Structural variation targets neurodevelopmental genes and identifies *SHANK2* as a tumor suppressor in neuroblastoma

Gonzalo Lopez[1,2+], Karina L. Conkrite[1,2+], Miriam Doepner[1,2], Komal S. Rathi[3], Apexa Modi[1,2,10], Zalman Vaksman[1-3], Lance M. Farra[1,2], Eric Hyson[1,2], Moataz Noureddine[1,2], Jun S. Wei[4], Malcolm A. Smith[9], Shahab Asgharzadeh[5,6], Robert C. Seeger[5,6], Javed Khan[4], Jaime Guidry Auvil[8], Daniela S. Gerhard[8], John M. Maris[1-2,10,11], Sharon J. Diskin[1-3,10,11*]

[1]Division of Oncology, Children’s Hospital of Philadelphia, Philadelphia, PA, USA.

[2]Center for Childhood Cancer Research, Children’s Hospital of Philadelphia, Philadelphia, PA, USA.

[3]Department of Biomedical and Health Informatics, Children’s Hospital of Philadelphia, Philadelphia, PA, USA.

[4]Oncogenomics Section, Genetics Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA

[5]Division of Hematology, Oncology and Blood and Marrow Transplantation, Keck School of Medicine of the University of Southern California, Los Angeles, CA, USA.

[6]The Saban Research Institute, Children’s Hospital of Los Angeles, Los Angeles, CA, USA.

[7]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

[8]Office of Cancer Genomics, National Cancer Institute, Bethesda, MD, USA.

[9]Cancer Therapy Evaluation Program, National Cancer Institute, Bethesda, MD, USA.

[10]Genomics and Computational Biology, Biomedical Graduate Studies, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

[11]Abramson Family Cancer Research Institute, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA.

[+] Equal contribution.
[*] Corresponding Author: Sharon J. Diskin, Ph.D. (diskin@email.chop.edu)

Abstract

Neuroblastoma, like many childhood cancers, exhibits a relative paucity of somatic single nucleotide variants (SNVs). Here, we assess the contribution of structural variation (SV) in neuroblastoma using a combination of whole genome sequencing (WGS; n=135) and single nucleotide polymorphism (SNP) genotyping (n=914) of matched tumor-normal pairs. Our study design provided means for orthogonal validation of SVs as well as validation across genomic platforms. SV frequency, type, and localization varied significantly among high-risk tumors, with *MYCN* non-amplified tumors harboring an increased SV burden overall ($P=1.12x10^{-5}$). Genes disrupted by SV breakpoints were enriched in neuronal lineages and autism spectrum disorder. The postsynaptic adapter protein-coding gene *SHANK2*, located on chromosome 11q13, was disrupted by SVs in 14% and 10% of *MYCN* non-amplified high-risk tumors based on WGS and SNP array cohorts, respectively. Forced expression of *SHANK2* in neuroblastoma cell models resulted in significant growth inhibition ($P=2.62x10^{-2}$ to $3.4x10^{-5}$) and accelerated neuronal differentiation following treatment with all-trans retinoic acid ($P=3.08x10^{-13}$ to $2.38x10^{-30}$). These data further define the complex landscape of structural variation in neuroblastoma and suggest that events leading to deregulation of neurodevelopmental processes, such as inactivation of *SHANK2*, are key mediators of tumorigenesis.

Neuroblastoma is a cancer of the developing sympathetic nervous system that most commonly affects children under 5 years of age, with a median age at diagnosis of 17 months[1]. Approximately 50% of cases present with disseminated disease at the time of diagnosis, and despite intense multi-modal therapy, the survival rate for this high-risk subset remains less than 50%[1]. Recent whole genome and exome sequencing studies of neuroblastoma have revealed relatively few recurrent protein-coding somatic mutations including single nucleotide variations (SNVs) and small (<50b) insertion/deletions (indels)[2-5].

Large-scale structural variations (SVs) such as deletions, insertions, inversions, tandem duplications and translocations can arise from mutational processes that alter chromosome structure and evade innate mechanisms of maintaining genomic stability. These diverse SVs are commonly acquired somatically and act as driver mutations[6].

A plethora of approaches have been applied to detect SVs across large cancer datasets[6-9]. First, methods that identify copy number variations (CNVs) can be applied to intensity log R ratios from genotyping and comparative genomic hybridization (CGH) arrays as well as read-depth measures from next generation sequencing. Different segmentation algorithms have been applied to either platform in order to obtain copy number gain and loss calls, which range from a few hundred base-pair size to whole chromosomal alterations. These methods are dosage sensitive, allowing numerical quantification of amplifications and homozygous deletions. Analysis of CNVs in neuroblastoma primary tumor and matched blood samples has led to identification of recurrent somatically acquired DNA copy number alterations (SCNA), such as *MYCN* amplification, gain of chromosome 17q, and deletion of chromosomes 1p and 11q; these events are associated with an undifferentiated phenotype, aggressive disease, and poor survival to the disease[10-22]. In addition, focal deletions cause deleterious loss of function in the chromatin remodeler gene *ATRX*[23,24], implicated in the alternative lengthening of telomeres (ALT) phenotype, and other tumor suppressor genes such as *PTPRD*[25], *ARID1A* and *ARID1B*[26].

Second, NGS technologies have profoundly expanded our understanding of the impact of SVs in cancer[7]. DNA sequencing methods focus on discordantly aligned reads and read-pairs to the reference genome. As such, alignment based approaches do not rely on dosage quantification and cannot quantify numerical changes of deletions and tandem-duplications; however, they provide information about inversions, translocations and transposable elements, which are elusive for CNV callers. In addition, alignment based approaches offer single base pair resolution and genome-wide coverage in the case of WGS. Recent studies using alignment based detection of SVs from WGS profiles from primary neuroblastomas revealed structural rearrangements as key oncogenic drivers mediating focal enhancer amplification or enhancer hijacking, influencing telomere maintenance through activation of telomerase reverse transcriptase gene (*TERT*)[24,27,28] or by deregulating the *MYC* oncogene[29]. Despite the demonstrated importance of somatic CNVs and other SVs in neuroblastoma, studies systematically integrating CNV analysis and alignment based approaches are lacking; hence the global landscape and mechanisms of pathogenicity of many of these events remain poorly understood.

Here, we studied the role of somatic SVs in the largest available dataset to date, including 997 distinct primary neuroblastoma tumors and integrating whole genome sequencing (WGS) from 135 tumor-normal pairs and 914 single nucleotide polymorphism (SNP) arrays obtained at diagnosis. We considered alternative approaches for SV detection from both datasets, which overlap in a subset of 52 cases. As such, our study provides orthogonal as well as cross-platform validation of SV breakpoints. Furthermore, we explored the functional impact of SVs by integrating overlapping transcriptomic profiles and gene fusions from 153 RNA-sequencing samples and expression data from 247 HumanExon arrays in a combined subset of 361 tumor samples. Finally, *in vitro* studies demonstrated the functional relevance *SHANK2*, a newly identified tumor suppressor gene disrupted by SVs. Altogether, our dissection of multi-omic datasets together with patient clinical profiles and biological experimentation, expands the genomic landscape of neuroblastoma.

Results

Patient characteristics and multi-omic datasets for the study of structural variations. To establish the landscape of SVs in neuroblastoma, we first sequenced the genomes of 135 primary diagnostic tumors and matched normal (blood leukocyte) DNA pairs through the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative (https://ocg.cancer.gov/programs/target). Samples were obtained through the Children's Oncology Group (COG) and included 106 patients with high-risk tumors (29 *MYCN*-amplified and 77 non-*MYCN*-amplified), 14 with intermediate-risk tumors and 15 with low-risk tumors (**Fig. 1a, Supplementary Tables 1 and 2**). Whole genome sequencing (WGS) was performed by Complete Genomics[30] to a median average depth of 76x (**Supplementary Fig. 1a**) and primary data was processed via the Complete Genomics pipeline version 2.0. This pipeline reports small somatic variants (SNVs, small indels, and substitutions)[31], larger SVs, and read-depth coverage across the genome used to infer copy number segmentation profiles (**Online methods**).

To augment the WGS data, and to provide independent replication, we genotyped and analyzed 914 patient tumor samples using Illumina SNP platforms (**Fig. 1a**, **Supplementary Tables 1 and 2**). This cohort comprised 696 high-risk (239 *MYCN*-amplified and 457 non-*MYCN*-amplified), 70 intermediate-risk and 145 low-risk tumors (**Fig. 1a**, **Supplementary Tables 1 and 2**); 488 of these samples were previously released[32] and reanalyzed here. Copy number segmentation was obtained using the *SNPrank* algorithm implemented by the NEXUS® software platform (**Online Methods**).

To further assess the functional impact of SVs, we integrated additional data types generated through the TARGET initiative, we obtained transcriptional profiles from RNA sequencing (N=153) and Affymetrix HumanExon arrays (HuEx, N=247). In addition, the RNA-seq dataset was studied with three available gene fusion pipelines (STAR-fusion[33], fusionCATCHER[34] and DeFUSE[35]).

Patient clinical covariates were organized by the Children's Oncology Group (COG) (**Fig. 1a**, **Supplementary Table 1 and 2;** https://ocg.cancer.gov/programs/target/data-matrix; phs000218.v4.p1).

Along this study, we examined disease risk groups as defined by the COG and the International Neuroblastoma Risk Group (INRG)[36]. Specifically, the following subtypes were considered: **LOWR**: low-risk neuroblastomas; **INTR:** including those with intermediate-risk disease; **MNA**: high-risk neuroblastomas with amplification of the *MYCN* oncogene, and **HR-NA**: high-risk neuroblastomas without *MYCN* amplification.

Identification of novel regions of recurrent DNA copy number gain and loss. WGS-derived copy number profiles were compared with those obtained from the larger SNP array dataset. SCNAs were visualized with Integrative Genome Viewer (IGV) and confirmed well-established patterns of large SCNAs in neuroblastoma that differed between the tumor clinical subtypes (**Fig. 1b, c**)[11,37]. We further analyzed CNV segmentation profiles within neuroblastoma subtypes using GISTIC2.0[38]. As expected, LOWR and INTR tumors harbored few focal or large SCNAs, although aneuploidy was observed (**Fig. 1b, Supplementary Fig. 2a, b**). Consistent with clinical records and previous reports, the MNA and HR-NA subsets shared highly frequent 17q gains and *PTPRD* deletions (9p23) and differ in 2p24 (*MYCN* locus) and prevalence in deletions of 1p, 3p, 4p and 11q (**Fig. 1d, e, Supplementary Fig. 2c-e**). We also observed less frequently reported variants in HR-NA group, including deletions at 16q24.3[39] and segmental gains of the q-arm of chromosome 7, a region recently suggested to exhibit oncogenic potential in neuroblastoma[40] (**Fig. 1d, Supplementary Fig. 2e**).

CNV profiles derived from WGS are deemed to have higher resolution and returned peaks not found in SNP arrays. These SCNAs involved, focal gains at chromosome 5p15.33 (Q-value=1.42×10^{-3}) harboring the telomerase reverse transcriptase (*TERT*) gene) (**Fig. 1d**), intragenic deletions of the ATRX chromatin remodeler gene at Xq21.1, (Q-value=3.76×10^{-3}). Moreover, we observed a novel region of recurrent deletions at 10p15.3 (Q-value=6.16×10^{-2}, **Fig. 1e**).

Orthogonal detection to SV identification: sequence junction, read-depth and copy number breakpoint analyses. To strengthen our findings, we considered three approaches to SV identification

(**Table 1**). We integrated alignment-based SV calls and read-depth CNVs from WGS as well as intensity-based CNV calls from genotyping arrays (**Table 1**, **Online Methods**), and subsequently assessed to extent to which SV breakpoints overlapped between alternative methods and across WGS and SNP datasets. First, we obtained alignment-based SVs reported by CGI somatic pipeline, which provides information about SV boundaries, size and the type of variant in every sample; including deletions (>500b), tandem-duplications (>40b), inversions (>30b), translocations, inversions and complex events (**Supplementary Fig. 1c-e**). We applied additional filters by removing likely artifacts including duplicate junctions across samples and common germline variants found in the database of genomic variants (**Online Methods**)[41]. This resulted in a total of 7,366 (**Supplementary Table 3**); SV calls distributed heterogeneously across neuroblastoma subtypes (**Fig. 2a**). These SVs were defined by sequence junctions delimited by two breakpoints in the genome, which will be referred to as sequence junction breakpoints (SJ-BP). We next mapped copy number dosage breakpoints derived from WGS read-depth segmentation profiles, hereafter referred as read-depth breakpoint (RD-BP, **Online Methods**). A total of 2836 RD-BPs were identified (μ=21) unevenly distributed across samples (**Fig. 2b**). Finally, analogous to the RD-BPs, we mapped copy number breakpoints from segmentation profiles derived from the larger SNP cohort, referred to as copy number breakpoints (CN-BP, **Online Methods**); a total of 6,241 CN-BPs were identified across 914 samples (μ=6.8); As expected from previous reports[11,42], we observed increased number of events in high-risk compared to intermediate and low-risk tumors when studied as SJ-BPs (**Fig. 2a**), RD-BPs (**Fig. 2b**) and CN-BP (**Fig. 2c**).

We further studied the co-localization of breakpoints derived from alternative measurements. First, we compared SJ-BPs and RD-BPs in each of the 135 WGS samples; overall, 30.5% of SJ-BPs co-localize with a RD-BPs (**Fig. 2d**) whereas 62% RD-BPs matched with SJ-BPs (**Fig. 2e**). The lower overlap in SJ-BPs is expected since not all SVs necessarily involve a change in copy number dosage (i.e. inversions and translocations). We next evaluated the co-localization of breakpoints across WGS and SNP platforms within the subset of 52 overlapping samples. 50.2% of CN-BPs from SNP arrays co-localized with SJ-BPs from the WGS dataset (**Supplementary Fig.3a**) whereas only 8.2% of SJ-BPs co-localize with CN-BPs (**Supplementary Fig. 3b**). Furthermore, when comparing dosage based breakpoints across platforms (RD-BP and CN-BP), 23.6% RD-BPs where found co-localizing CN-BP (**Supplementary Fig. 3c**) whereas 66.6% CN-BP co-localized with RD-BPs (**Supplementary Fig. 3d**). Overall, SNP arrays display reduced the number of breakpoints compared read-depth based profiles; we attribute these differences to a narrower dynamic range and lower probe density of the platform.

Finally, we performed a randomized test by sample shuffling (N_i=1000) in order to evaluate whether each of the co-localization percentages listed above could arise by chance or due to recurrence of structural variants across samples. All randomized percentage distributions range between 0.7% and 2.3%; in all cases the null hypothesis was discarded ($p\text{-value} < 0.001$, **Supplementary Fig. 3e-j**). Taken together, alternative breakpoint detection methods returned consistent results even when derived from different platforms providing means for both orthogonal and cross-platform validation of SVs. However, certain types of SVs can only be detected using alignment-based methods.

