Pan-cancer Cell-level Expression of Alternative Stop Codon Transcripts from Single Cell RNA Sequencing Data

by Zhe Yu

M.D. in Clinical Medicine, July 2004, Medical College, Dalian University, China
Ph.D. in Neurobiology, July 2013, Institute of Neuroscience, Chinese Academy of Sciences, China

A Thesis submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Master of Science

August 31, 2023

Thesis directed by

Anelia Horvath
Research Professor of Biochemistry and Molecular Medicine

© Copyright 2023 by Zhe Yu
All rights reserved

Dedication

I would like to dedicate this dissertation to my husband, who has been supportive and encouraging during the challenges of graduation, and to my daughter, who has made me stronger and better than I could have ever imagined.

Acknowledgments

I would like to express my sincere gratitude to my thesis advisor and mentor, Dr. Anelia Horvath. Thank you for all of your guidance, support, and instruction throughout my M.S. program. I would also like to thank Dr. Paul Marvar, my current supervisor, for his support in my studies. Additionally, I am grateful to the members of the Horvath lab, Hovhannes Arestakesyan, Allen Kim, and Kai Saito, for their assistance.

Abstract of Thesis

Pan-cancer Cell-level Expression of Alternative Stop Codon Transcripts from Single Cell RNA Sequencing Data

Background: Alternative stop codons can result from mutations creating a new or destroying an existing stop codon (stop codon mutations). Stop-codon mutations have been associated with numerous genetic disorders, including cancer. Disease phenotypes caused by stop codon mutations are often more severe, as these mutations frequently result in a complete loss of protein function. In recent studies, most alternative stop codons have been analyzed from bulk sequencing datasets, which can obscure cell-level information. The current study aims to establish a methodology for exploring the expression of alternative stop codon variants at single-cell resolution in cancer and normal scRNA-seq datasets generated on the widely used platform 10x-Genomics.

Method: Twenty-eight publicly available cancer and normal single-cell RNA-sequencing datasets were used to demonstrate our pipeline. The SCExecute tool, in conjunction with GATK and Strelka2, was used for cell-level variant calling. After annotation, the alignments at the stop codon locus were visually examined using the Integrative Genomics Viewer (IGV) package. Potential effects of the stop-codon variants on their harboring genes were assessed using Deseq2. Downstream analyses on pathway and gene-network alterations associated with the stop-codon variants were performed using the String package, and the PANTHER classification system.

Results: We identified a large number of expressed alternative stop codon transcripts more than 70% of which were novel (not previously described). We exemplify our analyses with a novel stop gain mutation W109* expressed in a gene encoding

ubiquitin conjugating enzyme E2 C gene (*UBE2C*). *UBE2C*^{W109*} is found in different cancer and normal embryo samples and shows significantly higher cell frequency and more pronounced effects in cancer samples such as cholangiocarcinoma.

Conclusion: Our study uncovered a previously unrecognized collection of cell-level expressed stop codon variants. Some of these variants show potential pathogenic effects on their harboring gene, as well as on other related genes. Analyses as the one demonstrated here are essential for better understanding of the stop codon mutations effects in normal and cancer tissues.

Table of Contents

Dedication……...……...iii

Acknowledgements……...iv

Abstract of Thesis……...…...v

List of Figures…….....viii

List of Tables…….ix

Chapter 1: Introduction…….…...1

Chapter 2: Materials and methods…….……10

Chapter 3: Results…….…….14

Chapter 4: Discussion……...25

Chapter 5: Conclusions …….28

References……..29

List of Figures

Figure 1……2

Figure 2……4

Figure 3……..……..9

Figure 4……..12

Figure 5……..15

Figure 6……..16

Figure 7……..17

Figure 8……..18

Figure 9……..19

Figure 10……20

Figure 11……21

Figure 12……22

Figure 13……23

List of Tables

Table 1……..…….10

Chapter 1: Introduction

1.1 Stop codon(s)

A stop codon is a sequence of three nucleotides in DNA or mRNA that signals the termination of the protein translation (1). In most organisms, there are three different termination codons: UAA, UGA, and UAG in mRNA, or TAA, TGA, and TAG in genomic sequence (2). TAG and UAG are called amber stop codons; TAA and UAA are known as ochre stop codons, and TGA and UGA are given the name of opal stop codons. Different stop codons have different efficiencies of termination: UAA>UAG>UGA. Typically, the first nucleotide 3′ of the stop codon acts as an important efficiency determinant (3).

During protein translation, the translating ribosome decodes coding sequence and recruits cognate amino-acylated tRNA to build the growing polypeptide amino acid chain. When the translation reaches the stop codon location, stop codons are recognized by and become bound to the class I release factor. The binding between the release factor and stop codon begins a cascade leading to polypeptide release via the action of the class II release factor, and thus terminates the translation process **[Figure 1]** (2). Stop codons and the stop codon recognition process constitute the last essential stage of translation and ensure the formation of full-sized proteins (4).

1.2 Expressed alternative stop codons

To date, alternative stop codons have been assessed mostly at DNA level, where genetic variants generating a premature stop codon (stop-gain, nonsense mutations) or destroying an existing stop codon (stop-loss, nonstop mutations) have been excessively

studied. There are multiple types of DNA-variants that can lead to an alternative stop codon, one of the most common being a single nucleotide substitution (Single Nucleotide Variant, SNV). At RNA-level, additional mechanisms – such as RNA-editing or transcriptional infidelity - can lead to the generation of alternative stop codons. At the level of the transcript, the effects of these variants on the protein synthesis are expected to be principally indistinguishable from DNA-originating stop-codon variants (5). However, at the level of the cell these effects may differ due to different proportion of transcripts undergoing the RNA-editing process. In addition, the transcription process perse is vulnerable to a wide range of potential errors that can result in alternative stop codons (6).

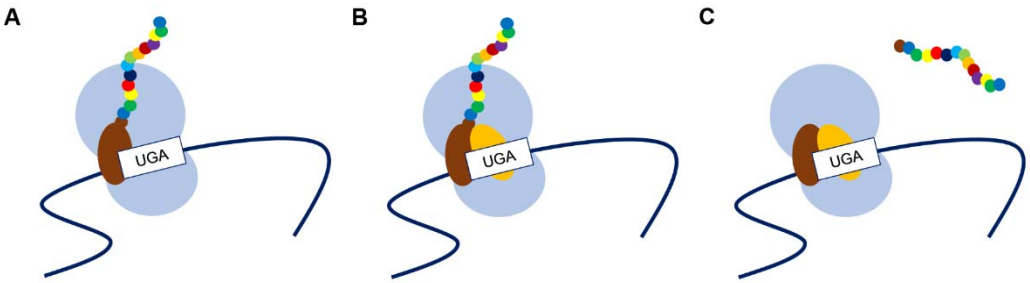

Figure 1: The basic mechanism of stop codon recognition (2).

1.2.1 Stop gain (nonsense) variant

If an SNV changes an amino acid codon into a premature stop codon, a stop gain variant is generated, and can cause early termination of the translation and releasing of unfinished (truncated) protein product [**Figure 2**]. The truncated protein can potentially exhibit dangerous effects on the cells, such as dominant negative interactions or deleterious gain-of-function effects. These effects are hindered by a mechanism called

nonsense-mediated mRNA decay (NMD), which results in the degradation of mRNA transcripts containing the premature stop codon. NMD was discovered due to the unexpectedly low concentrations of mRNAs carrying nonsense mutation (7). A popular model of the NMD involves up-frameshift (UPF) factors and exon-exon junction complexes (EJC). EJC binds to the mRNA after splicing and can be removed by ribosome during translation. If the mRNA contains premature stop codon, the translation will terminate early and the EJC located downstream to the premature stop codon will not be removed from the mRNA because the ribosome is released before reaching them. NMD is activated if any of the EJC remains bound to the mRNA after the first round of translation. The remaining EJC promotes recruitment or activation of UPF factors, and with the help of other phosphorylation proteins and enzymes, the mRNA will be finally degraded (8,9).

As a result of NMD, the steady-state level of cytoplasmic mRNA is reduced. However, not all the nonsense mutations are subjected to NMD. For example, in vertebrates, whether a mRNA will be subjected to NMD is often determined by the location of the last EJC relative to the stop codon. If the premature stop codon is downstream of or within about 50 nucleotides of the final EJC, NMD is not operational (10) [**Figure 2**].

