Insights into Recurring Intragenic Rearrangements in High-Grade Serous Carcinoma

by

Matthew Wexler

B.S. Microbiology, Michigan State University, 2015

Submitted to the Graduate Faculty of

the School of Medicine in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

School of Medicine

This thesis was presented

by

Matthew Wexler

It was defended on

May 14, 2018

and approved by

Adrian Lee, Professor, Department of Pharmacology and Chemical Biology

Hun-Way Hwang, Assistant Professor, Department of Pathology

Thesis Director: Xiaosong Wang, Associate Professor, Department of Pathology

Insights into Recurring Intragenic Rearrangements in High-Grade Serous Carcinoma

Matthew Wexler

University of Pittsburgh, 2018

High grade serous ovarian cancer (HGSC) accounts for 80% of ovarian cancer mortality, and no targeted therapies are available for this disease. Although gene fusions have been a significant focus of cancer genetics studies, other types of genomic rearrangements are known to be cancer-driving, including intragenic genetic rearrangements (IGRs) that result in exons within genes being duplicated or deleted. Some IGRs have been reported to drive growth in tumors, such as *EGFR* and *ERBB2* exon rearrangements, which are known to cause activation of these kinases, but IGRs have not been systematically analyzed in cancer.

Here, we performed a comprehensive bioinformatics analysis to identify potential recurrent intragenic rearrangements in high-grade serous carcinoma. Of note, we identified a potential intragenic duplication of exon 4-5 in *EPHA3* based on copy number data analysis; we term this duplication EPHA3d45. We examined the presence of EPHA3d45 transcript by reverse-transcription PCR and capillary sequencing in positive cell lines identified by RNA-seq, and we showed that the siRNA designed to specifically target the duplication junction inhibits the growth of a duplication-positive cell line, TOV112D, but not in a duplication-negative cell line. Additional analysis will be needed to understand the origin and oncogenic potential of EPHA3d45, and further *in silico* validation, followed by experimental confirmation, could lead to the identification of other novel recurring IGRs in HGSC.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my PI, Xiaosong Wang, for the opportunity to work in his laboratory and for his multidisciplinary guidance. I would also like to thank the members of the Wang laboratory for their mentorship and expertise. In particular, thanks to Xu Chi for helping me to learn bioinformatics and to Grace Hu and Xian Wang for their assistance in cell and molecular biology experiments.

I would also like to offer my gratitude to the staff and technicians at the University of Pittsburgh, the Department of Pathology, and IBGP for their continual assistance, both scientific and logistic, throughout my graduate career.

Thanks also to the Pittsburgh Fencing Association for helping to keep me sane, and to my friends and family for their continued support.

Special thanks to Nala for being the cutest dog ever.

Finally, a note of appreciation for Fall Out Boy for creating music that, while tolerable at best in most contexts, is strangely perfect to listen to while coding.

1.0 INTRODUCTION

Since the discovery of the Philadelphia Chromosome in 1960, it has been widely recognized that genomic structural rearrangements can cause cancer. The first such rearrangements characterized, like the Philadelphia Chromosome, are inter-chromosomal rearrangements that lead to the creation of fusion genes in hematologic malignancies [1]. At the time such rearrangements were characterized, available technologies were best suited to the discovery of these 'large' translocations; rearrangements within chromosomes were generally beyond the sensitivity of available detection methods.

In recent years, the advent of sequencing technologies has exponentially increased the sensitivity with which genomic rearrangements can be detected, and several important gene fusions have been found to lead to tumor growth [2-5]. However, the legacy of the first intergenic rearrangements discovered persists because the field of cancer genomics is generally preoccupied with oncogenic gene fusions. Although these were the first clinically important rearrangements discovered, this preoccupation has led to other kinds of rearrangements being overlooked.

One important kind of genomic structural rearrangement is intragenic rearrangements (IGRs). Such rearrangements occur when one or more exons within a gene are duplicated or deleted. In the literature, there are sporadic reports of such rearrangements in cancer. The best-known example is likely EGFRvIII, a variant of *EGFR* in which exons coding for protein regulatory domains are deleted, leading to constitutive, oncogenic activation of the gene in

glioblastoma [6, 7]. An oncogenic intragenic duplication in *EGFR* has also been reported in lung cancer [8], and activating exonal rearrangements of *ERBB2* have been reported in breast cancer [9]. A number of IGRs have also been reported in tumor suppressors such as *BRCA1-2*; these rearrangements disrupt the open reading frame of the gene, causing it to be functionally deleted [10-18]. Despite these interesting results, to the best of our knowledge, IGRs have not been the subject of focused, systemic investigation.

In this study, we used a bioinformatics approach to systematically identify recurring IGRs in high-grade serous carcinoma (HGSC). HGSC accounts for up to 80% of ovarian cancer mortality, and overall survival has not been improved for decades [19]. Profiling of HGSC genomic alterations by The Cancer Genome Atlas (TCGA) revealed that HGSC genomes are dominated by genomic rearrangements, while recurrent point mutations (with the exception of P53) are rare [20]. Although HGSC is known to have high genomic instability [21], investigations searching for cancer-driving genomic rearrangements—primarily gene fusions—have identified a paucity of "actionable" aberrations [22-27]. We reasoned that a study of IGRs in HGSC might identify recurrent genomic rearrangements that have been previously been overlooked and could be pathological events in HGSC. Our investigation identified several interesting IGRs; this report will focus on the intragenic duplication of exons 4 and 5 of the gene *EPHA3* (also known as *ETK1*; the gene harboring the duplication is hereafter referred to as "EPHA3d45").

2.0 MATERIALS AND METHODS

Bioinformatics Analysis of Genomic Datasets

Segmented tumor and paired normal copy number data was obtained from TCGA (https://cancergenome.nih.gov/, version 161227, downloaded on 12/28/2016); cell line copy number and RNA-seq data was obtained from CCLE (https://portals.broadinstitute.org/ccle, downloaded on 12/3/2013 and 2/24/2018, respectively); HGSC tumor RNA-seq data was obtained from Dr. Adrian Lee (unpublished dataset); GTEX RNA-seq data was downloaded from GTEX (https://www.gtexportal.org, version 7, downloaded on 4/5/2018).