Patterns of SV and SNV mutational burden differ across neuroblastoma subtypes. High-risk tumors (MNA and HR-NA) presented considerably higher SV mutational burden than low- and intermediate-risk cases (INTR and LOWR), across SJ-BP, RD-BP and CN-BP measures (**Fig. 2a-c**)[11,37]. Comparison of MNA vs. HR-NA tumors revealed these high-risk subsets differed in SV type and genomic location (**Fig. 2f-i, Supplementary Fig. 4a-d**). MNA tumors harbored more SVs on chromosome 2 (Wilcox $P=1.6 \times 10^{-14}$; **Fig. 2f**), largely confined to complex junctions at the *MYCN* amplicon at chromosome 2p24 (**Supplementary Fig. 4a**). However, nearly all chromosomes displayed a higher frequency of SVs in HR-NA than MNA (**Fig. 2f**). Specifically, HR-NA tumors harbored more tandem-duplications in all chromosomes except chromosome 2 (**Fig. 2g**). Inter-chromosomal events were also more frequent in HR-NA tumors and overlapped with regions of known SCNAs other than chr2, including chromosome 3p ($P=1.8 \times 10^{-3}$), chromosome 4p ($P=9.1 \times 10^{-6}$) and chromosome 11q ($P=1.9 \times 10^{-8}$), but not chromosome 1p and 17q (**Fig. 2h**). In contrast, complex events showed no overall differences between high-risk groups with the exception of the aforementioned chr2 (**Fig. 2i**). Finally, RD-BP and CN-BP frequencies followed a similar pattern across chromosomes as that of SJ-BPs; MNA tumors harbored increased number of breakpoints in chromosomes 2 ($P_{RD\text{-}BP}=2.4 \times 10^{-9}$, **Fig. 2j**; $P_{CN\text{-}BP}=4.2 \times 10^{-83}$, **Fig. 2k**) while HR-NA harbored increased frequencies in most other chromosomes and in particular, chromosome 11 ($P_{RD\text{-}BP}=2.0 \times 10^{-8}$, **Fig. 2j**; $P_{CN\text{-}BP}=4.0 \times 10^{-25}$, **Fig. 2k**).

We next studied overall differences in mutational burden and chromosomal instability across subtypes; we posit that the densities of breakpoints (SJ-BP, RD-BP and CN-BP) throughout the genome represent a *bonafide* measure of chromosomal instability. We also obtained measures of somatic SNV density. In order to avoid skewing of results due to the *MYCN* amplicon in MNA and regions exhibiting chromothripsis[43], we implemented an SNV and SJ-BP tumor burden measure robust against outliers. To this end, the genome was divided into 41 sequence mapped chromosome arms and the density of SVs per Mb was measured; then, for each sample, the interquartile mean (IQM) was derived from the 41 arm measurements (**Supplementary Fig. 4e,f**). Similarly, we obtained IQM density measurements from RD-BP and CN-BP chromosomal burdens (**Fig. 2n,o**). As expected, LOWR and INTR tumors carried very low mutational burden (**Fig. 2l-o**)[11,37]. We observed increased CIN (SJ-BP, RD-BP and CN-BP) in HR-NA compared to MNA (Wilcoxon rank test: $P_{SJ\text{-}BP}=4.5 \times 10^{-5}$, **Fig. 2m**; $P_{RD\text{-}BP}=1.3 \times 10^{-2}$, **Fig. 2n**; $P_{CN\text{-}BP}=4.6 \times 10^{-8}$, **Fig. 2o**), similar to previous reports[44]. In contrast, MNA and HR-NA did not differ in their average SNV burden (Wilcoxon rank test: P=0.29, **Fig. 2h**). These results confirm that HR-NA has increased chromosomal instability[37,44] and supports the observation that small SNVs and larger SVs arise from different mutational processes.

Chromothripsis associates with major neuroblastoma oncogenic mechanisms. Previous studies have reported chromothripsis to occur in up to 18% of high-risk neuroblastomas[43] and identified associations between chromothripsis and key neuroblastoma oncogenes *MYCN* and *TERT*[27,28]. We therefore sought to leverage our large dataset to further explore the oncogenic associations of chromothripsis in neuroblastoma. We first identified alterations of major neuroblastoma oncogenes (*MYCN*, *TERT* and *ALK*) in our WGS and SNP cohorts. Rearrangements near *TERT locus* were confirmed in 23 HR-NA samples and 2 MNA from the WGS dataset as well as 15 cases (14 HR-NA and 1 MNA) from the SNP dataset, one sample (PAPUTN) was present in both datasets (**Fig 3a**); 11 cases from the WGS set with available DNA were validated using Sanger sequencing (**Supplementary Fig. 5**). We confirmed that *TERT* expression was increased in those samples as well as in MNA tumors in accordance with previous reports (**Supplementary Fig. 6**) [27,28]. In addition, CN-BPs were found near *TERT* in 15 HR-NA samples from the SNP array dataset (**Fig. 3a**); highlighting the capability of SNP arrays for detecting this type of event using the breakpoint analysis approach introduced in this study. *MYCN* amplification was determined diagnostically by FISH experiments in 29 samples from the WGS dataset (**Supplementary Table 2**). IGV visualization of segmentation data of 7Mb region surrounding *MYCN* confirms the clinical records (**Supplementary Fig. 7a**). We also explored events affecting the *ALK* gene, which can co-occur with amplification of *MYCN*. Two out of four rearrangements found near *ALK* involved also *MYCN*, (**Supplementary Fig. 7b**); these events were validated via Sanger sequencing (**Supplementary Fig. 7c**).

Next, chromothripsis was characterized by clustered somatic rearrangements and alternating copy number states in defined chromosome regions[45]. We identified candidate chromothripsis events at chromosome arms with unusual high breakpoint densities (> 2σ of each sample's breakpoint burden distribution) and a minimum of 6 breakpoints (both SJ-BPs and RD-BPs) in 27 regions (**Online Methods, Supplementary Table. 4**) involving 20 distinct high-risk tumors (19%). Chromothripsis was observed in chromosome 2 in a total of 8 samples (**Fig. 3b,d,e; Supplementary Fig. 8**); those samples showed enrichment in samples harboring *MYCN* amplification (MNA) (7/8 samples, Binomial test P $=7.4 \times 10^{-4}$) (**Fig. 3**). Among them, two samples (PARETE and PATESI) involved co-amplification of *ALK* with *MYCN* (**Supplementary Fig. 7b,c**). In addition, 9 tumors harbored shattered chromosome 5p with strong enrichment in samples with rearrangements near *TERT* (8/9, Binomial test $P = 7.3 \times 10^{-5}$) (**Fig. 3c,d**; **Supplementary Fig. 9**). Two samples (PAPSRJ and PAPUTN, **Fig. 3d**) included interchromosomal events associating the *MYCN* and *TERT* gene loci and co-amplification of both oncogenes. Other chromosomes involved included chromosome 1, 10, 11 and X in a female sample (**Supplementary Fig. 10**). Chromothripsis in most cases (15/20) was localized to a single chromosome involving either the whole chromosome (i.e. PATBMM, **Fig. 3b**) or local regions (i.e. PATESI, **Fig. 3c**). Multiple chromosomes were involved in 5/20 (25%) of cases with chromothripsis. One sample (PARIRD) harbored an event involving chromosomes 2, 17 and 22, while PANRVJ involved large regions of chromosomes 1 and 2 (**Supplementary Fig. 8**).

We next sought further confirmation of our results in the larger SNP array dataset. In the absence of sequence junction information, we focused on unusual high density (>2σ) of CN-BPs (**Fig. 3f).** We observed high-breakpoint density on chromosome 2 enriched in MNA samples (46/46, P ~ 0). We also observed enrichment of high breakpoint density on chromosome 5 involving cases in tumors harboring rearrangements or CN-BPs near *TERT* (7/11, P = 3.01 x 10^{-8}). In addition, chromosome X high breakpoint density was enriched in female patients (6/7, P = 4.7 x 10^{-2}), although no specific oncogenic associations were determined. Overall, SNP array analysis of high CN-BP density supports and replicates observations based on the WGS dataset.

Identification of genes recurrently altered by SVs in high-risk neuroblastomas. In order to identify genes affected by recurrent somatic SVs in neuroblastoma we generalized the approach described previously for *TERT*, *MYCN* and *ALK* genes (**Fig 3a, Supplementary Fig. 7**); SVs were assigned to different categories according to the inferred impact on the exonic sequence of known RefSeq genes (**Fig. 4a,b**). Sequence junctions (SJ-BPs) provide more detailed information including the type of SV and the two genomic breakpoint locations involved; with this knowledge we classified SVs into: a) “Coding”; SVs that modify the exonic sequence of known genes including whole gene copy number alterations (duplications and deletions, size up to 2Mb) and b) “Non-coding”: SV that do not modify the exonic sequences but might have an impact on regulatory regions proximal to known genes (100Kb upstream and 25Kb downstream) as well as intronic regions (**Fig. 4a**). In contrast to SJ-BP junctions obtained from discordantly aligned mate read pairs; dosage-based breakpoints (RD-BP and CN-BP) cannot identify their counterpart location in the genome. Therefore, events such as translocations and inversions cannot be defined. Conversely, read-depth and array intensity based copy number inform about multiple dosage gains and losses. With this in mind, we assumed the impact as a) "Coding": breakpoints within the transcription "start" and "end" positions of known genes and b) "Non-coding", breakpoints located on proximal upstream and downstream regions (**Fig. 4b**). In addition, we localized copy number variants involving amplification ($CN_{WGS} > 8$; $CN_{SNP} > 4.5$, **Online Methods**) and deep deletions ($CN_{WGS} < 0.5$; $CN_{SNP} < 0.9$, Online Methods) (**Fig. 4b**).

Based on the aforementioned definitions, we ranked recurrently altered genes according to the number of samples harboring "coding" and "non-coding" SVs for each of the 3 alternative breakpoint analyses (SJ-BPs, RD-BPs and CN-BPs; **Fig. 4c-h and Supplementary Table 5**). Overall, recurrently altered genes by 'coding' events return highly concordant results across the three approaches; *MYCN* neighbor genes occupying top ranks followed by known neuroblastoma altered genes, *PTPRD* and *ATRX* and novel genes including *SHANK2* and *DLG2* located at chr11.q13 and chr11.q14 respectively, and others such as *AUTS2* at chr7.q11 and CACNA2D3 at chr3.p14 (**Fig. 4c,e,g**). On the other hand, non-coding recurrent alterations consistently reflect as top ranking genes, *MYCN* and *TERT* and their respective neighbor genes at chr2.p24 and chr5.p15 (**Fig. 4d,f,h**).

In order to provide an integrated overview of the landscape of altered genes, we combined WGS based methods (SJ-BP and RD-BP) into a ranking of recurrently altered genes with co-localizing breakpoints, hence orthogonally validated. A total of 77 genes have at least 1 co-localizing SJ-BP and RD-BP breakpoint (**Fig. 4i, Supplementary Table 6**); in addition, we overlaid likely pathogenic SNV calls (**Supplementary Table 7**). Many altered genes cluster in specific regions associated with known oncogenes such as chr2.p24 near *MYCN* (11 genes) and chr5.p15 near *TERT* (7 genes). The ranking is led by *MYCN* with 37 samples harboring variants, which orthogonal validation was obtained in 26 cases by co-localizing SJ-BP and RD-BP; those include 29 MNA and 8 HR-NA tumors. Interestingly, 11 HR-NA samples harbor alterations of *MYCN* (8 SVs and 3 SNVs) supporting the pathogenic role of

MYCN in non-amplified tumors. *TERT* rearrangements were identified in 25 samples; orthogonal validation of breakpoints was observed in 12 cases (Fig. **3a**). *PTPRD* was found altered in 20 samples, 11 of which were orthogonally validated (**Supplementary Fig. 11a**) [25,49]. We found *ATRX* (N^{orth}=5; N^{tot}=12) intragenic deletions and one tandem-duplication in HR-NA tumors (**Supplementary Fig. 11b**)[50]. The *SHANK2* gene was found disrupted in 11 HR-NA tumors; 3 samples involved gene fusions that didn't appear in frame. *DLG2*, a newly described tumor suppressor in osteosarcoma[46,47], was found disrupted in 10 samples based on SJ-BP and RD-BP analyses from a total of 14 samples, two of which involved gene fusion events. Both *SHANK2* and *DLG2* are located on chromosome 11q and play a role in the formation of postsynaptic density (PSD)[48]. Other novel candidate altered genes include *AUTS2* with frequent intragenic deletions at chr7q (N^{orth}=3; N^{tot}=18, **Supplementary Fig. 12a**) and the calcium channel *CACNA2D3* (N^{orth}=4; N^{tot}=11), which represent a frequent breakpoint associated with 3p deletion at chr3.p14.3 (**Supplementary Fig. 12b**). A region proximal to LINC00910 lncRNA in chromosome 17 suffered rearrangements in 13 tumors (**Supplementary Fig. 13a**). Finally, the list includes additional genes with known roles in neuroblastoma and cancer: *ALK,* which somatic SNVs were found in 18 samples, also harbors rearrangements (N^{orth}=1; N^{tot}=5; **Supplementary Fig. 7b**)[23,51] in a combined set of 23 samples (17% of all neuroblastomas). Also, *CDKN2A* and *CDKN2B* deletions were found in 3 tumors (**Supplementary Fig. 13b**).

Systematic validation of SVs in high-risk neuroblastoma patient samples. Along this study, we produced extensive validation via Sanger sequencing of variant junctions, provided the samples had available DNA supply in our tumor bank; The validations focused on key genes and included 12 proximal *TERT* SVs (**Supplementary Fig. 5**), 4 *ATRX* deletions (**Supplementary Fig. 14**), 4 proximal *ALK* variants (**Supplementary Fig. 7c**), 11 *SHANK2* translocation events, 9 of which involved chromosome 17q (**Supplementary Fig. 15**) and 12 *DLG2* variants (**Supplementary Fig. 16**). In a total we validated 45 SVs (**Supplementary Table 8**). The original CGI cancer pipeline classifies SVs into high and low confidence variants depending on the number of read pairs supporting the evidence

(N_{reads} threshold = 10); our pipelines rescue many cases below that threshold; specifically 6 out of 45 SVs (13.3%, 3 *ATRX* and 3 *DLG2*) returned positive Sanger validation.

SVs have a regional transcriptional effect. To gain further understanding of the functional relevance of SVs, we performed an expression quantitative trait loci (eQTL) analysis for each of the recurrent SV associated genes (**Supplementary Fig. 13**). The analysis, which was replicated in the two available transcriptional datasets (RNA-seq and HuEx array), reported consistent up-regulation of *MYCN* and *TERT* including their neighbor genes. We also observed up-regulation of the lncRNA *LINC00910* ($P_{RNA} = 7.0 \times 10^{-3}$) at chr17.q21, a region with frequent inter-chromosomal translocations (**Supplementary Fig. 17a**). On the other side, *CDKN2A* was down-regulated ($P_{HuEx}= 4.7 \times 10^{-2}$; $P_{both} = 2.5 \times 10^{-2}$) by focal deletions. Finally, *PLXDC1* at chr17.q12 was also down-regulated ($P_{HuEx}= 4.7 \times 10^{-2}$; $P_{both} = 2.5 \times 10^{-2}$) in association with 17q gain breakpoints.

