Recurrent structural variations of lncRNA gene *CCDC26* in diffuse intrinsic pontine glioma

Lihua Zou (lihuazou@gmail.com)

Northwestern University

Chicago, IL, USA

Abstract

We report recurrent somatic structural variations (SVs) involving long noncoding RNA (lncRNA) *CCDC26* in 13% of Diffuse Intrinsic Pontine Glioma (DIPG) patients. We validate our findings using whole genome sequencing data from two independent patient cohorts. *CCDC26* SVs cause increased expression of *CCDC26* gene in patients. In addition, *CCDC26* expression is associated with elevated expression of *MYC* and proliferation signature. Our findings identify *CCDC26* as a novel significantly mutated gene in DIPG and highlight the importance of structural variations in pediatric brain cancer.

Main

Diffuse intrinsic pontine glioma (DIPG), the most frequent brainstem tumor in pediatric patients, is one of the most devastating childhood cancers, and virtually all DIPG patients die within two years after diagnosis. Current standard of care, chemotherapy followed by radiation, yields no improvements in survival. There is an unmet need for the identification of molecular mechanisms and efficacious therapeutic agents to improve treatment outcomes for DIPG patients. The discovery of somatic histone gene mutations, resulting in replacement of lysine 27 by methionine (K27M) in the encoded histone H3 proteins, in DIPG has dramatically improved our understanding of disease pathogenesis and stimulated the development of novel therapeutic approaches for the treatment of DIPG[1,2]. Past sequencing analyses of DIPG were largely focused on somatic point mutation or chromosome copy

number alteration[3]. Structural variation (SV) is another class of mutations that can lead to duplication, deletion or reordering of DNA at scales ranging from single genes to entire chromosomes. The role of SVs in DIPG is poorly understood. To address this issue, we analyzed the whole genome sequences of matched tumor and normal pairs from 60 DIPG patients. These patients were from two independent cohorts: one from the CBTTC OpenDIPG project (*n* = 45) and a second from PNOS project (*n* = 15) (**Methods**). These data enable a gene-centric approach to detect SVs in DIPG.

Strikingly, we discovered recurrent SV mutations at lncRNA *CCDC26* in 13% of DIPG patients (8 out of 60) from the combined cohorts (**Fig.1a**). The role of *CCDC26* in cancer is not well understood. It was implicated in childhood acute myeloid leukemia (AML) because of altered chromosome copy numbers in AML patients[4]. We examined 8 patients with recurrent *CCDC26* SVs sample by sample. We found 6 out of 8 patients have *CCDC26* amplified through tandem duplication. To delineate critical functional region of *CCDC26*, we overlaid sequences altered by SVs at *CCDC26* locus and pinpointed a common 140kb amplicon at chr8:129,523,594-129,662,989 on hg38 reference genome (**Fig.1b**). Notably, the common amplicon is next to a germline SNP (rs4295627) associated with 1.3-fold increased risk in glioma development[5]. We observed three samples have tandem duplication breakpoints intersecting with the exon region of two neighboring genes *GSDMC* and *FAM49B* (**Fig.1b**). We searched for gene fusions involving *CCDC26* and *GSDMC* or *FAM49B* using RNA-Seq from matched samples. In patient BS_1Q524P3B, we observed a gene fusion joining transcripts from *CCDC26* exon 1 and *GSDMC* exon 7-14 (**Fig.1c**). *GSDMC* encodes Gasdermin-C, a protein coding gene which may be acting by homooligomerizing within the membrane and forming pores. We observed two patients (BS_CBMAWSAR and BS_FKQ7F6D1) displaying highly amplified DNA segments involving many breakpoints (>=100) proximal to *CCDC26*. The high copy numbers changes linked across multiple distant chromosome segments suggest ecDNA (e.g. double minute, neochromosome) as underlying structure (**Fig.1e; Fig.2c**)[6–8]. ecDNA has been reported as a

mechanism which can lead to amplification of driver oncogene under selection pressure[8]. Consistent with this, we observed co-amplified breakpoints involving *CCDC26* on chromosome 8 and *EGFR* on chromosome 7 In patient BS_CBMAWSAR (**Fig.1f**). We also observed co-amplified breakpoints involving *CCDC26* and *MYC* on chromosome 8 in patient BS_FKQ7F6D1 (**Fig.2d**).