To identify IGRs from copy number data, we developed a novel pipeline called "Exon CNV Mapper." Conceptually, this tool identifies exons within genes that have a copy number value that is greater or less than the median value for all exons in that gene by a cutoff value. The pipeline identifies exons by matching exonal coordinates to the corresponding location in the segmented copy number data (genomic build hg38 was used), then assigns a copy number value to each exon based on the segmented copy number data. Importantly, these copy number values are not determined experimentally for each exon; rather, they are assigned based on the exon's location in larger copy number-assigned segments. Due to the relatively small size of exons, all exons were able to be assigned only one copy number value. Exons within the same segment would all be assigned the same copy number value, whereas exons between segments will not be assigned a value. Copy number values for all exons within a given gene are then used to calculate the median.

After assigning values to each exon, Exon CNV Mapper determines whether consecutive exons within each gene with the same copy number value which is higher or lower than the median

copy number value for all exons in the gene. In order to determine a meaningful difference, setting a cutoff is the standard method used in the field. Similar studies in the field have used copy number value differences ranging from 0.2 to 0.3 as the cutoff [2, 28-30]; therefore, we chose a copy number value difference 0.25 as our cutoff value. As an example of how Exon CNV Mapper works, suppose that *gene x* has 10 exons, and the copy number value for exons 6-8 is determined to be 0.6, while the copy number value for the remaining exons is determined to be 0.1. In this case, Exon CNV Mapper will annotate a duplication of exons 6-8 in *gene x*.

To identify aberration transcripts with exon duplications or deletions from RNA-seq datasets, we first applied the splice aligner TopHat-Fusion to align the RNA-seq Fastq files against human genome sequences (genome build hg38) [31]. The following parameters were used: --fusion-search --keep-fasta-order --b2-very-sensitive --no-coverage-search -r 112 --mate-std-dev 150 --fusion-min-dist 1000 --fusion-anchor-length 18. TopHat-Fusion identifies aberrant exon junctions supported by chimeric reads and discordant mate-pairs. Junction reads occur when exons adjoin each other in abnormal ways. For example, a duplication in exons 6-8 of *gene x* would produce a transcript containing a novel junction from exon 8 to exon 6. The result from TopHat-Fusion analysis were parsed using our IGR sensor pipeline. This pipeline matches the fusion junctions detected by TopHat2-fusion to human exons and determines the consequence of these potential rearrangements on gene structures. Based on how these fusion junctions map to human genes, exonal duplications or deletions are annotated as in **Figure 1**.

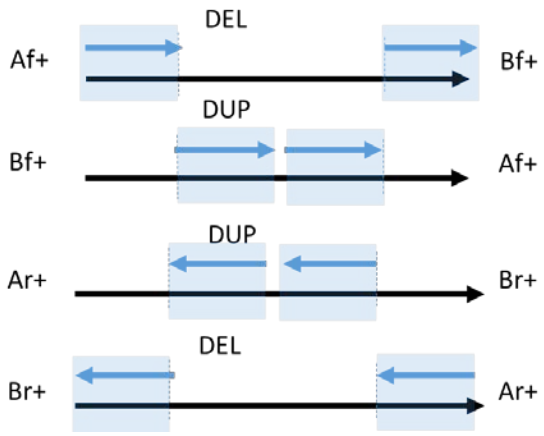


Figure 1: Schematic for characterization of IGRs using the IGV sensor pipeline.

Representations of duplications (DUP) and deletions (DEL) are shown in the figure. Blue arrows represent duplication junctions detected by TopHat2, black arrows represent the genomic placements of human genes, and dashed lines represent the breakpoints. Labels on either side are representative of TopHat2 output as follows: 5' or 3' junction partner (A or B, respectively), forward or reverse orientation (f or r, respectively), and the strand of human genes (positive or negative). Please note that all cases shown are for the human gene of positive strand; the same logic also applies to IGRs of genes located in the negative strand.

Tissues, Cell Lines, and Culture Method

100 ng RNA from each sample of paired primary/recurrent HGSC tumors was provided by Dr. Adrian Lee. Normal pooled human RNAs were purchased from Clontech. TOV-112D, OVCAR3, and NCIH446 were purchased from ATCC (https://www.atcc.org/). TOV-112D cells were cultured in a 1:1 mixture of MCDB105 and Medium 199 (ATCC) supplemented with 15% FBS. OVCAR3 cells were cultured in RPMI (ATCC) supplemented with 10% FBS and 0.01 mg/mL human insulin. NCIH446 cells were cultured in DMEM (ATCC) supplemented with 10% FBS. OVKATE cells were a gift from Dr. Steffi Oesterreich; DOV13 cells were a gift from Dr. Robert Bast; NCIH460 and NCIH23 were gifts from Dr. Laura Stabile, and these cell lines were cultured in DMEM with 10% FBS. All cell lines were cultured at 37°C and 5% CO_2.

Reverse Transcriptase PCR (RT-PCR)

RT for cell lines was performed with the Transcriptor First Strand cDNA Synthesis kit (Roche). 1µg cell line RNA was combined with 1.5 µL of both random oligo primers (50 µM) and hexamers (600 µM) and water to a final volume of 13 µL, then heated to 65°C for 10 minutes before being placed on ice. Then, 4 µL TRT Reaction buffer, 0.5 µL Protector RNase Inhibitor, 2 µL DNT mix and 0.5 µL Transcriptor Reverse Transcriptase was added. The mixture was briefly centrifuged, followed by heating at 25°C for 10 minutes, 50°C for 50 minutes, and 85°C for 5 minutes. The cDNA was then diluted 1:2 and stored at -20°C. For all RT reactions, 20 µL undiluted cDNA was generated.

RT for tissues was performed using the Superscript IV kit (Invitrogen). 100ng RNA (HGSC tumors) or 500ng RNA (pooled human normal tissues) were combined with 1.5 µL DNT mix, 1.5 µL each of random oligo primers and hexamers, and water to a total volume of 13 µL, then heated at 65°C for 5 minutes before being placed on ice. Then, 4 µL First-strand buffer, 1 µL 0.1M DTT, 1 µL RNaseOUT (Thermo Fisher), and 1 µL Superscript IV enzyme were added. The mixture was briefly centrifuged, followed by heating at 25°C for 10 minutes, 50°C for 55 minutes, and 85°C for 5 minutes. 1 µL RNase H (Invitrogen) was added and the mixture was heated at 37°C for 20 minutes. The cDNA was stored at -20°C.