In addition to changes in overall gene expression by eQTL, translocations may lead to the expression of gene fusion transcripts; we explored RNA-seq samples with three available gene fusion methods (STAR-fusion[33], fusionCATCHER[34] and DeFUSE[35], **Supplementary Fig. 17b**). We then refined the list of fusion transcripts to those confirmed by the presence of translocations in the WGS SV calls (**Supplementary Table 9**), which dramatically increases the overall agreement across the three gene-fusion detection methods (**Supplementary Fig. 17c**). The most frequent gene fusion event with both RNA and DNA evidence involved *SHANK2*; the three *SHANK2* fusion events involved 17q genes: *EFTUD2*, *MED1* and *FBXL20* (**Fig 6a**). *DLG2* exhibited gene fusion events in two samples involving *SEMA6C* and *MYCBP2* at chromosomes 12 and 13 respectively (**Fig 6b**). However, none of the *SHANK2* and *DLG2* fusion transcripts appeared to be in-frame, suggesting the fusion transcripts may not be biologically relevant and that these are more likely loss of function events. Conversely, we found an in-frame fusion transcript and translocation involving *FOXR1:DDX6* (**Supplementary Table 9**), which oncogenic fusion events have previously been described in neutoblastoma[52].

Neurodevelopmental genes are recurrently disrupted by structural variations in neuroblastoma. In order to identify pathways targeted by SVs we considered recurrently altered genes from each the coding (N>2) and non-coding (N>3) altered gene lists (#genes: SJ-BP^{coding}=109, SJ-BP$^{non\text{-}coding}$=36, RD-BP^{coding}=76, RD-BP$^{non\text{-}coding}$=27, CN-BP^{coding}=77 And CN-BP$^{non\text{-}coding}$=88, **Fig. 4c-h, Supplementary Fig. 18, Supplementary Table 10**). We tested each gene list for enrichment across Gene Ontology, pathway and disease gene classes using ToppGene[53] (**Supplementary Table 10**). Genes with coding sequences altered showed consistent results across the three breakpoint mappings, revealing strong enrichment in genes involved in autism spectrum disorder susceptibility ($P_{SJ\text{-}BP} = 2.8 \times 10^{-9}$; $P_{RD\text{-}BP} = 2.9 \times 10^{-5}$; $P_{CN\text{-}BP} = 2.7 \times 10^{-9}$) and other neurodevelopmental disorders (NDD) as well as protein localization to synapse ($P_{SJ\text{-}BP} = 1.2 \times 10^{-5}$; $P_{RD\text{-}BP} = 1.1 \times 10^{-7}$; $P_{CN\text{-}BP} = 2.4 \times 10^{-6}$) and other neuronal related classes (**Fig. 5a-c**; **Supplementary Table 10**). The gene sets with 'non coding' alterations were more variable across the alternative breakpoint analyses, but were dominated by events involving *MYCN* and *TERT* in association with the disease class "stage, neuroblastoma" ($P_{SJ\text{-}BP} = 1.9 \times 10^{-6}$; $P_{RD\text{-}BP} = 2.5 \times 10^{-5}$; $P_{CN\text{-}BP} = 9.2 \times 10^{-5}$, **Supplementary Fig. 18 and Supplementary Table 10**).

Recurrently disrupted neurodevelopmental genes are down-regulated in high-risk neuroblastoma. To further characterize the clinical relevance of recurrently altered genes in neuroblastoma, we studied their differential expression between high-risk subtypes and low-risk (Stage 1 and 4s) groups (**Fig. 5d,e, Supplementary Fig. 19**). We first used gene set enrichment analysis (GSEA)[54] to confirm the directionality of the regulation of gene classes enriched in recurrently altered genes; we observed down-regulation of both neuronal and synaptic genes ($P_{HuEx} = 1.09 \times 10^{-9}$) and autism disorder susceptibility genes ($P_{HuEx} = 6.38 \times 10^{-7}$) in high-risk tumors when compared to stage 1 low-risk tumors (**Fig. 5d-f**). We then focused on differential expression of genes with recurrent SVs in high-risk subtypes (**Fig. 5g**). As expected, known oncogenes including *TERT* and *ALK* are up-regulated in both MNA and HR-NA while *MYCN* is up-regulated only in MNA tumors. Known neuroblastoma tumor suppressor genes including *CAMTA1* and *RERE* from the 1p chromosome region and *PTPRD* are down-regulated in both subtypes. Most genes with a role in autism disorder predisposition and those involved in neuron parts and synapse formation, are down-regulated in both high-risk subtypes; in particular expression was significantly reduced for *SHANK2* ($P_{MNA} = 2.15 \times 10^{-11}$;$P_{HR-NA} = 1.05 \times 10^{-8}$) and *DLG2* ($P_{MNA} = 2.1 \times 10^{-8}$;$P_{HR-NA} = 4.86 \times 10^{-8}$) in high-risk compared with stage 1 low-risk tumors (**Fig. 6c**) and compared to stage 4S low-risk tumors ($P_{MNA} = 1.41 \times 10^{-3}$;$P_{HR-NA} = 1.82 \times 10^{-5}$ and $P_{MNA} = 1.09 \times 10^{-4}$; $P_{HR-NA} = 2.72 \times 10^{-4}$ respectively).

Neurodevelopmental genes *SHANK2* and *DLG2* are frequently disrupted by chromosome 11 translocation events. High-risk neuroblastomas without *MYCN* amplification frequently exhibit deletion of chromosome 11q and this event is associated with a poor outcome[16,44]. The most frequent breakpoints observed in this study were located at chromosome 11q.13 and 11q.14 disrupting the *SHANK2* and *DLG2* gene loci respectively (**Supplementary Fig. 19**). *SHANK2* translocation partners involved chromosome 17q in 10/11 WGS cases, in addition we identified 49 samples from the SNP dataset (10.7%) with breakpoints in *SHANK2* (**Fig 6a**). In contrast, *DLG2* translocation partners include multiple chromosomes; breakpoints were also identified in *DLG2 locus* in 28 samples from SNP dataset (**Fig 6b**).

SHANK2 is a scaffold protein in the postsynaptic density (PSD) with two known coding isoforms (long: NM_012309; short: NM_133266). We therefore studied the expression pattern of *SHANK2* at the exon level using both HumanExon arrays (**Fig. 6c**) and RNA-seq (**Supplementary Fig. 20**) data. Clustering analysis of *SHANK2* exon expression revealed two distinct clusters corresponding to the two known coding isoforms. Expression of both isoforms was decreased in high-risk tumors compared to INTR and LOWR as observed from RNA-seq (**Fig. 6d**) and HuEx expression analysis (**Supplementary Fig. 21a, b**). Finally, in a large independent cohort[55], reduced expression of the long isoform (NM_012309) was associated with increased tumor stage ($P=1.62 \times 10^{-22}$, **Supplementary Fig. 21c**) and poor survival ($P=7.21 \times 10^{-13}$, **Supplementary Fig. 21d**). Consistent with the *SHANK2* expression pattern, we observed decreased activation of PSD genes based on GSEA in high-risk compared to low-risk neuroblastomas in multiple prognostic signatures (**Supplementary Fig. 22**). We decided to further study the long isoform of *SHANK2* (NM_012309) given that nearly all SVs uniquely disrupt this splice variant, leaving the short isoform (NM_133266) intact.

***SHANK2* expression inhibits cell growth and viability of neuroblastoma cells.** To further elucidate the role of *SHANK2* in neuroblastoma, three neuroblastoma cell lines with low or no endogenous SHANK2 expression (**Supplementary Fig. 23**), including SY5Y (*MYCN* Non-amplified), Be(2)C (*MYCN*-amplified), and NGP (*MYCN* amplified), were stably transduced to constitutively overexpress *SHANK2* long isoform or an empty vector control. SHANK2 expression was confirmed by Western blot (**Fig. 7a-c**). When maintained in selection media and grown alongside empty vector controls, the SHANK2-expressing cells consistently exhibited decreased cell growth and viability as measured by RT-CES cell index (**Fig. 7d-f**) as well as CellTiter Glo assay (**Fig. 7g-i**). For SY5Y, when control reached confluence, the comparable cell indexes of the SHANK2 overexpressing lines were reduced by 75% ($P=3.4 \times 10^{-5}$; **Fig. 7d**), Be(2)C cell index reduced by 62% ($P=3.16 \times 10^{-4}$; **Fig. 7e**), and NGP showed a 14% reduction ($P=2.62 \times 10^{-2}$; **Fig. 7f**). We also observed decreased cell viability in SHANK2-expressing cells at both 4- and 7-day endpoints using an ATP-dependent CellTiter Glo assay. Specifically, viability of SY5Y SHANK2-expressing cells was reduced to 65.51% ($P=1.34 \times 10^{-18}$) and 52.64% ($P=4.72 \times 10^{-26}$) of controls (**Fig. 7g**). This was reinforced in the similar results for Be(2)C SHANK2-expressing cells (49.21% and 44.26%, $P=5.76 \times 10\text{-}^{28}$ and 5.74×10^{-15}; **Fig. 7h**) and NGP (90.63% and 74.01%, $P=5.11 \times 10^{-3}$ and 6.01×10^{-13}) (**Fig. 7i)**.

***SHANK2* expression accelerates differentiation of neuroblastoma cells exposed to all-trans retinoic acid (ATRA).** We next investigated the role of *SHANK2* in neuronal differentiation in Be(2)C and SY5Y cells exposed to ATRA. In the presence of ATRA, overexpression of *SHANK2* accelerated differentiation as measured by presence and length of neurites compared to cell body (**Fig 7j-o**; **Supplementary Fig. 24a-d**). While decreases in growth can be measured even without drug application, once ATRA is applied, cells overexpressing *SHANK2* develop neurites more quickly, and those neurites extend further than empty vector controls (**Supplementary Fig. 24c,d**). In Be(2)C cells, significant differences in neurite outgrowth normalized to cell-body area was seen at 72 hours post treatment with 1 uM ATRA (**Fig. 7j-l**) with *SHANK2* cells exhibiting a 1.6-fold increase over controls ($P=3.09 \times 10^{-13}$), and the difference increased at 96 hours to 2.76-fold ($P=2.37 \times 10^{-30}$). Even with vehicle alone, *SHANK2* cells had more neurite outgrowth per cell body compared to their empty vector counterparts at both 72 and 96 hours post treatment ($P=1.02 \times 10^{-5}$, $P=1.25 \times 10^{-13}$, respectively). In SY5Y, though differentiation takes longer and both *SHANK2* cells and controls eventually reach 100% confluence with vehicle alone, SHANK2 overexpression still led to a decreased confluence in samples ($P=1.69 \times 10^{-6}$, **Supplementary Fig. 24b**). In analyzing total neurite outgrowth without normalization for cell body area, SY5Y ATRA-treated *SHANK2* cells outpaced controls starting at hour 144 post-treatment and continued to lead until the experiment end, with a total neurite measurement 1.55-fold increased over controls ($P=1.62 \times 10^{-35}$; **Supplementary Fig. 24d**). Once normalized, *SHANK2* cells have higher measured outgrowth starting at 75 hours post treatment, hour 96, and maintain from there. At 195 hours past treatment, *SHANK2* cells treated with 5uM ATRA displayed neurites at 1.71-fold increase over their empty vector controls ($P=2.36 \times 10^{-17;}$ **Fig. 7m-o**). Taken together, these data suggest *SHANK2* is a newly identified haplo-insufficient tumor suppressor in high-risk neuroblastoma that is disrupted by recurrent somatic structural variation in the *MYCN* non-amplified subset of cases.

DISCUSSION

Sequencing studies of neuroblastoma tumors have revealed a relatively low SNV burden and limited mutational landscape, leaving aneuploidy and large segmental chromosomal alterations as the main candidate driver mutations in many tumors[23]. Structural variants (including insertions, deletions duplications and translocations) may also function as potent cancer drivers, as demonstrated with the discovery of rearrangements near *TERT* driving aberrant *TERT* expression in many high-risk neuroblastomas[27,28]. In this study, we have substantially expanded the landscape of structural variation in neuroblastoma and revealed their functional impact in the disease through integrative genomic analysis of a large cohort of patient samples profiled by whole genome sequencing and SNP arrays together with additional transcriptional data. To the best of our knowledge, we presented here the largest fully integrated genome wide survey of structural variation in neuroblastoma, combining alignments based (SJ-BP) and copy number based (RD-BP and CN-BP) breakpoint analyses. Despite the rich landscape of structural variations described in this study, we didn't identify any oncogenic gene fusion events, with the exception of the previously reported case of FOXR1[52], which appears to occur preferentially in intermediate risk tumors within our cohort.

Overall, we showed that structural variation considerably increases the genetic complexity of high-risk neuroblastomas. This complexity is most evident in high-risk tumors without amplification of *MYCN* (HR-NA), which show increased chromosomal instability[44], as confirmed by structural variation and breakpoint burden analyses. Moreover, this subset harbors more SVs in known cancer genes as well as novel genes. Interestingly, the SNV burden is very similar between MNA and HR-NA groups. As shown in pan-cancer studies, the underlying mechanisms potentiating chromosomal instability (CIN) and somatic SNV burden may differ[56]**.** Nonetheless, loss of *TP53* function by deleterious mutations, associated with increased CIN in pan-cancer studies, is largely absent in primary neuroblastomas and drivers of the observed increased chromosomal instability in neuroblastoma remain unknown. Despite lesser burden, MNA tumors also present widespread structural variation but often associated to the *MYCN locus*. Indeed, translocations involving *MYCN* complex events may have a broader effect throughout the genome as observed in cases where *MYCN* appears co-amplified with *ALK* and *TERT*.

Chromothripsis is a well-documented genetic alteration in neuroblastomas, reported in as many as 18% high-stage tumors[57]. Similarly, in the current study, 19% of high-risk tumors from the TARGET cohort exhibited chromothripsis (N=20/105) involving a total of 27 chromosomal regions. These events largely overlap with amplification of *MYCN* (as well as some *ALK* cases) on chromosome 2p and *TERT* on chromosome 5p, suggesting an important role of chromothripsis followed by purifying selection as an underlying cause of those alterations. We also observed high-breakpoint density in the X chromosome of females based on the SNP data, which could be explained by higher tolerance to chromothripsis of diploid regions. Altogether, the prevalence and oncogenic role of chromothripsis in neuroblastomas is confirmed; future studies need to address whether it represents a therapeutic intervention opportunity.

