International Journal of Biochemistry and Cell Biology 108 (2019) 51–60

Contents lists available at ScienceDirect

International Journal of Biochemistry and Cell Biology

journal homepage: www.elsevier.com/locate/biocel

Exploiting single-cell RNA sequencing data to link alternative splicing and cancer heterogeneity: A computational approach

Ichcha Manipur[1], Ilaria Granata[1,*], Mario Rosario Guarracino

High Performance Computing and Networking Institute, National Research Council, Italy

ARTICLE INFO

Keywords:
Alternative splicing
Isoforms
Single-cells
Tumor heterogeneity
Molecular classification

ABSTRACT

Cell heterogeneity studies using single-cell sequencing are gaining great significance in the era of personalized medicine. In particular, characterization of tumor heterogeneity is an emergent issue to improve clinical oncology, since both inter- and intra-tumor level heterogeneity influence the utility and application of molecular classifications through specific biomarkers. Majority of studies have exploited gene expression to discriminate cell types. However, to provide a more nuanced view of the underlying differences, isoform expression and alternative splicing events have to be analyzed in detail.

In this study, we utilize publicly available single cell and bulk RNA sequencing datasets of breast cancer cells from primary tumors and immortalized cell lines. Breast cancer is very heterogeneous with well defined molecular subtypes and was therefore chosen for this study. RNA-seq data were explored in terms of genes, isoforms abundance and splicing events. The study was conducted from an average based approach (gene level expression) to detailed and deeper ones (isoforms abundance/splicing events) to perform a comparative analysis, and, thus, highlight the importance of the splicing machinery in defining the tumor heterogeneity. Moreover, here we demonstrate how the investigation of gene isoforms expression can help to identify the appropriate *in vitro* models. We furthermore extracted marker isoforms, and alternatively spliced genes between and within the different single cell populations to improve the classification of the breast cancer subtypes.

1. Introduction

1.1. Opportunities from new technologies

The emergence of new sequencing methods and analysis approaches allows deepening of the study into nucleic acids, concerning sequence, interaction, and abundance. In the era of personalized medicine, the primary purpose of research is the possibility to investigate and characterize biological phenomena, taking into account the heterogeneity among and within individuals, both in health and disease conditions. The need for such detailed information is particularly urgent in the case of highly heterogeneous diseases, such as cancer. Tumor heterogeneity can be classified as inter-tumor (tumor by tumor) and intra-tumor (within a tumor) heterogeneity. Tumor heterogeneity is the major contributing factor for the refractory nature of many cancers (Dagogo-Jack and Shaw, 2018). Traditional approaches applied to the study of the transcriptome can be considered as average-based methods and can lead to the loss of significant information. Specifically, bulk level molecular phenotyping represents the outcome phenotyping of a large number of cells and does not take into account factors such as clonal evolution, tissue hierarchies, rare cells and dynamic cell states. As opposed to this, single-cell RNA sequencing (scRNA-Seq) allows us to analyze gene expression variability at the single-cell level (Deng et al., 2014; Wang and Navin, 2015) and thus also to investigate the heterogeneity among cells. ScRNA-Seq is becoming more popular year after year in spite of the higher cost and has been exploited in a wide range of research topics, including studies of circulating tumor cells (Ramskö et al., 2012; Deng et al., 2014), breast cancer (Nguyen et al., 2018; Chung et al., 2017; Savas et al., 2018), prostate cancer (Horning et al., 2017), transcriptional dynamics (Trapnell et al., 2014), cell cycle (Kowalczyk et al., 2015), tissue heterogeneity (Achim et al., 2015) and many others. Algorithms, pipelines, and methods for analyzing data coming from single-cell sequencing represent an exciting and challenging issue for bioinformatics. Gene-level abundance estimation can also be considered as an average-based method of investigating the regulation of transcription machinery since gene expression is the result of different isoform contributions. Alternative splicing (AS) considerably expands the functional repertoire of eukaryotic genomes.

* Corresponding author.
E-mail address: ilaria.granata@icar.cnr.it (I. Granata).
[1] Equal contributor.

https://doi.org/10.1016/j.biocel.2018.12.015
Received 12 October 2018; Received in revised form 21 December 2018; Accepted 24 December 2018
Available online 08 January 2019

Transcriptional isoforms are mRNA molecules originating from the same locus but having different length and exon composition, and, as a consequence, they give rise to multiple forms of the corresponding protein. This diversity can derive from different transcriptional starting or polyadenylation sites, or mostly from alternative splicing mechanisms (Black, 2003; Matlin et al., 2005). There are numerous studies of alternative splicing using bulk-cell RNA-Seq, but, to date, there are relatively few studies that are focused on characterization of isoforms expression at the single-cell level (Song et al., 2017; Vu et al., 2018; Faigenbloom et al., 2015). The characterization of transcriptional heterogeneity at the single cell level is a powerful resource to fulfill different tasks and in the current study, we address some of them.

1.2. Primary vs immortalized cell lines

Despite rapid scientific progress, cell line-based assays still represent an essential tool for pharmaceutical, chemical, medical and cosmetic industries. The lower costs, easiness of handling of culture methods, and high reproducibility ensure their extensive use. However, the relevance of cell lines as tumor models strongly depends on the type of experimental approach and on how close their molecular landscape is to that of tumor tissue (Gillet et al., 2013). Thus, the investigation and the definition of this closeness is a critical issue and might lead to better use of the *in vitro* models. Many works have highlighted a weak correlation between cell lines and tumors in terms of CNV, mutation, gene, and protein expression (Jiang et al., 2016; Vincent et al., 2015; Ahmed et al., 2013; Ince et al., 2015; Qiu et al., 2016). To the best of our knowledge, there are no studies which have considered transcriptional isoforms to compare immortalized and primary cell lines. We addressed this issue in previous work and highlighted the presence of alternative splicing events and possible causative nucleotide variants which likely determine the distance between hepatocellular carcinoma cells and HepG2 cell line (Tripathi et al., 2017). Nonetheless, in our opinion, increasing the resolution of the analysis can help identify the right model for the specific condition, rather than weaken the possibility to use such widespread models. Breast cancer is one of the most heterogeneous tumors and greatly differs among patients (inter-tumor heterogeneity) and even within each tumor (intra-tumor heterogeneity) (Badve, 2016).