1.2.2 Stop loss variants

Another type of stop codon variant constitutes the variants that change an existing stop codon to an amino-acid coding codon. Such variants cause the translation to continue into the 3'UTR until the next stop codon (11) and are also known as nonstop

variants. The polypeptides produced from a gene with nonstop variants are abnormally long and may lose their function or acquire new functions due to variety of processes, including misfolding [**Figure 2**].

Similar as stop gain mutation induced NMD, aberrant mRNA transcripts that do not contain a proper stop codon can also be identified and degraded by a cellular pathway called non-stop decay (NSD) (12). The occurrence of NSD dependents on the distance between the mutated stop codons and the next alternative in-frame stop codon downstream. Nonstop mutations with alternative stop codons in the close proximity to the original one frequently escapes from NSD, resulting in the mRNAs are translation into proteins of near-normal length and biological function. In contrast, alternative stop codons at greater distance from the original frequently trigger NSD, thereby decreasing the production of the protein and often yielding a clinical phenotype (11).

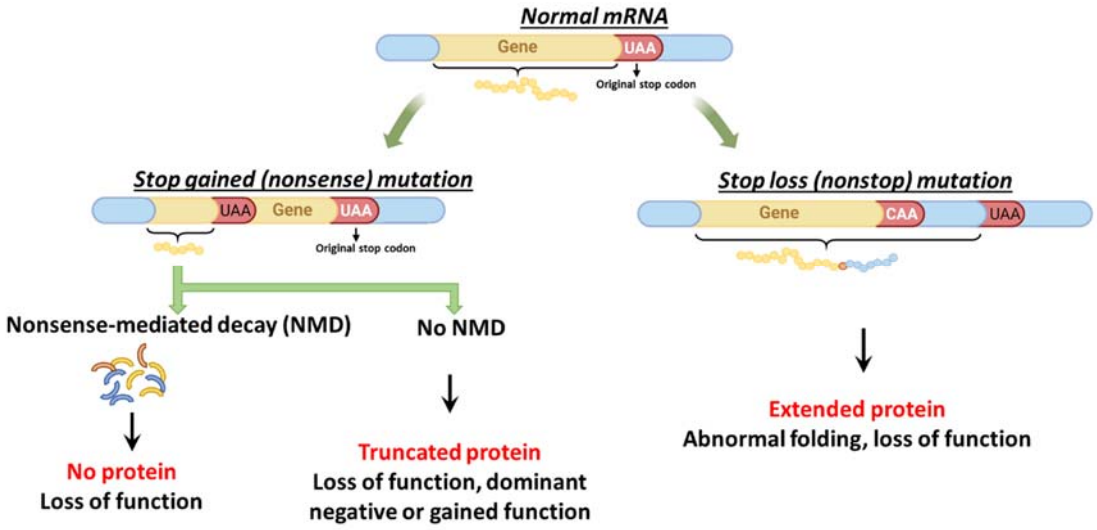

Figure 2: Stop gained mutation and stop loss mutation.

1.3 Stop codon variants and diseases

Both protein truncation caused by stop gain variants and the protein elongation caused by stop loss variants have been linked with diseases. Nonsense mutations can cause rare genetic diseases such as Duchenne muscular dystrophy, cystic fibrosis, and hemophilia, and also frequent diseases such as cancers, metabolic disorders, and neurological disorders (13). It is estimated that 12% of all mutations reported in Human Gene Mutation Database are single base-pair substitutions that result in nonsense mutation. Adding the other types of alternative stop generating variants – such as deletions, insertions, and splice variants, alternative stop codon mutation may be responsible for almost one-third of all inherited genetic disorders, including cancer (7).

Meanwhile, nonstop mutations have also been linked to many diseases. For example, a point mutation in type II 3β- hydroxysteroid dehydrogenase (3β-HSD) convert the stop codon into Cysteine, and the structurally lengthy mutated 3β-HSD enzyme due to this nonstop mutation was relatively detrimental in intact cells, causing the nonclassical phenotype of 3β-HSD deficiency congenital adrenal hyperplasia (14). Also, a mutation in the transcription factors *FOXE3* gene substitutes the stop codon for a leucine residue, predicting the addition of 72 amino acids to the C-terminus of *FOXE3*. *FOXE3* gene is a lens-specific transcription factor involved in the formation of the lens and surrounding structures, and mutation in this gene interrupts eye development and causes anterior segment dysgenesis (15).

Overall, unlike the missense variants that only result in a substitution with a different amino acid at a particular position, stop codon variants often result in a complete loss of protein function, and the disease phenotypes caused by stop codon

variants are frequently more severe. As mentioned earlier, stop codon variants originating at from RNA (i.e. not transcribed from DNA-encoded genetic variants) are expected to have transcript-level effects similar to the DNA-originating stop variants. However, in contrast to DNA variants, RNA variants have been rarely assessed, especially at cell resolution. For all the above reasons, identifying and characterizing stop codon variants, including RNA-originating ones, is of great importance for furthering our understanding of the relationship between gene structure and function.

1.4 Strategies to study expressed stop codon variants

The advent of high-throughput sequencing technologies and associated methodologies has facilitated the accumulation of information pertaining to stop codon mutations. This has led to a more profound understanding of the relationship between these mutations and various diseases. However, most studies on expressed variation have been conducted using bulk RNA sequencing, which measures the average expression level of genes across hundreds to millions of cells. Although variation analyses from single-cell RNA sequencing have started to emerge, the so far existing ones are typically based on variant calls from bulk or pseudo-bulk (pooled across the cells) sequencing data. Such analyses may perform post-call annotation of the variants to particular cells (barcodes). However, our previous studies have shown that variant calls on bulk and pseudo-bulk sequencing data typically miss the vast majority of variants present (i.e. expressed) in a small proportion of the cells, which are likely to include somatic variants as well as RNA-originating variance.

We have developed a novel approach to study expressed variation from 10x Genomics scRNA-seq data, that employs cell-level variant call on computationally extracted "on the fly" individual cellular alignments (binary alignment map files, .bam files) using our new tool SCExecute (14,15). This approach addresses the following challenges in the existing approaches to study cell-level genetic variants:

1) Identification of low (cellular) frequency variants, which can include pre- or early- somatic events, as well as RNA-originating variance.
2) Automatic annotation of variant to cell ID (cellular barcode).
3) Direct quantification of the variant expression (estimated as cellular expressed Variant Allele Frequency (ceVAF), calculated as the proportion of the sequencing reads harboring the variant allele over all the sequencing reads covering the variant. Naturally, ceVAF can obtain values in the range [0,1]. The quantification of ceVAF, in turn, allows quantitative analyses on the variant cellular distribution and expression and downstream analysis of association of the variant to additional cellular features, such as gene expression, or cell-level correlations between allele- and gene-expression.

Here, we demonstrate the utility of the above approach to study expressed stop-codon altering variants from 10x Genomics scRNA-seq datasets. For all identified variants, we assessed their distribution and the cellular expression. Based on this, we have outlined a set of potentially functional variants, that show non-random cell expression distribution. After individual curation, we have selected a subset of high priority variants, for which we performed downstream analysis on the effects of the

variant on the harboring gene and its gene network. We demonstrate this analysis on the novel stop gain SNV W109*, in *UBE2C* (*UBE2C*^{W109*}).

1.5 Scope and aims of the study:

In this study we sought to explore the expression of alternative stop codons at single cell level in cancer and normal scRNA-seq datasets generated on the 10xGenomics platform. Our specific aims are:

1) Develop methods and a pipeline to analyze alternative stop codon transcripts at cell level

2) Explore transcripts harboring alternative stop codons across different cancer types, and compare with normal tissues

3) Identify alternative stop codon variants driving changes in cell functions and developmental trajectories

4) Characterize *in silico* the gene- and cell-level effects of alternative stop codon transcripts.

The processing pipeline is illustrated in **Figure 3**.