Along this study, we report a common genetic repertoire of altered genes between neuroblastomas and neurodevelopmental disorders (NDD) such as autism. Linkage between cancer and autism has been previously established in PTEN-associated germline syndromes[58]. Furthermore, multiple autism susceptibility genes also have a known role in cancer[59]. Certain germline deletions with NDD associations such as 10p15[60] and 16p24.3[61] are reported here to occur somatically in neuroblastoma. We hypothesize that structural variants in *SHANK2* and *DLG2* genes coding proteins of the postsynaptic density (PSD) comprise novel neuroblastoma candidate tumor suppressors involved in neuronal differentiation; additional candidate altered genes with a role in neurotransmission and synapsis and involvement in autism include *AUTS2*, *CNTNAP2*, *NRXN1*, *CTNND2*[62]. These alterations are more prevalent in high-risk tumors without amplification of *MYCN*, which is itself a potent driver of dedifferentiation[63]. Transcriptomic analyses have shown that neural lineage pathways are commonly down-regulated in high-risk neuroblastomas compared to low-risk signatures[64]. Synaptogenesis is a key process in neuronal differentiation, and mutations in genes involved in the formation of synapses have frequently been implicated in NDD (also termed shankopathies)[65]. Furthermore, *DLG2* has been recently described as a tumor suppressor in osteosarcoma[46,47]. We propose that the dysregulation of *SHANK2* and *DLG2* synaptic genes is involved in maintaining the undifferentiated state of the neuroblastic cancer cell. In particular, here we show that *SHANK2* expression reduces cell growth and increases neurite outgrowth in human derived neuroblastoma cell lines in the presence of ATRA. The sensitizing effect of *SHANK2* expression to ATRA treatment reveals the importance of understanding the mechanisms of differentiation disrupted in neuroblastoma. Retinoids are currently utilized as maintenance therapy in high-risk neuroblastoma standard of care[66,67]; subsequent studies with larger cohorts should evaluate the contribution of alterations in neurodevelopmental genes to retinoic acid treatment response as a maintenance therapy.

Despite the background genetic heterogeneity of neuroblastoma subtypes, SVs systematically target telomere maintenance mechanisms and neurodevelopmental pathways influencing differentiation. While *MYCN*-amplified tumors are largely sustained by the oncogene's strong effect, other high-risk tumors suffer recurrent hits in both key pathways. Altogether, we depict a new landscape of structural variation in neuroblastoma and provide mechanistic insight into the neuronal development abrogation hallmark of the high-risk form of this pediatric disease.

Supplementary Data

Supplementary data include 10 tables and 24 figures.

Acknowledgements

This work was supported in part by NIH grant R01-CA124709 (SJD) and the Roberts Collaborative Forefront Award (GL). This project was also funded in part by a supplement to the Children's Oncology Group Chair's grant CA098543 and with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E to S.J.D and Complete Genomics.

Author Contributions

S.J.D designed the experiment. G.L., K.L.C and S.J.D. drafted the manuscript. G.L. and S.J.D. performed analyses of SVs from WGS. G.L. performed RNA data analysis. G.L. and K.L.H. performed telomere analyses and allele-specific expression studies. G.L. and A.M. performed *de novo* transcript analyses. G.L. and K.S.R. performed fusion transcript analyses. K.L.C. and M.D. performed Sanger sequencing. K.L.C., M.D., L.M.F., and E.H. performed *SHANK2* experiments. Z.V. assisted with sequence data analysis. J.S.W. and J.K. generated RNA sequencing data. S.A. and R.C.S. generated array-based expression data. H.S. and P.W.L generated methylation array data. All authors commented on or contributed to the current manuscript.

Analytical pipeline	**Platform**	**Number of genomic breakpoints**	**Resolution**	**Specific event types**
Sequence Junction (SJ)	WGS	2 x SJ-BP	1 bp (variable coverage)	SV type: translocation, inversion, deletion, tandem-duplication
Read-Depth (RD)	WGS	1 x RD-BP	2 Kbp (variable coverage)	Dosage information: gain/amplification and loss/deep deletion
Copy Number (CN)	SNP array	1 x CN-BP	~5 Kbp (variable probe density)	

Table 1. Alternative breakpoint analyses for the identification of structural variants

FIGURE LEGENDS

Figure 1: Novel somatic DNA copy number alterations (SCNAs) revealed by whole genome sequencing (WGS) of neuroblastoma tumors. (a) Survey of available samples, clinical information and data types used throughout this study (See also Supplementary Tables 1 & 2). (b-c) Integrated Genome Viewer (IGV) visualization of DNA copy number gains (red) and losses (blue) across neuroblastoma subtypes in the WGS (b) and SNP (c) datasets. (d-e) GISTIC q-value plots showing significant regions of gain (d) and deletion (e) in HR-NA samples in 77 samples derived from WGS dataset.

Figure 2: Somatic mutation burden differs between neuroblastoma subtypes by quantity, type and genomic location. (a) Stacked bar chart of discordant mate pair derived structural variants by type and neuroblastoma subtype in WGS dataset. **(b)** Bar plot representing the number of read-depth breakpoint (RD-BP) per sample across subtypes in the WGS dataset. **(c)** Bar plot representing the number of copy number breakpoint (CN-BP) per sample across subtypes in the SNP dataset. **(d)** Co-localization of RD-BPs into SJ-BP across WGS dataset samples and overall co-localization percentage (right bar). **(e)** Co-localization of SJ-BPs into RD-BP across WGS dataset samples and overall co-localization percentage (right bar). **(f-k)** By chromosome comparison between MNA and HR-NA of the inter-quantile average number of SVs including all SJ-BP variant types (f), duplications (g), interchromosomal (h), complex (i) as well as RD-BP and CN-BP. A Wilcoxon test is obtained for every chromosome and the p-value is represented by asterisk (*** = $p < 0.001$, ** = $P < 0.01$, * = $P < 0.05$; asterisk color indicates what group has the highest IQM, red=MNA and orange=HR-NA). Mutation burden analysis plot across neuroblastoma subtypes representing the burden of SNVs **(l)**, SJ-BPs **(m),** RD-BPs(n) and CN-BP **(o)**.

Figure 3: Incidence of chromothripsis and oncogenic associations in high-risk neuroblastoma. (a) Identification of rearrangements near *TERT* in 25 WGS samples based on RD-BPs combined with

SJ-BPs and 15 SNP samples based on CN-BPs. Nearest upstream *TERT* junctions derived by sequencing are highlighted in text; 'S' at the left of the panel indicates positive validation by Sanger sequencing. **(b-d)** Circos representation of regions suffering chromothripsis; outer track represents read-depth based copy number (red=gain; blue=loss); inner track represents structural variant junctions (red = inter-chromosomal; blue = intra-chromosomal) **(b)** Examples of chromothripsis in chromosome 5 with *TERT* structural variant association. **(c)** Examples of chromothripsis in chromosome 2 with *MYCN* amplification. **(d)** Chromothripsis involving both TEET and MYCN loci. **(e-f)** Association between chromothripsis and genomics and clinical variables; **(e)** Incidence of chromothripsis by chromosome across 135 WGS samples. **(f)** Incidence of high breakpoint density chromosomes across 914 SNP array samples. Binomial test p-value: *** $p < 0.001$; * $p< 0.05$

Figure 4 Identification of recurrently altered genes in neuroblastoma by breakpoint analyses. (a) Schematic representation of structural variants derived from junction breakpoints (SJ-BPs)) classified according to their impact on known genes. **(b)** Schematic representation of read-depth and copy number breakpoints (RD-BPs, CN-BPs) classified according to their impact on known genes. **(c-h)** Recurrently altered genes rankings based on different breakpoint analyses and mode of impact: (c) gene coding sequences with recurrent SJ-BPs, (d) gene proximal and intronic sequences with recurrent SJ-BPs, (e) gene proximal sequences with recurrent RD-BPs (f) gene coding sequences with recurrent RD-BPs, (g) gene coding sequences with recurrent CN-BPs and (h) gene proximal sequences with recurrent CN-BPs. (i) Oncoprint based on the WGS dataset recurrently altered genes by structural variants detected through orthogonal approaches (SJ-BP and RD-BP) as depicted in bar plot (right). The oncoprint aggregates three tracks per gene representing different breakpoint analysis (upper =SJ-BP, center=RD-BP) and recurrent pathogenic SNVs (lower).

Figure 5. Neurodevelopmental genes are recurrently targeted by structural variations in neuroblastoma. (a-c) Function enrichment analysis bar plots for genes recurrently altered based on

breakpoint analyses of (a) SJ-BPs, (b) RD-BPs and (c) CN-BPs. Analysis includes gene sets associated with diseases (green), Gene Ontology (purple) and Pathways (red). (d-e) Gene Set Enrichment Analysis across the signature of high- versus low-risk tumors from the HumanExon array show enrichment of **(d)** neuronal and synapse part and **(e)** autism disorder predisposition genes. (f) Volcano plot showing differential expression between high- and low-risk highlighting genes with recurrent SVs and their functional classification **(g)** Subtype specific high- versus low-risk differential expression analysis of 77 recurrently altered genes from Fig 4i shown as scatter plot (MNA = x-axis, HR-NA = y axis). (d-g) Analysis replicated in two datasets: HuEx arrays (here) and RNA-seq (Supplementary Fig. 19 a-d).

Figure 6: Neuronal genes *SHANK2* and *DLG2* are frequently disrupted by translocation events involving chromosome 11*.* (a-b) Copy number, breakpoint location and types of SVs at genomic regions harboring rearrangements that span **(a)** *SHANK2 and* **(b)** *DLG2* loci; ‘S’ at the left of the panel indicates positive validation by Sanger sequencing for *SHANK2* (Supplementary Fig ?) and *DLG2* (Supplementary Fig ?). Associated gene fusion events obtained from RNA-seq indicated in purple text. **(c)** Clustering analysis of *SHANK2* exon level FPKM from RNA-seq data. The heatmap (left) shows higher exon expression level in S4s compared to MNA and HR-NA samples. The correlation matrix (right) shows two well-defined clusters associated with the two known coding isoforms of the gene. Exons are color coded according to their isoform span.

Figure 7: *SHANK2* reduces cell growth and promotes differentiation in neuroblastoma cell line models. **(a-c)** Western blots confirming overexpression of SHANK2 in all tested neuroblastoma cells: **(a)** SY5Y, **(b)** Be(2)C and **(c)** NGP. **(d-f)** Decreased proliferation in all 3 lines overexpressing SHANK2 (red) compared to controls (green), as measured by RT-CES. **(h-i)** Decreased viability in SHANK2 overexpressing cells (red) versus controls (green) as measured by ATP-dependent CellTiter Glo Assay.

(j,k) Incucyte images Be(2)C cells for **(j)** vector control and **(k)** SHANK2 expressing cells at 78 hours post treatment with 1 uM ATRA. Neurite extensions masked in pink; cell bodies masked in blue. **(l)** Neurite length normalized to cell body area starting immediately after ATRA application corresponding to Be(2)C cells images at different time points. **(m, n)** SY5Y images from Incucyte at day 9 post ATRA treatment (5uM). **(o)** Neurite outgrowth normalized to cell body area in corresponding to SY5Y cells images at different time points.

ONLINE METHODS

Description of the dataset and data availability

The whole genome and RNA-sequencing data was downloaded directly from NCBI dbGaP (https://www.ncbi.nlm.nih.gov/gap) with study-id phs000218 and accession number: Neuroblastoma (NBL)-phs000467[23]; The HumanExon arrays and Methylation arrays were obtained from the TARGET data NCI data matrix (https://target-data.nci.nih.gov/), In addition to the different TARGET datasets we used SEQC RNA-seq dataset (Gene Expression Omnibus: GSE62564) for survival analysis. The primary dataset in this study comprises 128 whole genomes sequenced tumor/blood matched pairs by Complete Genomics (CGI). CGI short read sequencing uses commercial software for processing, aligning to reference genome (hg19) and variant calling (Cancer Pipeline 2.0; http://www.completegenomics.com/documents/DataFileFormats_Cancer_Pipeline_2.0.pdf); Processed variant calls from CGI are available through TARGET data NCI pages (https://target-data.nci.nih.gov/). Additional genomic profiles from neuroblastoma available cell lines included: SNP genotyping arrays (Gene Expression Omnibus: GSE89968) and RNA-seq from 39 neuroblastoma cell lines (Gene Expression Omnibus: GSE89413)[68].

Copy number segmentation, visualization (IGV) and recurrence analysis

CGI “somaticCnvDetailsDiploidBeta” files provide information on estimated ploidy and tumor/blood coverage ratio for every 2 kb along the genome. We used in home scripts to reformat coverage data to be processed with “copynumber” R bioconductor pakage[69]. Data used Winsorization (winsorize) data smoothing and segmentation with piecewise constant segmentation (pcf) algorithm with attributes kmin=2 and gamma=1000. Segmented data was visualized with IGV. Segmented data was used as GISTIC2.0[38] input; GISTIC attributes v 30 -refgene hg19 -genegistic 1 -smallmem 1 -broad 1 -twoside 1 -brlen 0.98 -conf 0.90 -armpeel 1 -savegene 1 -gcm extreme -js 2 -rx 0.

Filtering of CGI Structural Variants calls

The CGI Cancer Pipeline 2.0 is a full report including quality control, variant calling and CNV analyses: somaticAllJunctionsBeta files provides information for individual junctions detected in tumor genome that were absent in the normal genome**.** The “highConfidenceJunctionsBeta” files contains a filtered subset of the junctions reported in “somaticAllJunctionsBeta” file. This subset comprises junctions that likely resulted from a true physical connection between the left and right sections of the junctions. To obtain the junctions reported in this file the following filter criteria was applied to the junctions in the somaticAllJunctionsBeta Include the junction if DiscordantMatePairAlignments ≥ 10 (10 or more discordant mate pairs in cluster) AND Include the junction if JunctionSequenceResolve = Y (local de novo assembly is successful) AND Exclude interchromosomal junction if present in any genomes in baseline samples (FrequencyInBaseline > 0) AND Exclude the junction if overlap with known underrepresented repeats (KnownUnderrepresentedRepeat = Y): ALR/Alpha, GAATGn, HSATII, LSU_rRNA_Hsa, and RSU_rRNA_Hsa AND Exclude the junction if the length of either of the side sections is less than 70 base pairs. Additional filtering of high confidence variants was initially applied to the whole TARGET repertoire of tumor datasets including ALL, AML, NBL, OS, CCSK and RT in order to remove duplicate junctions indicative of common variations. We added additional filters to remove rare/common germline variants that passed CGI filters as well as artifacts and low confidence variants. To this end we used the Database of Genomic Variants (DGV v. 2016-05-15, GRCh37) in order to remove SVs which reciprocal overlap with DGV annotated common events was higher than 50%. We only filtered variants which type matched in both CGI SV set and DGV database.

Annotation of Structural variants from WGS

We used RefSeq gene definitions for hg19 downloaded from UCSC (10/31/2018) in order to map Structural Variant calls to nearby genes. First, we used two approaches to map variants: numerical changes (tandem-duplications and deletions size <2Mb) containing whole genes were used to define copy number alterations. Second, we mapped breakpoints relative to gene exonic coordinates; SVs were considered 'disrupting' when either one of the breakpoints localized between transcription start and ends of any of a genes isoform and proximal when localized at within 100Kb upstream or 25Kb downstream the most distal isoform of each gene, a graphical description is represented in Figure 4b. In addition we considered 'intronic' SVs those in which both breakpoints mapped to the same intron.

Processing Copy Number segmentation from WGS

We processed cnvDetailsDiploidBeta files from the Cancer Pipeline 2.0 (CGI©) containing average normalized read-depth coverage values at every 2Kb sliding window throughout the genome. Then, tumor/blood normalized ratios were subject to piecewise constant segmentation algorithm[69] implemented in the 'copynumber' R package. The processed segmentation file is available through the TARGET data matrix (https://target-data.nci.nih.gov/).