1.3. Tumor heterogeneity

Characterization of molecular signatures is an indicator of genetic tumor heterogeneity, which can lead to improved stratification for personalized therapy. The intra-tumor heterogeneity occurs at the morphologic, genomic, transcriptomic, and proteomic levels, thus determining new diagnostic and therapeutic challenges. Understanding the players and mechanisms underlying tumor heterogeneity has become crucial. The overall knowledge of tumor heterogeneity has drastically increased, and theories based on cancer stem cells have become very popular (Batlle and Clevers, 2017), but, still, there are only limited advancements in diagnostic, prognostic, or predictive strategies for breast cancer. Discovery and validation of biomarkers aim to maximize patient eligibility for targeted therapy, but the intra-tumor heterogeneity is rarely considered in these cases. Molecular classification of breast cancer is not implemented in routine clinical practice. New in-depth analyses are required to manage and get insight from the vast amount of data continuously produced. Genetic expression patterns divide breast cancer into four major molecular subtypes with prognostic and therapeutic implications: Luminal A (lum A), Luminal B (lum B), HER2-enriched (HER2+), and basal-like (Dai et al., 2015). The lum A and lum B subtypes have better survival than HER2+ and Basal-like subtypes. Each of the four subtypes is nicely mapped to an immunohistochemical-defined subtype. Both luminal subtypes express ER, but the lum B tumors are characterized by increased expression of proliferation-associated genes and have a worse prognosis than lum A tumors. The HER2+ subtype is characterized by increased expression of HER2 and proliferation genes and includes ER−/PR−/HER2+ and ER+/PR+/HER2+ tumors. The basal-like subtype is enriched for genes expressed in basal epithelial cells, of which 70% are triple-negative (TN) with ER−/PR−/HER2− profile. Several studies have been aimed at identifying gene expression signatures of these subtypes with various numbers of genes included. For instance, Hu et al. (2006) found a 306 genes signature that can distinguish these subtypes with significant differences observed on relapse-free and overall survival. Parker et al. reported a 50-gene classifier (PAM50) which contains mostly hormone receptor and proliferation-related genes, and genes exhibiting myoepithelial and basal features. The subtypes can be assessed using a multiplexed gene-expression profiling technology (NanoString Technologies; Seattle, WA, USA) which has significant prognostic and predictive values on breast tumors and can be widely applied in the clinical setting (Parker et al., 2009). Although containing different genes, the signatures identified by different studies should belong to the same pathways, thus not generating a divergent classification of samples. Unfortunately, this is still not possible due to the lack of stringent standardization of the methodology and breast cancer intrinsic subtype definition. Immunohistochemical panels are often used to improve the accuracy of PAM50 array based classification (Allott et al., 2018). Misclassification can be due to the presence of normal breast tissue or stroma contamination into the samples (Elloumi et al., 2011). From this perspective, the use of single-cell sequencing comes to aid, since the discrimination of single cells allows to remove the non-tumor cells from the downstream analyses. It is our opinion that, also, a different regulation of the splicing machinery can strongly be implicated in heterogeneity mechanisms. Here we present a comparative study to highlight the importance and the influence of splicing variants in defining the intra- and inter-tumor heterogeneity, and how they can help in improving the accuracy of subtypes classification.

2. Materials and methods

2.1. Data and sequencing files pre-processing

RNA sequencing data of single cells were obtained from publicly available datasets. Since our study was focused on isoforms and splicing event analysis, we searched for all available single cell datasets obtained from full length sequencing protocols as UMI based data is not suitable for such studies. The first single cell public dataset was obtained from patients with different subtypes of breast cancer: Lum A (BC01, BC02), Lum B (BC03), HER2+ (BC04–BC06) and TN (BC07–BC11). Single cells from metastatic lymph nodes were also obtained from two patients BC03LN and BC07LN. 317 single cells which were identified as epithelial breast cancer cells after performing tumor purity estimation by Chung et al were downloaded from the Gene Expression Omnibus (GEO) portal (GSE75688). Out of these samples 75 single cells from patient BC05 were excluded as this patient had undergone neoadjuvant chemotherapy and Herceptin treatment, while the other patients had not undertaken treatment prior to mastectomy.

Sequencing data of 96 single cells of MDA-MB-231 and T47D cell lines were obtained from the NCBI Sequence Read Archive: PRJNA419090, while data of 68 cells from MCF7 and SKBR3 were downloaded from the EBI European Nucleotide Archive (ERP022266). Bulk sequencing data were obtained from the GEO portal: primary samples (GSE71651), MCF7 (GSE80537), MDA-MB-468 (GSE90519) and BT549 (GSE112365) and EBI ENA: SKBR3 (ERP022266). FastQC (Andrews et al., 2010) was used for quality control of reads and Trim Galore (Krueger and Galore, 2015) was used for trimming reads and library adapters.

2.2. Genes and isoforms abundance estimation

Reads from both bulk and single cell samples were aligned to the

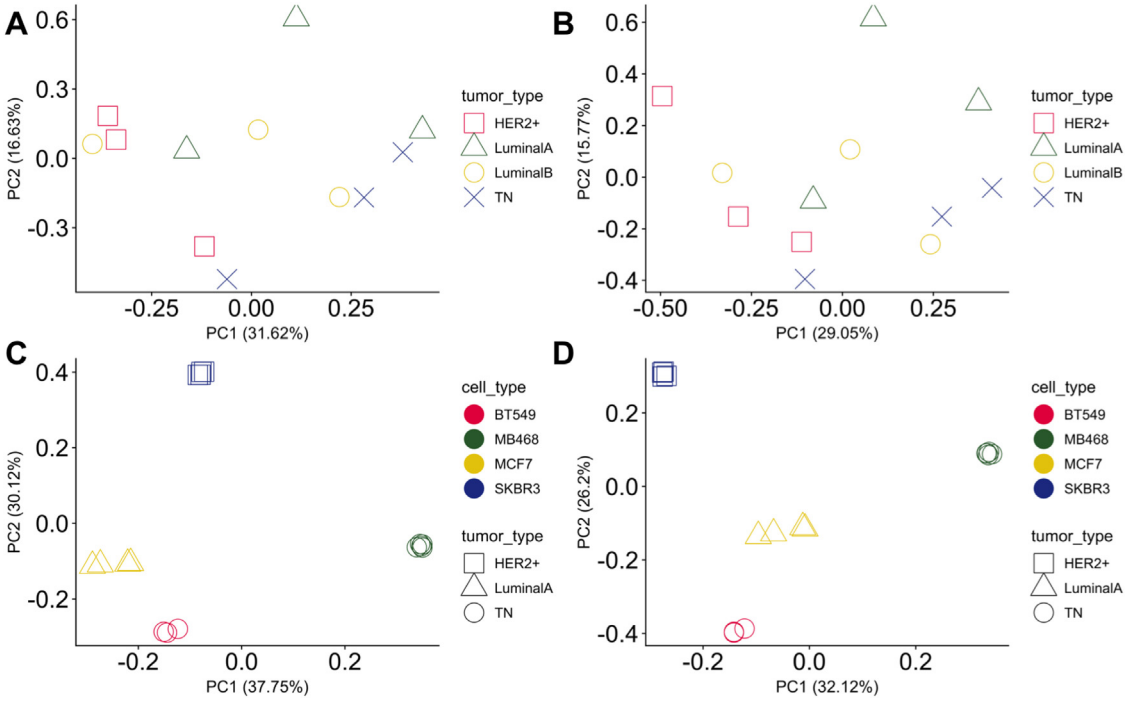

Fig. 1. PCA of bulk sequencing data. PCA was performed on gene-level (A) and isoform-level (B) expression of primary tumor samples, as well as on gene-level (C) and isoform-level (D) expression of immortalized cell lines, models for the subtypes indicated in the relative legend.

hg38v27 human reference genome using STAR (Dobin et al., 2013) in the two-pass mode. TPM normalized gene and isoform counts were obtained using RSEM (v1.3.0) (Li and Dewey, 2011). Single cells with a mapping rate ≥60% and ≥1 million reads were retained. Cells were filtered based on the total number of reads, the number of genes detected and the percentage of mitochondrial genes. Finally, 351 single cells passed the QC criteria and were used for downstream analyses. Genes/isoforms expressed in ≤8 cells were removed and TPM expression values ≤5 were replaced with zero. Seurat (Macosko et al., 2015) R package was used to log normalize single cell expression data and for the detection of highly variable genes/isoforms. We also used Seurat to identify the presence of batch effects by regressing out the different experimental batches. On visualization with Principal component analysis (PCA), we identified no effects of batch and proceeded without any correction. TPM counts of bulk samples were log normalized and highly variable genes were detected using the genefilter R package (Gentleman et al., 2018).

PCA and t-Distributed Stochastic Neighbor Embedding (t-SNE) were used for gene/isoform expression visualization of bulk and single cell samples, as well as primary tumors and cell lines. To study the clustering pattern of single cells belonging to primary tumors and cell lines, clustering based on isoforms expression was performed using Seurat. Seurat performs clustering by constructing a Shared Nearest Neighbor (SNN) graph and then optimizing the modularity function (Waltman and Van Eck, 2013). We set different resolution parameters ranging from 0.5 to 2 to assess optimal cell clustering.

