

RESEARCH ARTICLE

Single cell transcriptome analysis of MCF-7 reveals consistently and inconsistently expressed gene groups each associated with distinct cellular localization and functions

Yih-Shien Chiang¹, Yu-Feng Huang¹, Mohit K. Midha^{1,2}, Tzu-Han Chen¹, Hsin-Chieh Shiau¹, Kuo-Ping Chiu^{1,2,3*}

1 Genomics Research Center, Academia Sinica, Taipei, Taiwan, **2** Institute of Biochemistry and Molecular Biology, National Yang-Ming University, Taipei, Taiwan, **3** College of Life Science, National Taiwan University, Taipei, Taiwan

* chiukp@gate.sinica.edu.tw



OPEN ACCESS

Citation: Chiang Y-S, Huang Y-F, Midha MK, Chen T-H, Shiau H-C, Chiu K-P (2018) Single cell transcriptome analysis of MCF-7 reveals consistently and inconsistently expressed gene groups each associated with distinct cellular localization and functions. PLoS ONE 13(6): e0199471. <https://doi.org/10.1371/journal.pone.0199471>

Editor: Aamir Ahmad, University of South Alabama Mitchell Cancer Institute, UNITED STATES

Received: April 26, 2018

Accepted: June 7, 2018

Published: June 19, 2018

Copyright: © 2018 Chiang et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The sequence data have been deposited in ArrayExpress database (Accession no. E-MTAB-609).

Funding: This work was supported by Summit Project provided by Academia Sinica and PI funding provided by Genomics Research Center, Academia Sinica. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Single cell transcriptome (SCT) analysis provides superior resolution to illustrate tumor cell heterogeneity for clinical implications. We characterized four SCTs of MCF-7 using 143 housekeeping genes (HKGs) as control, of which lactate dehydrogenase B (*LDHB*) expression is silenced. These SCT libraries mapped to 11,423, 11,486, 10,380, and 11,306 RefSeq genes (UCSC), respectively. High consistency in HKG expression levels across all four SCTs, along with transcriptional silencing of *LDHB*, was observed, suggesting a high sensitivity and reproducibility of the SCT analysis. Cross-library comparison on expression levels by scatter plotting revealed a linear correlation and an 83–94% overlap in transcript isoforms and expressed genes were also observed. To gain insight of transcriptional diversity among the SCTs, expressed genes were split into consistently expressed (CE) (expressed in all SCTs) and inconsistently expressed (IE) (expressed in some but not all SCTs) genes for further characterization, along with the 142 expressed HKGs as a reference. Distinct transcriptional strengths were found among these groups, with averages of 1,612.0, 88.0 and 1.2 FPKM for HKGs, CE and IE, respectively. Comparison between CE and IE groups further indicated that expressions of CE genes vary more significantly than that of IE genes. Gene Ontology analysis indicated that proteins encoded by CE genes are mainly involved in fundamental intracellular activities, while proteins encoded by IE genes are mainly for extracellular activities, especially acting as receptors or ion channels. The diversified gene expressions, especially for those encoded by IE genes, may contribute to cancer drug resistance.

Introduction

Cancer is known to result from progressive accumulation of genomic and epigenomic alterations, leading to dysregulated cell growth [1, 2]. The process of carcinogenesis and subsequent cancer development is strongly enhanced by chromosome instability, a hallmark of cancer,

Competing interests: The authors have declared that no competing interests exist.

causing the formation of diversified cancer genomes in the cancer mass, each of which drives a cell-specific transcriptome. Through an evolutionary process, cancer cells carrying lethal mutations undergo apoptosis or necrosis, leaving the rest to continue the journey of cancer development in an uncontrolled manner [3, 4]. Due to the advances in Next-Generation Sequencing (NGS) technologies, cancer genomes and associated transcriptomes can be analyzed at single cell level [5, 6]. Due to the fact that it is able to provide a superior resolution as enhanced by NGS, single cell transcriptome (SCT) analysis has a strong potential to facilitate our understanding of cancer evolution and its transcriptional heterogeneity [7, 8], which is believed to be associated with acquired drug resistance [9]. However, although with a great potential [10], SCT data from cancerous cells remain very limited.

Here, we conducted SCT sequencing on MCF-7 breast cancer cell line to reveal their transcriptional diversity and potential clinical implications. MCF-7 is one of the most widely used metastatic breast cancer cell line. It retains several characteristics of differentiated mammary epithelium (<https://www.atcc.org/Products/All/HTB-22.aspx#characteristics>). On the other hand, breast cancer cells such as those characterized as HER2-positive and TNBC (Triple Negative Breast Cancer) have already experienced intensive selection process to abolish the expressions of some key genes, which may further affect the expressions of many other genes to make the transcriptome more subtype-specific. We prefer to use MCF-7 due to the fact that it represents a more general form of breast cancer and has a better potential to express both consistently expressed and inconsistently expressed genes. The analysis employs RNA-Seq approach [11], in conjunction with Tophat and Cufflinks software encompassed in UCSC Galaxy pipeline. To evaluate the reliability of our approach, we used 143 housekeeping genes (HKGs) agreed by two independent groups [12, 13], as control. HKGs, originally defined as genes with constitutive transcriptional activation in all cells [14], have frequently been used as reference to estimate gene expression levels [13, 15]. Moreover, *LDHB* gene, encoding for lactate dehydrogenase B, is silenced in MCF-7 by promoter hypermethylation and thus can serve as another layer of control [16, 17].

We split the expressed genes into consistently expressed (CE) (i.e., expressed in all SCTs) and inconsistently expressed (IE) (i.e., expressed only in some but not all SCTs) genes and conducted serial analyses on these two groups together with HKGs. Results indicated a bipartite transcriptional pattern in MCF-7 single cells, with consistently expressed genes mainly coding for proteins involved in intracellular activities and inconsistently expressed genes coding for extracellular proteins, including receptors and ion channels. We suspect that diversified expression of IE genes may act as a frontline protection of cancer cell from the attack of cancer drugs, due to the fact that it is essential for cancer drugs to interact with extracellular and/or membrane-bound proteins and the absence in expression in some cells would render the cancer as a whole to survive cancer drug treatment.

Materials and methods

Cell culture and single cell isolation

MCF-7 (BCRC 60436, Lot-01337) was purchased from Bioresource Collection and Research Center of Taiwan and cultured in DMEM medium (GIBCO 11955–065) supplemented with 10% (v/v) fetal bovine serum (GIBCO 26140–079) and 5 mg/ml insulin (Sigma I0516), in the presence of penicillin (100 unit/ml) and streptomycin (100 µg/ml). Cell culture was maintained inside a 37°C humidified incubator containing 5% CO₂. Culture media were changed every 3 days. Single cells were picked up by mouth pipetting under a microscope and transferred into 0.2 ml thin-walled PCR tubes, each containing 4.0 µl of cell lysis buffer. Volume of solution carryover was kept below 0.5 µl to minimize its impact on experimentation.

Sequencing library construction and sequencing

Sequencing libraries were constructed using ABI protocol [7], but with some modifications: 1) SuperScript II, instead of SuperScript III, was used as the reverse transcriptase during reverse transcription with UP1 primer; 2) incubation conditions were changed from “50°C for 30 min” to “42°C for 50 min”; and 3) duration time of the first PCR cycle was extended from 6 min to 8 min. PCR products were displayed by 1.5% agarose gel electrophoresis and 0.5–3 kb cDNA molecules were isolated for sequencing library construction. All SCT libraries were sequenced with SOLiD 3 for 35 ligation cycles (eq. 35 bp in length).