The following primers were used for PCR of wild-type EPHA3: Forward: TCCCTGGTGGAGGTTAGAGG; Reverse: TGTTCGTCCCATATCCAGCG. The following primers were used for PCR of EPHA3d45: Forward: TGTGAGGCGGGCACTTAGCA; Reverse: GGAGTTGGCCCCTGGACACA. GAPDH primers were reported previously [2]. All PCRs were performed using the High Fidelity Platinum Taq kit (Invitrogen) according the manufacturer's instructions. For all PCR reactions, 2 µL cDNA (generated as described above) was used, and all

reactions included 6% DMSO. Total PCR reaction volume was 50 µL, and 1 µL of Platinum Taq was used. Primers were used at final concentrations of 10 µM. For all primers, PCR reaction mixtures were heated to 94°C for 2 minutes, followed by variable cycles (see below) of 94°C for 15 seconds, 58°C for 30 seconds, and 68°C for 1 minutes; cycles were followed by heating at 68°C for 7 minutes, then samples were cooled to 4°C until they were loaded onto the gel. GAPDH was subjected to 25 cycles of amplification; all other targets were subjected to 35 (for cell lines) or 37 (for human tissue) cycles of amplification. PCR products were run by electrophoresis on 1.2% agarose gel containing SYBR Safe DNA gel stain (Invitrogen), and gels were imaged with a Chemidoc Touch Imaging System (Bio-Rad).

EPHA3d45 bands were purified from gels using the Zymoclean Gen DNA Recovery Kit (Zymo Research) according to the manufacturer's instructions. Purified PCR products were sent to the University of Pittsburgh Genomics Core for sequencing by capillary sequencing, according to Core guidelines.

Knockdown Experiments

siRNA specific to the duplication junction of EPHA3d45 was designed and ordered from Dharmacon (sequence: AATCAGGCTGCTTGTCGACC). Scrambled siRNA used as a negative control was also purchased from Dharmacon. Transfections were performed using Lipofectamine RNAiMAX (Thermo Fischer) according to the manufacturer's instructions; cells were transfected with a final concentration of 10nM siRNA.

Clonogenic and MTS assays were performed as previously described [2, 32]. For colonogenic assays, 15,000 cells/well were seeded in 6-well plates on day 0. Cells were transfected with siRNA on days 1 and 8, with media changes the day prior to transfection. On day 11, colonies

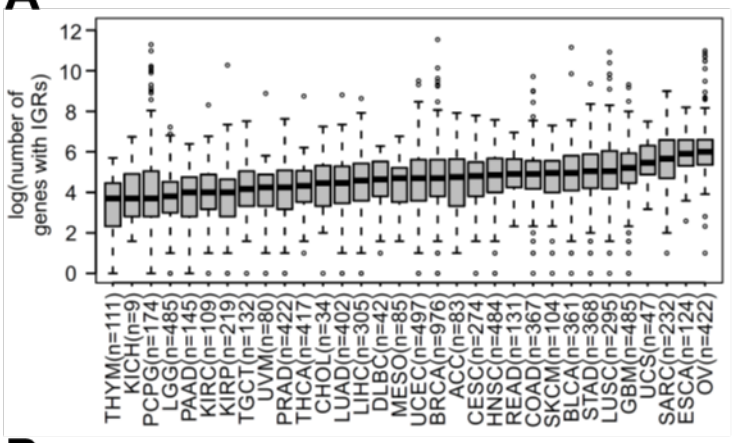
were stained with 0.1% crystal violet for 1 hour, then washed with PBS and left to dry overnight. Then, images were taken with the Chemidoc Touch Imaging System. MTS reagent was prepared by dissolving 1.9 mg/mL CellTiter 96 Aqueous MTS Reagent Powder (Promega) in PBS, then adding 1:1000 PES and applying a 0.2 µM filter. For MTS assays, 5000 cells/well were seeded in 96-well plates on day 0. Measurements were taken by incubating cells with the appropriate complete media containing 20% MTS reagent for 2 hours at 37°C and 5% CO_2, then measuring OD490 with an iMark Microplate Reader (Bio-Rad) Cells were transfected on day 1, and measurements were taken every 2-3 days starting on day 1.

3.0 RESULTS

Systematic identification of recurrent IGRs through analysis of TCGA copy number data

To identify IGRs from TCGA copy number data, we have developed a novel pipeline called "Exon CNV Mapper" (see Methods). Analysis of TCGA Pan-cancer copy number data revealed that HGSC exhibits a higher frequency of unbalanced IGR events than all other cancers (Figure 2A). More interestingly, this analysis identified candidate IGRs in established oncogenes, such as ERBB2 and FGFR1. These rearrangement hotspots in specific exons suggest the potential to be oncogenic (Figure 2B).

A



B

Gene	Cancer uniConSig Score*	Exon(s)	IGRs	TCGA Somatic Incidence (n=418)	TCGA Germline Incidence (n=418)	HapMap Incidence (n=982)
ETK1	0.59	4-5	Dup	8.29%	0.24%	0.00%
ERBB2	0.87	14-15	Dup	4.03%	0.47%	0.92%
FGFR1	0.79	6	Dup	3.32%	0.24%	0.10%

Figure 2: Unbalanced IGRs identified by analysis of copy number from TCGA.

A. Number of genes harboring recurrent IGRs in all TCGA tumor types, sorted by median number of IGRs per sample. HGSC (OV) harbors the greatest number of IGRs per sample. B. Table showing the top

candidate exon duplications and their frequency in TCGA HGSC tumors, as well as HapMap normal cohorts. *A cancer uniConSig score >0.45 indicates a high functional relevance underlying cancer calculated based on the signature molecular concepts characteristic of known cancer genes. Dup, exon duplication. The segmented copy number data used in the above analyses were downloaded from TCGA commons on 12/28/2016.

Identification of EPHA3d45 based on analysis of TCGA copy number data

Of the unbalanced IGRs identified in HGSC, the most frequent one is the duplication of exons 4-5 in *EPHA3 (*Figure 2B, 3A). *EPHA3* is an ephrin-receptor protein-tyrosine kinase; it contains an intracellular kinase domain and an extracellular region containing a ligand binding domain, a Cys-rich domain, and 2 fibronectin type III repeats. *EPHA3* has been proposed as a therapeutic target in hematological malignancies and glioblastoma multiforme [33, 34]. Copy number data from TCGA suggests that this duplication may be present in 8.3% of HGSC tumors (Figure 2B). Based on copy number data of paired normal blood from the TCGA, this duplication may also present as a rare germline event in up to 0.24% of the patients in the TCGA HGSC dataset (Figure 2B). This duplication has also been reported as a rare germline event in 1 out of 29084 healthy individuals by Coe BP. Et al. based on copy number data [35].