Differential isoforms expression analysis of primary tumor subtypes was performed using the Wilcoxon rank-sum test available in Seurat. 131 marker isoforms with an adjusted *p*-value ≤0.05 and ≥2 average fold change were detected and gene set enrichment analysis was performed using the Molecular Signature Database (MSigDB) (Subramanian et al., 2005; Liberzon et al., 2011, 2015).

2.3. Alternative splicing events detection

Alternative splicing analysis of single cells was performed using BRIE, the Bayesian Regression for Isoform Estimation tool (Huang and Sanguinetti, 2017). Brie outputs two different outputs: (1) table of the genes with alternative splicing events and the associated FPKM values per each sample; (2) a list of differentially spliced events ranked by the number of cell pairs in which the splicing events are detected. We used both the output files for two different purposes: (1) to detect the inter-tumor heterogeneity we extracted FPKM normalized isoform estimates from the BRIE output for different tumor subtypes. We then applied the Wilcoxon rank-sum test in Seurat and identified 95 alternatively spliced genes (adjusted *p*-value ≤0.05 and ≥2 average fold change); (2) to investigate the differential splicing within groups. We chose a threshold of events detected in at least in ≥30% of total cells. This translates to a cell pair threshold of ≥30% of possible pairwise cell comparisons given by nCr where n is the number of cells and r is equal to 2.

2.4. Subtypes classification

Classification was performed using the sequential minimal optimization (SMO) (Platt, 1999) algorithm, an implementation of support vector machine (SVM) classification available in WEKA (Hall et al., 2009). We used three lists of genes/isoforms for comparison of subtype classification. (A) Top 50 expressed marker isoforms; (B) top 50 alternatively spliced marker genes; and (C) the Pam50 gene set. We used the Poly kernel followed by a ten-fold cross-validation to assess the accuracy of the three models in classifying cells into the four different tumor subtypes. The number of correct and incorrect predictions is summarized with count values in the confusion matrix.

3. Results and discussion

3.1. Isoform-level abundance of single cells better discriminates breast cancer subtypes

Both single cell and bulk sequencing data were subjected to dimensionality reduction techniques to investigate the similarity among the groups under study. Principal component analysis (PCA) was applied to bulk data, while t-distributed Stochastic Neighbor Embedding (t-SNE) was used for single cell since it can capture complex non-linear

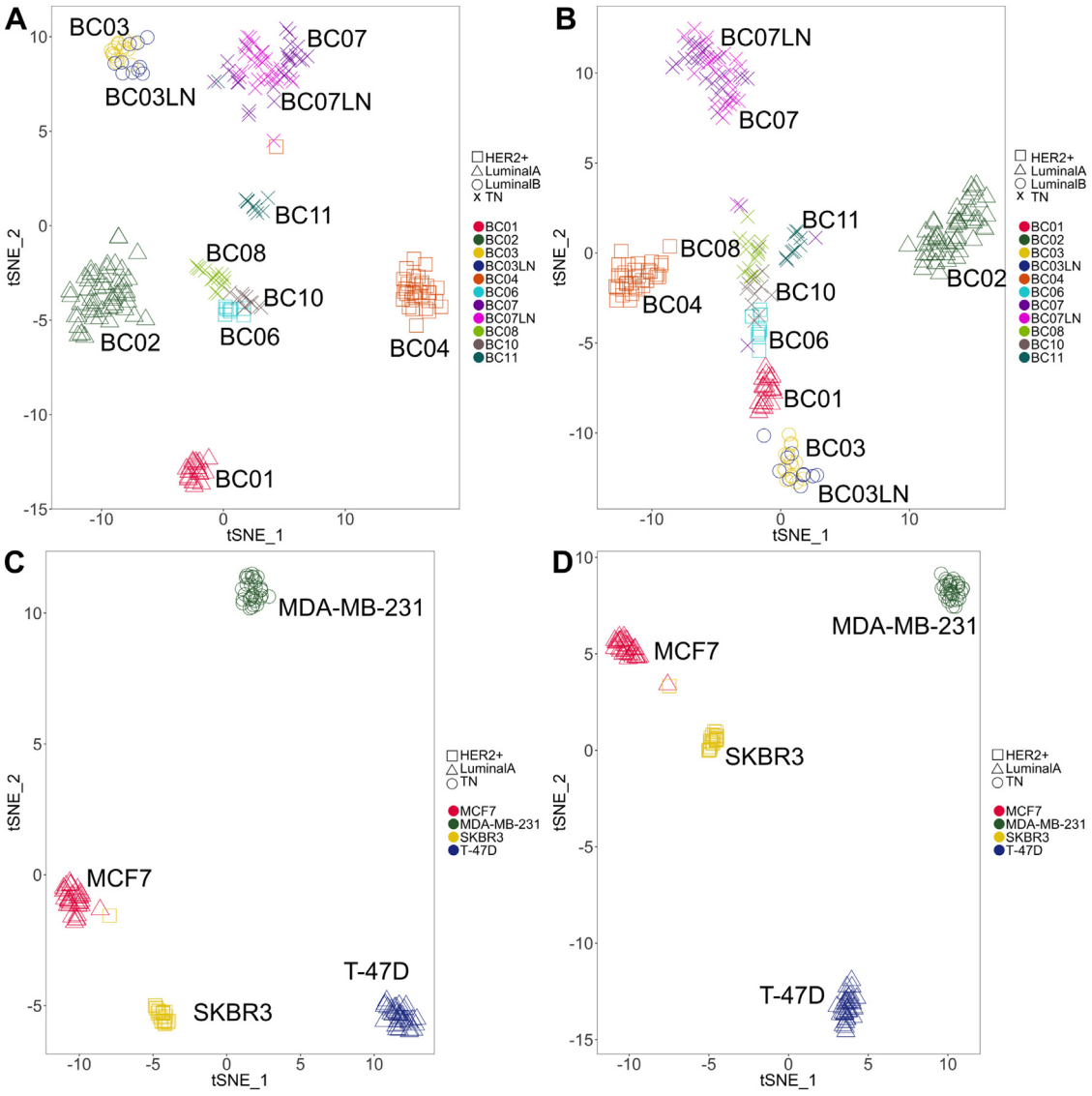

Fig. 2. t-SNE of single-cell sequencing data. t-SNE was performed on gene-level (A) and isoform-level (B) expression of primary tumor samples, as well as on gene-level (C) and isoform-level (D) expression of immortalized cell lines, models for the subtypes indicated in the relative legend.

structures better than PCA and it is suitable to find cell-types. Gene and isoform level abundances were used separately to perform the analysis and to compare the discriminative power of the two approaches. As shown in Fig. 1A and B, the bulk sequencing of primary cells does not show any difference between the gene and isoform expression, indeed, in both cases, the samples belonging to the same subtype showed dis-homogeneity. Even for the immortalized cell lines, gene and isoform abundance provided the same visualization but with an expected higher homogeneity (Fig. 1C and D).