RNA-Seq data processing and analysis

Galaxy pipeline was used for SCT data analysis. Data processing comprised the following steps: quality sequence extraction, decontamination, mapping with Tophat, and gene annotation with Cufflinks [18]. In more detail, we used a cutoff value of 5 in the ‘remove reads containing color quality below this value’ parameter selection box to select quality reads. Sequence reads containing vector sequences or homopolymers (≥ 9 bp) were discarded. Quality reads were mapped to hg19 with Tophat originally designed for short read alignment for RNA-Seq experiments. We allowed two mismatches for each 35 bp read. ‘Minimum isoform fraction: filter out junctions supported by too few alignments’ was set to zero, and all other parameters were retained as default. Sequence reads with ≤ 10 hits were used and subjected to Cufflinks for annotation. To maximize sensitivity, both ‘min-isoform fraction’ and ‘pre-mRNA fraction’ were set to 0.0, and the parameter for ‘max-intron-length’ was changed from 300,000 to 500,000. To exclude singleton transfrags (transcribed fragments), we changed ‘min-frags-per-transfrag’ from 10 to 2 in program coding. All the other Cufflinks parameters were retained as default.

Cross-library comparison and gene level analysis

The expression levels of all expressed genes were compared between libraries by scatter plotting and the degree of overlap in transcript isoforms, gene IDs, and gene symbols were analyzed based on results generated by Cufflinks. For overlap analysis, we checked the presence of the same transcript isoform, gene ID, and gene symbol for all two-library combinations, and calculated the percentage of common objects over the total number of objects for each library in every combination. Expressed genes, including those in reference HKGs, were divided into CE and IE groups.

Gene ontology analysis

We employed Gene Ontology to analyze CE and IE genes. Biological Process (BP), Cellular Component (CC) and Molecular Function (MF) are the three fundamental domains of Gene Ontology. The controlled vocabularies were built as a directed acyclic graph (DAG), with the top hierarchy of GO tree defined more general terms and the more specific terms were obtained when traversed down the tree.

Before the analysis, both gene_ontology.1_2.obo file (http://www.geneontology.org/ontology/obo_format_1_2/gene_ontology.1_2.obo) and gene2go.gz (<ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz>) were downloaded from Gene Ontology website and NCBI FTP site, respectively. The gene_ontology.1_2.obo file was parsed into three individual domain files and gene2go file was served as reference for Gene Symbol-GO ID correlation. All RefSeq genes (13,764 total) were associated with corresponding GO terms in various domains based on the gene_ontology.1_2.obo and gene2go.

To perform statistical test, the number of genes in an annotation file (i.e., BP, CC and MF), the number of genes annotated to GO term, a list of genes from user input and genes annotated to each GO term are needed. For each GO term, we performed hypergeometric test to obtain significant GO term based on p-value. Then, we ranked each GO term based on its smallest p-value. For each selected group (CE or IE), the hypergeometric test was performed to obtain the top GO terms from three different domains, and the GO terms from CE and IE groups were compared with each other.

Pathway analysis

We chose BioCarta, which provides clear overview of pathways in suitable scale, for pathway analysis. It associates genes with pathways and we used those genes as our sampling for further calculation. We annotated CE genes to pathways and calculated the gene coverage rate for each pathway, shown as “percent” in table. Based on hypergeometric distribution, which is a discrete probability distribution, we get the “p-value” from formula

$$f(k; N, m, n) = \sum_{k+1}^m \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}}$$

, where N equals total number of Refseq genes in BioCarta, n equals total number of CE genes in BioCarta, m equals the number of genes in each pathway, and k equals the number of CE genes matched to the pathway. To identify the most significantly annotated pathways, we set percentage > 80% and p-value < 10⁻³.

Results and discussion

Library statistics

A total of five single cells (sC4, sC5, sC6, sC7 and sC8) were initially isolated from culture for experimentation. Among those, sC7 was found to have much lower P2#/P1# ratio (48%) compared to the rest (78–86%) and was excluded, so to minimize bias [19].

As described in SOLiD 3 Instrument Operation Guide (page 174) provided by Applied Biosystems (<http://tools.thermofisher.com/content/sfs/manuals/4407430b.pdf>), P2#/P1# ratio is defined as “the frequency of template-positive beads (P2#) relative to total beads (P1#) deposited on the slide; this metric is also referred to as “% P2 Positive” value”.

In more detail, SOLiD systems use two different adaptors, P1 and P2, to ligate to the ends of target DNA fragments, which are subsequently denatured into single-stranded templates for sequencing library construction. To make fluorescent signal detectable, clusters are built from adaptor-ligated single-stranded templates on magnetic beads using a procedure called “emulsion PCR” which contains millions of PCR reactions (each contains all PCR reagents, with P2 adaptor in excess and with none or limited number of beads, as following Poisson distribution) taking place in micro-scale aqueous droplets separated by mineral oil. In each PCR reaction, the P1 region of single-stranded template is annealed to the oligos pre-anchored on bead, while the P2 region is managed to reside on the other end. After emulsion PCR, beads with clusters are enriched, while beads without clusters are discarded. Since the presence of P1 is essential for making clusters on beads, P1# is used to represent total beads. On the other hand, although enriched, some beads may not have P2 on the other end (e.g., it may have P1 instead). As such P2#/P1# ratio is a metric to indicate the quality of template construction, where P2# represents the desired template-positive beads. High and balanced P2#/P1# ratios are desired for reliable analysis.

Table 1. Library statistics.

Description	sC4	sC5	sC6	sC8
#Original raw reads	47,563,622	48,815,843	53,363,798	45,762,729
#Quality reads	25,981,470 (54.6%)	27,144,452 (55.6%)	26,920,303 (50.4%)	26,926,581 (58.8%)
#Unique reads (uReads)	8,758,507	8,484,167	8,112,309	8,362,464
T2U ratio	3.0	3.2	3.3	3.2
#Mapped reads (1 ≤ #Mapped locations ≤ 10)	17,940,246 (69.1%)	20,349,372 (75.0%)	18,768,185 (69.7%)	18,815,190 (69.9%)
#Mapped uReads (1 ≤ #Mapped locations ≤ 10)	4,379,627 (50.0%)	5,249,695 (61.8%)	4,555,764 (56.2%)	4,482,719 (53.6%)
#Genes (singleton included)	12,178	12,721	11,745	13,522
#Genes (singleton excluded)	11,423	11,487	10,380	11,306
#Genes mapped by singletons	755	1,234	1,365	2,216
#Novel clusters	47,544	70,099	52,711	110,671

<https://doi.org/10.1371/journal.pone.0199471.t001>

Prior to analysis, a series of examinations were conducted to estimate the degree of balance between SCT libraries. The sequencer generated 45.7–53.4 million 35 bp reads per library (Table 1). Over half of these reads passed quality checking, accounting for 8.1–8.8 million unique reads (uReads) per library. Since PCR increases the copy number of uRead, we calculated “total count-to-uRead” (T2U) ratio and used it as an indicator to estimate the degree of PCR-induced copy number increase. The T2U ratios fell within a narrow range between 3.0–3.3, indicating a balanced PCR amplification among these libraries. Both numbers of mapped reads and numbers of mapped uReads also fell within a narrow range. Reads with 1–10 mapped locations were then used for further analysis. They were annotated to 10,380–11,487 RefSeq genes after the exclusion of those mapped by singletons. These results indicate that preparations of these SCT libraries were well-balanced.