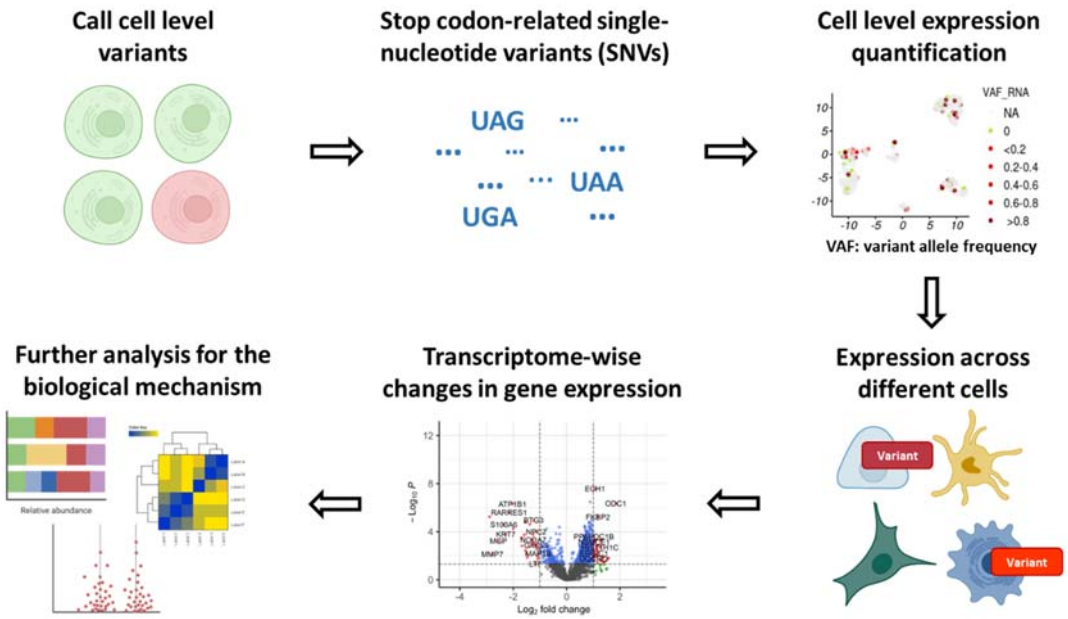

Figure 3: Processing and analytical workflow for studying expressed SNVs from 10x Genomics scRNA-seq datasets.

Chapter 2: Materials and methods

2.1 Datasets

Twenty-eight publicly available cancer and normal scRNA-seq datasets, including prostate cancer, non-small cell lung carcinoma, cholangiocarcinoma, neuroblastoma, normal fetal adrenal and normal embryo, were used in this study (16–18). The number of samples from each study is shown in **Table 1**.

Samples	Sample number
Prostate cancer	2
Cholangiocarcinoma	6
Neuroblastoma	11
Non-small cell lung cancer	3
Normal fetal adrenal	4
Normal embryo	2

Table 1: Datasets and samples used in this study.

2.2 Cell level variant call and alternative stop codon annotation

Detailed method for variant call was previously described (19). In brief, the raw scRNA sequencing reads were aligned to the human reference genome assembly GRCh38.79 using the STARsolo module of STAR (20). Transcriptome-wide variant calling was performed on each single cell alignment using SCExecute in conjunction with GATK and Strelka2 (21–23). Read-count matrices with the raw gene counts per cell generated by STARsolo were used to estimate cell-level gene expression.

Alternatively stop codon variants were identified after annotation and variant function prediction through SeattleSeq (37). The annotation through DbSNP (24) was used to outline novel (not previously described) SNVs.

The quality of the alignments for each SNV of interest was visually examined using the Integrative Genomics Viewer (IGV), and only the SNVs in regions with high quality alignment were retained for further analyses.

2.3 Screening for alternative stop codon SNVs with non-random cell-level distribution

We employ the assumption that SNVs with non-random cell-level distribution across cell clusters (defined by cell expression), are likely to exhibit functional effects on the gene expression. These functional effects can consist of effects on the expression of the variant harboring gene (cis effects), as well as direct effects on the expression of other genes interacting with the harboring gene (trans effects).

To screen for SNVs with non-random cell-level distribution, cells were first clustered using methods implemented in the popular package Seurat (25), and then classified as cell types using SingleR (26). Pseudotime trajectory inference was calculated using Slingshot (27).

Then the cell expression of each SNV was overlayed with the overall gene expression across cells and clusters and non-randomly expressed SNVs were outlined using F-test. SNVs with non-random distribution were quantitatively visualized in the reduced dimensionality space using UMAP on the gene-expression defined clusters

[**Figure 4**]. The SNVs with non-random cell expression distribution were used for downstream analyses.

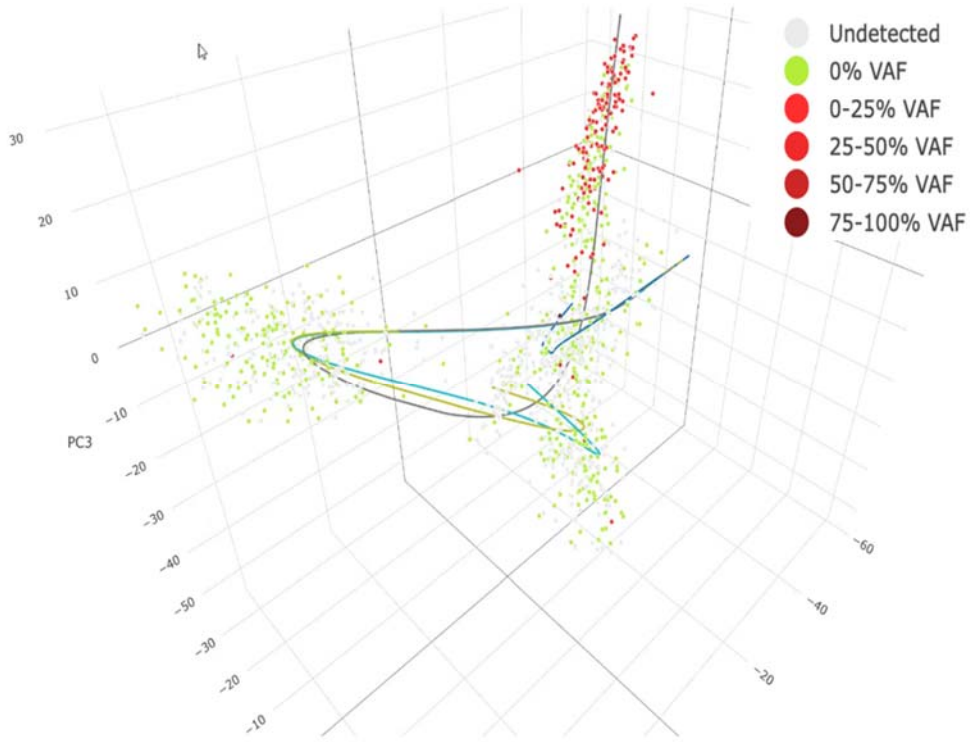

Figure 4: Example of preferentially expressed SNV. *Grey dots*, cells that do not express the gene expressing the SNV; *Green dots*, cells that express only the reference allele of the gene; *Red dots*, cells expressing the variant allele of the SNV, where the relative proportion of the expressed variant genes is indicated with the intensity of the red color.

2.4 Transcriptome-wise changes in gene expression analysis

To detect the transcriptome-wise changes in gene expression correlated with the expressed SNVs, differential gene expression between cells expressing the SNV variant stop codon allele and those expressing only the referent allele was performed using Deseq2 (28). Correction for multiple trials was performed using the module incorporated

in Deseq2, and genes with adjusted *p*-value below 0.05 were considered significantly deregulated.

2.5 Network and pathway analysis

Different pathway and gene network databases were used to identify cellular mechanisms potentially affected by the expression of the alternative stop codon variant. Differentially expressed (DE) gene-sets between the calls expressing the variant and the cells expressing only the reference allele were inputted in the "multiprotein" module of the String network database (29). In parallel, the same DE gene-sets were analyzed using the PANTHER (Protein ANalysis THrough Evolutionary Relationships) classification System (30). PANTHER was also used to identify pathways enriched with genes differentially expressed in relation with the alternative stop codon SNV.

Chapter 3: Results

Using the methods described above, we were able to uncover a wealth of information about alternatively expressed stop codons that would not have been possible using bulk-based variant callers. The analysis of the 20:45812900_G>A variant expressed in the ubiquitin conjugating enzyme E2 C gene (*UBE2C*) was used as an example to illustrate our pipeline.