Single cell data allowed us to go into details of the inter-tumor and intra-tumor heterogeneity. As shown in Fig. 2A and B, there was a good separation of the different samples. The overall distribution highlights some differences between the gene- (A) and isoform (B) level, suggesting that the isoform expression, rather than genes, contributes to the intra-tumor heterogeneity. TN cells presented more spread at isoform-level and there is more closeness with lum A BC01 and lum B BC03 along the t-SNE 1. The Lum A patients BC01 and BC02 were close at gene-level but the isoforms determined their separation along both t-SNE 1 and t-SNE 2. The HER2+ patient BC06 showed in both cases a great similarity with TN cells rather than with the patient of the same subtype BC04. The gene expression placed close the patients having lymph node metastasis, BC03/BC03LN and BC07/BC07LN, even if they belonged to Lum B and TN respectively. This closeness was instead not present in the case of the isoform expression. It is also worthy of notice that the metastatic site cells showed the same signature of the primary tumor. Moreover, the sample BC07, which is classified as TN, was distant from the samples belonging to the same subtype, likely due to the presence of metastasis, but at isoform-level, some of its cells were closer to the other TN samples, suggesting an intra-tumor heterogeneity made of primary cells and cells with metastatic potential, which was not recognizable at gene-level. Lum A BC01 and Lum B BC03 patients showed isoform level similarity and gene level divergences. Regarding the cell lines, no differences were evaluable between gene and isoform level expression and the group of cells were well defined and homogeneous in both cases (Fig. 2C and D).

3.2. Clustering based on isoforms expression reveals the inter- and intra-tumor heterogeneity

Unsupervised clustering was performed on isoforms expression of samples to investigate the tumor heterogeneity and the different cell types. The clusters were then plotted in the space of the t-SNE for visualization. Clustering of all primary cells (Fig. 3) highlighted that the two Lum A patients showed such a different pattern of expression to fall

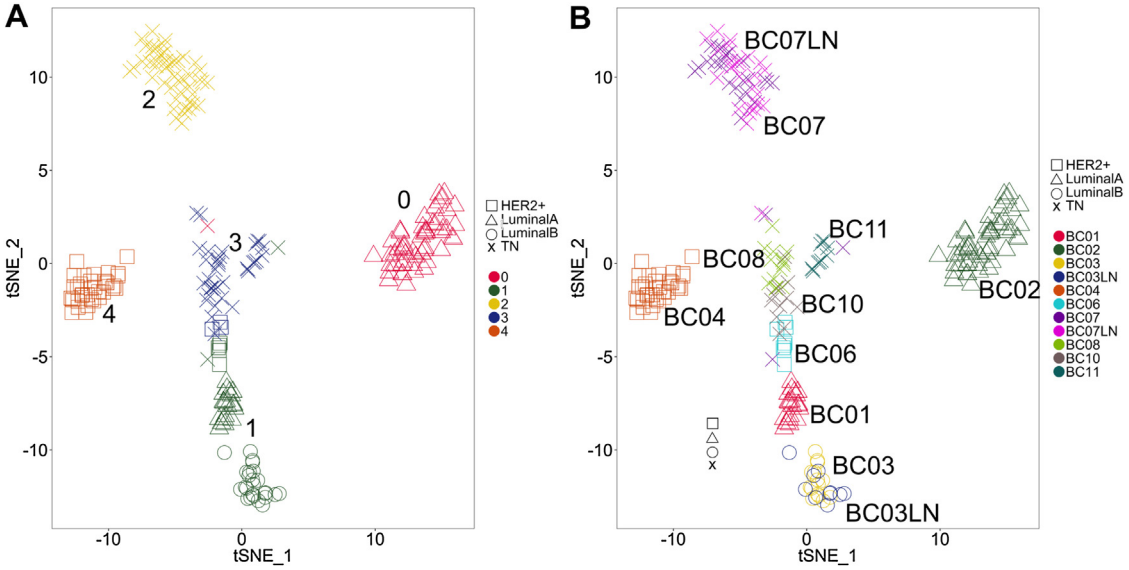

Fig. 3. Clustering of primary cells (A) plotted into the space of t-SNE (B).

into two different clusters. One of them, BC01, was closer to Lum B patient BC03. At the same cluster belonged the HER2+ sample BC06 as well, which seemed to be divided between this cluster and the one containing the vast majority of TN samples, but distant from the cluster containing the other HER2+ sample, BC04, which instead formed one cluster alone. TN samples, as already seen by previous analysis, were divided into metastatic and not metastatic tumor, even though the metastatic tumor patient, BC07, had some cells belonging to the cluster of non-metastatic samples.

In order to better interpret the unexpected localization of some cells, we plotted the expression of HER2 expression at gene and isoform level (Fig. 4).

The most predominant isoforms were ENST00000541774.5 (protein-coding) and ENST00000583038.5 (no protein due to retained intron). The protein-coding isoform was highly expressed in most of the cells of HER2+ sample BC04, less in BC06, where some cells showed low or no expression of this isoform and higher expression of the no-protein isoform. The other HER2 enriched subtype, Lum B, showed higher expression of HER2 in lymph node metastasis than in the primary tumor. The rest of the samples had predominantly the no-protein isoform. The gene-level, as expected, showed an average expression of the two isoforms. This result confirms that the molecular subtype distinction, based essentially on the expression of ER, PgR and HER2, is not valid for all the cells of one tumor tissue. Cluster analysis was also

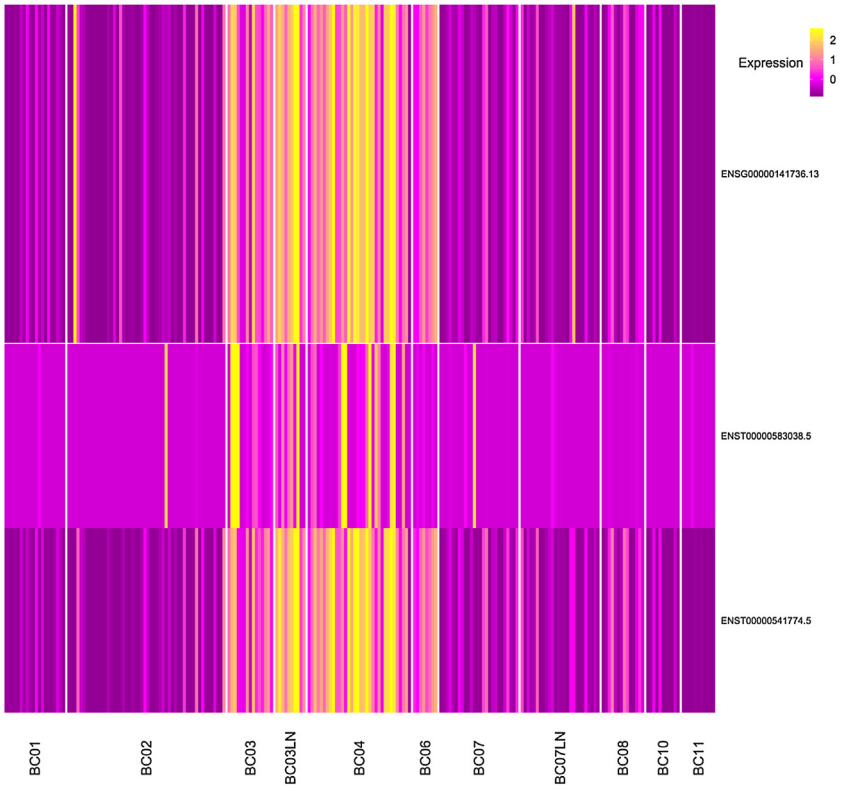

Fig. 4. HER2 gene-level and isoform-level expression in the single cells of patients under study.