Expressions of the 143 reference housekeeping genes

Expressions of the 143 reference HKGs served as a reference to evaluate the reliability of the SCT analysis (S1 Table). Although HKGs may be redefined in cancer cells, their expressions, in general, are expected to be maintained within a fair range in single cells, due to the fact that their expressions are essential for maintaining cellular structure and functions.

Among the 143 HKGs, 140 were found consistently expressed (and thus included in the CE group), while *LDHB* was not expressed for all libraries as expected and *MSN* and *PIMI* were expressed at barely detectable levels in some but not all single cells (and thus included in the IE group) (Table 2). *MSN* encodes moesin protein that mediates the association of cell membrane and actin, and the interaction with extracellular matrix, while *PIMI* is a proto-oncogene encoding a serine/threonine kinase.

The expression levels of each reference HKG in four SCT libraries were compared and used to calculate the coefficient of variation (CV) which represents the degree of fluctuation in

Table 2. Expression of the 143 reference housekeeping genes.

Libraries \ HKGs	# expressed HKGs (143 total)	Unexpressed HKGs
sC4	141	<i>LDHB, MSN</i>
sC5	142	<i>LDHB</i>
sC6	140	<i>LDHB, MSN, PIMI</i>
sC8	142	<i>LDHB</i>

<https://doi.org/10.1371/journal.pone.0199471.t002>

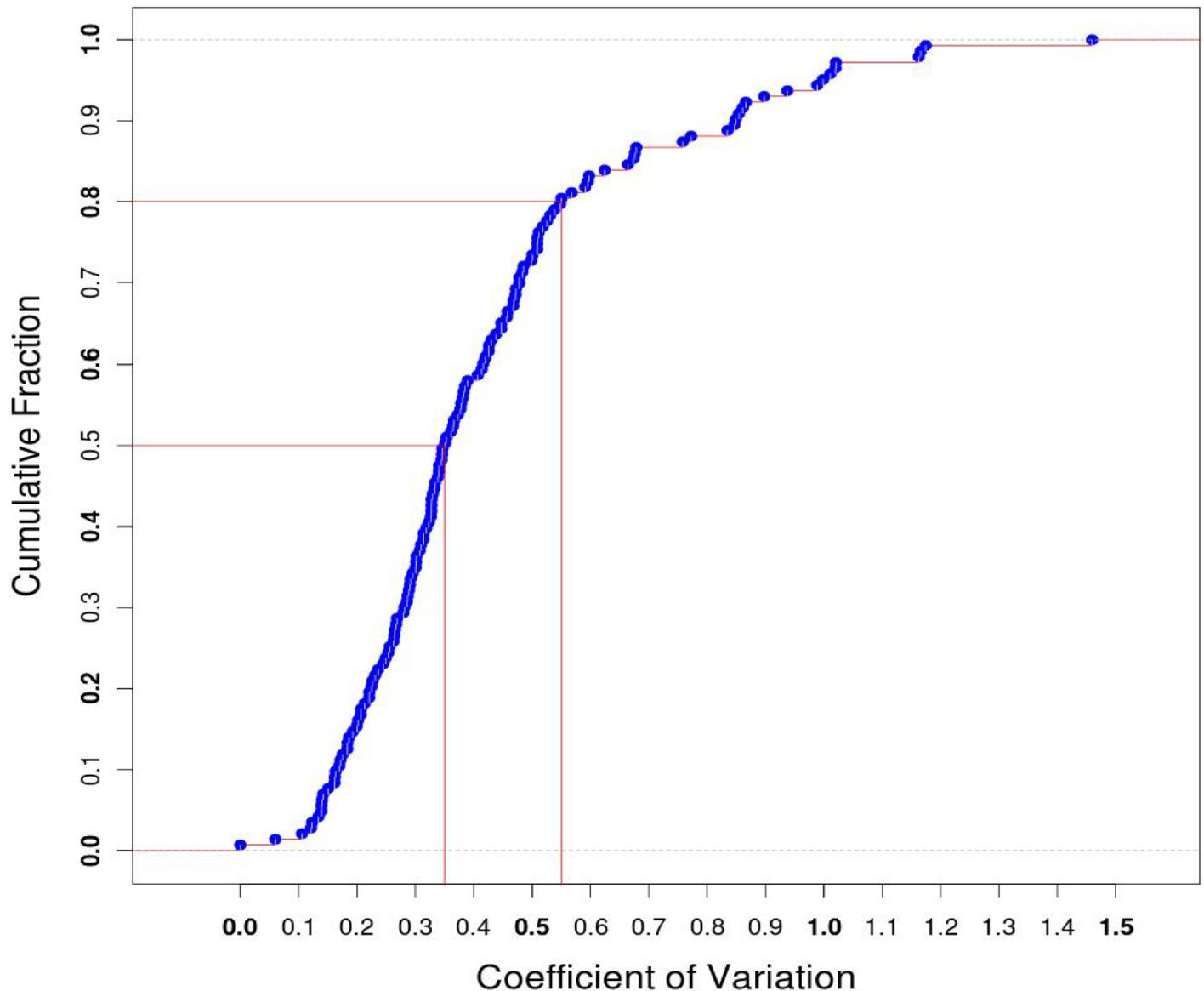


Fig 1. Distribution of coefficient of variation (CV) of the 143 reference HKGs. The coefficient of variation was calculated for the expressions of each 143 reference HKGs in four SCT libraries and then plotted against their corresponding percentage. As shown in the figure, about 80% of the CVs were less than 0.55.

<https://doi.org/10.1371/journal.pone.0199471.g001>

expression. About 80% of the reference HKGs had CV values less than 0.55 (Fig 1), indicating a stable expression of these HKGs among single cells.

Noticeably, the expressions of the *GAPDH* gene, which encodes glyceraldehyde-3-phosphate dehydrogenase and which is one of the HKGs most frequently used for the quantification of mRNA and protein expressions, fell within a narrow range between 5,078.3–6,658.2 FPKM, suggesting a strong reliability of the SCT approach.

Transcriptome similarities among SCT libraries

Due to the fact that these single cells were all derived from the same cell line, we expected a high degree of similarity among their SCTs. The similarities among these SCTs were first exemplified graphically prior to the characterization of transcriptional heterogeneity (Fig 2).