Further, analysis of RNA-seq data for 18 paired primary and recurrent HGSC tumors (provided by Dr. Adrian Lee) detected this duplicated transcript at a comparable frequency in primary tumors (2 out of 18) but a higher frequency in recurrent tumors (5 out of 18; Figure 3B). In addition, RNA-seq data from the Cancer Cell Line Encyclopedia (CCLE) detected this duplication in the ovarian cell line TOV-112D and three lung cancer cell lines, including NCI-H446 (Figure 3C). In addition, this duplication has been previously reported as an expressed duplication in a breast cancer cell line, HCC38, and the precise genomic fusion points have been identified (chr3:89364714-89393561) [36].

Duplication of exon 4-5 in *EPHA3* is predicted to generate a protein which has an extra fibronectin III (FN3) domain (Figure 3D), which is frequently present in oncogenes [37]. Interestingly, the FN3 domains of *EPHA3* are known to bind to the ligand-binding domain of the protein. Some have suggested that this acts to inhibit *EPHA3* signaling by preventing it from binding to ligands, whereas others have proposed that *EPHA3* proteins bound to each other via FN3 domains would functionally oligomerize, resulting in increased signaling [38].

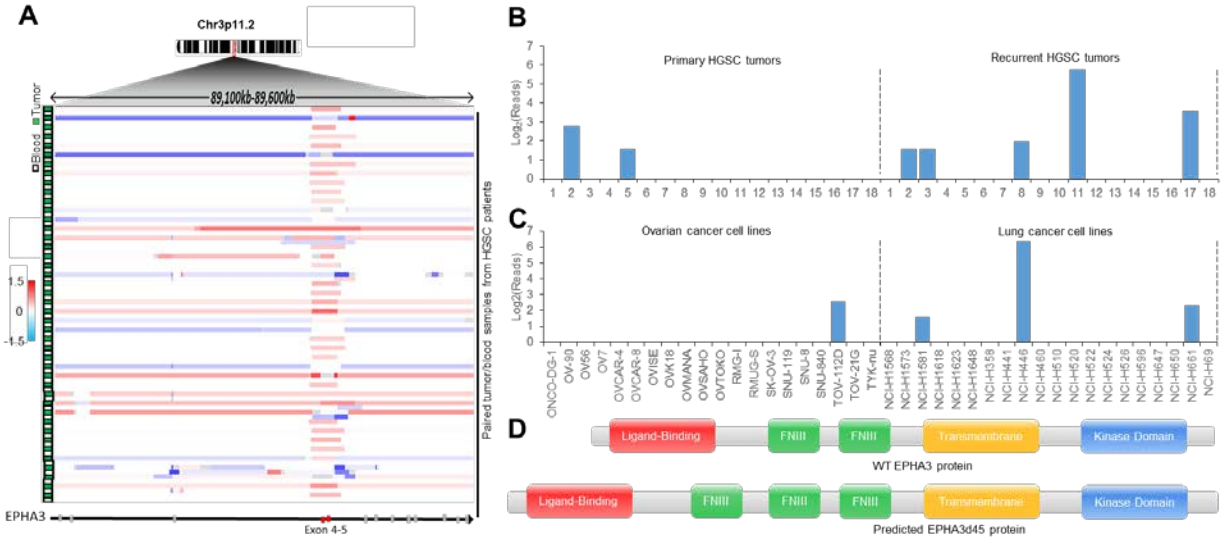


Figure 3: Identification of EPHA3d45 in HGSC tumors and cancer cell lines from copy number and RNA-seq data.

A. TCGA copy number data for paired tumor and blood samples from EPHA3d45-positive patients. Data are visualized with IGV. B-C. Detection of EPHA3d45-specific reads by RNA-seq in paired primary and recurrent HGSC tumors (B) and in a panel of ovarian and lung cancer cell lines from CCLE (C). D. Domain schematic of the wild-type EPHA3 protein and the predicted protein structure of EPHA3d45.

Confirmation of EPHA3d45 transcript expression in cancer cell lines

We selected two cell lines which were positive for EPHA3d45 by RNA-seq for experimental validation: the ovarian cancer cell line TOV-112D and the lung cancer cell line NCI-

H446. An 'internal' PCR strategy was developed to specifically amplify the duplicated EPHA3, but not the wild-type (Figure 4A). We successfully amplified EPHA3d45 bands from both positive cell lines but none of the negative control cell lines (Figure 4B-C). It is worth noting, however, that EPHA3d45 transcript-positive cell lines showed higher expression level of the wild-type EPHA3 transcript than EPHA3d45-negative cell lines (Figure 4B-C). The identity of all EPHA3d45 bands were verified by capillary sequencing (Figure 5). It is necessary to note that this sequence verifies a junction between exons 5 and 4, but this does not definitively demonstrate a duplication of only exons 4 and 5 just once—for example, if exons 4 and 5 were duplicated more than once, the same results would be obtained. Therefore, further characterization of the exact transcript structure will be necessary in future work.

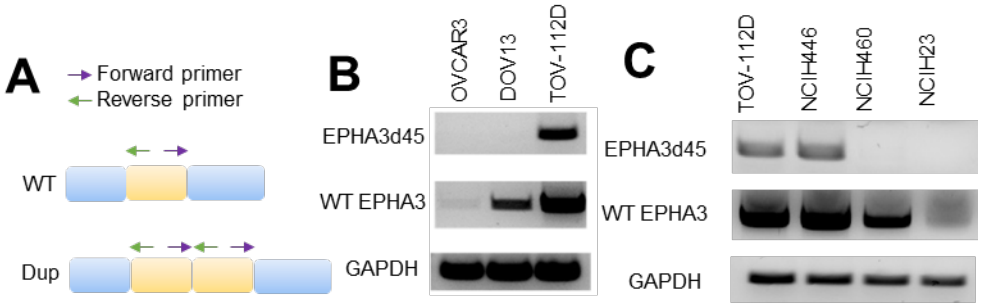


Figure 4: Validation of EPHA3d45 transcript expression by RT-PCR in ovarian and lung cancer cell lines.

A. Schematic showing RT-PCR strategy to specifically detect the duplication junction. A PCR product is only formed when the region of interest (shown in yellow) is duplicated. B-C. RT-PCR validation of EPHA3d45 in ovarian (B) and lung (C) cancer cell lines. TOV-112D and NCIH446 are positive for the duplication. Results shown are representative of at least two independent experiments.