3.1 *UBE2C*^{W109*} variant

UBE2C is a member of the E2 ubiquitin-conjugating enzyme family, which plays a critical role in the ubiquitination process (31,32). UBE2C protein in both the nucleus and the cytoplasm and functions as an integral part of the Anaphase Promoting Complex/Cyclosome (APC/C) (33). The protein is required for the destruction of mitotic cyclins and cell cycle progression, and is also involved in cancer progression (33,34).

UBE2C is composed of 179 amino acids, and the *UBE2C*^{W109*} SNV is located at chromosome 20:45815900 (20:45812900_G>A). This variant is a stop gain substitution that converts a Tryptophan codon TGG, which encodes the 109th amnio acid of the protein, into the stop codon TGA in the same position. This alternative stop codon terminates UBE2C translation early and is predicted to result in reduction of 40% of the length (on the 3'-end) of the protein. Given the importance of UBE2C in the ubiquitin-proteasome pathway, the variant may potentially affect the process of cell ubiquitination.

The quality of the alignment of the reads covering the variant position was first visually examined through the IGV [**Figure 5**], which was consistent with high quality of the variant call. The IGV examination suggested high confidency for the G>A variant in

position 20:45812900, as evidenced by multiple unique high-quality reads harboring the variant nucleotide.

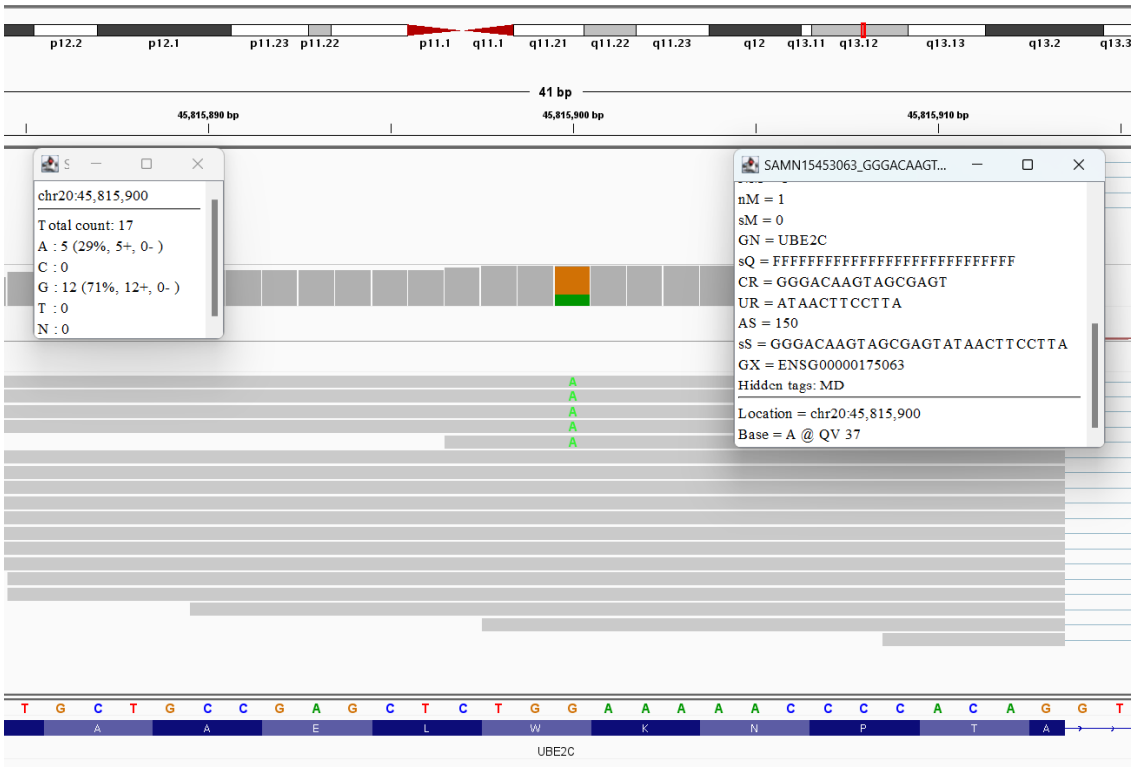

Figure 5: IGV visualization results showing the high-quality sequencing reads and alignment.

3.2 Preferential expression of the *UBE2C*^{W109*} mutation

To test whether the expression of *UBE2C*^{W109*} variant was non-randomly expressed across the cells we applied F-test statistics as described above. This analysis showed preferential expression of the variant allele in a subpopulation of the cells [**Figure 6**].

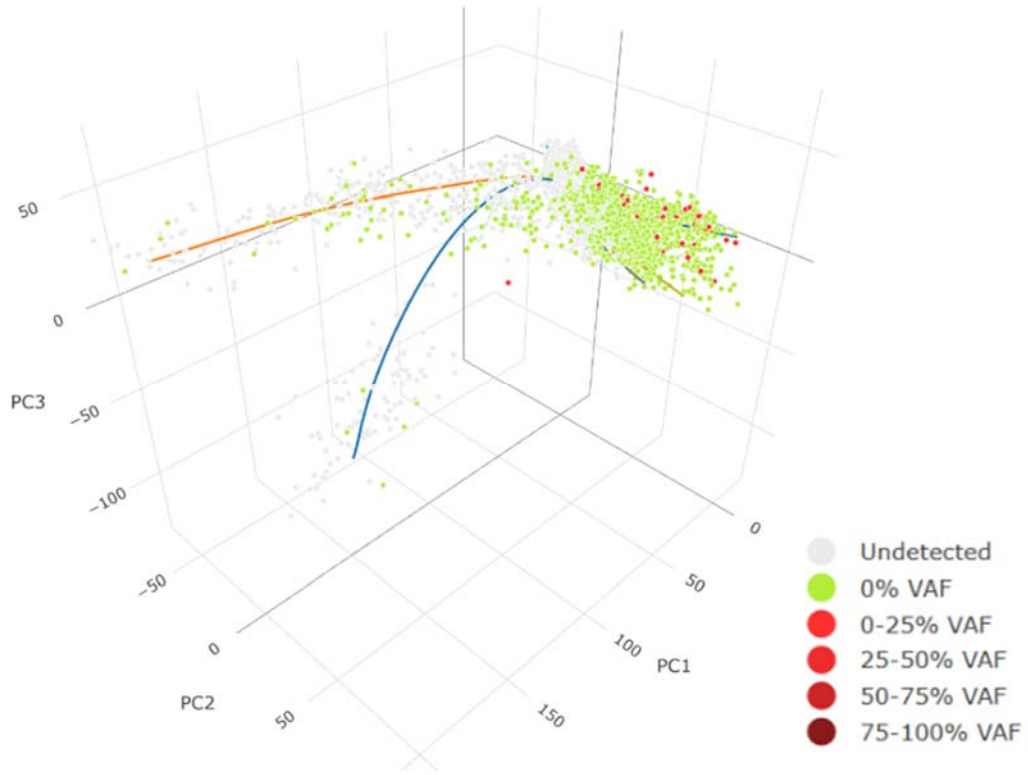

Figure 6: *UBE2C*^{W109*} variant distribution after PCA dimensionality reduction and pseudotime trajectory inference in one of the neuroblastoma samples. Only a subset of the *UBE2C* expressing cells express the variant allele (green dots vs red dots).

3.3 *UBE2C*^{W109Stop} expression across the 28 datasets

Next, we assessed the expression of *UBE2C*^{W109*} across the 28 publicly available cancer and normal datasets, included in our study (see Datasets and **Table 1**). Among the 28 datasets tested, *UBE2C*^{W109*} variant was found in cholangiocarcinoma, neuroblastoma and normal embryo.

In the cholangiocarcinoma dataset, cells were clustered into three groups, epithelia cells, fibroblasts and monocytes, and *UBE2C*^{W109*} was preferentially expressed in epithelia cells [**Figure7**, left panel]. Neuroblastoma cells were grouped into four clusters and the majority of cells that express *UBE2C*^{W109*} were neurons [**Figure7**, middle panel]. Given the fact that cholangiocarcinoma arises from epithelial cells (35)

and the origin of neuroblastoma is mainly neuronal cells (36), the expression of *UBE2C*^{W109*} in cholangiocarcinoma epithelial cells and neuroblastoma neurons suggested the variant might play a role in critical cell types in both cancers. Furthermore, *UBE2C*^{W109*} was found in the epithelial cells and fibroblasts of normal embryo sample [**Figure7**, right panel].