To study the functional impact of *CCDC26* SVs, we compared *CCDC26* expression between the samples in the presence and absence of SVs (*CCDC26*-SV vs. *CCDC26*-WT) using RNA-Seq from matched samples. *CCDC26*-SV samples have significantly higher *CCDC*26 expression than *CCDC26*-WT samples (**Fig.2a**; t-test $p<0.05$) indicating a functional role of *CCDC26* in DIPG. To identify differential genes and pathways between *CCDC26*-SV and *CCDC26*-WT, we performed differential gene expression analysis using DESeq2[9] and found *MYC* up-regulated in *CCDC26*-SV samples ($p<0.005$). We performed Gene Set Enrichment Analysis (GSEA)[10] and found proliferation signatures and MYC targets enriched in *CCDC26*-SV samples (FDR $q<0.1$). In addition, *CCDC26* expression is correlated with *MYC* expression in our cohorts (**Fig.2b;** Pearson correlation $p<0.005$).

Taken together, our findings identify *CCDC26* as a frequently mutated lncRNA gene in DIPG. Expression analysis suggests *CCDC26* is associated with *MYC* and proliferation pathways.The detailed molecular mechanism of *CCDC26* remains to be elucidated. Our findings highlight the importance of studying genome structural rearrangements in this deadly disease.

Methods

Cohort description

The 60 DIPG specimens used in our study are composed of radiologically diagnosed DIPG from Children's Brain Tumor Tissue Consortium (CBTTC) and the Pediatric Pacific Neuro-oncology Consortium (PNOC). The raw whole genome sequencing and RNA-seq data can be downloaded from the Gabriella Miller Kids First Data Resource Center (KF-DRC). The CBTTC is a collaborative, multi-institutional research program dedicated to the study of childhood brain tumors. The Pacific Pediatric Neuro-Oncology Consortium (PNOC) is an international consortium dedicated to bringing new therapies to children and young adults with brain tumors. PNOC collected blood and tumor biospecimens from newly diagnosed DIPG patients as part of the clinical trial PNOC003/NCT02274987.

Whole-genome sequencing analysis

Paired-end DNA-Seq reads were aligned to hg38 (patch release 12) reference genome using BWA-MEM[11]. Duplicates were marked using Samblaster[12]. BAMs were merged and processed using Broad's Genome Analysis Toolkit (GATK)[13]. For WGS variant calling, Strelka2[14] was used to call Indels and Mutect2[15] was used to call SNVs using default parameters. The final Strelka2 and Mutect2 VCFs were filtered for PASS variants for downstream analysis. For structural variant (SV) calls, Manta[16] was used using hg38 as reference genome. Manta SV output was annotated using AnnotSV[17]. The docker image of Whole Genome Sequence Analysis Workflow can be found in the KidsFirst GitHub repository.

Gene expression analysis

Paired-end RNA-Seq reads were aligned using ENSEMBL's GENCODE 27 as the reference genome. Transcript- and gene-level expression values were calculated using RSEM[18]. Data normalization and

differential gene expression were done by DESeq2[9]. Gene set enrichment analysis (GSEA)[10] was used to find groups of enriched genes between different groups of samples.

Gene fusion analysis

Gene fusions were called using Arriba[19]. Gene fusion calls from Arriba were annotated using FusionAnnotator (https://github.com/FusionAnnotator) followed by filtering of recurrent fusion artifacts and transcripts present in normal tissue using a blacklist file bundled with Arriba.

Figures and Legends

Figure 1. Recurrent structural variants of CCDC26 in DIPG a) Gene-centric table showing frequency of *CCDC26* SVs along with previously reported driver mutations in DIPG in the combined cohorts; b) Common *CCDC26* amplicon (indicated by grey bar in the top panel) altered by SVs; supporting reads at breakpoints of *CCDC26* and *GSDMC* are shown in the bottom panel; c) SV and copy number changes associated with *CCDC26-GSDMC* in patient BS_1Q524P3B; d) Structure of gene fusion *CCDC26-GSDMC* identified in RNA-Seq of patient BS_1Q524P3B; e) Circos plots showing clustering of breakpoints at chromosome 7 and 8 in patient BS_CBMAWSAR; f) SVs (top panel) and associated high copy number changes (bottom panel) on chromosome 7 and 8 in patient BS_CBMAWSAR. Colored curves in the top panel encode different types of SVs (*red*: tandem duplication; *blue*: deletion; *green*: 5'Inversion; *orange*: 3'Inversion; *purple*: translocation).