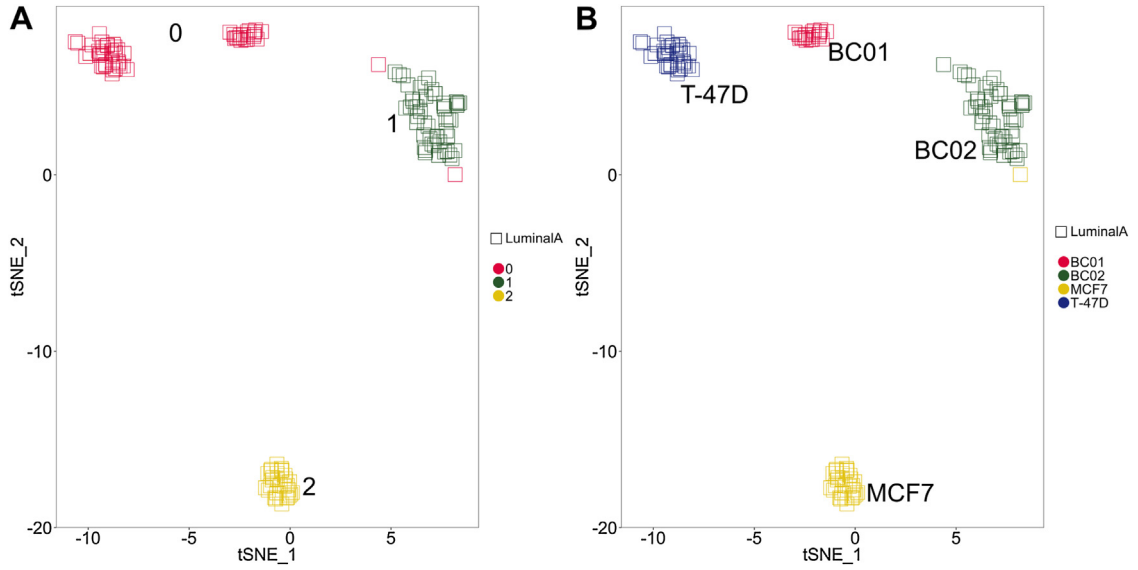

Fig. 5. Clustering of Lum A primary cells and its specific *in vitro* models (A) plotted into the space of t-SNE (B).

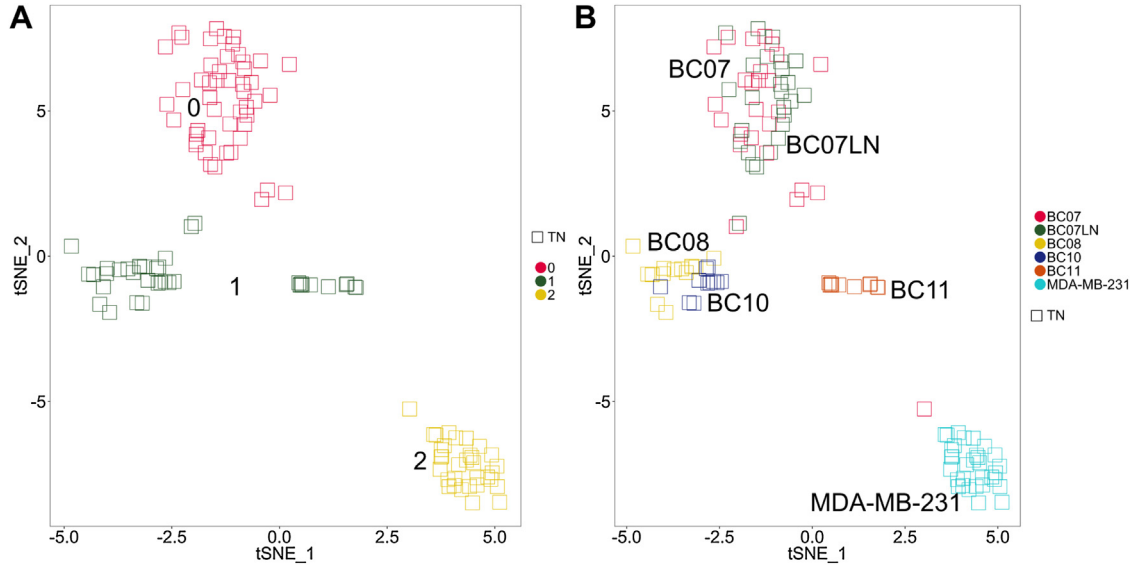

Fig. 6. Clustering of TN primary cells and its specific *in vitro* models (A) plotted into the space of t-SNE (B).

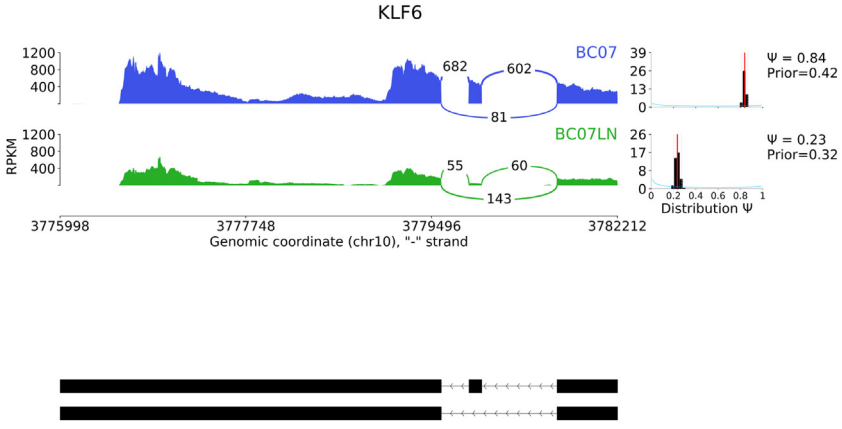

Fig. 7. Sashimi plot of alternative splicing of IP6K2 depicting the skipping of exon 3 in BC02 cells. BC01 presents some cells having predominantly the isoform with the exon included (top) and some having both the isoforms (middle), highlighting the intra-tumor heterogeneity related to this gene. The left panel shows the sashimi plot of the read density and the number of junction reads. The right panel shows the prior distribution (blue curve) and the histogram of the posterior distribution, learned by BRIE. The red line in the histogram represents the mean. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

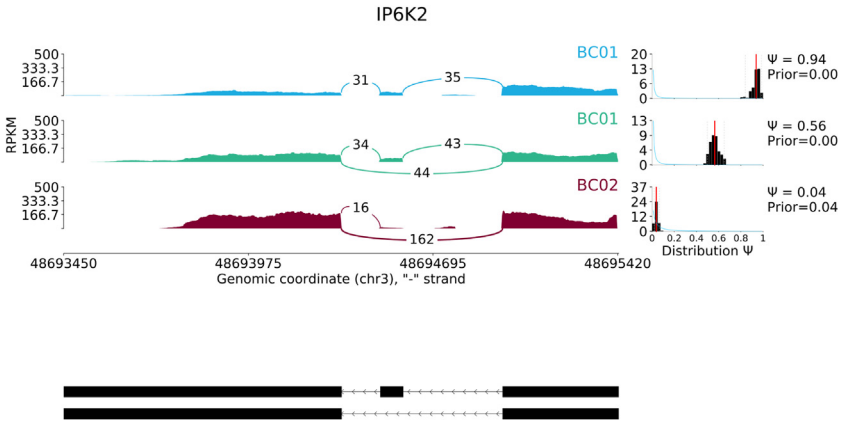

Fig. 8. Sashimi plot of alternative splicing of KLF6 depicting the skipping of exon 3 in BC07 LN cells (bottom). The left panel shows the sashimi plot of the read density and the number of junction reads. The right panel shows the prior distribution (blue curve) and the histogram of the posterior distribution, learned by BRIE. The red line in the histogram represents the mean. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Confusion Matrices for Best Accuracy Results on Breast cancer subtypes by PAM50, isoform markers (this study) and alternatively spliced genes (this study) signatures. The confusion matrix has been obtained by 10-fold cross validation of the best classification accuracies. The value of the classification accuracy obtained using the different gene lists is shown at the top of the tables. The number of correct and incorrect predictions is summarized with count values. The values having the same label on rows and columns represents the number of corrected prediction for that particular class, the others are the counts of incorrected predictions.