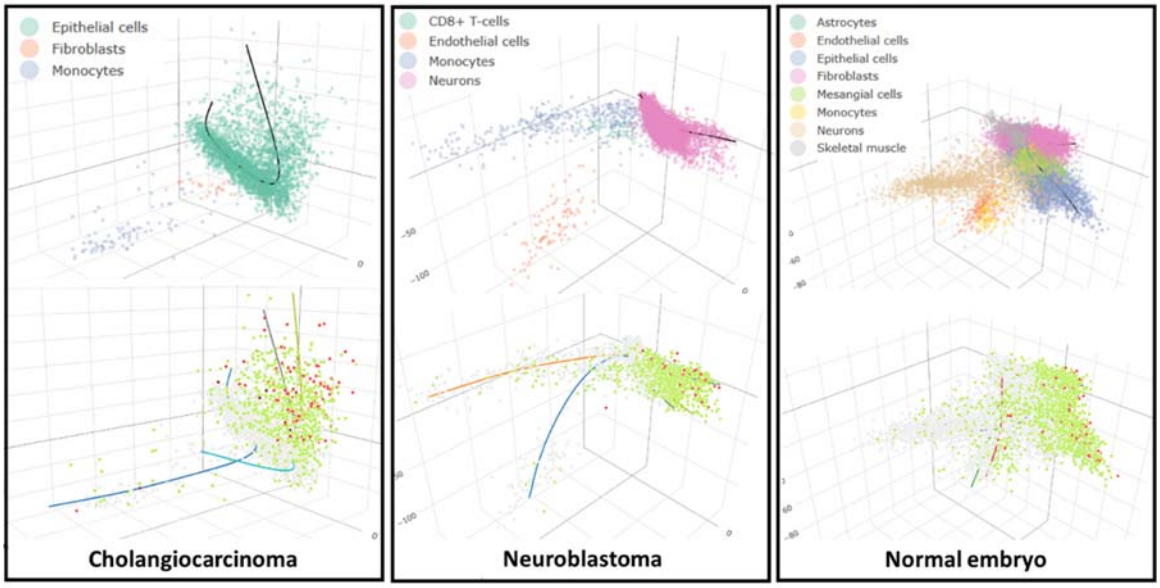

Figure 7: *UBE2C*^{W109*} was found in cholangiocarcinoma, neuroblastoma and normal embryo.

3.4 *UBE2C*^{W109*} is frequently observed in cholangiocarcinoma

Information form **Figure 7** showed that *UBE2C*^{W109*} was expressed both in the epithelial cells of cholangiocarcinoma and normal embryo samples, but it's possible that the expression level might be different between the two types of samples. Using two parameters, % variant expressing cells and % variant expressing reads, the expression level of the variant between the samples was examined. As shown in **Figure 8**,

UBE2C^{W109*} expressing cells and reads were all higher in cholangiocarcinoma compared to the other samples, which is an observation that requires further investigation.

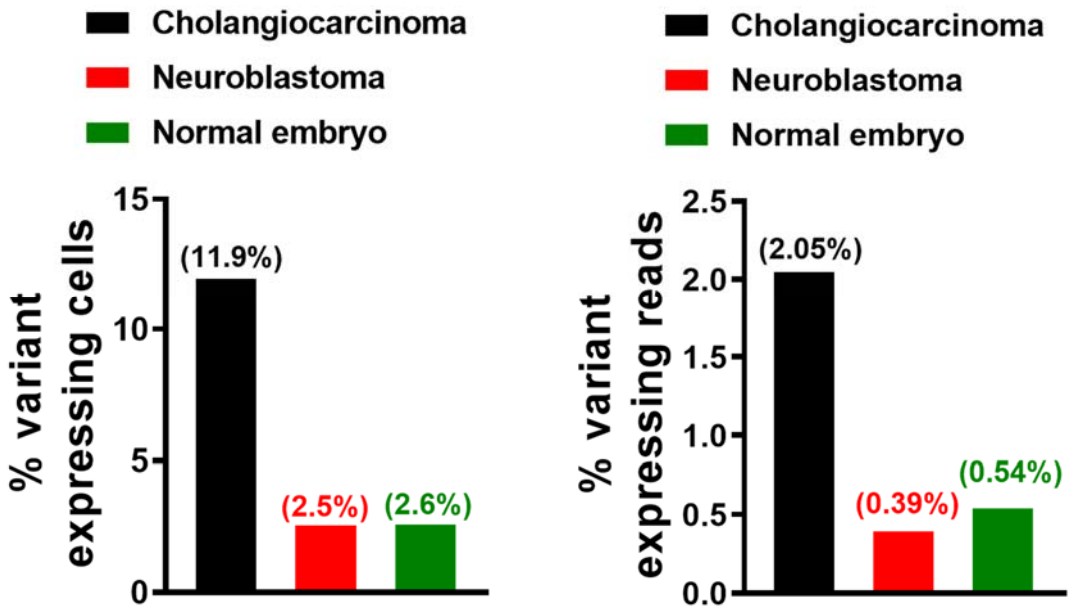

Figure 8: *UBE2C*^{W109Stop} expression differences between samples.

3.5 The effects of *UBE2C*^{W109*} on the gene expression are more pronounced in cancer

To test if *UBE2C*^{W109*} is correlated with changes in the gene expression, DE analysis was performed between the cells expressing the variant and the cells expressing only the reference allele. On **Figure 9**, differentially expressed genes from cholangiocarcinoma epithelial cells, neuroblastoma neurons, and normal embryo epithelial cells and fibroblasts were represented by blue dots, while the genes with high *p* values and fold changes were highlighted by red. Most of the highlighted genes were on the right side of the X-axis, indicating they were up-regulated genes. Furthermore, the two cancer samples had more up-regulated genes than the normal embryo. Interestingly,

the most up-regulated gene in all samples is the variant harboring gene *UBE2C* (red arrows in **Figure 9**).

The numbers of the dysregulated genes in the samples were also compared in **Figure 10** using adjusted *p* value of 0.05 as a cutoff line. Based on the results, cholangiocarcinoma epithelial cells had 92 differentially expressed genes, which is the highest one among the four cell clusters. These data suggested that the effects of *UBE2C*^{W109*} on the gene expression were more pronounced in cancer, especially in cholangiocarcinoma. Of note, this last observation can be affected by the sample size (in this case, cell numbers), and cell types (i.e. different number of genes expressed by certain cell types)