Processing Copy Number from SNP arrays

We genotyped 914 matched patient tumor and normal samples using Illumina SNP arrays. This cohort comprised 488 of these samples were previously reported[32] and reanalyzed here. The complete datasets comprises three different genotyping Illumina chip architectures: HumanHap550, Human610-Quad HumanOmniExpress. We processed for further segmentation the common set of 316210 probes Copy number segmentation was obtained using the *SNPrank* algorithm implemented by the NEXUS® software platform for tumor samples. As a result we obtained segmentation file with 914 unique samples available at the TARGET data matrix.

Identification and Annotation of copy number amplifications, deep deletions and breakpoints from WGS and SNP datasets

Breakpoints were called from segmentation profiles. We observed frequent artifact at subtelomeric and pericentromeric regions, those regions were filtered out from the analyses. A breakpoint is called when the absolute value of the copy number log-ratio difference between contiguous segments is higher than 0.152 for SNP arrays and 0.304 for WGS; both cutoffs account for 10% and 20% copy number change respectively (i.e for diploid regions, $\Delta CN = 0.2$ and $\Delta CN = 0.4$ respectively). We called for amplifications to segments from WGS dataset with CN >=8 and deep deletions CN <= 0.5; for SNP arrays we used less stringent cutoffs for amplification (CN >= 4.5) and deep deletions (CN <= 0.9). We used different thresholds since SNP array dynamic range is narrower than that of sequencing platforms and the resolution is also lower due to large regions with low probe density. We used RefSeq gene definitions for hg19 downloaded from UCSC (October 31st, 2018) version from UCSC in order to map copy number alterations variants and breakpoints to nearby genes. Genes were considered amplified or deep deleted when all isoforms were contained within the altered segment boundaries. Breakpoints were considered 'disrupting' when the breakpoint localized between transcription start and ends of any of a genes isoform and proximal when localized at within 100Kb upstream or 25Kb downstream the most distal isoform of each gene, a graphical description is represented in Figure 4b.

Tumor mutational burden analyses

We obtained measures representative of the burden of different mutation types under study (including SNVs, SVs and BPs). To this end, the density of mutations of every type is calculated as the average number of mutations in a given sample per sequence window (10Mb for SVs and BPs and 1Mb for SNVs). Instead of a single density value per sample we measure mutational densities for each chromosomal arm, excluding short arms with very low mappability (13p, 14p, 15p, 21p, 22p and Y chromosome). The remaining 41 chromosomal arms in each sample represent single sample distributions of mutational densities from which quantiles are obtained. We used the interquartile mean (IQM) since it offered a measure robust against outliers while conserving the variability across samples even in low-density breakpoint samples.

Analysis of recurrent structural variants

We studied SV recurrence as a means of their overlap and proximity to known genes defined by RefSeq database, hg19 genome version from 10/31/2018 obtained from UCSC. Calculations of genomic overlaps among SVs between SVs and with coding genes was performed using the R package GenomicRanges [70]. Overlapping variants were considered those were junction and/or spanning segment overlap with known genes start and end coordinates. Proximal variants were defined as those affecting regions upstream and downstream of known genes; the high-confidence set considered variants at 50Kb upstream the transcription start site and 50 Kb downstream the transcription end. While the extended set incorporated variants up to 200Kb in upstream the transcription start site.

Filtering and annotation of somatic SNVs

The CGI cancer pipeline 2.0 provides somatic variant calls for SNVs and small indels. Given the gapped nature of CGI reads which leads to high noise to signal ratio we incorporated SNV filtering. We first annotated CGI SNV calls using Variant Effect Predictor (VEP) pipeline[71]. Our filter follows two steps: 1) collect high quality somatic non-synonymous coding variants (Phred like Fisher's exact test $P<0.001$) annotated as having a moderate or high functional impact; this set of variants was combine with COSMIC catalogue of pathogenic variants (release v84). 2) Hot-spot analysis of variants from our combined catalogue (step 1) to identify both clonal and low allele frequency pathogenic variants.

Gene fusion analysis

Gene fusion analysis from RNA-seq data was studied using three available tools: (STAR-fusion[33], fusionCATCHER[34] and DeFUSE [35]). We then collected fusion events that matched interchromosomal events from the CGI structural variation calls. While the three methods return ($N_{\text{STAR-fusion}}=24{,}837$, $N_{\text{fusionCATCHER}}=6{,}898$, $N_{\text{DeFUSE}}=22{,}837$) events with an overlap of 68 cases (0.1%) the subset of fusion events with matching translocation from WGS comprise 66 events ($N_{\text{STAR-fusion}}=45$, $N_{\text{fusionCATCHER}}=36$,

N_{DeFUSE}=44) with the three methods overlapping 26 (40%) events. This data shows that DNA/RNA combined evidence verification returns high precision gene fusion events (see Supplementary Fig 13b,c).

Sanger Sequence Validation

From alignment and variant calls provided by Complete Genomics, structural variation breakpoints were mapped and junction sequence was assembled using public UCSC Genome browser. The assembled sequence was then submitted into Primer3 to engineer PCR primers to bridge the translocation. Resultant primers were then checked against BLAT as well as an internal algorithm for binding specificity. PCR reactions were then carried out on 25 ng of DNA using optimized conditions for each reaction. Products were checked via gel electrophoresis for specificity to expected size and uniqueness. If the product had multiple bands, the entire remaining sample would be run out then bands of interest excised and the DNA extracted using MinElute Gel Extraction Kit from Qiagen. Products with single bands were cleaned up and prepared for sequencing using the MinElute PCR Purification Kit (Qiagen). Samples were then sequenced at a core facility with 2 picomoles of the same primer used to create amplicon. Resultant sequences were then aligned to the expected sequence assembled from CGI results and mapped to chromosomes using BLAT at UCSC genome browser.

Cell Culture

Cells were grown in RPMI-1640 with HEPES, L-glutamine, and phenol red (cat # 22400-089), supplemented with 10% Fetal Bovine Serum, 1% antibiotic-antimycotic (cat # 15240-062), and 1% L-glutamine (cat # 25-005-CI) in 5% CO_2 at 37°C in the dark. Transduced cells also had the appropriate concentration of puromycin in media for selection.

LentiVirus Infection

Lentiviral vector plasmid for the long isoform of *SHANK2* (NM_012309) was obtained commercially from GeneCopoeia (EX-H5274-Lv105). Empty vector control plasmid pLv105 was originally from

GeneCopoeia. Creation of the virus media was accomplished using Lipofectamine 3000 ™ applied to 293TN cells with packaging plasmid psPAX2, envelope plasmid pMD2.g, and the Lentiviral backbone plasmid containing the ORF for NM_012309 or empty vector. Infectious viral media was pooled over 2 days then filtered through 0.45µm nitrocellulose and combined with polybrene at 8 µg/mL media and applied to cells. Following infection, transduced cells were selected with puromycin in line-dependent concentrations.

Growth and Proliferation Assay using RT-CES

Cells were plated in 96-well RTCES microelectronic sensor arrays (ACEA Biosciences, San Deigo, CA, USA). Density measurements were made every hour. Cell densities were normalized to 5 hours post-plating.

Cell Viability Assays

Cells were plated in clear-bottomed, 96-well plates in 200 µL media and allowed to grow under normal conditions for either 4 or 7 days. Before reading, 100 µL media was replaced with equal volume of CellTiter Glo ® reagent and read on a GloMax Multi-detection instrument (Promega). Arbitrary luminescence units were normalized to empty vector-transduced controls and results expressed as percentages of control levels from the same assay.

ATRA-Induced Differentiation

Cells were plated in normal media at optimized densities for each parental line in 96-well plates and allowed 24-48 hours to firmly attach to plates. Media was then switched for low-serum media containing either 1% or 3% FBS and allowed 24 hours to equilibrate, after which it was replaced with low-serum media supplemented with varying concentrations of ATRA (all-*trans*-retinoic-acid, Sigma, R2625) or vehicle (DMSO) alone, in volume corresponding to the highest concentration of ATRA for each experiment. Plates were then left in normal growth conditions and protected from light. RA media was refreshed every 72 hours to prevent oxidation. Plates were placed in an IncuCyte ZOOM™ instrument to utilize live cell imaging. Each well was imaged every four hours and the “NeuroTrack” software module to quantify neurite outgrowth.

Protein Isolation and Western Blotting

Whole cell lysates were created by applying denaturing lysis buffer containing protease/phosphatase inhibitors (Cell Signaling Technology, 5872) to cells on ice and allowing lysis for 30 minutes. The total sample was sonicated for 5 seconds and spun at max speed in a microcentrifuge for 15 minutes at 4°C before collecting supernatant to clean tube. Quantification of protein was done using the Pierce BCA Protein Assay Kit (Thermo, 23227). Protein was loaded on 4-12% Tris-Glycine gels, transferred to PVDF membrane, and probed with antibodies in 5% milk in TBST. Antibody stripping used Restore ™ Stripping buffer (Thermo, 21059). Detection of HRP-conjugated secondary antibodies used SuperSignal™ West Femto Maximum Sensitivity Substrate (Thermo, 34096).

