Y.; Bild, A.H.; Piccolo, S.R. Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. *Bioinformatics* **2015**, *31*, 3666–3672. [[CrossRef](#)]
18. Eisenberg, E.; Levanon, E.Y. Human housekeeping genes, revisited. *Trends Genet.* **2013**, *29*, 569–574. [[CrossRef](#)]
19. Risso, D.; Ngai, J.; Speed, T.P.; Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **2014**, *32*, 896–902. [[CrossRef](#)]
20. Risso, D.; Schwartz, K.; Sherlock, G.; Dudoit, S. GC-content normalization for RNA-Seq data. *BMC Bioinform.* **2011**, *12*, 480. [[CrossRef](#)]
21. Wang, Q.; Armenia, J.; Zhang, C.; Penson, A.V.; Reznik, E.; Zhang, L.; Minet, T.; Ochoa, A.; Gross, B.E.; Iacobuzio-Donahue, C.A.; et al. Unifying cancer and normal RNA sequencing data from different sources. *Sci. Data* **2018**, *5*, 180061. [[CrossRef](#)]
22. Wang, Q.; Armenia, J.; Zhang, C.; Penson, A.V.; Reznik, E.; Zhang, L.; Ochoa, A.; Gross, B.E.; Iacobuzio-Donahue, C.A.; Betel, D.; et al. Enabling cross-study analysis of RNA-Sequencing data. *BioRxiv* **2017**. [[CrossRef](#)]
23. Gong, T.; Hartmann, N.; Kohane, I.S.; Brinkmann, V.; Staedtler, F.; Letzkus, M.; Bongiovanni, S.; Szustakowski, J.D. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS ONE* **2011**, *6*, e27156. [[CrossRef](#)] [[PubMed](#)]
24. Liberzon, A.; Birger, C.; Thorvaldsdóttir, H.; Ghandi, M.; Mesirov, J.P.; Tamayo, P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **2015**, *1*, 417–425. [[CrossRef](#)] [[PubMed](#)]
25. Frycz, B.A.; Murawa, D.; Borejsza-Wysocki, M.; Wichtowski, M.; Spychała, A.; Marciniak, R.; Murawa, P.; Drews, M.; Jagodziński, P.P. mRNA expression of steroidogenic enzymes, steroid hormone receptors and their coregulators in gastric cancer. *Oncol. Lett.* **2017**, *13*, 3369–3378. [[CrossRef](#)] [[PubMed](#)]
26. Gan, L.; He, J.; Zhang, X.; Zhang, Y.-J.; Yu, G.-Z.; Chen, Y.; Pan, J.; Wang, J.-J.; Wang, X. Expression profile and prognostic role of sex hormone receptors in gastric cancer. *BMC Cancer* **2012**, *12*, 566. [[CrossRef](#)] [[PubMed](#)]
27. Falco, G.; Stanghellini, I.; Ko, M.S.H. Use of Chuk as an internal standard suitable for quantitative RT-PCR in mouse preimplantation embryos. *Reprod. Biomed. Online* **2006**, *13*, 394–403. [[CrossRef](#)]
28. Lee, Y.-M.; Chang, W.-C.; Ma, W.-L. Hypothesis: Solid tumours behave as systemic metabolic dictators. *J. Cell Mol. Med.* **2016**, *20*, 1076–1085. [[CrossRef](#)] [[PubMed](#)]
29. Izawa, M.; Inoue, M.; Osaki, M.; Ito, H.; Harada, T.; Terakawa, N.; Ikeguchi, M. Cytochrome P450 aromatase gene (CYP19) expression in gastric cancer. *Gastric Cancer* **2008**, *11*, 103–110. [[CrossRef](#)]
30. Saitoh, Y.; Sasano, H.; Naganuma, H.; Ohtani, H.; Sasano, N.; Ohuchi, A.; Matsuno, S. De novo expression of aromatase in gastric carcinoma. Light and electron microscopic immunohistochemical and immunoblot study. *Pathol. Res. Pract.* **1992**, *188*, 53–60. [[CrossRef](#)]
31. Colaprico, A.; Silva, T.C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T.S.; Malta, T.M.; Pagnotta, S.M.; Castiglioni, I.; et al. TCGAAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **2016**, *44*, e71. [[CrossRef](#)]
32. Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **2010**, *11*, R106. [[CrossRef](#)]
33. Zhou, X.; Lindsay, H.; Robinson, M.D. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res.* **2014**, *42*, e91. [[CrossRef](#)]
34. Kolesnikov, N.; Hastings, E.; Keays, M.; Melnichuk, O.; Tang, Y.A.; Williams, E.; Dylag, M.; Kurbatova, N.; Brandizi, M.; Burdett, T.; et al. ArrayExpress update—Simplifying data submissions. *Nucleic Acids Res.* **2015**, *43*, D1113–D1116. [[CrossRef](#)]

35. Zhao, C.-M.; Hayakawa, Y.; Kodama, Y.; Muthupalani, S.; Westphalen, C.B.; Andersen, G.T.; Flatberg, A.; Johannessen, H.; Friedman, R.A.; Renz, B.W.; et al. Denervation suppresses gastric tumorigenesis. *Sci. Transl. Med.* **2014**, *6*, 250ra115. [[CrossRef](#)]
36. Dunning, M.J.; Smith, M.L.; Ritchie, M.E.; Tavaré, S. beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics* **2007**, *23*, 2183–2184. [[CrossRef](#)]
37. Krijthe, J.H. Rtsne: T-Distributed Stochastic Neighbor Embedding Using Barnes-Hut Implementation. 2015. Available online: <https://github.com/jkrijthe/Rtsne> (accessed on 18 July 2019).
38. Kanehisa, M.; Sato, Y.; Furumichi, M.; Morishima, K.; Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **2019**, *47*, D590–D595. [[CrossRef](#)]
39. Yu, G.; Wang, L.-G.; Han, Y.; He, Q.-Y. clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **2012**, *16*, 284–287. [[CrossRef](#)]
40. Luo, W.; Brouwer, C. Pathview: An R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **2013**, *29*, 1830–1831. [[CrossRef](#)]
41. Hänzelmann, S.; Castelo, R.; Guinney, J. GSEA: Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform.* **2013**, *14*, 7. [[CrossRef](#)]
42. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019; Available online: <http://www.R-project.org> (accessed on 18 July 2019).



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).