Figure 2. Impact of CCDC26 SVs on gene expression. a) Expression of *CCDC26* in *CCDC26*-SV vs. *CCDC26*-WT; *y*-axis indicate log2-transformed RSEM gene expression value measured by RNA-Seq; b) Correlation between *CCDC26* and *MYC* expression (Pearson correlation; $p<0.005$); *x*-axis and *y*-axis indicate log2-transformed RSEM gene expression value measured by RNA-Seq; c) Circos plots showing clustering of breakpoints on chromosome 8 in patient BS_FKQ7FD1; d) Detailed

view of SV breakpoints and high copy number changes involving *CCDC26* and *MYC* on chromosome 8 in patient BS_FKQ7FD1.

References

1. Schwartzentruber, J. *et al.* Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature* **482**, 226–231 (2012).
2. Wu, G. *et al.* Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. *Nat. Genet.* **44**, 251–253 (2012).
3. Mackay, A. *et al.* Integrated Molecular Meta-Analysis of 1,000 Pediatric High-Grade and Diffuse Intrinsic Pontine Glioma. *Cancer Cell* **32**, 520–537.e5 (2017).
4. Storlazzi, C. T. *et al.* Identification of a commonly amplified 4.3 Mb region with overexpression of C8FW, but not MYC in MYC-containing double minutes in myeloid malignancies. *Hum. Mol. Genet.* **13**, 1479–1485 (2004).
5. Shete, S. *et al.* Genome-wide association study identifies five susceptibility loci for glioma. *Nat. Genet.* **41**, 899–904 (2009).
6. Zhang, C.-Z. *et al.* Chromothripsis from DNA damage in micronuclei. *Nature* **522**, 179–184 (2015).
7. Turner, K. M. *et al.* Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* vol. 543 122–125 (2017).
8. Garsed, D. W. *et al.* The architecture and evolution of cancer neochromosomes. *Cancer Cell* **26**, 653–667 (2014).
9. Love, M. I. *et al.* Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
10. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for

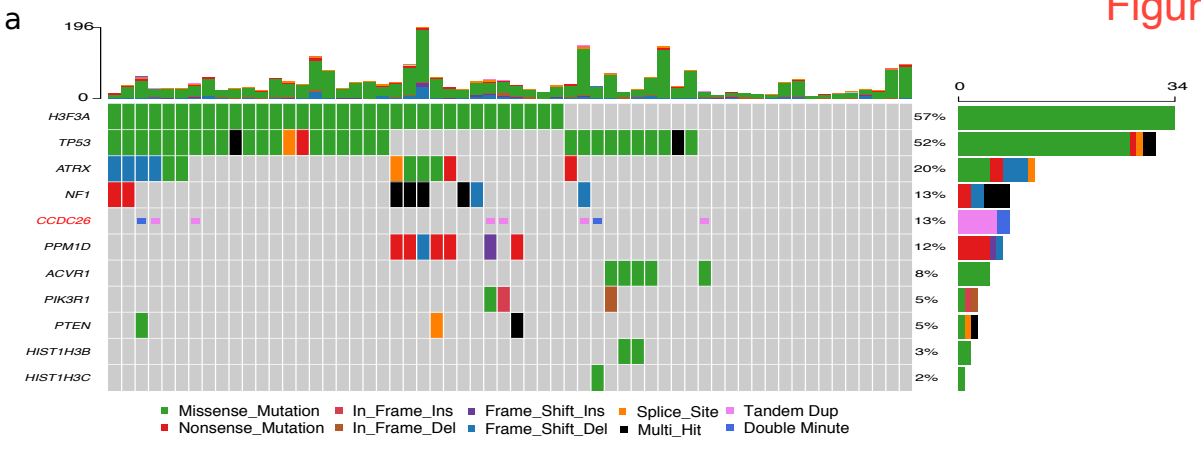
interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