PAM50 – accuracy 91.63%

Lum A	HER2+	Lum B	TN	
71	0	0	1	Lum A
0	34	0	8	HER2+
1	0	24	0	Lum B
3	6	0	79	TN

Isoforms' markers (Seurat) – accuracy 98.24%

Lum A	HER2+	Lum B	TN	
71	0	0	1	Lum A
0	41	0	1	HER2+
0	0	25	0	Lum B
0	2	0	86	TN

Alternatively spliced genes (BRIE) – accuracy 98.68%

Lum A	HER2+	Lum B	TN	
71	0	0	1	Lum A
0	41	0	1	HER2+
0	0	25	0	Lum B
0	1	0	87	TN

performed to verify the closeness of the primary to the *in vitro* model subtypes. We collected data of two *in vitro* models of Lum A subtype, namely MCF7 and T47D. The literature recognizes T47D as one of the cell lines mostly similar to primary tumors (Jiang et al., 2016). Our analysis partially confirmed this assumption, since MCF7 appears to be distant from both the patients, while T47D represent a good model for BC01 belonging to the same cluster. Our analysis suggests that the choice of the right model should contemplate a deeper evaluation of tumor heterogeneity. Indeed, the plot shows that BC02 is not well represented by none of the two models, but only two cells clustered with T47D (Fig. 5).

The same was done for TN cells. We could analyze only one immortalized cell line model, MDA-MB-231, and it resulted to be a bad model of study for all the patients of the dataset (Fig. 6).

3.3. Isoforms' markers are involved in immune system response

131 significant (Bonferroni adjusted p-value ≤ 0.05) isoforms' markers were detected through Wilcoxon rank-sum test. The Gene Set Enrichment Analysis (GSEA) based on Gene Ontology Biological Process (GO-BP) returned 10 significant overlapping terms, all regarding immune system response, response to external stimuli and defense response (Supplementary File 1). Particularly, interferon involvement came out from the hallmark genes enrichment. It has been demonstrated that the immune system plays a dual role in tumor initiation and progression, capable of both inhibiting and promoting tumor expansion. It is also considered for further classification of TN breast cancer, based on whether the immune system is immunosuppressed or activated with a different prognostic indication (Nagarajan and McArdle, 2018). The immune system response, coming from the interaction between cancer and infiltrating immune cells, should therefore always be considered for a better understanding of subtypes and their prognosis.

3.4. Alternative splicing events contribute to the inter- and intra-tumor heterogeneity

Splicing events detection is generally limited to bulk data, and as discussed previously, the variability between single cells is often pursued at the gene level. Methods to analyze splicing in single cells are still in development, but it is well known that algorithms suitable for bulk data are not easily adaptable to single-cells for several reasons, such as minute amounts of starting material, low cDNA conversion efficiency, low coverage, and high technical noise. Due to this, we used an ad hoc tool named BRIE (Huang and Sanguinetti, 2017) which is reported to be strongly outperforming compared to other methods, as RSEM (Li and Dewey, 2011), Cufflinks (Trapnell et al., 2010), Kallisto (Bray et al., 2016), rMATS (Shen et al., 2014), etc. The output of BRIE is the estimation of an approximate posterior distribution on the values (exon inclusion ratio) as well as the learned regression weights. For each isoform (exon inclusion and exclusion) the number of cell pairs in which it is alternatively spliced is reported. For each comparison we made, we considered a different threshold of cell pairs based on the total number of cells contributing to the comparison. We kept an event if it was present in at least 30% of the cells. The lists of the filtered events are reported in Supplementary File 2. Most of the alternatively spliced genes are involved in splicing regulation, indicating a different regulation of the splicing machinery that contributes to the overall heterogeneity. Among all, it is worth noticing the presence of the heterogeneous nuclear ribonucleoprotein (HNRNP) family genes, in Lum A, TN, and HER2+ subtypes. Particularly, HNRNP C seems to regulate the stability and/or translation of the proteins BRCA1/2 (Anantha et al., 2013). Furthermore, it has been shown that HNRNP C presence influences the activation of interferon response (Wu et al., 2018). This result

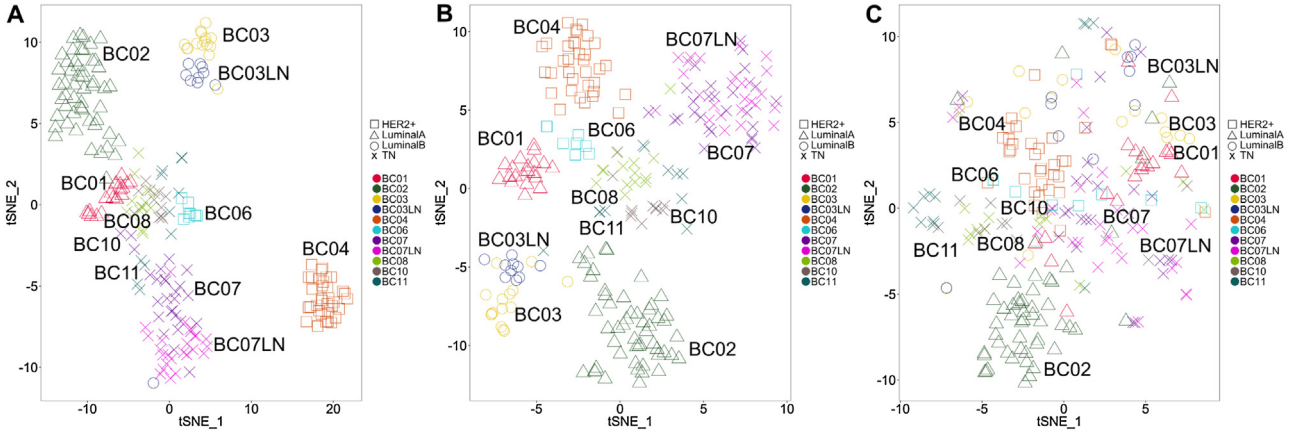

Fig. 9. t-SNE on single primary cells performed using the isoform markers (A), the alternatively spliced genes (B) and the PAM50 (C) signatures. A clear better separation of cancer subtypes is obtained by the first two lists, especially by alternatively spliced genes.

was in agreement with what we got from the enrichment of markers, indicating a possible connection among splicing events, isoform expression regulation, and tumoral heterogeneity. The role of HNRNP family members in cancer progression and metastasis has been investigated in several studies (Geuens et al., 2016; Han et al., 2013; Gallardo et al., 2016; Zhou et al., 2013; Capaia et al., 2018; Ferrari et al., 2017). As a confirmation of the prognostic differences found within the subtypes, it is worth noticing that among the alternatively spliced genes in Lum A patients BC01 and BC02 we found IP6K2, which is a well known factor involved in risk, survival, and prognosis of several types of cancer (Tan et al., 2015; Rao et al., 2015), as also reported by Kaplan–Meier plots deposited in Human Protein Atlas database (Uhlen et al., 2010). The sashimi plots (Fig. 7) show the presence of two different IP6K2 isoforms in the two Lum A patients and in BC02 an exon skipping event has been detected.

Among the alternatively spliced genes found within the TN cells, Kruppel-like factor 6 (KLF6) showed an exon skipping event (Fig. 8) in lymph node metastasis of BC07 patient. The resulting isoform is already known as SV3 variant. This gene is a transcriptional activator, and functions as a tumor suppressor. The SV3 variant localizes into the nucleus as the full-length isoform but functional studies have not been performed, although all the splicing variants have been found increased in malignant tissues (Chiam et al., 2013; Narla et al., 2005).