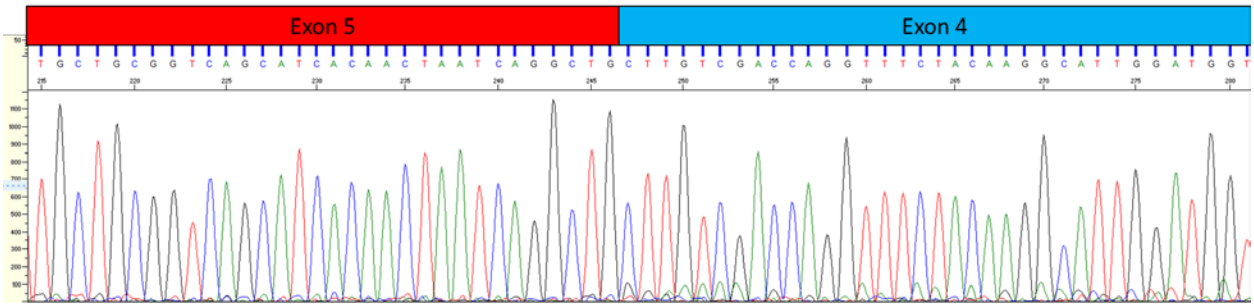


Figure 5: Representative chromatograph showing the junction sequence of EPHA3d45.

The junction between exon 5 (left) and exon 4 (right) is shown. This same sequence was detected for all EPHA3d45 PCR products for which sufficient PCR product could be purified for sequencing.

Knockdown of EPHA3d45 in TOV-112D

We designed siRNA targeting the junction of exons 5 and 4 in EPHA3d45. Transfection with this siRNA knocked down the expression of EPHA3d45 in TOV-112D cells, but had no observable effect on the wild-type transcript (Figure 6A). We subjected EPHA3d45-knockdown and control (transfected with scrambled siRNA) TOV-112D cells to clonogenic and MTS assays. EPHA3d45-knockdown TOV-112D cells formed fewer colonies than control cells and occupied less of the well area ($p<.001$ by unpaired T test; Figure 6B). EPHA3d45-knockdown TOV-112D cells also showed diminished metabolic activity over time as assessed by MTS assay, which we interpret to be indicative of diminished cell growth. Importantly, the EPHA3d45-negative cell line OVKATE did not show the same trend, supporting the specificity of this knockdown against the intended target (Figure 6C). These results suggest that EPHA3d45 expression may potentially promote cancer cell growth in TOV112D cell line.

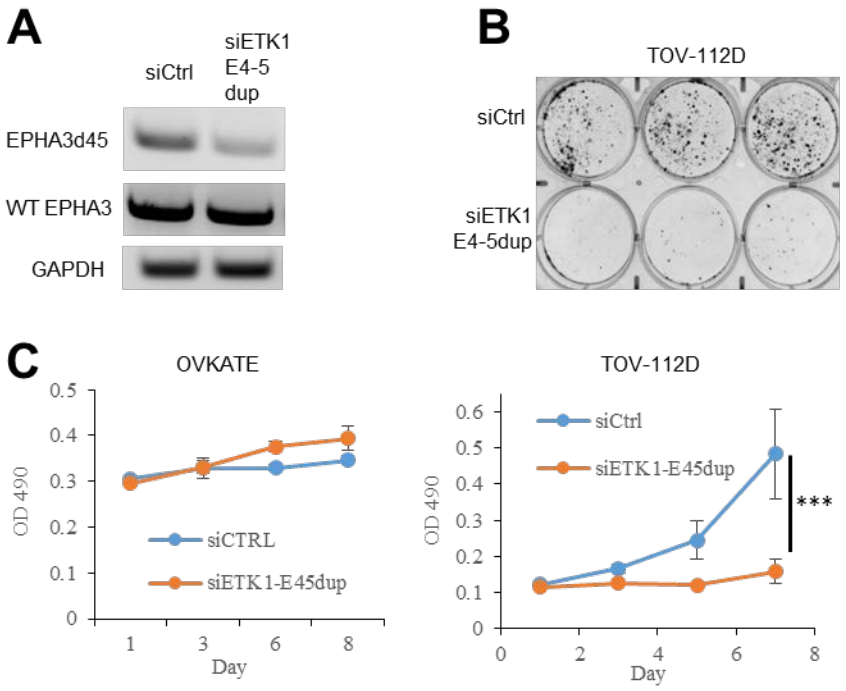


Figure 6: Knockdown of EPHA3d45 in TOV112D.

A. RT-PCR from TOV-112D cells transfected with scrambled siRNA (siCTRL) or siRNA specific to the junction of exon 5 and exon 4 in EPHA3d45 (siETK1-E4-5dup). B. Clonogenic assay of TOV-112D cells transfected as in A. C. MTS assay of the EPHA3d45-negative ovarian cell line OVKATE and EPHA3d45-positive ovarian cell line TOV-112D, transfected as in A. **p<.001 by unpaired T test. All results are representative of at least two independent experiments.

EPHA3d45 Transcript Expression in Human Samples

We performed RT-PCR on a subset of the primary and recurrent HGSC tumors for which RNA-seq data are available. EPHA3d45 bands were detected in most samples expressing the wild-type transcript (Figure 7A). Interestingly, the expression of EPHA3d45 appeared to correlate with expression of the wild-type. The EPHA3d45 bands were verified by capillary sequencing (Figure 5), though importantly this does not fully characterize the transcript structure, as discussed above. We then investigated the expression of EPHA3d45 in pooled normal human

ovary, uterus, and breast tissues from healthy Caucasian women (Clontech). RT-PCR detected EPHA3d45 in pooled normal human ovarian and uterine tissue (Figure 7B). To further assess the expression of this duplication in normal individuals, we analyzed the RNA-seq data for 87 normal ovary samples and 6 Fallopian tube samples of healthy individuals from the Genotype-Tissue Expression Project (GTEX) [41]. This analysis however, did not detect EPHA3d45 reads in these samples. Nonetheless, we cannot exclude the possibility that this dataset may not be sensitive enough to detect relative lower level expression of EPHA3d45 or that the cohort may not be comprehensive enough to capture germline expression.