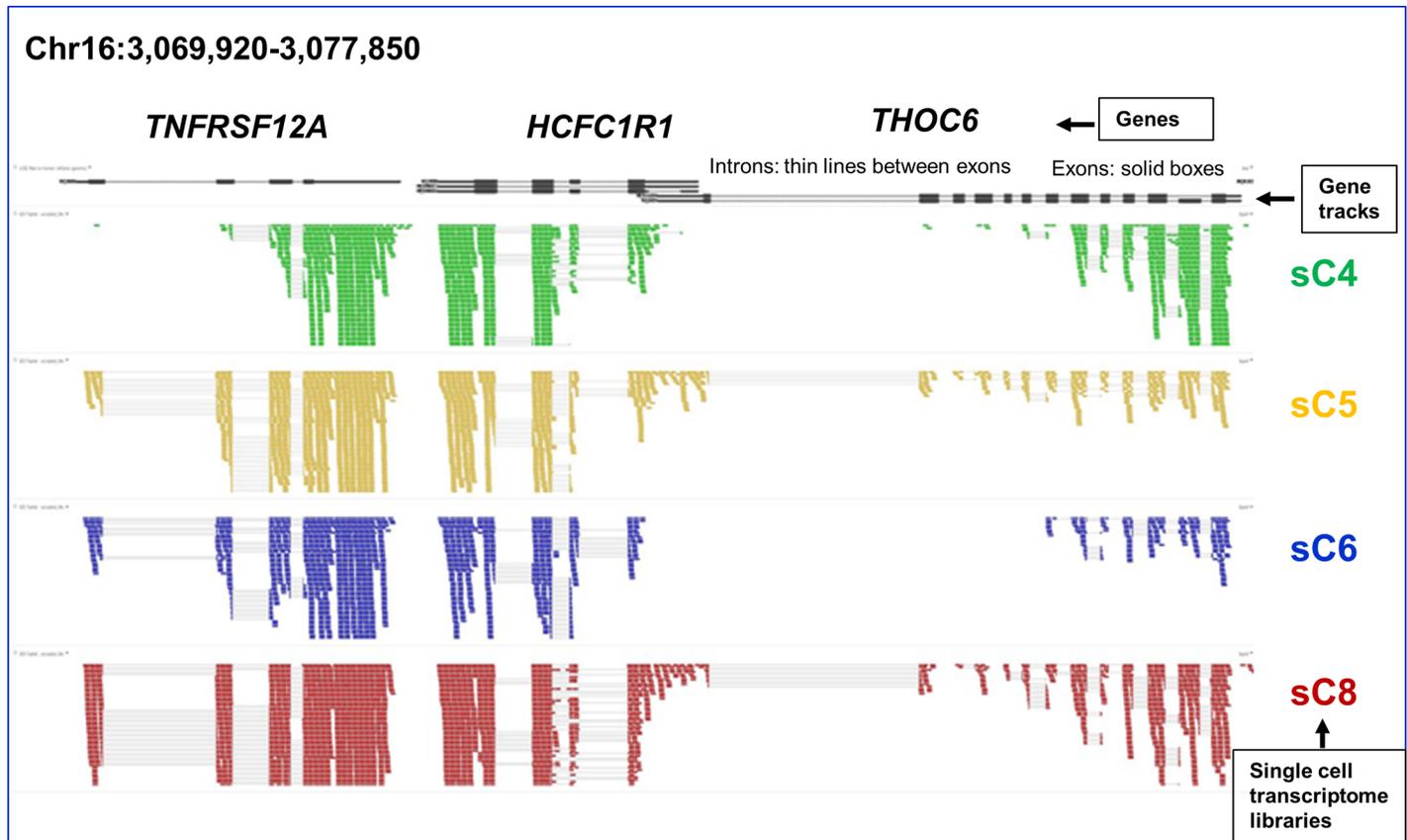


Fig 2. An example of single cell transcriptome profiles produced by Galaxy. Genome-wide RNA-Seq shotgun sequences of sC4, sC5, sC6 and sC8 single cell transcriptomes (highlighted in different colors) were analyzed by Galaxy pipeline which adopts Tophat for mapping (against hg19) and Cufflinks for transcript isoform identification. The height of each cluster represents its relative expression level. Some peaks are not shown completely due to space constraint. Introns (thin grey lines) between exons and intergenic regions (thick grey lines) are barely seen between exons.

<https://doi.org/10.1371/journal.pone.0199471.g002>

The similarity is not only shown in the state of ON or OFF of transcription, but also in expression level (shown as heights in the figure).

We then employed scatter plotting to analyze the correlation in gene expression levels between SCTs (see [S1 Table](#)). Result indicates a linear correlation in gene expression between any two-library combination of the SCT libraries ([Fig 3](#)).

To gain more insight of the similarity among these SCTs, we then analyzed the overlap in transcript isoforms among the SCTs. For the analysis, gene IDs and transcript isoforms were identified by Cufflinks. Gene IDs were then assigned to their corresponding gene symbols using UCSC database. Results indicated a high degree of similarity in transcript isoforms, ranging between 83–94, 83–94 and 82–93 for the overlaps in transcripts, gene IDs and gene symbols, respectively ([Fig 4](#)), further indicating a high degree of reliability of the experimental and analytical procedures.

Thus, both linear scatter plots and high percentage of overlaps in transcripts indicate a high degree of similarity among the SCTs.

The consistently expressed genes, inconsistently expressed genes and the 142 expressed reference housekeeping genes are expressed at distinct levels

To analyze transcriptional diversity among the SCTs, expressed genes were first split into consistently expressed (8,653 total, including 140 consistently expressed HKGs), inconsistently