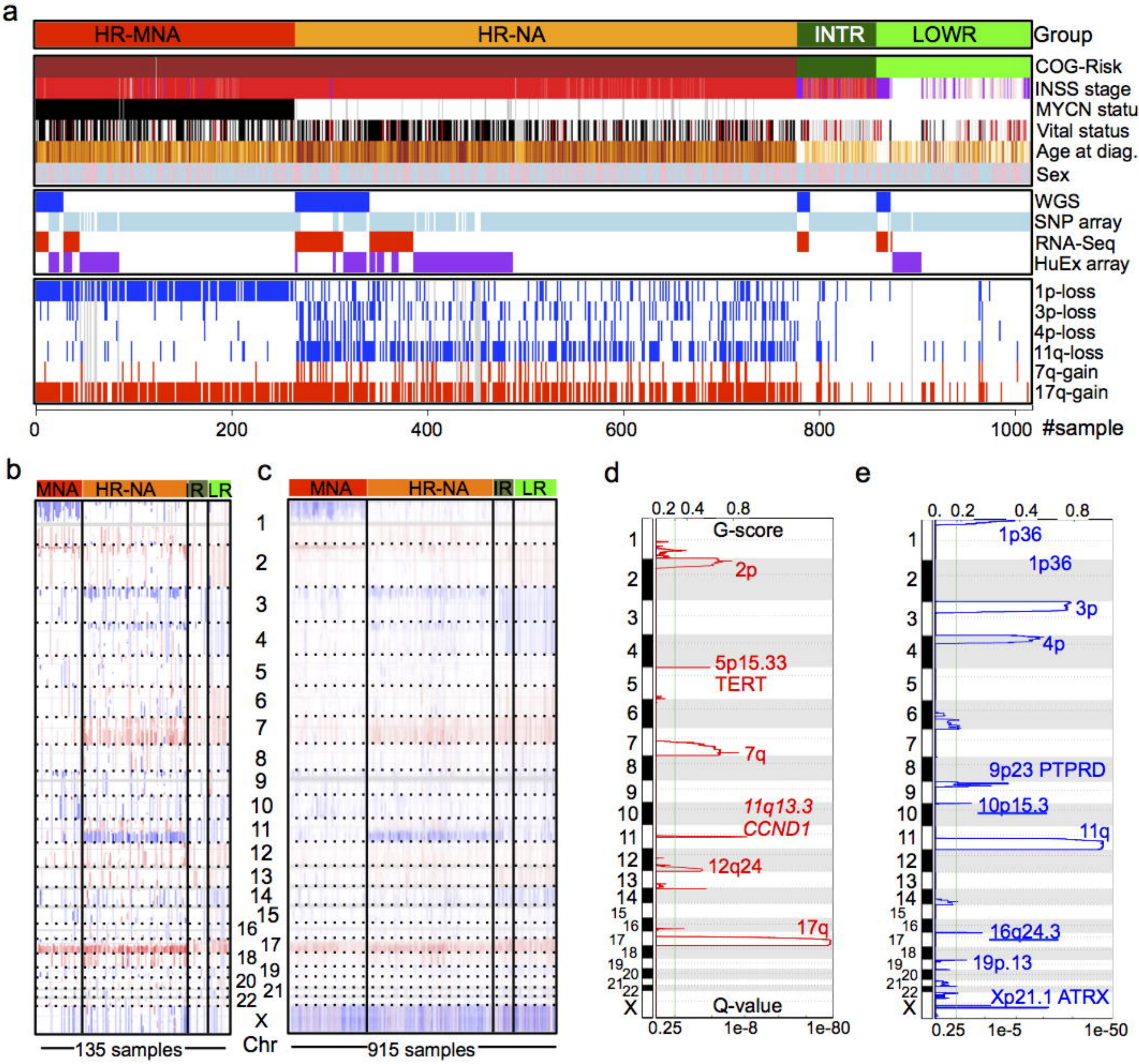
REFERENCES

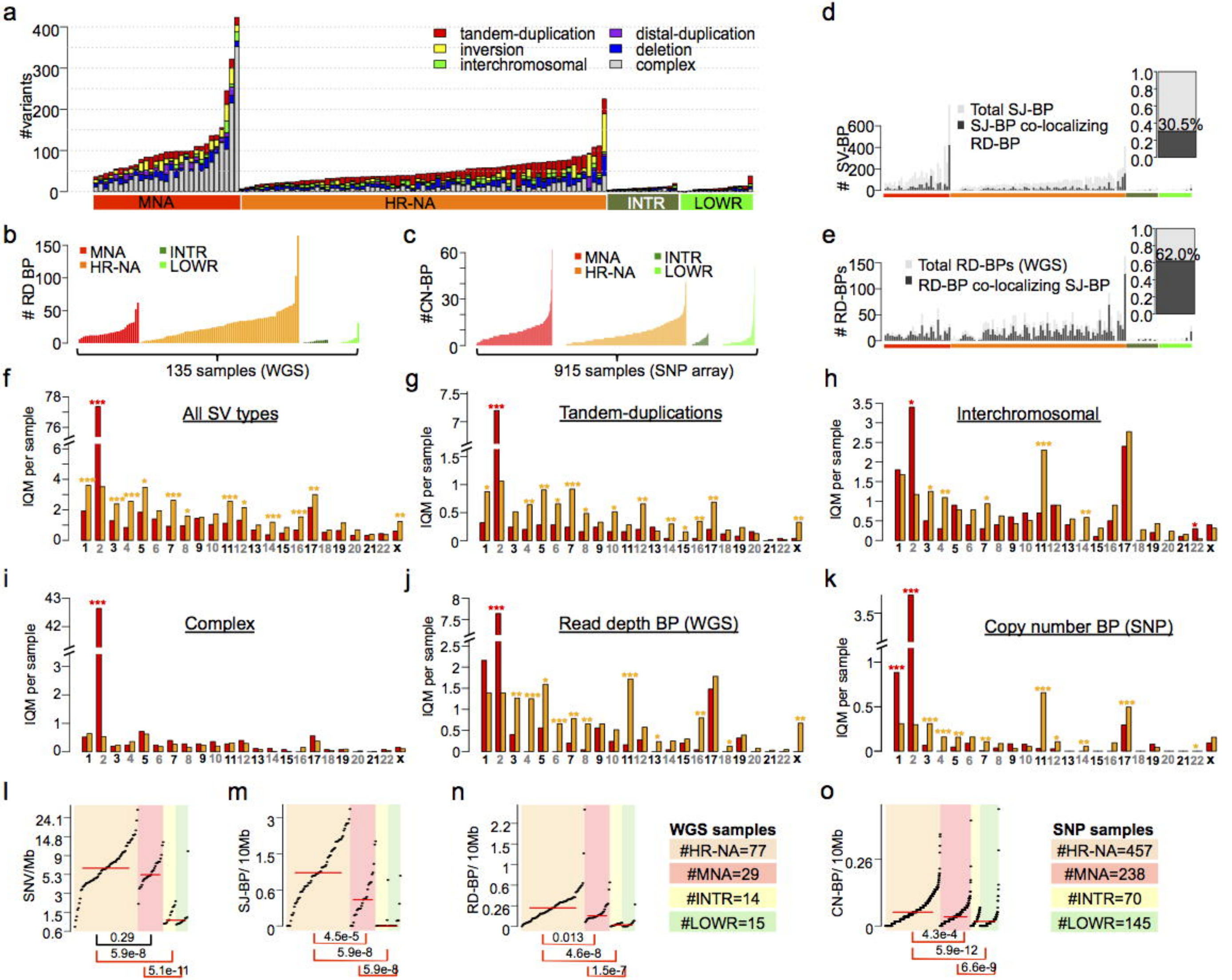
1. Maris, J.M. Recent advances in neuroblastoma. *N Engl J Med* **362**, 2202-11 (2010).
2. Cheung, N.K. *et al.* Association of age at diagnosis and genetic mutations in patients with neuroblastoma. *JAMA* **307**, 1062-71.
3. Molenaar, J.J. *et al.* Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature*.
4. Pugh, T.J. *et al.* The genetic landscape of high-risk neuroblastoma. *Nat Genet* (2013).
5. Sausen, M. *et al.* Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. *Nat Genet* **45**, 12-7.
6. Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919-29 (2013).
7. Macintyre, G., Ylstra, B. & Brenton, J.D. Sequencing Structural Variants in Cancer for Precision Therapeutics. *Trends Genet* **32**, 530-542 (2016).
8. Tubio, J.M. Somatic structural variation and cancer. *Brief Funct Genomics* **14**, 339-51 (2015).
9. Alkan, C., Coe, B.P. & Eichler, E.E. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**, 363-76 (2011).
10. Deyell, R.J. & Attiyeh, E.F. Advances in the understanding of constitutional and somatic genomic alterations in neuroblastoma. *Cancer Genet* **204**, 113-21.
11. Michels, E. *et al.* ArrayCGH-based classification of neuroblastoma into genomic subgroups. *Genes Chromosomes Cancer* **46**, 1098-108 (2007).
12. Attiyeh, E.F. *et al.* Chromosome 1p and 11q deletions and outcome in neuroblastoma. *The New England journal of medicine* **353**, 2243-2253 (2005).
13. Łastowska, M. *et al.* Breakpoint position on 17q identifies the most aggressive neuroblastoma tumors. *Genes, Chromosomes & Cancer* **34**, 428-436 (2002).
14. Maris, J.M. *et al.* Comprehensive analysis of chromosome 1p deletions in neuroblastoma. *Medical and Pediatric Oncology* **36**, 32-36 (2001).
15. Bown, N. *et al.* Gain of chromosome arm 17q and adverse outcome in patients with neuroblastoma. *The New England Journal of Medicine* **340**, 1954-1961 (1999).
16. Guo, C. *et al.* Allelic deletion at 11q23 is common in MYCN single copy neuroblastomas. *Oncogene* **18**, 4948-57 (1999).
17. Plantaz, D. *et al.* Gain of chromosome 17 is the most frequent abnormality detected in neuroblastoma by comparative genomic hybridization. *The American Journal of Pathology* **150**, 81-89 (1997).
18. Caron, H. *et al.* Allelic loss of chromosome 1p as a predictor of unfavorable outcome in patients with neuroblastoma. *The New England Journal of Medicine* **334**, 225-230 (1996).
19. Gehring, M., Berthold, F., Edler, L., Schwab, M. & Amler, L.C. The 1p deletion is not a reliable marker for the prognosis of patients with neuroblastoma. *Cancer Research* **55**, 5366-5369 (1995).
20. Seeger, R.C. *et al.* Association of multiple copies of the N-myc oncogene with rapid progression of neuroblastomas. *The New England journal of medicine* **313**, 1111-1116 (1985).

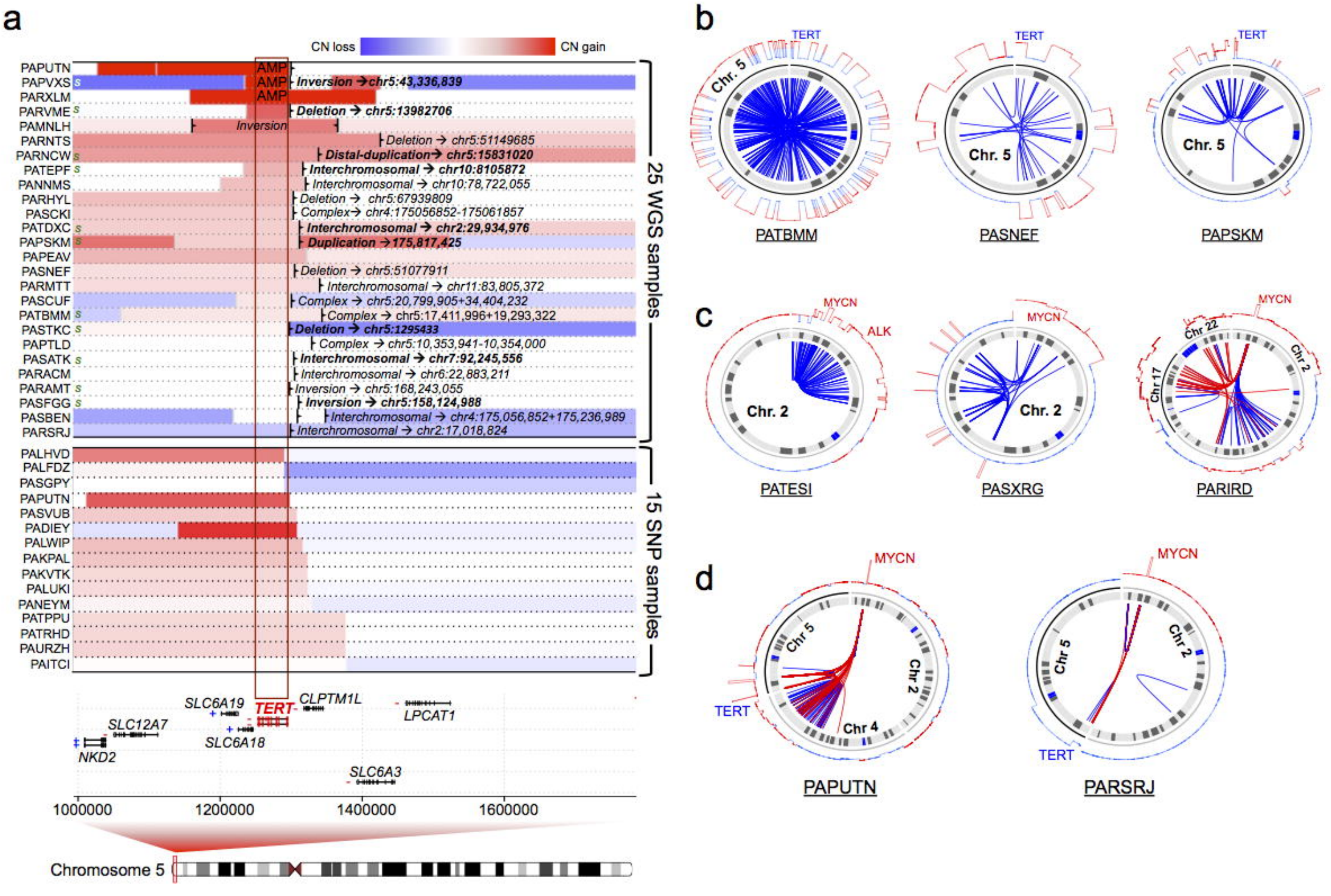
21. Brodeur, G.M., Seeger, R.C., Schwab, M., Varmus, H.E. & Bishop, J.M. Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. *Science (New York, N.Y.)* **224**, 1121-1124 (1984).
22. Gilbert, F. *et al.* Human Neuroblastomas and Abnormalities of Chromosomes 1 and 17. *Cancer Research* **44**, 5444-5449 (1984).
23. Pugh, T.J. *et al.* The genetic landscape of high-risk neuroblastoma. *Nat Genet* **45**, 279-84 (2013).
24. Kawashima, M., Kojima, M., Ueda, Y., Kurihara, S. & Hiyama, E. Telomere biology including TERT rearrangements in neuroblastoma: a useful indicator for surgical treatments. *J Pediatr Surg* **51**, 2080-2085 (2016).
25. Stallings, R.L. *et al.* High-resolution analysis of chromosomal breakpoints and genomic instability identifies PTPRD as a candidate tumor suppressor gene in neuroblastoma. *Cancer Res* **66**, 3673-80 (2006).
26. Sausen, M. *et al.* Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. *Nat Genet* **45**, 12-7 (2013).
27. Peifer, M. *et al.* Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature* **526**, 700-4 (2015).
28. Valentijn, L.J. *et al.* TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors. *Nat Genet* **47**, 1411-4 (2015).
29. Zimmerman, M.W. *et al.* MYC Drives a Subset of High-Risk Pediatric Neuroblastomas and Is Activated through Mechanisms Including Enhancer Hijacking and Focal Enhancer Amplification. *Cancer Discov* **8**, 320-335 (2018).
30. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81 (2010).
31. Carnevali, P. *et al.* Computational techniques for human genome resequencing using mated gapped reads. *J Comput Biol* **19**, 279-92 (2012).
32. Attiyeh, E.F. *et al.* Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. *Genome Res* **19**, 276-83 (2009).
33. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
34. Daniel Nicorici, M.S., Henrik Edgren, Sara Kangaspeska, Astrid Murumagi, Olli Kallioniemi, Sami Virtanen, Olavi Kilkku. FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv* (2014).
35. McPherson, A. *et al.* deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol* **7**, e1001138 (2011).
36. Cohn, S.L. *et al.* The International Neuroblastoma Risk Group (INRG) classification system: an INRG Task Force report. *J Clin Oncol* **27**, 289-97 (2009).
37. Wang, Q. *et al.* Integrative genomics identifies distinct molecular classes of neuroblastoma and shows that multiple genes are targeted by regional alterations in DNA copy number. *Cancer Res* **66**, 6050-62 (2006).
38. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).
39. Mosse, Y.P. *et al.* High-resolution detection and mapping of genomic DNA alterations in neuroblastoma. *Genes Chromosomes Cancer* **43**, 390-403 (2005).

40. Bosse, K.R. *et al.* Identification of GPC2 as an Oncoprotein and Candidate Immunotherapeutic Target in High-Risk Neuroblastoma. *Cancer Cell* **32**, 295-309 e12 (2017).
41. MacDonald, J.R., Ziman, R., Yuen, R.K., Feuk, L. & Scherer, S.W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* **42**, D986-92 (2014).
42. Schleiermacher, G. *et al.* Segmental chromosomal alterations have prognostic impact in neuroblastoma: a report from the INRG project. *Br J Cancer* **107**, 1418-22 (2012).
43. Molenaar, J.J. *et al.* Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* **483**, 589-93 (2012).
44. Caren, H. *et al.* High-risk neuroblastoma tumors with 11q-deletion display a poor prognostic, chromosome instability phenotype with later onset. *Proc Natl Acad Sci U S A* **107**, 4323-8 (2010).
45. Maher, C.A. & Wilson, R.K. Chromothripsis and human disease: piecing together the shattering process. *Cell* **148**, 29-32 (2012).
46. Shao, Y.W. *et al.* Cross-species genomics identifies DLG2 as a tumor suppressor in osteosarcoma. *Oncogene* **38**, 291-298 (2019).
47. Smida, J. *et al.* Genome-wide analysis of somatic copy number alterations and chromosomal breakages in osteosarcoma. *Int J Cancer* **141**, 816-828 (2017).
48. Kaizuka, T. & Takumi, T. Postsynaptic density proteins and their involvement in neurodevelopmental disorders. *J Biochem* **163**, 447-455 (2018).
49. Clark, O., Schmidt, F., Coles, C.H., Tchetchelnitski, V. & Stoker, A.W. Functional analysis of the putative tumor suppressor PTPRD in neuroblastoma cells. *Cancer Invest* **30**, 422-32 (2012).
50. Cheung, N.K. *et al.* Association of age at diagnosis and genetic mutations in patients with neuroblastoma. *JAMA* **307**, 1062-71 (2012).
51. Mosse, Y.P. *et al.* Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature* **455**, 930-5 (2008).
52. Santo, E.E. *et al.* Oncogenic activation of FOXR1 by 11q23 intrachromosomal deletion-fusions in neuroblastoma. *Oncogene* **31**, 1571-81 (2012).
53. Chen, J., Bardes, E.E., Aronow, B.J. & Jegga, A.G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* **37**, W305-11 (2009).
54. 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**, 15545-50 (2005).
55. Wang, C. *et al.* The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat Biotechnol* **32**, 926-32 (2014).
56. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* **45**, 1127-33 (2013).
57. Molenaar, J.J. *et al.* Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* **483**, 589-593 (2012).
58. Goffin, A., Hoefsloot, L.H., Bosgoed, E., Swillen, A. & Fryns, J.P. PTEN mutation in a family with Cowden syndrome and autism. *Am J Med Genet* **105**, 521-4 (2001).
59. Crawley, J.N., Heyer, W.D. & LaSalle, J.M. Autism and Cancer Share Risk Genes, Pathways, and Drug Targets. *Trends Genet* **32**, 139-146 (2016).