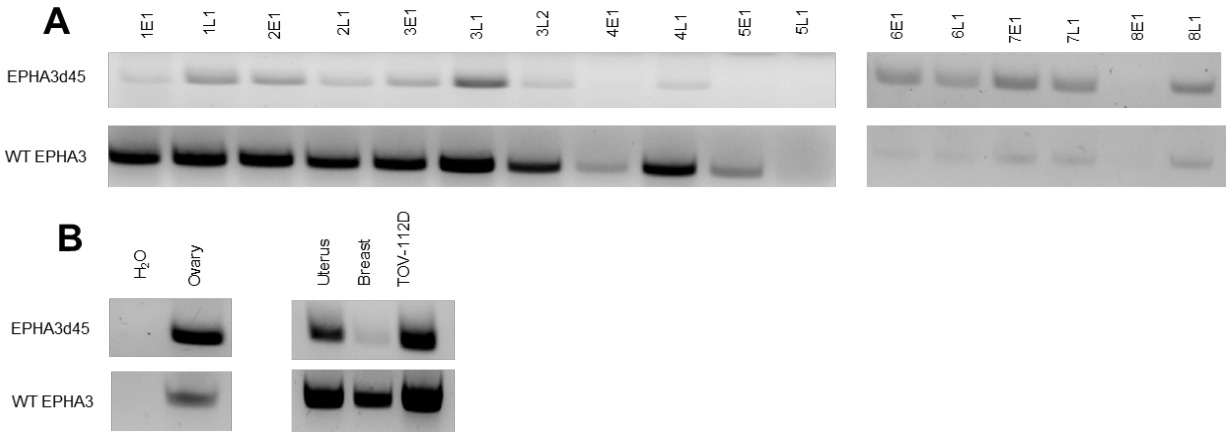


Figure 7: RT-PCR of EPHA3d45 in human tissue.

A. RT-PCR of EPHA3d45 and wild-type EPHA3 in early and recurrent HGSC tumor samples. Tumors are annotated as patient number, stage (E=early/primary disease; L=late/recurrent disease), sample number. Samples from patients 1-5 and from patients 6-8 were analyzed in separate experiments. Due to limited tumor RNA availability, results shown are representative of only one experiment. B. RT-PCR of EPHA3d45 and wild-type EPHA3 using RNA from pooled normal human tissue or water. The two panels are representative of separate experiments, and results for normal tissue are representative of at least two independent PCR experiments from the same batch of cDNA products.

4.0 DISCUSSION

Intragenic rearrangements represent an understudied area of cancer genetics because they have been overlooked by the field at large. Our initial systematic investigation into IGRs in HGSC based on copy number data revealed many potential IGRs in HGSC (Figure 1). It is noteworthy that this analysis was based on copy number data, which has certain limitations in detecting IGRs. First, balanced IGRs cannot be detected by copy number analysis because there is no gain or loss in the total amount of genetic material. Second, IGRs represent a class of cryptic genomic rearrangements; thus, the detection of unbalanced IGRs can be greatly limited by the resolution and density of the genomic profiling array. Third, a gain/loss of a small region of a gene in the copy number dataset does not necessarily indicate a contiguous deletion or duplication in that gene. It is our hope that these results will be corroborated in the future using other types of datasets, such as high-resolution whole-genome sequencing data.

Here, we report the nomination of a potential intragenic duplication of exons 4-5 of the gene *EPHA3*, termed EPHA3d45, which we detected by copy number and RNA-seq datasets (Figure 1-2). Although the transcript of this duplication remains to be completely characterized (see Results), this duplication is predicted to generate an additional FNIII domain in the kinase. This domain is known to bind to the ligand-binding domain of EPHA3 and other ephrin receptors [38]. Interestingly, the effect of this would have on protein function is disputed: On the one hand, binding of a non-ligand to a kinase receptor's ligand-binding domain would intuitively be expected to block ligand-mediated signaling. On the other hand, it has been proposed that this type of interaction could bring ephrin receptors physically closer on the cell membrane, leading to oligomerization and, therefore, activation [38, 39]. Given what is known about the FNIII domain

in ephrin receptors, and that EPHA3 itself has been reported as an oncogene in some cancer types, such as brain and blood cancers [33, 34], and a tumor suppressor in others, such as lung cancer [40], EPHA3d45 could conceivably promote cell growth by suppressing or activating EPHA3 signaling. Whether and how EPHA3d45 contributes to cancer growth will require detailed future studies to elucidate.

Our RT-PCR data show that strong expression of EPHA3d45 is detected in the positive cell lines detected by RNA-seq (Figure 2). We further showed that specific siRNA knockdown of EPHA3d45 inhibits cell growth (Figure 3). It is necessary to note that our knockdown experiments used a single siRNA. This siRNA was designed to specifically target the duplication junction of EPHA3d45 with minimal effect on wild-type EPHA3, but this design strategy necessitates using a very limited sequence from the duplication junction for siRNA design. Although BLAST against human reference RNA sequences did not reveal potential off-targets of the siRNA sequence we used, it is still conceivable that the phenotypes observed in our siRNA-mediated knockdown experiments could be due to an off-target effect against an unknown transcript. Further experiments, such as rescuing the siRNA effects through ectopic overexpression of EPHA3d45 or engineering the EPHA3d45 into duplication-negative cell lines, will be required to further assess and verify the oncogenic effects of EPHA3d45.

In RT-PCR of human tumor and normal tissues, PCR bands representing potential EPHA3d45 transcripts were detected in many samples with appreciable expression of the wild-type gene. This raises some questions about the origin of EPHA3d45 transcript expression. It is necessary to note first and foremost that this experiment was only performed once per tumor sample due to the limited availability of the materials, so the results of tumor samples should be viewed with appropriate skepticism. Additionally, although our RT-PCR results for pooled normal

RNA were repeatable, EPHA3d45 transcripts were not detected by RNA-seq data for 87 normal ovary samples and 6 Fallopian tube samples of healthy individuals from the Genotype-Tissue Expression Project [41]; however, we cannot exclude the possibility that this dataset may not be sensitive or comprehensive enough to detect expression of EPHA3d45. There are several likely explanations for these ambiguous results:

1. The weak ubiquitous bands observed could be experimental artifacts created during the reverse transcription and/or PCR process or simply PCR product contamination. Notably, RT for cell lines was performed with the Roche Transcriptor First Strand cDNA Synthesis kit, whereas RT for tissues was performed using the Invitrogen Superscript IV kit (see Methods). The former is a recombinant reverse transcriptase expressed in E. coli with RNase H activity, whereas the latter is a more efficient MMLV Reverse Transcriptase which is known to cause PCR artifacts [42]. In addition, PCR contamination is a major issue for this type of experiments, and while we have strived to avoid such contamination, we cannot definitively rule it out as a possible cause of the ubiquitous weak bands observed in HGSC tumors. Further, the end-point PCR at 35-37 cycles is not ideal for detecting differences in transcript expression as the PCR will exceed the linear amplification phase. Finally, these experiments are not quantitative, and as such it can be difficult to distinguish between strong and weak expression. In particular, the human normal tissues were transcribed with a highly efficient reverse transcriptase Super Script IV and used 5 times as much RNA as was used for paired HGSC tumor tissues, and the PCR was done with 37 cycles, which makes it impossible to directly compare the expression levels of EPHA3d45 in normal tissues and HGSC tumors. Thus well-designed quantitative RT-PCR will be required to assess the expression levels of the EPHA3d45 transcript in tumor and normal tissues.