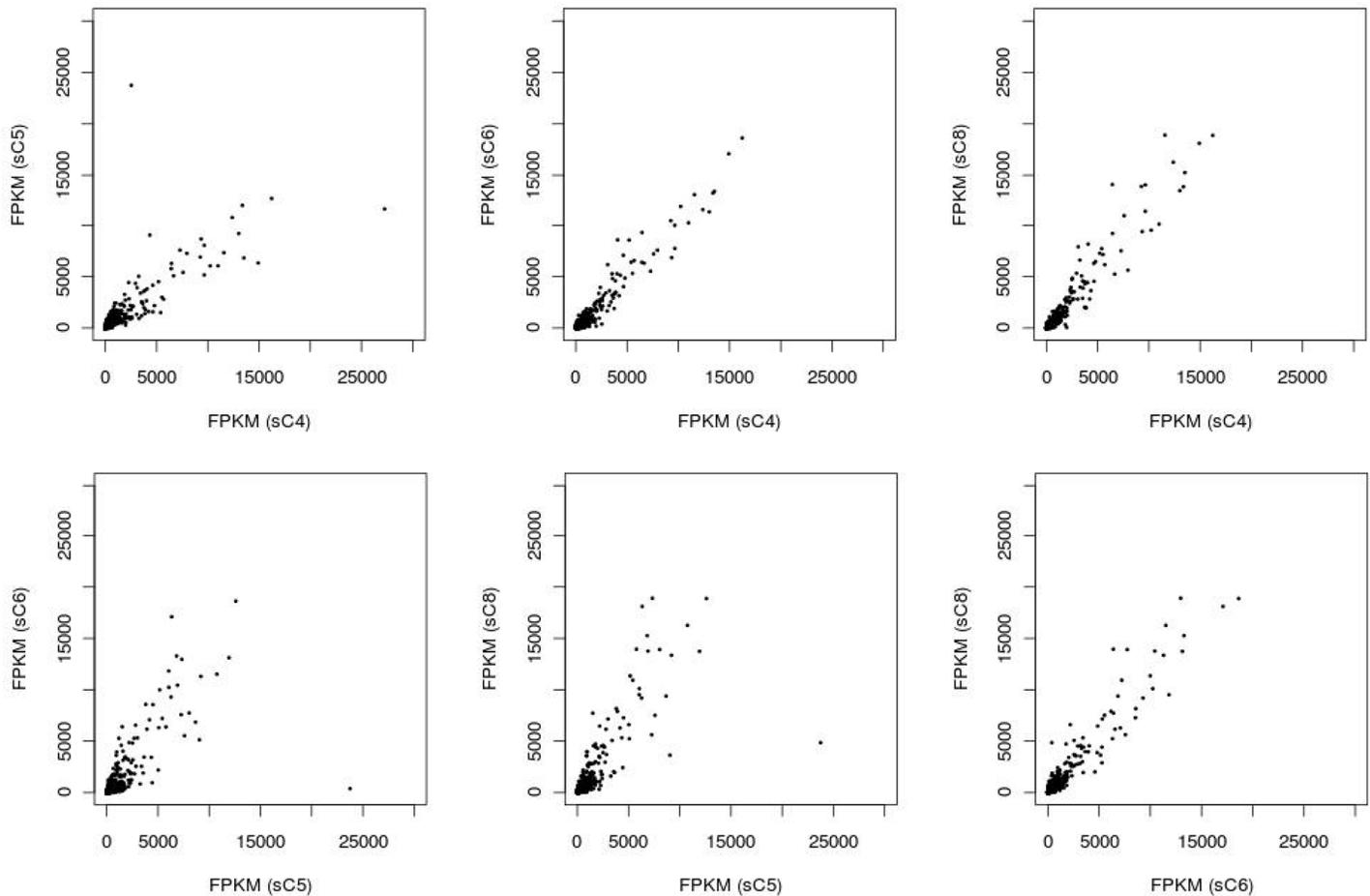


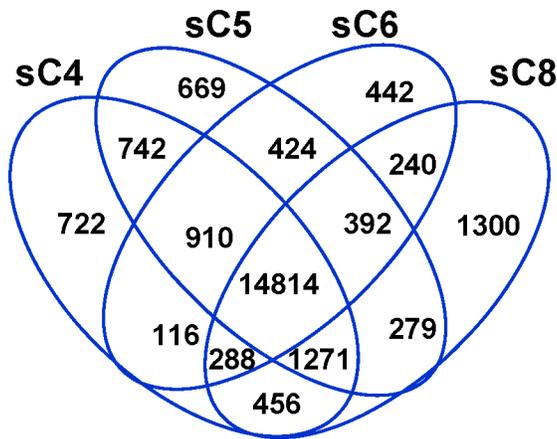
Fig 3. Scatter plots to show the correlation in gene expression levels for all two-library combinations of all four SCT libraries. The gene expression levels, in FPKM, of all expressed genes are plotted for any two-library combination of the SCT libraries.

<https://doi.org/10.1371/journal.pone.0199471.g003>

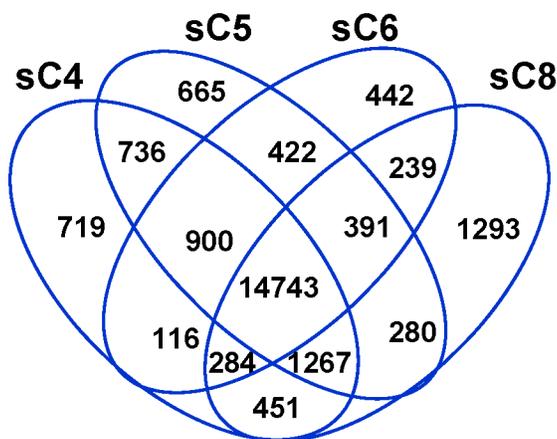
expressed genes (5,111 total, including 2 inconsistently expressed HKGs) and the 142 expressed HKGs as a whole. The expression levels of each gene in each group were averaged and sorted ascendingly. FPKM values at 5%, 15%, 25%, 35%, 45%, 55%, 65%, 75%, 85% and 95% were sampled and converted by \log_{10} and plotted side by side for comparison. Results demonstrate distinct expression levels (Fig 5), but with similar distribution patterns. Among these groups, 142 HKGs are expressed at the highest level, followed by CE, and then by IE. The mean values of expression levels of the 142 expressed HKGs, CE and IE are 1,612.0, 88.0 and 1.2, respectively, while the overall expression levels averaged at 54.6 FPKM. Thus, the mean value of CE gene expression levels is approximately 73 fold higher than that of the IE group.

The expression levels of consistently expressed genes vary much more significantly compared to that of inconsistently expressed genes and housekeeping genes

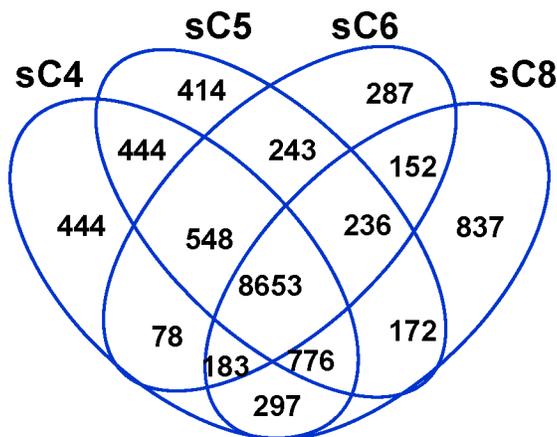
We then analyzed the variation in gene expression levels for every group in every single cell library. Interestingly, CV of the CE group was significantly higher than that of IE group and HKGs across all SCT libraries (Fig 6). We do not know the biological meaning of this phenomenon.



Tscp-overlap	sC4	sC5	sC6
sC4 (19,319)	-		
sC5 (19,501)	17,737 (91%, 92%)	-	
sC6 (17,626)	16,128 (92%, 83%)	16,540 (94%, 85%)	-
sC8 (19,040)	16,829 (88%, 87%)	16,756 (88%, 86%)	15,734 (83%, 89%)



GeneID-overlap	sC4	sC5	sC6
sC4 (19,216)	-		
sC5 (19,404)	17,646 (91%, 92%)	-	
sC6 (17,537)	16,043 (91%, 83%)	16,456 (94%, 85%)	-
sC8 (18,948)	16,745 (88%, 87%)	16,681 (88%, 86%)	15,657 (83%, 89%)



Gene Symbol-overlap	sC4	sC5	sC6
sC4 (11,423)			
sC5 (11,486)	10,421 (91%, 91%)		
sC6 (10,380)	9,462 (91%, 83%)	9,680 (93%, 84%)	
sC8 (11,306)	9,909 (88%, 87%)	9,837 (87%, 86%)	9,224 (82%, 89%)

Fig 4. Overlaps in gene transcript isoforms, gene IDs and gene symbols among the SCTs. The numbers (left) and percentages (right) of overlaps in transcript isoforms (top), gene IDs (middle), and gene symbols (bottom) among the SCT libraries are displayed with Venn Diagrams and tables. Tables on the right show the numbers of overlap in two-library combinations together with percentages for the library on the left (left within the parentheses) and the library on top (right within the parentheses).