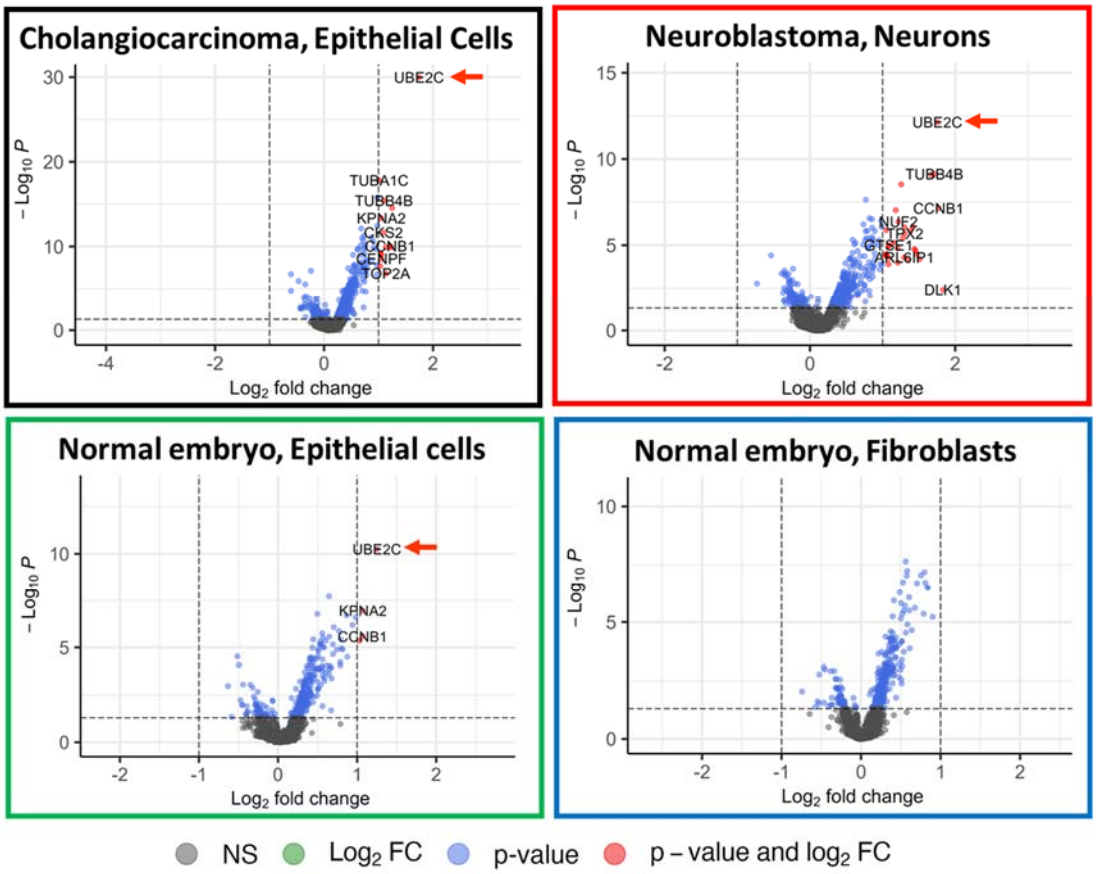

Figure 9: Differentially expressed genes between the cells with the variant or only the reference allele in regard to *UBE2C*^{W109*} in cholangiocarcinoma, neuroblastoma and normal embryo.

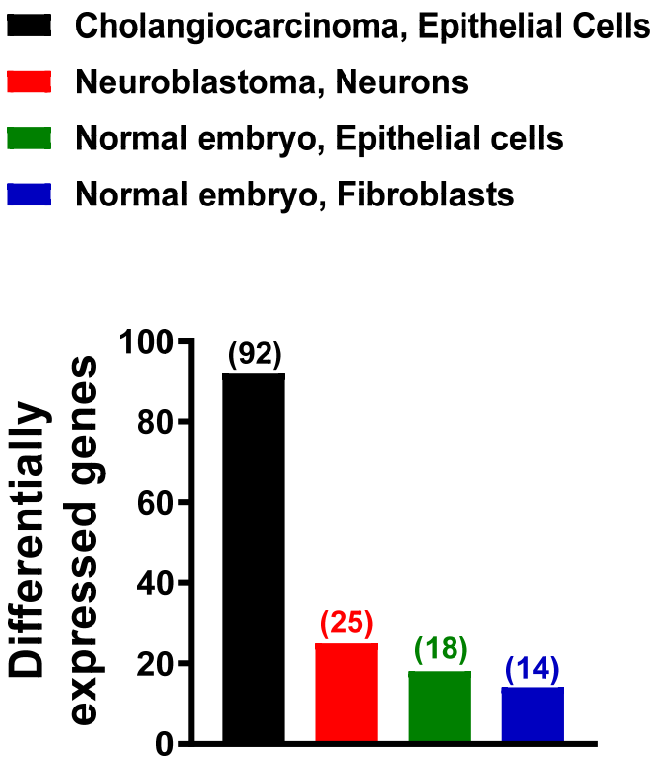

Figure 10: Number of differentially expressed genes in cholangiocarcinoma, neuroblastoma and normal embryo.

3.6 Differentially expressed genes in the two cancer samples belong to the same network

Next, we examined whether different samples had commonly regulated genes associated with the *UBE2C*^{W109*} variant. The overlay of the 92 differentially expressed genes in cholangiocarcinoma and the 25 genes in neuroblastoma showed that 16 genes were similarly deregulated in both of the cancer samples [**Figure 11**]. The interaction of

the common deregulated genes was then analyzed and visualized using String network analysis (29). As shown in **Figure 12**, most of the genes (13 out of 16) were predicted to be interacting with each other, indicating the genes influenced by the *UBE2C*^{W109Stop} mutation may belong to the same network.

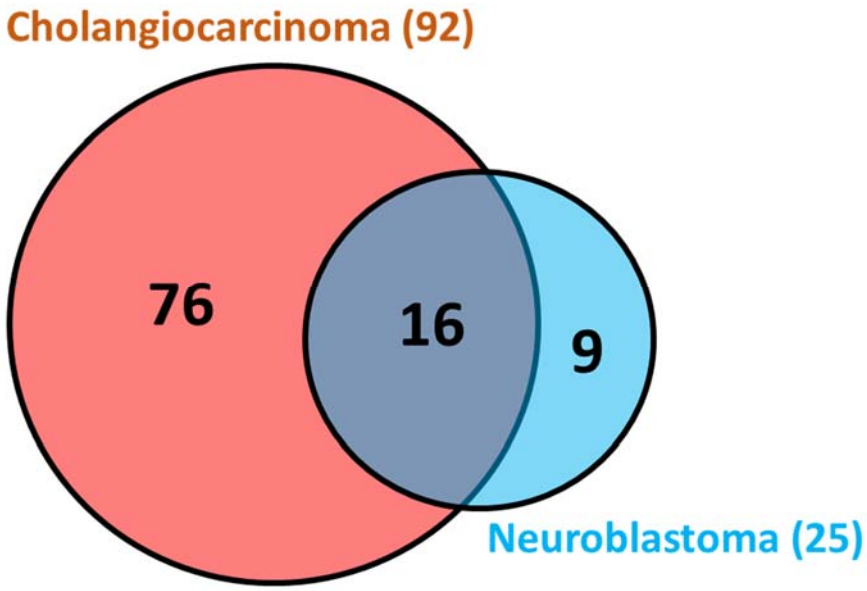

Figure 11: Commonly regulated genes between the cholangiocarcinoma and neuroblastoma samples due to *UBE2C*^{W109*} mutation.

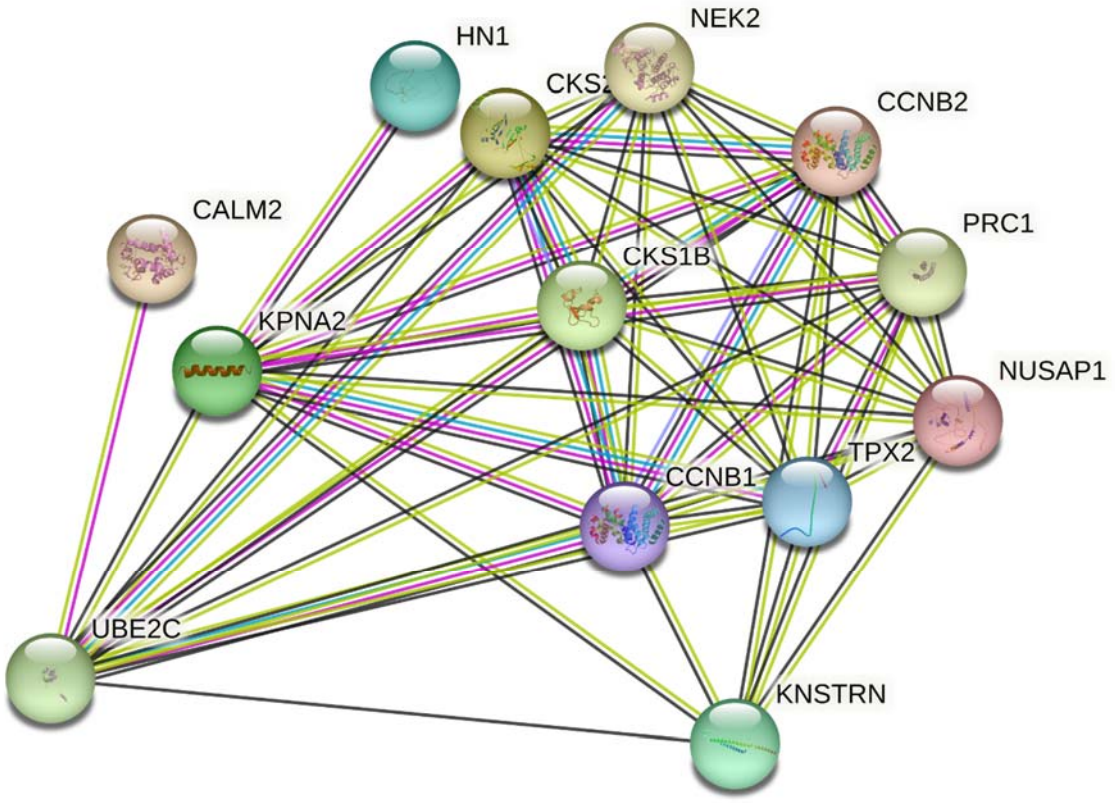

Figure 12: Predicted protein-protein interaction of genes enriched in *UBE2C*^{W109*} mutated cells.