60. DeScipio, C. *et al.* Subtelomeric deletion of chromosome 10p15.3: clinical findings and molecular cytogenetic characterization. *Am J Med Genet A* **158A**, 2152-61 (2012).
61. Willemsen, M.H. *et al.* Identification of ANKRD11 and ZNF778 as candidate genes for autism and variable cognitive impairment in the novel 16q24.3 microdeletion syndrome. *Eur J Hum Genet* **18**, 429-35 (2010).
62. Gai, X. *et al.* Rare structural variation of synapse and neurotransmission genes in autism. *Mol Psychiatry* **17**, 402-11 (2012).
63. Westermark, U.K., Wilhelm, M., Frenzel, A. & Henriksson, M.A. The MYCN oncogene and differentiation in neuroblastoma. *Semin Cancer Biol* **21**, 256-66 (2011).
64. Fredlund, E., Ringner, M., Maris, J.M. & Pahlman, S. High Myc pathway activity and low stage of neuronal differentiation associate with poor outcome in neuroblastoma. *Proc Natl Acad Sci U S A* **105**, 14094-9 (2008).
65. Sulong, S. *et al.* A comprehensive analysis of the CDKN2A gene in childhood acute lymphoblastic leukemia reveals genomic deletion, copy number neutral loss of heterozygosity, and association with specific cytogenetic subgroups. *Blood* **113**, 100-7 (2009).
66. Matthay, K.K. *et al.* Long-term results for children with high-risk neuroblastoma treated on a randomized trial of myeloablative therapy followed by 13-cis-retinoic acid: a children's oncology group study. *J Clin Oncol* **27**, 1007-13 (2009).
67. Matthay, K.K. *et al.* Treatment of high-risk neuroblastoma with intensive chemotherapy, radiotherapy, autologous bone marrow transplantation, and 13-cis-retinoic acid. Children's Cancer Group. *N Engl J Med* **341**, 1165-73 (1999).
68. Harenza, J.L. *et al.* Transcriptomic profiling of 39 commonly-used neuroblastoma cell lines. *Sci Data* **4**, 170033 (2017).
69. Nilsen, G. *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
70. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**, e1003118 (2013).
71. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).


a
HR-MNA
HR-NA
INTR
LOWR
Group
COG-Risk
INSS stage
MYCN status
Vital status
Age at diag.
Sex
WGS
SNP array
RNA-Seq
HuEx array
1p-loss
3p-loss
4p-loss
11q-loss
7q-gain
17q-gain
0
200
400
600
800
1000
#sample
b
MNA
HR-NA
IR
LR
135 samples
Chr
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
c
MNA
HR-NA
IR
LR
915 samples
d
0.2
0.4
0.8
G-score
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
2p
5p15.33
TERT
7q
11q13.3
CCND1
12q24
17q
Q-value
0.25
1e-8
1e-80
e
0.
0.2
0.4
0.8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
1p36
1p36
3p
4p
9p23 PTPRD
10p15.3
11q
16q24.3
19p.13
Xp21.1 ATRX
0.25
1e-5
1e-50


a
#variants
400
300
200
100
0
tandem-duplication
inversion
interchromosomal
distal-duplication
deletion
complex
MNA
HR-NA
INTR
LOWR
d
SV-BP
600
400
200
0
Total SJ-BP
SJ-BP co-localizing RD-BP
1.0
0.8
0.6
0.4
0.2
0.0
30.5%
b
RD BP
150
100
50
0
MNA
HR-NA
INTR
LOWR
135 samples (WGS)
c
#CN-BP
60
30
0
MNA
HR-NA
INTR
LOWR
915 samples (SNP array)
e
RD-BPs
150
100
50
0
Total RD-BPs (WGS)
RD-BP co-localizing SJ-BP
1.0
0.8
0.6
0.4
0.2
0.0
62.0%
f
All SV types
IQM per sample
78
76
6
4
2
0
g
Tandem-duplications
IQM per sample
7.5
7
1.5
1
0.5
0
h
Interchromosomal
IQM per sample
3.5
3
2.5
2
1.5
1
0.5
0
i
Complex
IQM per sample
43
42
3
2
1
0
j
Read depth BP (WGS)
IQM per sample
8
7.5
2.5
2
1.5
1
0.5
0
k
Copy number BP (SNP)
IQM per sample
3.5
1
0.5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
*
**
***
l
SNV/Mb
24.1
14.8
9
5.3
3
1.5
0.6
0.29
5.9e-8
5.1e-11
m
SJ-BP/ 10Mb
3
1.5
0.6
0
4.5e-5
5.9e-8
5.9e-8
n
RD-BP/ 10Mb
2.2
1.5
1
0.6
0.26
0
0.013
4.6e-8
1.5e-7
WGS samples
#HR-NA=77
#MNA=29
#INTR=14
#LOWR=15
o
CN-BP/ 10Mb
0.26
0
4.3e-4
5.9e-12
6.6e-9
SNP samples
#HR-NA=457
#MNA=238
#INTR=70
#LOWR=145

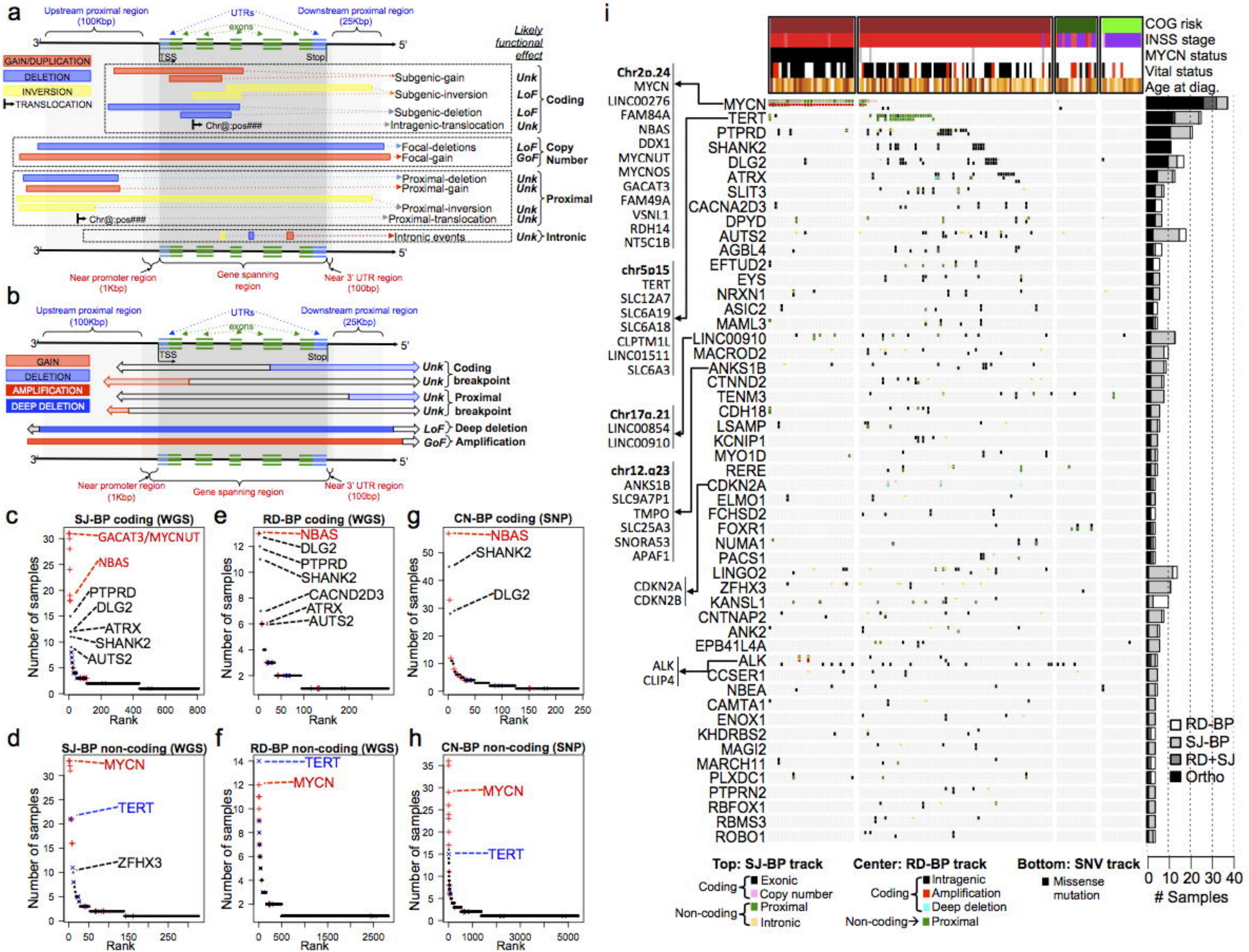


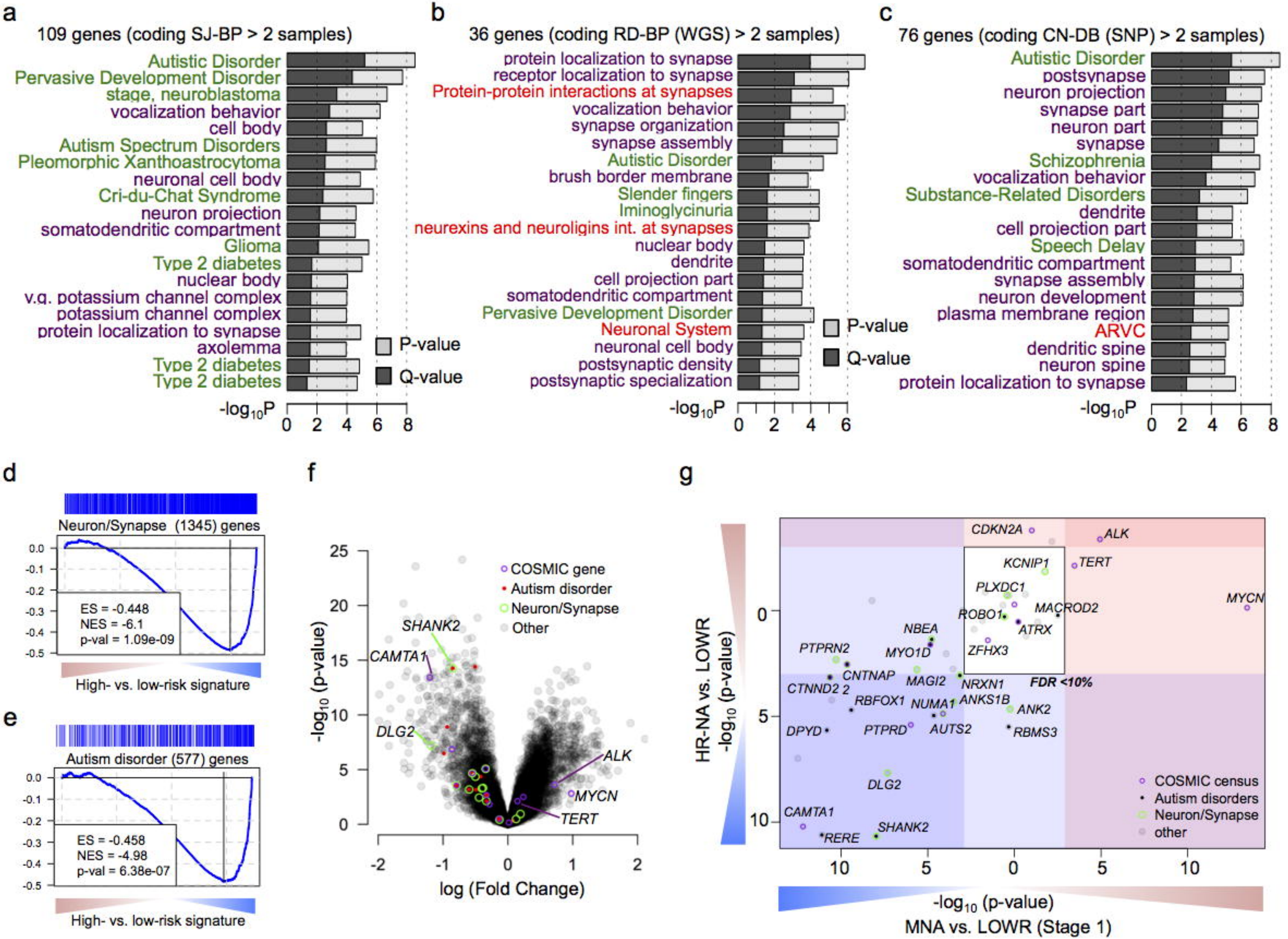
e

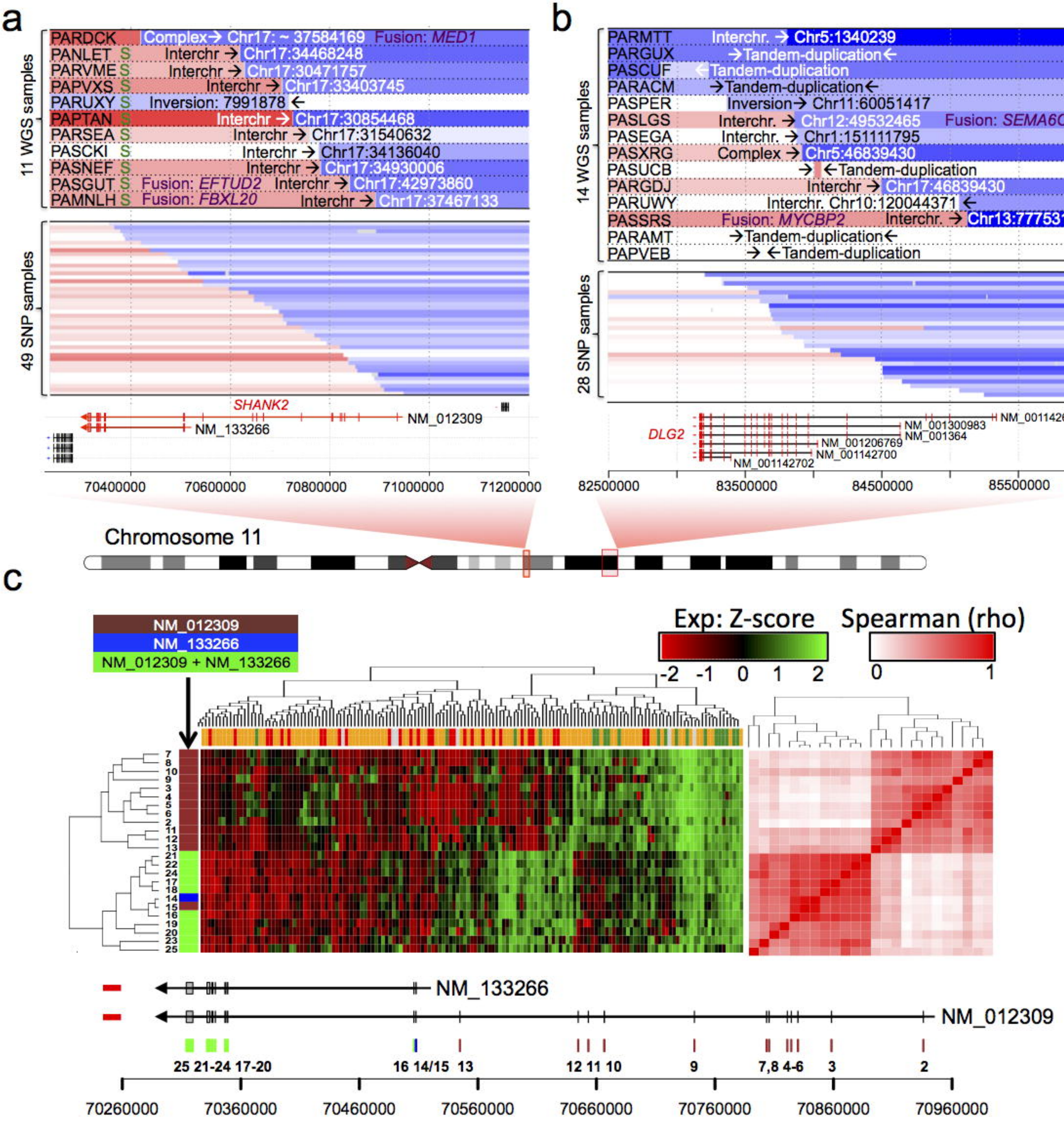
	Chr1	Chr2	Chr4	Chr5	Chr10	Chr11	Chr17	Chr19	Chr22	ChrX
HR_MNA (29)	0	7 ***	1	2	0	0	1	0	1	0
HR_NA (77)	2	1	0	7	1	2	0	1	0	1
TERT SV+(25)	1	2	1	8 ***	1	0	0	1	0	0
TERT WT (81)	1	6	0	1	0	2	1	0	1	1
Male (65)	1	7	1	4	1	1	1	1	1	0
Female (41)	1	1	0	4	0	1	0	0	0	1