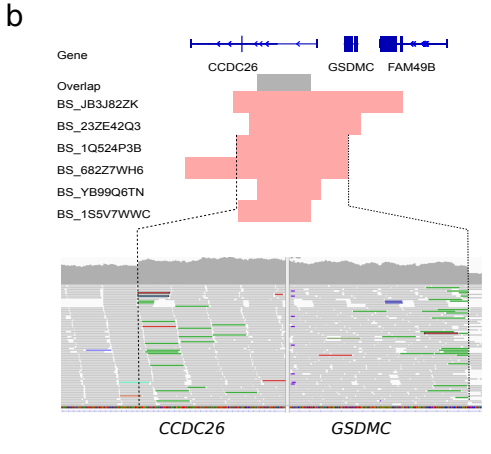
11. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
12. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
13. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
14. Saunders, C. T., Wong, W. S. W., Swamy, S. & Becq, J. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. (2012).
15. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
16. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
17. Geoffroy, V. *et al.* AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572–3574 (2018).
18. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
19. Uhrig, S., Fröhlich, M., Hutter, B. & Brors, B. PO-400 Arriba – fast and accurate gene fusion detection from rna-seq data. *Epigenetic Mechanisms* (2018) doi:10.1136/esmoopen-2018-eacr25.427.

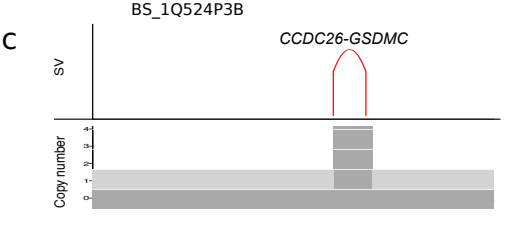
Acknowledgements

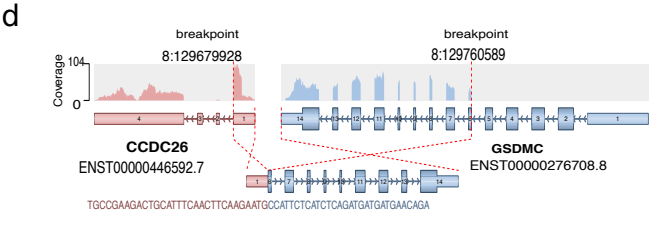
We thank members of Children's Brain Tumor Tissue Consortium (CBTTC) (www.cbttc.org) for their support of open access, biospecimen driven research.

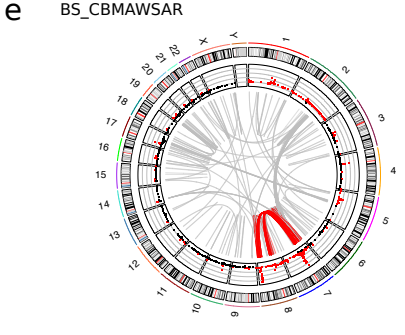
Figure 1


a
196
0
0
34
H3F3A
TP53
ATRX
NF1
CCDC26
PPM1D
ACVR1
PIK3R1
PTEN
HIST1H3B
HIST1H3C
57%
52%
20%
13%
13%
12%
8%
5%
5%
3%
2%
Missense_Mutation
Nonsense_Mutation
In_Frame_Ins
In_Frame_Del
Frame_Shift_Ins
Frame_Shift_Del
Splice_Site
Multi_Hit
Tandem Dup
Double Minute


b
Gene
CCDC26
GSDMC
FAM49B
Overlap
BS_JB3J82ZK
BS_23ZE42Q3
BS_1Q524P3B
BS_682Z7WH6
BS_YB99Q6TN
BS_1S5V7WWC
CCDC26
GSDMC


c
BS_1Q524P3B
CCDC26-GSDMC
SV
Copy number


d
breakpoint
8:129679928
breakpoint
8:129760589
Coverage
104
0
CCDC26
ENST00000446592.7
GSDMC
ENST00000276708.8
TGCCGAAGACTGCATTTCAACTTCAAGAATGCCATTCTCATCTCAGATGATGATGAACAGA