Wilcoxon rank-sum test was performed on the alternatively spliced genes and 95 significant markers were obtained (Supplementary File 2).

3.5. Alternatively spliced genes improve the molecular classification accuracy of breast cancer subtypes

In order to demonstrate the importance of our findings, we used the isoforms and the alternatively spliced genes to classify the BC subtypes and compared the accuracy of classification with the reference signature PAM50. Ranking our markers' lists by the highest average fold change we selected the first 50 to compare the same number of features. The results of the classification, in terms of accuracy and confusion matrices, using the three lists are shown in Table 1. The isoforms and the genes undergoing differential splicing events were capable of discriminating the subtypes better than PAM50. Indeed, by the PAM50 confusion matrix, we can see that 8 cells of the HER2+ subtype were classified as TN and that 9 cells of TN were misclassified, due to the intra-tumor heterogeneity we detected by clustering analysis. Taking into account the expression of genes alternatively spliced among subtypes, as well as the isoforms, which likely contribute to the heterogeneity, the classification accuracy was better (98.68% and 98.24% respectively). This result indicates that the splicing machinery is involved in differentiating the subtypes and that the differential expression of genes is not representative of all the cells in light of the strong heterogeneity among them.

In order to further show the discrimination power expressed by PAM50, isoform markers and alternatively spliced genes signatures, we performed a PCA on primary cells using the above cited gene lists. Fig. 9 shows a clear better separation of groups obtained with isoform markers (A) and even more with alternatively spliced genes (B) compared to PAM50 signature (C).

4. Conclusions

Our results emphasize the involvement of transcriptional regulation by the splicing machinery in determining tumor heterogeneity. Despite the emergence of single-cell sequencing and the well known role of splicing, many studies still rely on average-based methods to study highly heterogeneous tissues. We have demonstrated that the isoforms expression, as well as the prediction of splicing events, at the single-cell level, provide useful insights to better discriminate the tumor subtypes, commonly done through gene expression signatures. Our approach allowed us to identify potential markers capable of discriminating the nature of cancer cells and to ensure the success of precision medicine. Using the results of our investigation, in terms of isoforms and genes undergoing exon skipping events, we obtained a more accurate classification of breast cancer subtypes compared to the widely used PAM50 signature. Single cell investigation is furthermore a powerful source to identify the right *in vitro* model of study at sample level, thus to revalue use of cell lines, which present several advantages but have been demonstrated to be too divergent from primary cells. To the best of our knowledge, this is the first work which analyzes breast cancer heterogeneity with such detail using single cell data and investigating the alternative splicing events among single cells. We believe that our approach represents an evidence in support of the trend from patient-level to cell-level precision medicine.

Supplementary files

1. Supplementary File 1: Differentially expressed isoform markers – Sheet 1 contains the 131 differentially expressed markers of the four tumor subtypes identified by the Wilcoxon rank-sum test in Seurat. Sheets 2 and 3 contain the terms from the GO biological processes and the hallmark gene sets enriched by GSEA.
2. Supplementary File 2: Differentially spliced events within and between tumor subtypes – Sheet 1 contains the 95 differentially spliced events of the tumor subtypes detected as markers by performing the Wilcoxon rank-sum test on the BRIE isoforms estimate output. Sheet 2 contains the alternatively spliced events for Lum A,

HER2+, TN and Lum B breast tumors after applying a threshold on the percentage of cells involved in the pairwise comparisons (≥30%). The coordinates of skipped (exonAS) and flanking (exonC) exons are also reported.

Competing interests

The authors declare that they have no competing interests.

Funding

This work has been supported by MIUR PON02-00619, Interomics Italian Flagship Project and COFUND INCIPIT Project. The publication costs are funded by MIUR PON02-00619.

Authors' contributions

IM collected and processed the data, performed the analyses and generated the results. IG conceived the project, proposed the case study, organized and interpreted the results. MRG supervised the whole project. IG wrote the manuscript, which all authors read, edited and reviewed.

Acknowledgements

Authors would like to thank Giuseppe Trerotola and Gennaro Oliva for the administrative and technical support, respectively.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.biocel.2018.12.015.

References

Achim, K., Pettit, J.-B., Saraiva, L.R., Gavriouchkina, D., Larsson, T., Arendt, D., Marioni, J.C., 2015. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nat. Biotechnol. 33, 503.

Ahmed, D., Eide, P., Eilertsen, I., Danielsen, S., Eknaes, M., Hektoen, M., Lind, G., Lothe, R., 2013. Epigenetic and genetic features of 24 colon cancer cell lines. Oncogenesis 2, e71.

Allott, E.H., Geradts, J., Cohen, S.M., Khoury, T., Zirpoli, G.R., Bshara, W., Davis, W., Omilian, A., Nair, P., Ondracek, R.P., et al., 2018. Frequency of breast cancer subtypes among African American women in the AMBER consortium. Breast Cancer Res. 20, 12.

Anantha, R.W., Alcivar, A.L., Ma, J., Cai, H., Simhadri, S., Ule, J., König, J., Xia, B., 2013. Requirement of heterogeneous nuclear ribonucleoprotein C for BRCA gene expression and homologous recombination. PLOS ONE 8, e61368.

Andrews, S., et al., 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data.

Badve, S., 2016. Tumor heterogeneity in breast cancer. Molecular Pathology of Breast Cancer. Springer, pp. 121–132.

Batlle, E., Clevers, H., 2017. Cancer stem cells revisited. Nat. Med. 23, 1124.

Black, D.L., 2003. Mechanisms of alternative pre-messenger RNA splicing. Annu. Rev. Biochem. 72, 291–336.

Bray, N.L., Pimentel, H., Melsted, P., Pachter, L., 2016. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525.

Capaia, M., Granata, I., Guarracino, M., Petretto, A., Inglese, E., Cattrini, C., Ferrari, N., Boccardo, F., Barboro, P., 2018. A hnRNP K-AR-related signature reflects progression toward castration-resistant prostate cancer. Int. J. Mol. Sci. 19, 1920.

Chiam, K., Ryan, N.K., Ricciardelli, C., Day, T.K., Buchanan, G., Ochnik, A.M., Murti, K., Selth, L.A., BioResource, A.P.C., Butler, L.M., et al., 2013. Characterization of the prostate cancer susceptibility gene KLF6 in human and mouse prostate cancers. Prostate 73, 182–193.

Chung, W., Eum, H.H., Lee, H.-O., Lee, K.-M., Lee, H.-B., Kim, K.-T., Ryu, H.S., Kim, S., Lee, J.E., Park, Y.H., et al., 2017. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat. Commun. 8, 15081.

Dagogo-Jack, I., Shaw, A.T., 2018. Tumour heterogeneity and resistance to cancer therapies. Nat. Rev. Clin. Oncol. 15, 81.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., Shi, B., 2015. Breast cancer intrinsic subtype classification, clinical use and future trends. Am. J. Cancer Res. 5, 2929.

Deng, Q., Ramsköld, D., Reinius, B., Sandberg, R., 2014. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343, 193–196.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.

Elloumi, F., Hu, Z., Li, Y., Parker, J.S., Gulley, M.L., Amos, K.D., Troester, M.A., 2011. Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples. BMC Med. Genomics 4, 54.

Faigenbloom, L., Rubinstein, N.D., Kloog, Y., Mayrose, I., Pupko, T., Stein, R., 2015. Regulation of alternative splicing at the single-cell level. Mol. Syst. Biol. 11, 845.