2. EPHA3d45 can present as a true germline genomic event, as discussed above, which could present in the individuals that contributed to the pooled normal tissues. It is possible that the pooled normal tissue samples we analyzed could coincidentally include individual(s) that harbor this germline event, which could explain the detection of this duplication transcript from normal tissue RNA. While copy number data suggest this as a rare germline event, we cannot exclude the possibility that the true frequency of this germline duplication could be higher than it was detected by copy number datasets, which cannot detect balanced rearrangements, as discussed above. Future studies of paired HGSC tumors and contralateral fallopian tube tissues will be helpful to assess the somatic or germline nature of EPHA3d45. EPHA3d45 as a true genomic event is also supported by the detection of the duplication in the cell line HCC38 by a whole-genome sequencing study [36]; the duplication in this cell line is also positive by copy number. On the other hand, neither TOV-112D nor NCIH446, the cell lines positive for EPHA3d45 transcripts by RNA-seq and RT-PCR (Figure 3B and Figure 4) were positive by copy number; however, it is conceivable that the duplication could be the result of balanced rearrangements, as discussed above. In addition, cell lines that are positive for EPHA3d45 may not express wild-type EPHA3 and EPHA3d45 due to the lack of an active promoter. These complex aspects can result in inconsistency between different levels of data. Further studies such as genomic PCR will be needed to determine whether EPHA3d45 transcript-expressing cell lines harbor genomic rearrangements.

3. EPHA3d45 transcripts may be generated at the RNA level by physically regulated *trans*-splicing or back-splicing splicing events [43], which could lead to the expression of the EPHA3d45 transcript in normal tissues. Intragenic duplications generated by splicing of this nature have been previously reported [44-46]. However, the presence of EPHA3d45 as a splicing event does not necessarily preclude this aberration also existing as an oncogenic genomic event.

Many known oncogenic fusions, including BCR-ABL1, JAZF1-JJAZ1, and EML4-ALK, have been reported to be expressed at the transcript level in normal, healthy tissue [47-49]. For example, the EML4-ALK gene fusion has been matched with an effective oral drug with stunning clinical response; however, significant expression of this fusion has been detected in normal lung tissues by RT-PCR, which has led to a debate about the validity of this fusion as a driver event in lung cancer [48]. It has also been reported that JAZF1-JJAZ1, a neoplastic gene fusion in human endometrial stromal sarcoma, can be generated by physiologically regulated trans-splicing of RNAs in normal endometrium cells during the menstrual cycle [47]. The generally accepted view in the field is that such aberrant transcripts can be generated by trans-splicing in healthy tissues and their expression can be physically regulated, which can be 'locked in' by genomic rearrangements in cancer, ultimately leading to constitutive overexpression of the fusion transcripts. It is certainly possible, even probable, that a similar genomic process in HGSC may lead to constitutive overexpression of EPHA3d45 transcripts. The precise molecular underpinnings which lead to the expression of the EPHA3d45 transcript will need to be further studied to clarify whether the results presented here are caused by genomic rearrangements, splicing, or some combination thereof.

Regardless of the mechanism by which it is expressed, we have presented preliminary evidence that the expression of EPHA3d45 may potentially promote growth in ovarian cancer cells. The oncogenic potential of EPHA3d45, including confirmation of protein-level effects of the duplication, will call for future detailed studies. Thorough characterization of EPHA3d45, as well as other IGRs and even other understudied classes of genomic structural variants, may shed light on the molecular aberrations underlying HGSC initiation, progression, or recurrence.

WORKS CITED

1. Koretzky, G.A., *The legacy of the Philadelphia chromosome.* J Clin Invest, 2007. **117**(8): p. 2030-2.
2. Veeraraghavan, J., et al., *Recurrent ESR1-CCDC170 rearrangements in an aggressive subset of oestrogen receptor-positive breast cancers.* Nat Commun, 2014. **5**: p. 4577.
3. Veeraraghavan, J., et al., *Recurrent and pathological gene fusions in breast cancer: current advances in genomic discovery and clinical implications.* Breast Cancer Res Treat, 2016. **158**(2): p. 219-32.
4. Mitelman, F., B. Johansson, and F. Mertens, *The impact of translocations and gene fusions on cancer causation.* Nat Rev Cancer, 2007. **7**(4): p. 233-45.
5. Mertens, F., et al., *The emerging complexity of gene fusions in cancer.* Nat Rev Cancer, 2015. **15**(6): p. 371-81.
6. Wong, A.J., et al., *Structural alterations of the epidermal growth factor receptor gene in human gliomas.* Proc Natl Acad Sci U S A, 1992. **89**(7): p. 2965-9.
7. Fenstermaker, R.A. and M.J. Ciesielski, *Deletion and tandem duplication of exons 2 - 7 in the epidermal growth factor receptor gene of a human malignant glioma.* Oncogene, 2000. **19**(39): p. 4542-8.
8. Gallant, J.N., et al., *EGFR Kinase Domain Duplication (EGFR-KDD) Is a Novel Oncogenic Driver in Lung Cancer That Is Clinically Responsive to Afatinib.* Cancer Discov, 2015. **5**(11): p. 1155-63.
9. Castiglioni, F., et al., *Role of exon-16-deleted HER2 in breast carcinomas.* Endocr Relat Cancer, 2006. **13**(1): p. 221-32.
10. Wang, Y., et al., *Dystrophin is a tumor suppressor in human cancers with myogenic programs.* Nat Genet, 2014. **46**(6): p. 601-6.
11. The, B.E.D.S.G., *The exon 13 duplication in the BRCA1 gene is a founder mutation present in geographically diverse populations. The BRCA1 Exon 13 Duplication Screening Group.* Am J Hum Genet, 2000. **67**(1): p. 207-12.
12. *Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group.* Br J Cancer, 2000. **83**(10): p. 1301-8.
13. Risch, H.A., et al., *Prevalence and penetrance of germline BRCA1 and BRCA2 mutations in a population series of 649 women with ovarian cancer.* Am J Hum Genet, 2001. **68**(3): p. 700-10.
14. Gad, S., et al., *Significant contribution of large BRCA1 gene rearrangements in 120 French breast and ovarian cancer families.* Oncogene, 2002. **21**(44): p. 6841-7.
15. Walsh, T., et al., *Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer.* JAMA, 2006. **295**(12): p. 1379-88.
16. Sangha, N., et al., *Neurofibromin 1 (NF1) defects are common in human ovarian serous carcinomas and co-occur with TP53 mutations.* Neoplasia, 2008. **10**(12): p. 1362-72, following 1372.
17. Tedaldi, G., et al., *First evidence of a large CHEK2 duplication involved in cancer predisposition in an Italian family with hereditary breast cancer.* BMC Cancer, 2014. **14**: p. 478.
18. Torres, D., et al., *Prevalence and Penetrance of BRCA1 and BRCA2 Germline Mutations in Colombian Breast Cancer Patients.* Sci Rep, 2017. **7**(1): p. 4713.
19. Bowtell, D.D., et al., *Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer.* Nat Rev Cancer, 2015. **15**(11): p. 668-79.