e
BS_CBMAWSAR

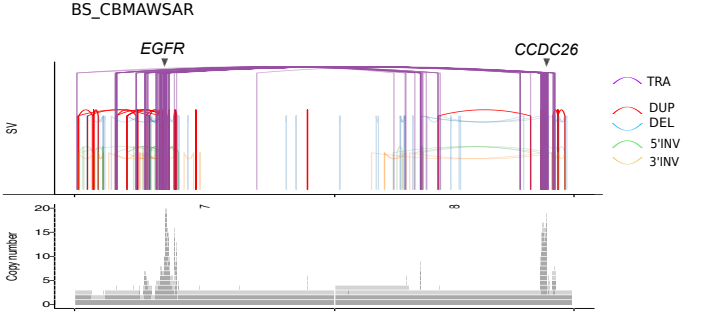
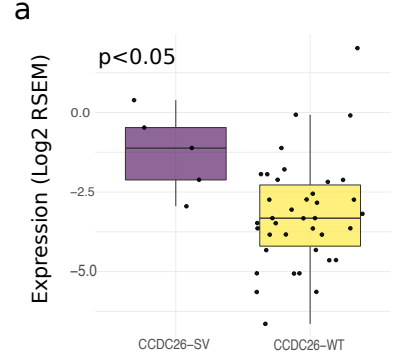
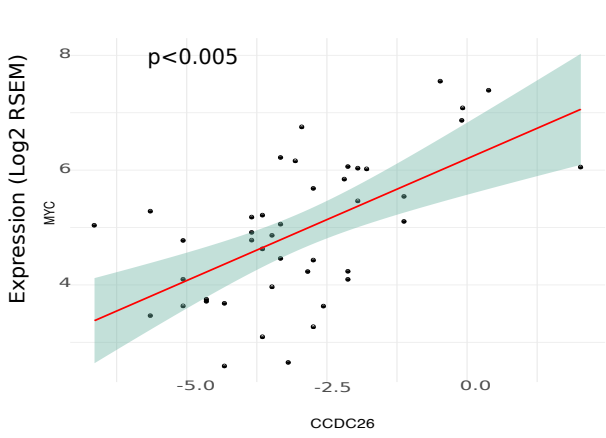
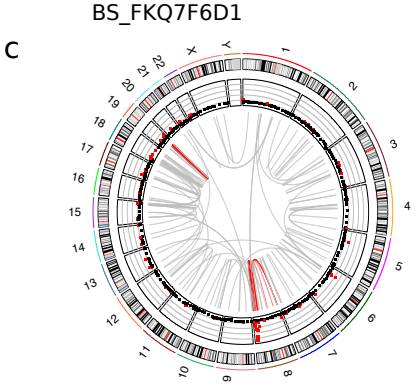
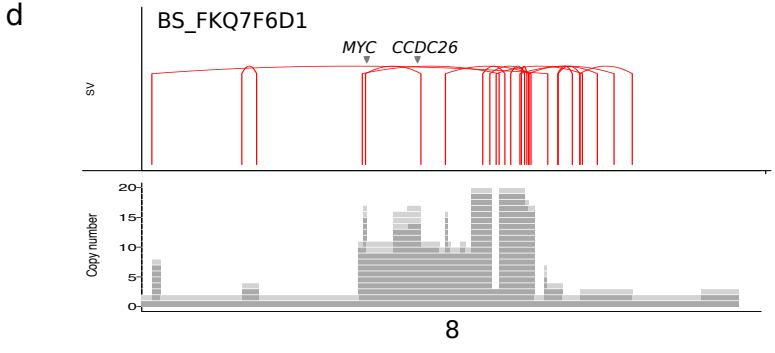

f
BS_CBMAWSAR
EGFR
CCDC26
SV
Copy number
TRA
DUP
DEL
5'INV
3'INV

Figure 2


a
p<0.05
Expression (Log2 RSEM)
0.0
−2.5
−5.0
CCDC26−SV
CCDC26−WT


b
p<0.005
Expression (Log2 RSEM)
MYC
8
6
4
-5.0
-2.5
0.0
CCDC26


BS_FKQ7F6D1
c


d
BS_FKQ7F6D1
MYC
CCDC26
SV
Copy number
20
15
10
5
0
8