Ferrari, N., Granata, I., Capaia, M., Piccirillo, M., Guarracino, M.R., Venè, R., Brizzolara, A., Petretto, A., Inglese, E., Morini, M., et al., 2017. Adaptive phenotype drives resistance to androgen deprivation therapy in prostate cancer. Cell Commun. Signal. 15, 51.

Gallardo, M., Hornbaker, M.J., Zhang, X., Hu, P., Bueso-Ramos, C., Post, S.M., 2016. Aberrant hnRNP K expression: all roads lead to cancer. Cell Cycle 15, 1552–1557.

Gentleman, R., Carey, V., Huber, W., Hahne, F., 2018. Genefilter: Methods for Filtering Genes from High-Throughput Experiments. R Package Version 1.62.0.

Geuens, T., Bouhy, D., Timmerman, V., 2016. The hnRNP family: insights into their role in health and disease. Hum. Genet. 135, 851–867.

Gillet, J.-P., Varma, S., Gottesman, M.M., 2013. The clinical relevance of cancer cell lines. J. Natl. Cancer Inst. 105, 452–458.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. ACM SIGKDD Explor. Newslett. 11, 10–18.

Han, N., Li, W., Zhang, M., et al., 2013. The function of the RNA-binding protein hnRNP in cancer metastasis. J. Cancer Res. Ther. 9, 129.

Horning, A.M., Wang, Y., Lin, C.-K., Louie, A.D., Jadhav, R.R., Hung, C.-N., Wang, C.-M., Lin, C.-L., Kirma, N.B., Liss, M.A., et al., 2017. Single-cell RNA-seq reveals a subpopulation of prostate cancer cells with enhanced cell cycle-related transcription and attenuated androgen response. Cancer Res canres-1924.

Hu, Z., Fan, C., Oh, D.S., Marron, J., He, X., Qaqish, B.F., Livasy, C., Carey, L.A., Reynolds, E., Dressler, L., et al., 2006. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics 7, 96.

Huang, Y., Sanguinetti, G., 2017. BRIE: transcriptome-wide splicing quantification in single cells. Genome Biol. 18, 123.

Ince, T.A., Sousa, A.D., Jones, M.A., Harrell, J.C., Agoston, E.S., Krohn, M., Selfors, L.M., Liu, W., Chen, K., Yong, M., et al., 2015. Characterization of twenty-five ovarian tumour cell lines that phenocopy primary tumours. Nat. Commun. 6 ncomms8419.

Jiang, G., Zhang, S., Yazdanparast, A., Li, M., Pawar, A.V., Liu, Y., Inavolu, S.M., Cheng, L., 2016. Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. BMC Genomics 17, 525.

Kowalczyk, M.S., Tirosh, I., Heckl, D., Rao, T.N., Dixit, A., Haas, B.J., Schneider, R.K., Wagers, A.J., Ebert, B.L., Regev, A., 2015. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res. 25, 1860–1872. https://doi.org/10.1101/gr.192237.115.

Krueger, F., Galore, T., 2015. A Wrapper Tool Around Cutadapt and FastQC to Consistently Apply Quality and Adapter Trimming to FastQ Files.

Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinform. 12, 323.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., Mesirov, J.P., 2011. Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., Tamayo, P., 2015. The molecular signatures database hallmark gene set collection. Cell Syst. 1, 417–425.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al., 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161, 1202–1214.

Matlin, A.J., Clark, F., Smith, C.W., 2005. Understanding alternative splicing: towards a cellular code. Nat. Rev. Mol. Cell Biol. 6, 386.

Nagarajan, D., McArdle, S.E., 2018. Immune landscape of breast cancers. Biomedicines 6, 20.

Narla, G., DiFeo, A., Yao, S., Banno, A., Hod, E., Reeves, H.L., Qiao, R.F., Camacho-Vanegas, O., Levine, A., Kirschenbaum, A., et al., 2005. Targeted inhibition of the KLF6 splice variant, KLF6 Sv1, suppresses prostate cancer cell growth and spread. Cancer Res. 65, 5761–5768.

Nguyen, Q.H., Pervolarakis, N., Blake, K., Ma, D., Davis, R.T., James, N., Phung, A.T., Willey, E., Kumar, R., Jabart, E., et al., 2018. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. Nat. Commun. 9, 2028.

Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al., 2009. Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. 27, 1160.

Platt, J.C., 1999. 12 fast training of support vector machines using sequential minimal optimization. Advances in Kernel Methods. pp. 185–208.

Qiu, Z., Zou, K., Zhuang, L., Qin, J., Li, H., Li, C., Zhang, Z., Chen, X., Cen, J., Meng, Z., et al., 2016. Hepatocellular carcinoma cell lines retain the genomic and transcriptomic landscapes of primary human cancers. Sci. Rep. 6, 27411.

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., et al., 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat. Biotechnol. 30, 777.

Rao, F., Xu, J., Fu, C., Cha, J.Y., Gadalla, M.M., Xu, R., Barrow, J.C., Snyder, S.H., 2015. Inositol pyrophosphates promote tumor growth and metastasis by antagonizing liver kinase B1. Proc. Natl. Acad. Sci. U. S. A. 112, 1773–1778.

Savas, P., Virassamy, B., Ye, C., Salim, A., Mintoff, C.P., Caramia, F., Salgado, R., Byrne, D.J., Teo, Z.L., Dushyanthen, S., et al., 2018. Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. Nat. Med. 1.

Shen, S., Park, J.W., Lu, Z.-x., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., Xing, Y., 2014. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. U. S. A. 111, E5593–E5601.

Song, Y., Botvinnik, O.B., Lovci, M.T., Kakaradov, B., Liu, P., Xu, J.L., Yeo, G.W., 2017. Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. Mol. Cell 67, 148–161.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550.

Tan, J., Yu, C.-Y., Wang, Z.-H., Chen, H.-Y., Guan, J., Chen, Y.-X., Fang, J.-Y., 2015. Genetic variants in the inositol phosphate metabolism pathway and risk of different types of cancer. Sci. Rep. 5, 8473.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., Rinn, J.L., 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. 32, 381.

Tripathi, K.P., Granata, I., Guarracino, M.R., 2017. A computational integrative approach based on alternative splicing analysis to compare immortalized and primary cancer cells. Int. J. Biochem. Cell Biol. 91, 116–123.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al., 2010. Towards a knowledge-based human protein atlas. Nat. Biotechnol. 28, 1248.

Vincent, K.M., Findlay, S.D., Postovit, L.M., 2015. Assessing breast cancer cell lines as tumour models by comparison of mRNA expression profiles. Breast Cancer Res. 17, 114.

Vu, T.N., Wills, Q.F., Kalari, K.R., Niu, N., Wang, L., Pawitan, Y., Rantalainen, M., Kelso, J., 2018. Isoform-level gene expression patterns in single-cell RNA-sequencing data. Bioinformatics 1, 9.

Waltman, L., Van Eck, N.J., 2013. A smart local moving algorithm for large-scale modularity-based community detection. Eur. Phys. J. B 86, 471.

Wang, Y., Navin, N.E., 2015. Advances and applications of single-cell sequencing technologies. Mol. Cell 58, 598–609.

Wu, Y., Zhao, W., Liu, Y., Tan, X., Li, X., Zou, Q., Xiao, Z., Xu, H., Wang, Y., Yang, X., 2018. Function of HNRNPC in breast cancer cells by controlling the dsRNA-induced interferon response. EMBO J. e99017.

Zhou, Z.-J., Dai, Z., Zhou, S.-L., Fu, X.-T., Zhao, Y.-M., Shi, Y.-H., Zhou, J., Fan, J., 2013. Overexpression of HnRNP A1 promotes tumor invasion through regulating CD44v6 and indicates poor prognosis for hepatocellular carcinoma. Int. J. Cancer 132, 1080–1089.