3.7 Association of the differentially expressed genes with predicted pathways

Genes differentially regulated in *UBE2C*^{W109*} mutated cells were also analyzed by PANTHER Classification System (30) to reveal their functional pathway associations. One of the most enriched pathways was the ubiquitin pathway (dark yellow in **Figure 13**), which was expected since *UBE2C*^{W109*} expressed on the ubiquitin-conjugating enzyme. Furthermore, 15% of the differentially expressed genes were related to the P53 pathway (boxed in **Figure 13**). The p53 pathway responds to stresses that can disrupt the fidelity of DNA replication and cell division (37), and is involved in multiple tumor

processes. The enrichment of the differentially regulated gene in this pathway suggested *UBE2C*^{W109Stop} mutation might play a critical role in cancer progression.

Figure 13: Genes enriched in *UBE2C*^{W109Stop} mutated cells associated with predicted pathways as identified by PANTHER.

3.8 Origin of *UBE2C*^{W109*}

In the absence of matched DNA, the origin of *UBE2C*^{W109*} is difficult to ascertain. However, several considerations can be applied to predict likely variation origin based on it cell and sample distribution and expression pattern. First, the *UBE2C*^{W109*} is seen in multiple samples, however, has never been described previously in major databases for germline or somatic variants (i.e. DbSNP, or COSMIC) (38). Both germline and somatic variants seen in multiple samples are likely to have been already catalogued in the databases, making the DNA origin of *UBE2C*^{W109*} unlikely. Comparatively, RNA origin

is more likely, for several reasons. First, RNA-originating variants exert more transient nature and short lifetime that ends with the degradation of the variant transcript. Nevertheless, multiple RNA variants resulting from editing or other types of post-transcriptional modifications have been catalogued (REDI, others, and cite), but *UBE2C*^{W109*} is not among them. The later, however, can be explained with the presence of the variation in only a small proportion of cells and not be easily captured by the bulk-based variant callers due to insufficient sensitivity. In addition to RNA-editing, possible RNA origin is transcriptional infidelity. Finally, sequencing errors might also account for some of these variants. However, the latter is less likely as errors are expected to be random, whereas the variations studied are typically seen in the same position across multiple (with different start/end) sequencing reads. Consistent with above, the Unique Molecular Identifiers (UMIs) employed by the 10xGenomics protocol enable robust removal of PCR duplicates. Based on the above reasoning, we suggest that the most likely origin of *UBE2C*^{W109*} and other variants with similar appearance is low (cellular) frequency RNA posttranscriptional modification.

Chapter 4: Discussion

In the current study, we developed methods and a pipeline for analyzing alternative stop codon variants at the single cell level. We identified a large number of novel (not described previously) expressed alternative stop codon transcripts. Using the W109* mutation on the *UBE2C* gene as an example, we characterize *in silico* cell-level effects of this stop gained mutation. $UBE2C^{W109*}$ is found in different cancer and normal embryo samples and shows significantly higher cell frequency and level of expression in cholangiocarcinoma. Furthermore, the effect of $UBE2C^{W109*}$ is more pronounced in cancer and genes deregulated in the cancer samples are in known networks with *UBE2C*.

Currently, most genetic variation is analyzed from bulk sequencing datasets, which averages the gene expression level from hundreds to millions of cells. In this case, variants that are carried by multiple cells can be easily revealed, but low cellular frequency variants only expressed in small clusters of cells or single cells are difficult to distinguish from artifacts. A low expression level of a variant within a cell does not necessarily imply that it is of lesser importance. In fact, low-frequency variants can be utilized to trace lineages and resolve sub-clonal tumor architecture. Furthermore, in addition to cancer research, variants identified in single cells are also valuable for genetic studies (39). Our study developed a computational framework for identification and analysis of Single Cell-specific Expressed Single Nucleotide Variants (SNV) from single cell RNA-sequencing data and identified a large amount of previously unreported stop codon variants, including stop gained (nonsense) and stop loss (nonstop) variants. Additionally, a significant proportion of these novel variants exhibit preferential expression in specific cell clusters and pseudo-time stages, suggesting the existence of

novel functions for the expressed stop codons, including in the context of cancer progression and cell fate determination.

As an example of our analyses, we reported a novel stop gained mutation, *UBE2C*^{W109Stop}. We demonstrated that *UBE2C*^{W109Stop} is preferentially expressed in a subpopulation of the cells in cancer and normal embryo samples, especially in the cholangiocarcinoma sample. The mutation is responsible for the dysregulation of gene expression within known networks associated with *UBE2C*. Interestingly, the majority of the deregulated genes are upregulated, with the gene carrying the mutation, *UBE2C*, being the most significantly upregulated. For most of the stop gained mutation, nonsense-mediated mRNA decay (NMD) will be activated to reduce the nonsense codons, causing a reduction in the steady-state level of cytoplasmic mRNA. The upregulation of *UBE2C* mRNA expression in *UBE2C*^{W109Stop}-positive cells suggests that nonsense-mediated decay (NMD) did not occur. This increased expression of *UBE2C* may represent a compensatory mechanism in response to the dysfunction of UBE2C enzyme caused by the mutation. Alternatively, because the variant appearance co-occurs with extremely high expression of the UBE2C gene, it is not impossible that the variant occurs to compensate for the very high gene expression. Importantly, such mechanisms have not been studies before and are worth further investigations.

To further investigate the contribution of the variant to the upregulation of *UBE2C* mRNA, it may be beneficial to examine the composition of the mRNA. This could help determine whether the upregulation is due to an increase in the expression of the normal gene or an accumulation of the mutated gene. Another experiment that may be useful is to investigate potential changes in protein structure caused by the variant.

Overall, the failure to trigger NMD in mutated cells may lead to the accumulation of non-functional truncated UBE2C protein, which could further induce abnormal ubiquitination processes that ultimately contribute to cancer development.

Of the stop codon mutations observed in this study, more than 70% have not been previously reported. The origins of these mutations are of great interest and warrant further investigation. The possible origins of the mutations could arise from genomic DNA variation, either germline or somatic, or from abnormal RNA editing, which involves post- or co-transcriptional modification of RNA nucleotides. The most effective way to determine the source of the mutations would be to compare RNA sequencing data with the corresponding genomic DNA sequence (40). However, the limited availability of DNA sequencing information from the dataset used in this study makes such a comparison impossible. Luckly, as the stunning progress has been made to maximize the potential of RNA-Seq data, multiple methodologies have been developed to identify, predict, assess the quality of and catalog RNA-editing events without the help of matching DNA information (41–44). With the aid of these tools and other approaches to exploring changes in RNA editing and transcriptional machinery, a greater understanding of the underlying mechanisms can be achieved. This can provide further insight into the pathological pathways of cancer and other diseases related to stop codon mutations.

Chapter 5: Conclusions

In this study, we reveal a previously unappreciated collection of stop codon mutations expressed at the single-cell level, and exemplify our analyses with a novel stop gained mutation *UBE2C*^{W109Stop}, whose expression level and effects are more pronounced in cancer. Our results expand upon the current understanding of the mechanisms and functions of expressed genetic variation. They can provide additional information on the pathological mechanisms underlying cancer dynamics. By examining the expression of genetic variation at the single-cell level, we are able to gain a more nuanced understanding of the complex processes involved in cancer development and progression. This knowledge can be used to inform the development of more effective therapies for cancer and other diseases associated with stop codon mutations.

References

1. Koonin EV, Novozhilov AS (2009): Origin and evolution of the genetic code: the universal enigma. *Iubmb Life* 61: 99–111.

2. Ho AT, Hurst LD (2022): Stop Codon Usage as a Window into Genome Evolution: Mutation, Selection, Biased Gene Conversion and the TAG Paradox. *Genome Biol Evol* 14: evac115.

3. Hull R (2014): Chapter 6 - Genome Composition, Organization, and Expression. In: Hull R, editor. *Plant Virology (Fifth Edition)*. Boston: Academic Press, pp 247–339.