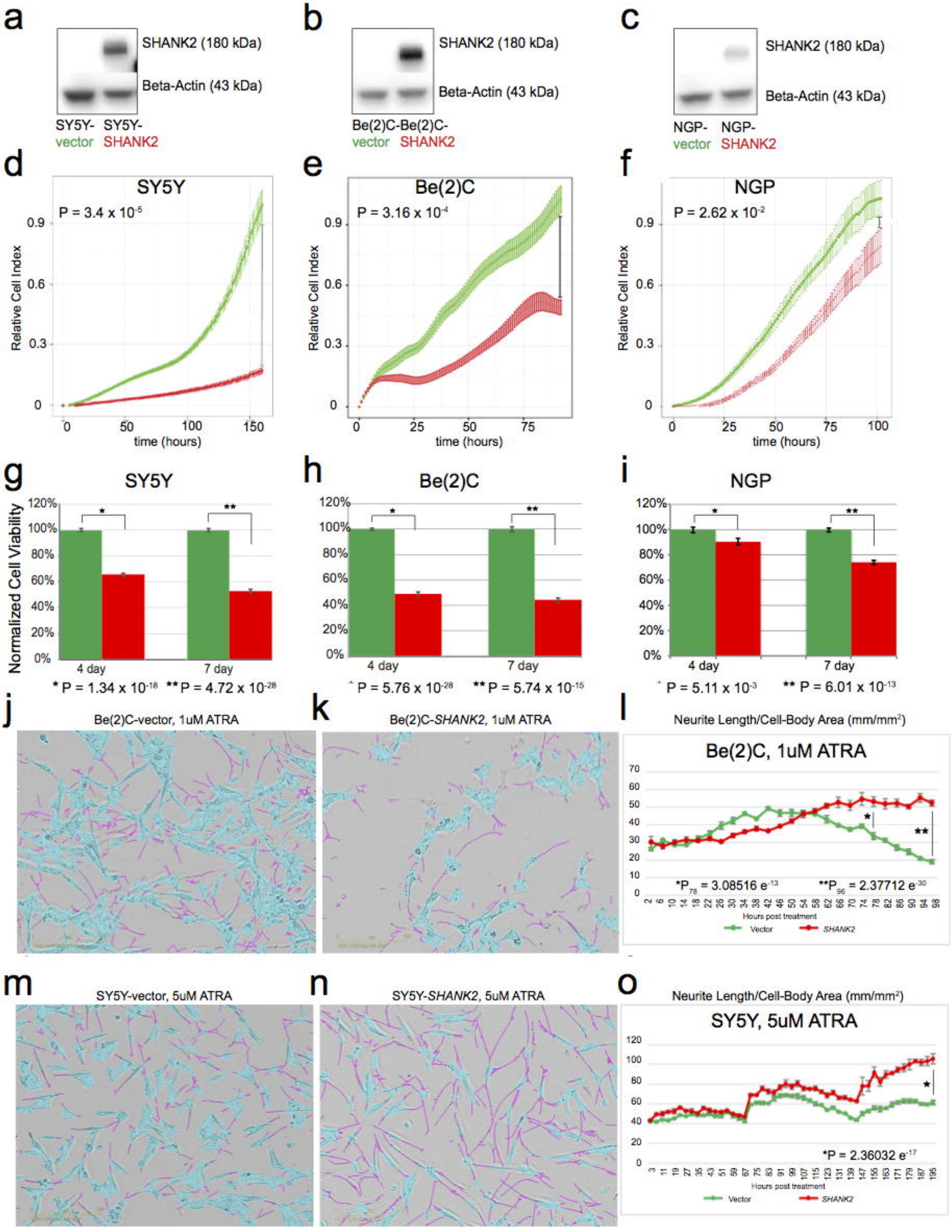
f

	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr11	chr12	chr13	chr15	chr17	chr18	chr19	chrX
HR_MNA (239)	2	46 ***	1	0	1	1	1	1	2	1	0	1	1	1	1	1
HR_NA (428)	2	0	3	4	10	2	2	1	5	5	2	1	2	1	3	6
TERT SV+(30)	0	0	0	1	7 ***	0	0	0	1	0	0	0	0	0	0	1
TERT Unk (670)	4	47	4	3	4	3	3	2	6	6	2	2	3	2	4	6
Male (401)	4	27	2	1	6	1	1	2	3	3	2	1	1	1	2	1 *
Female (299)	0	20	2	3	5	2	2	0	4	3	0	1	2	1	2	6


a
Upstream proximal region (100Kbp)
UTRs
exons
Downstream proximal region (25Kbp)
3'
5'
TSS
Stop
Likely functional effect
GAIN/DUPLICATION
DELETION
INVERSION
TRANSLOCATION
Subgenic-gain
Subgenic-inversion
Subgenic-deletion
Chr@:pos###
Intragenic-translocation
Unk
LoF
LoF
Unk
Coding
Focal-deletions
Focal-gain
LoF
GoF
Copy Number
Proximal-deletion
Proximal-gain
Proximal-inversion
Chr@:pos###
Proximal-translocation
Unk
Unk
Unk
Unk
Proximal
Intronic events
Unk
Intronic
Near promoter region (1Kbp)
Gene spanning region
Near 3' UTR region (100bp)
b
Upstream proximal region (100Kbp)
UTRs
exons
Downstream proximal region (25Kbp)
TSS
Stop
GAIN
DELETION
AMPLIFICATION
DEEP DELETION
Unk
Unk
Coding breakpoint
Unk
Unk
Proximal breakpoint
LoF
Deep deletion
GoF
Amplification
Near promoter region (1Kbp)
Gene spanning region
Near 3' UTR region (100bp)
c
SJ-BP coding (WGS)
Number of samples
Rank
GACAT3/MYCNUT
NBAS
PTPRD
DLG2
ATRX
SHANK2
AUTS2
d
SJ-BP non-coding (WGS)
MYCN
TERT
ZFHX3
e
RD-BP coding (WGS)
NBAS
DLG2
PTPRD
SHANK2
CACND2D3
ATRX
AUTS2
f
RD-BP non-coding (WGS)
TERT
MYCN
g
CN-BP coding (SNP)
NBAS
SHANK2
DLG2
h
CN-BP non-coding (SNP)
MYCN
TERT
i
COG risk
INSS stage
MYCN status
Vital status
Age at diag.
Chr2p.24
MYCN
LINC00276
FAM84A
NBAS
DDX1
MYCNUT
MYCNOS
GACAT3
FAM49A
VSNL1
RDH14
NT5C1B
chr5p15
TERT
SLC12A7
SLC6A19
SLC6A18
CLPTM1L
LINC01511
SLC6A3
Chr17q.21
LINC00854
LINC00910
chr12.q23
ANKS1B
SLC9A7P1
TMPO
SLC25A3
SNORA53
APAF1
CDKN2A
CDKN2B
ALK
CLIP4
MYCN
TERT
PTPRD
SHANK2
DLG2
ATRX
SLIT3
CACNA2D3
DPYD
AUTS2
AGBL4
EFTUD2
EYS
NRXN1
ASIC2
MAML3
LINC00910
MACROD2
ANKS1B
CTNND2
TENM3
CDH18
LSAMP
KCNIP1
MYO1D
RERE
CDKN2A
ELMO1
FCHSD2
FOXR1
NUMA1
PACS1
LINGO2
ZFHX3
KANSL1
CNTNAP2
ANK2
EPB41L4A
ALK
CCSER1
NBEA
CAMTA1
ENOX1
KHDRBS2
MAGI2
MARCH11
PLXDC1
PTPRN2
RBFOX1
RBMS3
ROBO1
RD-BP
SJ-BP
RD+SJ
Ortho
Samples
Top: SJ-BP track
Coding
Exonic
Copy number
Non-coding
Proximal
Intronic
Center: RD-BP track
Coding
Intragenic
Amplification
Deep deletion
Non-coding
Proximal
Bottom: SNV track
Missense mutation


a
109 genes (coding SJ-BP > 2 samples)
Autistic Disorder
Pervasive Development Disorder
stage, neuroblastoma
vocalization behavior
cell body
Autism Spectrum Disorders
Pleomorphic Xanthoastrocytoma
neuronal cell body
Cri-du-Chat Syndrome
neuron projection
somatodendritic compartment
Glioma
Type 2 diabetes
nuclear body
v.g. potassium channel complex
potassium channel complex
protein localization to synapse
axolemma
Type 2 diabetes
Type 2 diabetes
P-value
Q-value
-log$_{10}$P
0
2
4
6
8
b
36 genes (coding RD-BP (WGS) > 2 samples)
protein localization to synapse
receptor localization to synapse
Protein-protein interactions at synapses
vocalization behavior
synapse organization
synapse assembly
Autistic Disorder
brush border membrane
Slender fingers
Iminoglycinuria
neurexins and neuroligins int. at synapses
nuclear body
dendrite
cell projection part
somatodendritic compartment
Pervasive Development Disorder
Neuronal System
neuronal cell body
postsynaptic density
postsynaptic specialization
P-value
Q-value
-log$_{10}$P
0
2
4
6
c
76 genes (coding CN-DB (SNP) > 2 samples)
Autistic Disorder
postsynapse
neuron projection
synapse part
neuron part
synapse
Schizophrenia
vocalization behavior
Substance-Related Disorders
dendrite
cell projection part
Speech Delay
somatodendritic compartment
synapse assembly
neuron development
plasma membrane region
ARVC
dendritic spine
neuron spine
protein localization to synapse
-log$_{10}$P
0
2
4
6
8
d
Neuron/Synapse (1345) genes
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
ES = -0.448
NES = -6.1
p-val = 1.09e-09
High- vs. low-risk signature
e
Autism disorder (577) genes
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
ES = -0.458
NES = -4.98
p-val = 6.38e-07
High- vs. low-risk signature
f
COSMIC gene
Autism disorder
Neuron/Synapse
Other
SHANK2
CAMTA1
DLG2
ALK
MYCN
TERT
-log$_{10}$ (p-value)
25
20
15
10
5
0
-2
-1
0
1
2
log (Fold Change)
g
CDKN2A
ALK
TERT
KCNIP1
PLXDC1
MYCN
ROBO1
MACROD2
ATRX
NBEA
ZFHX3
PTPRN2
MYO1D
CNTNAP
MAGI2
NRXN1
CTNND2 2
FDR <10%
RBFOX1
NUMA1
ANKS1B
ANK2
DPYD
PTPRD
AUTS2
RBMS3
DLG2
CAMTA1
SHANK2
RERE
COSMIC census
Autism disorders
Neuron/Synapse
other
HR-NA vs. LOWR
-log$_{10}$ (p-value)
0
5
10
10
5
0
5
10
-log$_{10}$ (p-value)
MNA vs. LOWR (Stage 1)


a
11 WGS samples
PARDCK Complex→ Chr17: ~ 37584169 Fusion: MED1
PANLET S Interchr → Chr17:34468248
PARVME S Interchr → Chr17:30471757
PAPVXS S Interchr → Chr17:33403745
PARUXY S Inversion: 7991878 ←
PAPTAN S Interchr → Chr17:30854468
PARSEA S Interchr → Chr17:31540632
PASCKI S Interchr → Chr17:34136040
PASNEF S Interchr → Chr17:34930006
PASGUT S Fusion: EFTUD2 Interchr → Chr17:42973860
PAMNLH S Fusion: FBXL20 Interchr → Chr17:37467133
49 SNP samples
SHANK2
NM_012309
NM_133266
70400000
70600000
70800000
71000000
71200000
b
14 WGS samples
PARMTT Interchr. → Chr5:1340239
PARGUX →Tandem-duplication←
PASCUF ←Tandem-duplication
PARACM →Tandem-duplication←
PASPER Inversion→ Chr11:60051417
PASLGS Interchr. → Chr12:49532465 Fusion: SEMA6C
PASEGA Interchr. → Chr1:151111795
PASXRG Complex → Chr5:46839430
PASUCB → ←Tandem-duplication
PARGDJ Interchr → Chr17:46839430
PARUWY Interchr. Chr10:120044371 ←
PASSRS Fusion: MYCBP2 Interchr. → Chr13:77753129
PARAMT →Tandem-duplication←
PAPVEB → ←Tandem-duplication
28 SNP samples
DLG2
NM_001142699
NM_001300983
NM_001364
NM_001206769
NM_001142700
NM_001142702
82500000
83500000
84500000
85500000
Chromosome 11
c
NM_012309
NM_133266
NM_012309 + NM_133266
Exp: Z-score
-2 -1 0 1 2
Spearman (rho)
0 1
7
8
10
9
3
4
5
6
2
11
12
13
21
22
24
17
18
14
15
16
19
20
23
25
NM_133266
NM_012309
25 21-24 17-20
16 14/15 13
12 11 10
9
7,8 4-6
3
2
70260000
70360000
70460000
70560000
70660000
70760000
70860000
70960000


a
SHANK2 (180 kDa)
Beta-Actin (43 kDa)
SY5Y-vector
SY5Y-SHANK2
b
SHANK2 (180 kDa)
Beta-Actin (43 kDa)
Be(2)C-vector
Be(2)C-SHANK2
c
SHANK2 (180 kDa)
Beta-Actin (43 kDa)
NGP-vector
NGP-SHANK2
d
SY5Y
P = 3.4 x 10-5
Relative Cell Index
0
0.3
0.6
0.9
0
50
100
150
time (hours)
e
Be(2)C
P = 3.16 x 10-4
Relative Cell Index
0
0.3
0.6
0.9
0
25
50
75
time (hours)
f
NGP
P = 2.62 x 10-2
Relative Cell Index
0
0.3
0.6
0.9
0
25
50
75
100
time (hours)
g
SY5Y
Normalized Cell Viability
0%
20%
40%
60%
80%
100%
120%
4 day
7 day
*
**
* P = 1.34 x 10-18
** P = 4.72 x 10-28
h
Be(2)C
4 day
7 day
*
**
* P = 5.76 x 10-28
** P = 5.74 x 10-15
i
NGP
4 day
7 day
*
**
* P = 5.11 x 10-3
** P = 6.01 x 10-13
j
Be(2)C-vector, 1uM ATRA
k
Be(2)C-SHANK2, 1uM ATRA
l
Neurite Length/Cell-Body Area (mm/mm2)
Be(2)C, 1uM ATRA
0
10
20
30
40
50
60
70
*P78 = 3.08516 e-13
**P96 = 2.37712 e-30
2 6 10 14 18 22 26 30 34 38 42 46 50 54 58 62 66 70 74 78 82 86 90 94 98
Hours post treatment
Vector
SHANK2
m
SY5Y-vector, 5uM ATRA
n
SY5Y-SHANK2, 5uM ATRA
o
Neurite Length/Cell-Body Area (mm/mm2)
SY5Y, 5uM ATRA
0
20
40
60
80
100
120
*
*P = 2.36032 e-17
3 11 19 27 35 43 51 59 67 75 83 91 99 107 115 123 131 139 147 155 163 171 179 187 195
Hours post treatment
Vector
SHANK2