<https://doi.org/10.1371/journal.pone.0199471.g004>

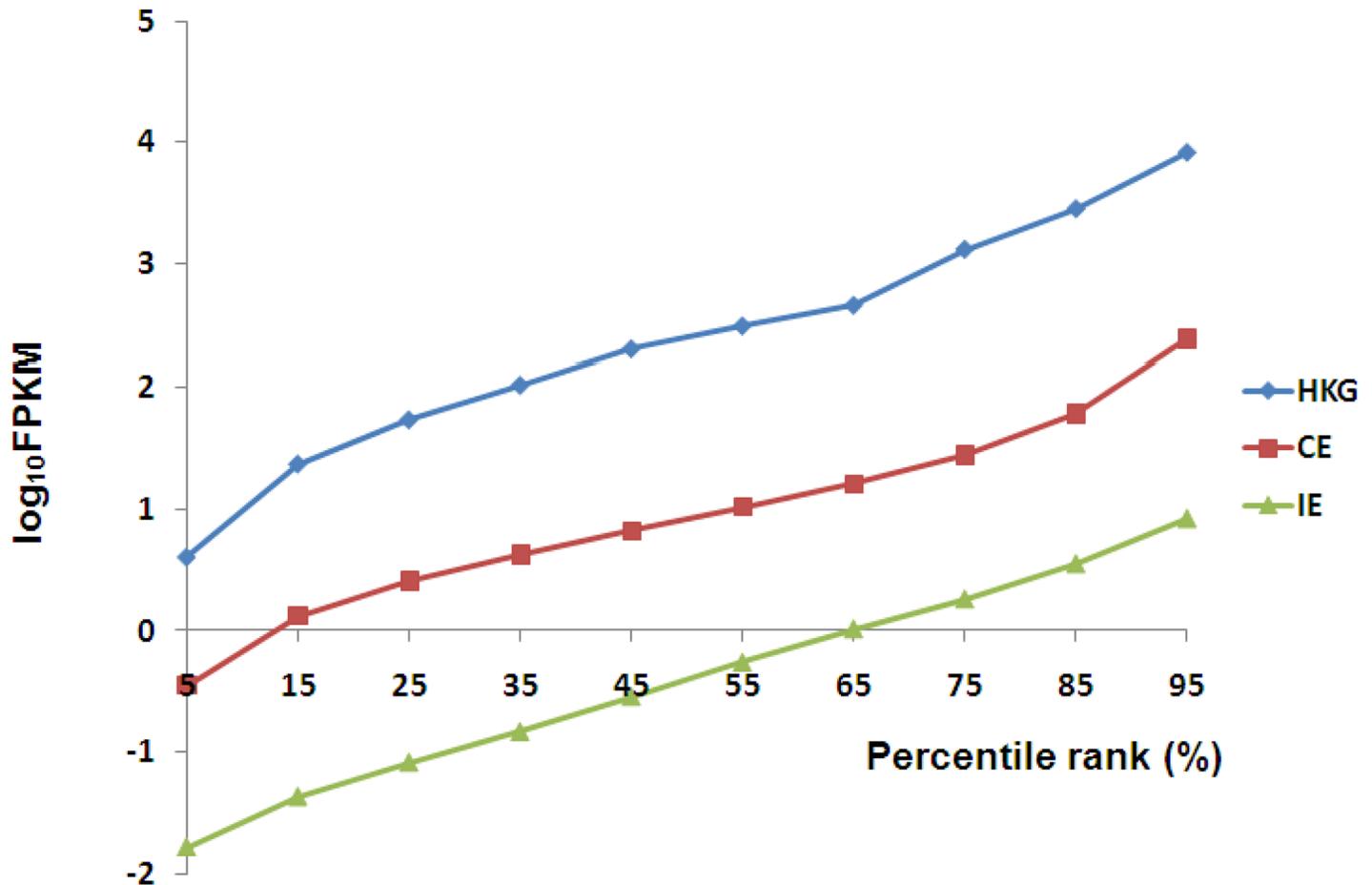


Fig 5. Expression level distributions for the 142 expressed reference HKGs, CE group, and IE group. For every gene in every group, the expression levels in all SCTs were averaged and sorted. FPKM values at 5%, 15%, 25%, 35%, 45%, 55%, 65%, 75%, 85% and 95% were sampled and converted by log₁₀ and plotted for cross-library comparison.

<https://doi.org/10.1371/journal.pone.0199471.g005>

Gene ontology analysis revealed distinguishable functional categories for consistently expressed genes and inconsistently expressed genes

We conducted GO analysis to gain insight of the biological functions of CE and IE genes. Results indicate that CE genes are more likely to encode proteins associated with intracellular structures such as membrane-bound organelles, organelle- and nuclear-associated structures, cytosol, cytoplasmic components, etc. (Fig 7). Their associated biological processes include translation, RNA (esp. mRNA) processing and metabolic activities, and their associated molecular functions are mainly related to protein and RNA binding and structural constituents of ribosome.

On the other hand, IE genes encoded proteins are more likely to be involved in extracellular structures such as cell periphery, plasma membrane, extracellular space and matrix, ion channel, etc. (Fig 8). Their associated biological processes include system process, multicellular organismal process, cell-cell signaling, etc., and their associated molecular functions are mainly related to receptor and channel protein activities (also see S2 and S3 Tables).

Thus, there is an evident bipartite feature in gene expression within every single cell and this bipartite feature is not limited to either healthy or diseased cells. However, since proteins encoded by IE genes are more involved in extracellular activities, inconsistent expression among single cells is expected to play a role in drug resistance in cancer.