4. Chavatte L, Kervestin S, Favre A, Jean-Jean O (2003): Stop codon selection in eukaryotic translation termination: comparison of the discriminating potential between human and ciliate eRF1s. *EMBO J* 22: 1644–1653.

5. Mango G, Osti N, Udali S, Vareschi A, Malerba G, Giorgetti A, *et al.* (2022): Novel protein-truncating variant in the APOB gene may protect from coronary artery disease and adverse cardiovascular events. *Atheroscler Plus* 49: 42–46.

6. Gordon AJE, Sivaramakrishnan P, Halliday JA, Herman C (2018): Transcription infidelity and genome integrity: the parallax view. *Transcription* 9: 315–320.

7. Keeling KM, Du M, Bedwell DM (2013): Therapies of Nonsense-Associated Diseases. *Madame Curie Bioscience Database [Internet]*. Landes Bioscience. Retrieved July 30, 2023, from https://www.ncbi.nlm.nih.gov/books/NBK6183/

8. Brogna S, Wen J (2009): Nonsense-mediated mRNA decay (NMD) mechanisms [no. 2]. *Nat Struct Mol Biol* 16: 107–113.

9. Lykke-Andersen S, Jensen TH (2015): Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes [no. 11]. *Nat Rev Mol Cell Biol* 16: 665–677.

10. Lewis BP, Green RE, Brenner SE (2003): Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A* 100: 189–192.

11. Hamby SE, Thomas NST, Cooper DN, Chuzhanova N (2011): A meta-analysis of single base-pair substitutions in translational termination codons ("nonstop" mutations) that cause human inherited disease. *Hum Genomics* 5: 241–264.

12. Vasudevan S, Peltz SW, Wilusz CJ (2002): Non-stop decay--a new mRNA surveillance pathway. *BioEssays News Rev Mol Cell Dev Biol* 24: 785–788.

13. Benhabiles H, Gonzalez-Hilarion S, Amand S, Bailly C, Prévotat A, Reix P, *et al.* (2017): Optimized approach for the identification of highly efficient correctors of nonsense mutations in human diseases. *PLOS ONE* 12: e0187930.

14. Pang S, Wang W, Rich B, David R, Chang YT, Carbunaru G, *et al.* (2002): A Novel Nonstop Mutation in the Stop Codon and a Novel Missense Mutation in the Type II 3β-Hydroxysteroid Dehydrogenase (3β-HSD) Gene Causing, Respectively, Nonclassic and Classic 3β-HSD Deficiency Congenital Adrenal Hyperplasia. *J Clin Endocrinol Metab* 87: 2556–2563.

15. Doucette L, Green J, Fernandez B, Johnson GJ, Parfrey P, Young T-L (2011): A novel, non-stop mutation in FOXE3 causes an autosomal dominant form of variable anterior segment dysgenesis including Peters anomaly. *Eur J Hum Genet* 19: 293–299.

16. Zhang M, Yang H, Wan L, Wang Z, Wang H, Ge C, *et al.* (2020): Single-cell transcriptomic architecture and intercellular crosstalk of human intrahepatic cholangiocarcinoma. *J Hepatol* 73: 1118–1130.

17. Ma X, Guo J, Liu K, Chen L, Liu D, Dong S, *et al.* (2020): Identification of a distinct luminal subgroup diagnosing and stratifying early stage prostate cancer by tissue-based single-cell RNA sequencing. *Mol Cancer* 19: 147.

18. Dong R, Yang R, Zhan Y, Lai H-D, Ye C-J, Yao X-Y, *et al.* (2020): Single-Cell Characterization of Malignant Phenotypes and Developmental Trajectories of Adrenal Neuroblastoma. *Cancer Cell* 38: 716-733.e6.

19. Dillard C, Ulianova E, Prashant N, Liu H, Edwards N, Horvath A (2022): *A Wealth of Novel Cell-Specific Expressed SNVs from Tumor and Normal ScRNA-Seq Datasets*. Genomics. https://doi.org/10.1101/2022.06.12.495797

20. Kaminow B, Yunusov D, Dobin A (2021): STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq dat. https://doi.org/10.1101/2021.05.05.442755

21. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, *et al.* (2010): The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20: 1297–1303.

22. Edwards N, Dillard C, Prashant NM, Hongyu L, Yang M, Ulianova E, Horvath A (2023): SCExecute: custom cell barcode-stratified analyses of scRNA-seq data. *Bioinforma Oxf Engl* 39: btac768.

23. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, *et al.* (2018): Strelka2: fast and accurate calling of germline and somatic variants [no. 8]. *Nat Methods* 15: 591–594.

24. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001): dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29: 308–311.

25. Hao Y, Stuart T, Kowalski M, Choudhary S, Hoffman P, Hartman A, *et al.* (2022, February 26): Dictionary learning for integrative, multimodal, and scalable single-cell analysis. bioRxiv, p 2022.02.24.481684.

26. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, *et al.* (2019): Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage [no. 2]. *Nat Immunol* 20: 163–172.

27. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, *et al.* (2018): Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 19: 477.

28. Love MI, Huber W, Anders S (2014): Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15: 550.

29. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ (2019): Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *J Proteome Res* 18: 623–632.

30. Mi H, Muruganujan A, Thomas PD (2013): PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res* 41: D377-386.

31. Hao Z, Zhang H, Cowell J (2012): Ubiquitin-conjugating enzyme UBE2C: molecular biology, role in tumorigenesis, and potential as a biomarker. *Tumor Biol* 33: 723–730.

32. Liu W, Tang X, Qi X, Fu X, Ghimire S, Ma R, *et al.* (2020): The Ubiquitin Conjugating Enzyme: An Important Ubiquitin Transfer Platform in Ubiquitin-Proteasome System. *Int J Mol Sci* 21: 2894.

33. Townsley FM, Aristarkhov A, Beck S, Hershko A, Ruderman JV (1997): Dominant-negative cyclin-selective ubiquitin carrier protein E2-C/UbcH10 blocks cells in metaphase. *Proc Natl Acad Sci U S A* 94: 2362–2367.

34. Kim JH, Lee KW, Yang HJ, Park JJ, Lee CH, Jeon YS, *et al.* (2023): The correlation between the expression of ubiquitin-conjugating enzyme 2C and prostate cancer prognosis. *J Cancer Res Clin Oncol* 149: 6351–6360.

35. Cardinale V, Carpino G, Reid L, Gaudio E, Alvaro D (2012): Multiple cells of origin in cholangiocarcinoma underlie biological, epidemiological and clinical heterogeneity. *World J Gastrointest Oncol* 4: 94–102.

36. Tsubota S, Kadomatsu K (2017): Origin and mechanism of neuroblastoma. *Oncoscience* 4: 70–72.

37. Harris SL, Levine AJ (2005): The p53 pathway: positive and negative feedback loops [no. 17]. *Oncogene* 24: 2899–2908.

38. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, *et al.* (2019): COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* 47: D941–D947.

39. Edwards N, Dillard C, Prashant NM, Hongyu L, Yang M, Ulianova E, Horvath A (2023): SCExecute: custom cell barcode-stratified analyses of scRNA-seq data. *Bioinforma Oxf Engl* 39: btac768.

40. Light D, Haas R, Yazbak M, Elfand T, Blau T, Lamm AT (2021): RESIC: A Tool for Comprehensive Adenosine to Inosine RNA Editing Site Identification and Classification. *Front Genet* 12. Retrieved June 18, 2023, from https://www.frontiersin.org/articles/10.3389/fgene.2021.686851

41. Guo Y, Yu H, Samuels DC, Yue W, Ness S, Zhao Y-Y (2019): Single-nucleotide variants in human RNA: RNA editing and beyond. *Brief Funct Genomics* 18: 30–39.

42. Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, Li JB (2013): Identifying RNA editing sites using RNA sequencing data alone [no. 2]. *Nat Methods* 10: 128–132.

43. Pinto Y, Levanon EY (2019): Computational approaches for detection and quantification of A-to-I RNA-editing. *Methods* 156: 25–31.

44. Ouyang Z, Liu F, Zhao C, Ren C, An G, Mei C, *et al.* (2018): Accurate identification of RNA editing sites from primitive sequence with deep neural networks [no. 1]. *Sci Rep* 8: 6005.

45. https://snp.gs.washington.edu/SeattleSeqAnnotation154/