20. Cancer Genome Atlas Research, N., *Integrated genomic analyses of ovarian carcinoma.* Nature, 2011. **474**(7353): p. 609-15.
21. Wang, Z.C., et al., *Profiles of genomic instability in high-grade serous ovarian cancer predict treatment outcome.* Clin Cancer Res, 2012. **18**(20): p. 5806-15.
22. Kannan, K., et al., *CDKN2D-WDFY2 is a cancer-specific fusion gene recurrent in high-grade serous ovarian carcinoma.* PLoS Genet, 2014. **10**(3): p. e1004216.
23. Kannan, K., et al., *Recurrent BCAM-AKT2 fusion gene leads to a constitutively activated AKT2 fusion kinase in high-grade serous ovarian carcinoma.* Proc Natl Acad Sci U S A, 2015. **112**(11): p. E1272-7.
24. Earp, M.A., et al., *Characterization of fusion genes in common and rare epithelial ovarian cancer histologic subtypes.* Oncotarget, 2017. **8**(29): p. 46891-46899.
25. Salzman, J., et al., *ESRRA-C11orf20 is a recurrent gene fusion in serous ovarian carcinoma.* PLoS Biol, 2011. **9**(9): p. e1001156.
26. Kumar-Sinha, C., S. Kalyana-Sundaram, and A.M. Chinnaiyan, *Landscape of gene fusions in epithelial cancers: seq and ye shall find.* Genome Med, 2015. **7**: p. 129.
27. Mittal, V.K. and J.F. McDonald, *Integrated sequence and expression analysis of ovarian cancer structural variants underscores the importance of gene fusion regulation.* BMC Med Genomics, 2015. **8**: p. 40.
28. Cancer Genome Atlas Research, N., *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.
29. Cooper, G.M., et al., *A copy number variation morbidity map of developmental delay.* Nat Genet, 2011. **43**(9): p. 838-46.
30. Zack, T.I., et al., *Pan-cancer patterns of somatic copy number alteration.* Nat Genet, 2013. **45**(10): p. 1134-40.
31. Kim, D. and S.L. Salzberg, *TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.* Genome Biol, 2011. **12**(8): p. R72.
32. Kim, J.A., et al., *Comprehensive functional analysis of the tousled-like kinase 2 frequently amplified in aggressive luminal breast cancers.* Nat Commun, 2016. **7**: p. 12991.
33. Keane, N., et al., *EPHA3 as a novel therapeutic target in the hematological malignancies.* Expert Rev Hematol, 2012. **5**(3): p. 325-40.
34. Day, B.W., et al., *EphA3 maintains tumorigenicity and is a therapeutic target in glioblastoma multiforme.* Cancer Cell, 2013. **23**(2): p. 238-48.
35. Coe, B.P., et al., *Refining analyses of copy number variation identifies specific genes associated with developmental delay.* Nat Genet, 2014. **46**(10): p. 1063-71.
36. Stephens, P.J., et al., *Complex landscapes of somatic rearrangement in human breast cancer genomes.* Nature, 2009. **462**(7276): p. 1005-10.
37. Liu, Q., et al., *Analyses of domains and domain fusions in human proto-oncogenes.* BMC Bioinformatics, 2009. **10**: p. 88.
38. Lisabeth, E.M., G. Falivelli, and E.B. Pasquale, *Eph receptor signaling and ephrins.* Cold Spring Harb Perspect Biol, 2013. **5**(9).
39. Choi, Y., et al., *Discovery and structural analysis of Eph receptor tyrosine kinase inhibitors.* Bioorg Med Chem Lett, 2009. **19**(15): p. 4467-70.
40. Zhuang, G., et al., *Effects of cancer-associated EPHA3 mutations on lung cancer.* J Natl Cancer Inst, 2012. **104**(15): p. 1182-97.
41. Carithers, L.J., et al., *A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project.* Biopreserv Biobank, 2015. **13**(5): p. 311-9.
42. Cocquet, J., et al., *Reverse transcriptase template switching and false alternative transcripts.* Genomics, 2006. **88**(1): p. 127-31.

43. Jeck, W.R. and N.E. Sharpless, *Detecting and characterizing circular RNAs.* Nat Biotechnol, 2014. **32**(5): p. 453-61.
44. Caudevilla, C., et al., *Natural trans-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver.* Proc Natl Acad Sci U S A, 1998. **95**(21): p. 12185-90.
45. Frantz, S.A., et al., *Exon repetition in mRNA.* Proc Natl Acad Sci U S A, 1999. **96**(10): p. 5400-5.
46. Rigatti, R., et al., *Exon repetition: a major pathway for processing mRNA of some genes is allele-specific.* Nucleic Acids Res, 2004. **32**(2): p. 441-6.
47. Li, H., et al., *A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells.* Science, 2008. **321**(5894): p. 1357-61.
48. Martelli, M.P., et al., *EML4-ALK rearrangement in non-small cell lung cancer and non-tumor lung tissues.* Am J Pathol, 2009. **174**(2): p. 661-70.
49. Ismail, S.I., et al., *Incidence of bcrabl fusion transcripts in healthy individuals.* Mol Med Rep, 2014. **9**(4): p. 1271-6.