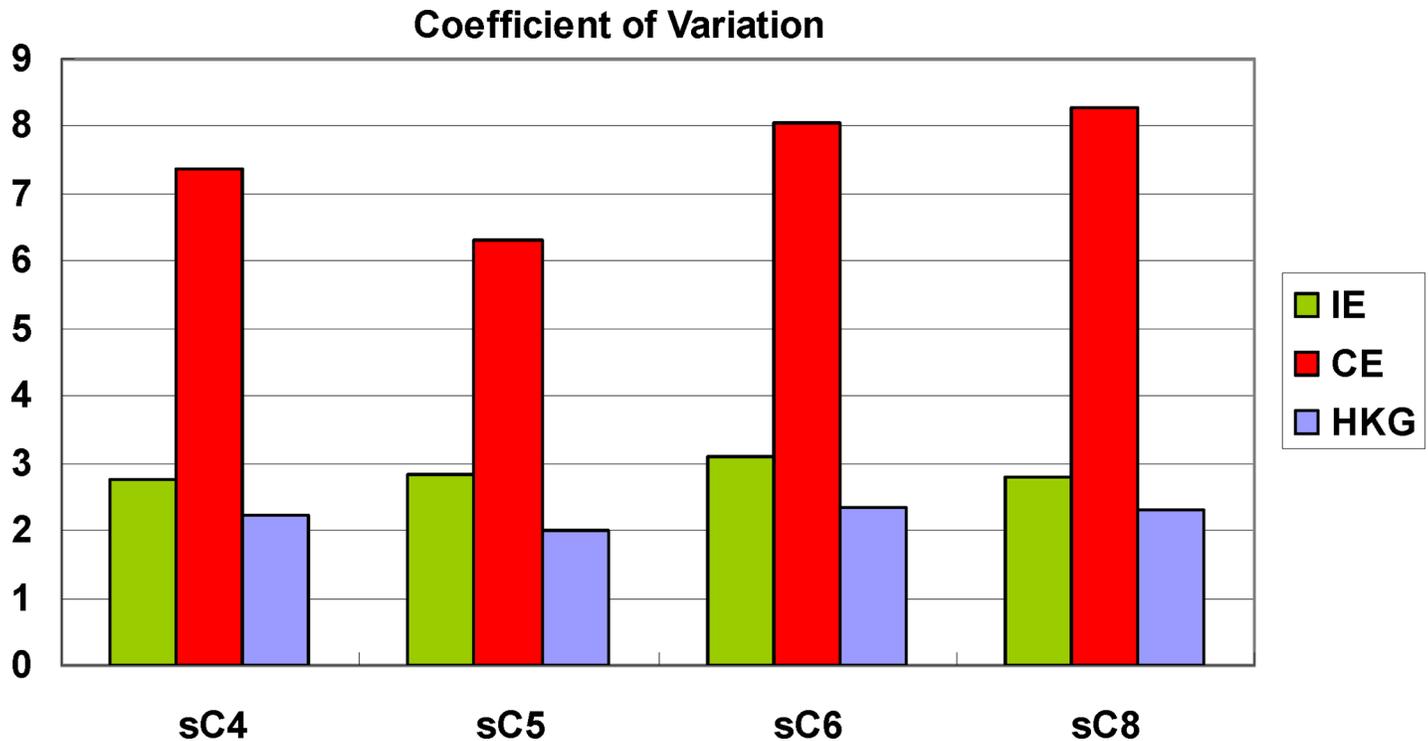


Fig 6. Comparison of coefficient of variation for HKGs, CE and IE genes. CVs were calculated for all three gene groups, including housekeeping genes, consistently expressed genes and inconsistently expressed genes and displayed separately for all single-cell library.

<https://doi.org/10.1371/journal.pone.0199471.g006>

Pathway analysis for the CE genes

Global pathway analysis using BioCarta pathway database indicated that many of the CE genes are participated in multiple key signaling pathways (including mTOR, MAPK, proteasome, EIF2, KREB, HIF, and RAS), together with glycolysis pathway, AKT pathway, and a number of growth factor/RTK pathways (including PDGF, EGF, IGF, VEGF, and insulin) (S3 Table). Many of these pathways (to name a few, HIF, VEGF, AKT, and RAS) are known to be tightly associated with cancer metabolism [20, 21]. We did not perform pathway analysis for IE genes because they are mainly extracellular structural proteins.

Discussion

Tumor cell heterogeneity is one of the key factors responsible for drug resistance, which in turn plays a key role for cancer therapy failure [9, 22]. Previous cancer investigations mainly used clinical tissues composed of mixed cancer cell populations, and thus the molecular interactions and interchange of proteins between individual cancer cells could not be well understood. Single cell sequencing would enhance our understanding of tumorigenesis and cancer treatment. Our study provides further information to explain part of the mechanisms associated with cancer drug resistance.

In an attempt to identify the transcriptional heterogeneity of protein-coding genes, we analyzed MCF-7 transcriptomes at single cell level. We split the expressed genes into consistently expressed genes and inconsistently expressed genes and studied the attributes of each group, followed by gene level and pathway level analyses. We show that HKGs, CE genes, and IE genes are expressed at distinguishable levels. Furthermore, in general, IE gene expressions showed less fluctuation than that of CE genes and this phenomenon was observed across all

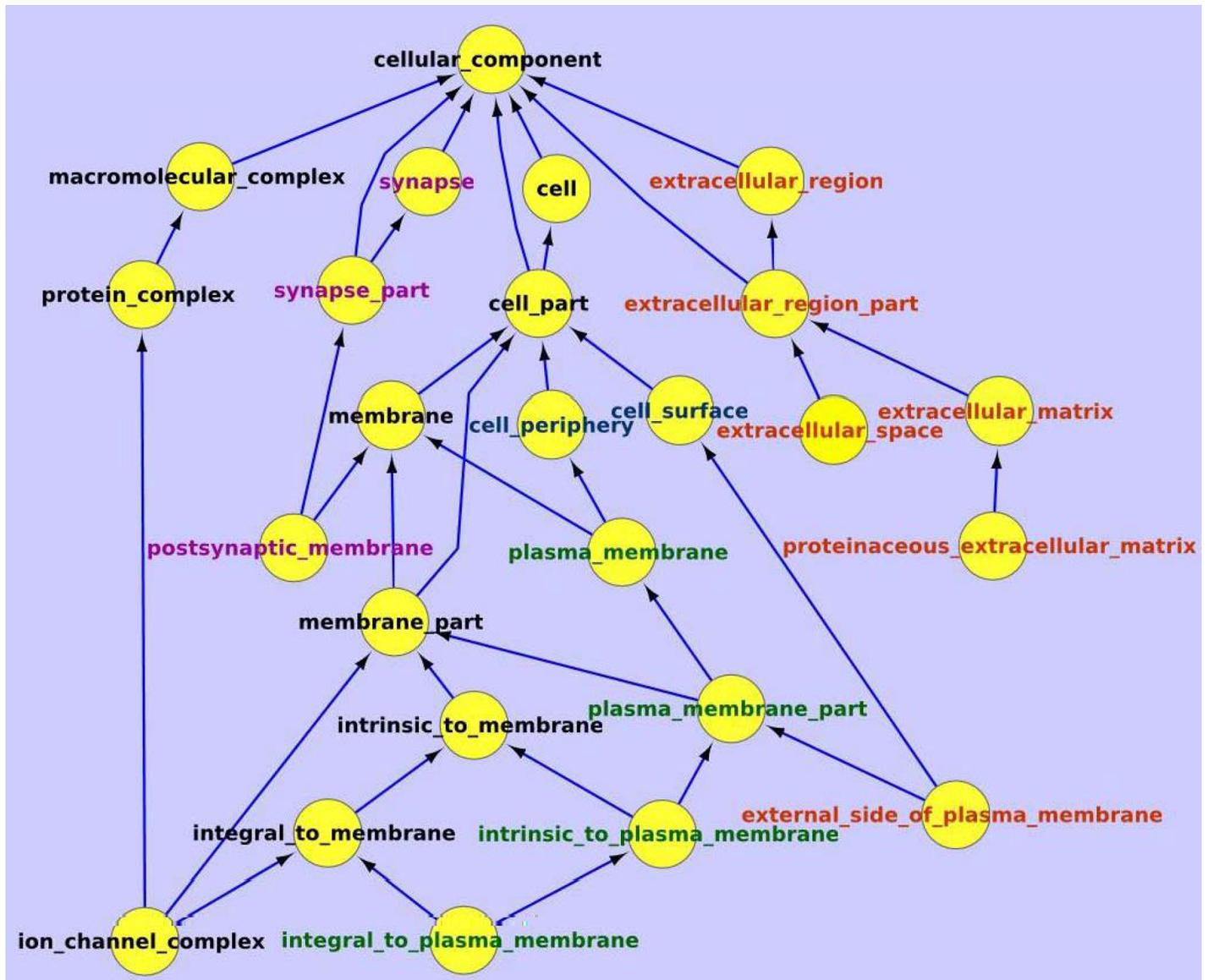


Fig 8. GO cellular component analysis for the inconsistently expressed (IE) genes. Genes inconsistently expressed in the four single cell transcriptome libraries are analyzed using Gene Ontology cellular component database. Components associated with major terms are highlighted with color, e.g., ‘synapse’ in purple, ‘plasma membrane’ in green, and ‘extracellular or external’ in red.

<https://doi.org/10.1371/journal.pone.0199471.g008>

and thus some can be omitted or facultatively expressed in some cells without causing lethal consequence. Cancer cells seem to take advantage of this feature to minimize the impact of drug treatment.

Supporting information

S1 Table. This Excel file lists all expressed genes (including CE, IE and 143 HKGs), their expression levels (in FPKM) in all four single cell transcriptome libraries, their averages, standard deviations and coefficient of variations (CVs).

(XLSX)

S2 Table. This file shows the results of GO analysis for CE and IE groups.
(PDF)

S3 Table. This file shows the results of BioCarta pathway analysis of CE genes. (Pathway analysis for IE is skipped because they are mainly extracellular proteins).
(PDF)

Acknowledgments

We thank Professor Alice L.T. Yu for generous material support.

Author Contributions

Conceptualization: Kuo-Ping Chiu.

Data curation: Yih-Shien Chiang, Hsin-Chieh Shiau.

Formal analysis: Yih-Shien Chiang, Yu-Feng Huang.

Funding acquisition: Kuo-Ping Chiu.

Investigation: Tzu-Han Chen.

Methodology: Yih-Shien Chiang, Mohit K. Midha, Tzu-Han Chen, Kuo-Ping Chiu.

Project administration: Tzu-Han Chen, Kuo-Ping Chiu.

Resources: Kuo-Ping Chiu.

Software: Yih-Shien Chiang.

Supervision: Kuo-Ping Chiu.

Validation: Yu-Feng Huang, Mohit K. Midha.

Writing – original draft: Kuo-Ping Chiu.

Writing – review & editing: Yih-Shien Chiang, Yu-Feng Huang, Mohit K. Midha, Hsin-Chieh Shiau, Kuo-Ping Chiu.

References

1. Ducasse M, Brown MA. Epigenetic aberrations and cancer. *Mol Cancer*. 2006; 5:60. <https://doi.org/10.1186/1476-4598-5-60> PMID: 17092350; PubMed Central PMCID: PMCPMC1636665.
2. Heng HH, Bremer SW, Stevens JB, Ye KJ, Liu G, Ye CJ. Genetic and epigenetic heterogeneity in cancer: a genome-centric perspective. *J Cell Physiol*. 2009; 220(3):538–47. <https://doi.org/10.1002/jcp.21799> PMID: 19441078.
3. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011; 144(5):646–74. <https://doi.org/10.1016/j.cell.2011.02.013> PMID: 21376230.
4. Kitano H. Biological robustness. *Nat Rev Genet*. 2004; 5(11):826–37. <https://doi.org/10.1038/nrg1471> PMID: 15520792.
5. Chiu KP. Applications of single cell sequencing in cancer. In: Tseng FG, Santra TS, editors. *Essentials of single-cell analysis: concepts, applications and future prospects*. Berlin Heidelberg: Springer; 2016. p. 347–61.
6. Navin N, Hicks J. Future medical applications of single-cell sequencing in cancer. *Genome Med*. 2011; 3(5):31. <https://doi.org/10.1186/gm247> PMID: 21631906; PubMed Central PMCID: PMCPMC3219072.
7. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*. 2009; 6(5):377–82. <https://doi.org/10.1038/nmeth.1315> PMID: 19349980.
8. Wang Y, Navin NE. Advances and applications of single-cell sequencing technologies. *Mol Cell*. 2015; 58(4):598–609. <https://doi.org/10.1016/j.molcel.2015.05.005> PMID: 26000845; PubMed Central PMCID: PMCPMC4441954.

9. Housman G, Byler S, Heerboth S, Lapinska K, Longacre M, Snyder N, et al. Drug resistance in cancer: an overview. *Cancers (Basel)*. 2014; 6(3):1769–92. <https://doi.org/10.3390/cancers6031769> PMID: 25198391; PubMed Central PMCID: PMC4190567.
10. Liu S, Trapnell C. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res*. 2016; 5. <https://doi.org/10.12688/f1000research.7223.1> PMID: 26949524; PubMed Central PMCID: PMC4758375.
11. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008; 5(7):621–8. Epub 2008/06/03. doi: nmeth.1226 [pii] <https://doi.org/10.1038/nmeth.1226> PMID: 18516045.
12. Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, et al. A compendium of gene expression in normal human tissues. *Physiol Genomics*. 2001; 7(2):97–104. Epub 2002/01/05. <https://doi.org/10.1152/physiolgenomics.00040.2001> [pii]. PMID: 11773596.
13. Eisenberg E, Levanon EY. Human housekeeping genes are compact. *Trends Genet*. 2003; 19(7):362–5. Epub 2003/07/10. doi: S0168952503001409 [pii]. [https://doi.org/10.1016/S0168-9525\(03\)00140-9](https://doi.org/10.1016/S0168-9525(03)00140-9) PMID: 12850439.
14. Watson J, Hopkins N.H., Roberts J.W., Steitz J.A., and Weiner A.M. *Molecular Biology of the Gene* (4th ed.) Benjamin/Cummings. 1987;vol. 1: p. 686.
15. de Jonge HJ, Fehrmann RS, de Bont ES, Hofstra RM, Gerbens F, Kamps WA, et al. Evidence based selection of housekeeping genes. *PLoS ONE*. 2007; 2(9):e898. Epub 2007/09/20. <https://doi.org/10.1371/journal.pone.0000898> PMID: 17878933; PubMed Central PMCID: PMC1976390.
16. Brown NJ, Higham SE, Perunovic B, Arafa M, Balasubramanian S, Rehman I. Lactate dehydrogenase-B is silenced by promoter methylation in a high frequency of human breast cancers. *PLoS One*. 2013; 8(2):e57697. <https://doi.org/10.1371/journal.pone.0057697> PMID: 23437403; PubMed Central PMCID: PMC3578788.
17. Maekawa M, Taniguchi T, Ishikawa J, Sugimura H, Sugano K, Kanno T. Promoter hypermethylation in cancer silences LDHB, eliminating lactate dehydrogenase isoenzymes 1–4. *Clin Chem*. 2003; 49(9):1518–20. PMID: 12928234.
18. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009; 25(9):1105–11. Epub 2009/03/18. doi: btp120 [pii] <https://doi.org/10.1093/bioinformatics/btp120> PMID: 19289445; PubMed Central PMCID: PMC2672628.
19. Ansorge WJ. Next-generation DNA sequencing techniques. *N Biotechnol*. 2009; 25(4):195–203. <https://doi.org/10.1016/j.nbt.2008.12.009> PMID: 19429539.
20. Hsu PP, Sabatini DM. Cancer cell metabolism: Warburg and beyond. *Cell*. 2008; 134(5):703–7. Epub 2008/09/09. doi: S0092-8674(08)01066-0 [pii] <https://doi.org/10.1016/j.cell.2008.08.021> PMID: 18775299.
21. Kaelin WG Jr., Ratcliffe PJ. Oxygen sensing by metazoans: the central role of the HIF hydroxylase pathway. *Mol Cell*. 2008; 30(4):393–402. Epub 2008/05/24. doi: S1097-2765(08)00292-X [pii] <https://doi.org/10.1016/j.molcel.2008.04.009> PMID: 18498744.
22. Gottesman MM. Mechanisms of cancer drug resistance. *Annu Rev Med*. 2002; 53:615–27. <https://doi.org/10.1146/annurev.med.53.082901.103929> PMID: 11818492.