nature genetics

The genetic landscape of high-risk neuroblastoma

Trevor J Pugh[1–3,19], Olena Morozova[4,5,19], Edward F Attiyeh[6–8], Shahab Asgharzadeh[9–11], Jun S Wei[12], Daniel Auclair[1], Scott L Carter[1], Kristian Cibulskis[1], Megan Hanna[1,3], Adam Kiezun[1], Jaegil Kim[1], Michael S Lawrence[1], Lee Lichenstein[1], Aaron McKenna[1], Chandra Sekhar Pedamallu[1,3], Alex H Ramos[1,2], Erica Shefler[1], Andrey Sivachenko[1], Carrie Sougnez[1], Chip Stewart[1], Adrian Ally[4], Inanc Birol[4], Readman Chiu[4], Richard D Corbett[4], Martin Hirst[4], Shaun D Jackman[4], Baljit Kamoh[4], Alireza Hadj Khodabakshi[4], Martin Krzywinski[4], Allan Lo[4], Richard A Moore[4], Karen L Mungall[4], Jenny Qian[4], Angela Tam[4], Nina Thiessen[4], Yongjun Zhao[4], Kristina A Cole[6–8], Maura Diamond[6–8], Sharon J Diskin[6–8], Yael P Mosse[6–8], Andrew C Wood[6–8], Lingyun Ji[9–11], Richard Sposto[9–11], Thomas Badgett[12], Wendy B London[2,13], Yvonne Moyer[14,15], Julie M Gastier-Foster[14,15], Malcolm A Smith[16], Jaime M Guidry Auvil[17], Daniela S Gerhard[17], Michael D Hogarty[6–8], Steven J M Jones[4], Eric S Lander[1], Stacey B Gabriel[1], Gad Getz[1], Robert C Seeger[9–11], Javed Khan[12], Marco A Marra[4,5], Matthew Meyerson[1–3] & John M Maris[6–8,18]

Neuroblastoma is a malignancy of the developing sympathetic nervous system that often presents with widespread metastatic disease, resulting in survival rates of less than 50%. To determine the spectrum of somatic mutation in high-risk neuroblastoma, we studied 240 affected individuals (cases) using a combination of whole-exome, genome and transcriptome sequencing as part of the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative. Here we report a low median exonic mutation frequency of 0.60 per Mb (0.48 nonsilent) and notably few recurrently mutated genes in these tumors. Genes with significant somatic mutation frequencies included *ALK* (9.2% of cases), *PTPN11* (2.9%), *ATRX* (2.5%, and an additional 7.1% had focal deletions), *MYCN* (1.7%, causing a recurrent p.Pro44Leu alteration) and *NRAS* (0.83%). Rare, potentially pathogenic germline variants were significantly enriched in *ALK, CHEK2, PINK1* and *BARD1*. The relative paucity of recurrent somatic mutations in neuroblastoma challenges current therapeutic strategies that rely on frequently altered oncogenic drivers.

Neuroblastoma is an embryonal malignancy of early childhood with a poor prognosis for individuals diagnosed at over 18 months of age with disseminated disease, accounting for 12% of all childhood cancer-related deaths[1,2]. Despite multimodal chemotherapeutic and immunotherapeutic strategies that have improved the survival of individuals with high-risk disease[3,4], a disproportionate number of these individuals will die or suffer profound treatment-related morbidity[5]. New therapeutic approaches are needed to improve cure rates while minimizing toxicity.

Highly penetrant, heritable mutations in *ALK* or *PHOX2B* account for the majority of familial neuroblastomas[6–9]. In individuals with sporadic disease, genome-wide association studies have identified multiple DNA polymorphisms influencing neuroblastoma susceptibility and clinical phenotype[10–15]. Somatically acquired amplification of *MYCN* and hemizygous deletions of chromosomes 1p and 11q are highly recurrent and associated with poor prognosis[16]. Although the chromosomal aberrations are useful as prognostic biomarkers of patient outcome, there remain few known oncogenic drivers of the malignant process.

Three recent studies have together reported genome or exome sequence analysis of 162 cases with neuroblastoma[17–19]. Molenaar and colleagues[17] reported an overall low coding somatic mutation count (12 per tumor), few recurrent mutations beyond those in *ALK* (7% of cases) and *TIAM1* (3%), a high frequency of chromothripsis in stage 3 and 4 tumors (18%) and frequent mutation of RAC-RHO

[1]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Harvard Medical School, Boston, Massachusetts, USA. [3]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [4]Genome Sciences Centre, British Columbia Cancer Agency, University of British Columbia, Vancouver, British Columbia, Canada. [5]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada. [6]Division of Oncology, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [7]Center for Childhood Cancer Research, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [8]Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, USA. [9]Division of Hematology/Oncology, The Children's Hospital Los Angeles, Los Angeles, California, USA. [10]Saban Research Institute, The Children's Hospital Los Angeles, Los Angeles, California, USA. [11]Keck School of Medicine, University of Southern California, Los Angeles, California, USA. [12]Pediatric Oncology Branch, Oncogenomics Section, Center for Cancer Research, National Institutes of Health, Gaithersburg, Maryland, USA. [13]Children's Hospital Boston/Dana-Farber Cancer Institute and Children's Oncology Group, Boston, Massachusetts, USA. [14]Biopathology Center, Nationwide Children's Hospital, Columbus, Ohio, USA. [15]The Ohio State University College of Medicine, Columbus, Ohio, USA. [16]Cancer Therapy Evaluation Program, National Cancer Institute, Bethesda, Maryland, USA. [17]Office of Cancer Genomics, National Cancer Institute, Bethesda, Maryland, USA. [18]Abramson Family Cancer Research Institute, Philadelphia, Pennsylvania, USA. [19]These authors contributed equally to this work. Correspondence should be addressed to M.A.M. (mmarra@bcgsc.ca), M.M. (matthew_meyerson@dfci.harvard.edu) or J.M.M. (maris@chop.edu).

Received 27 September 2012; accepted 20 December 2012; published online 20 January 2013; doi:10.1038/ng.2529

© 2013 Nature America, Inc. All rights reserved.

pathway genes regulating neuritogenesis. Cheung and colleagues[18] found *ATRX* loss-of-function mutations and deletions associated with neuroblastoma in adolescents and young adults. Sausen and colleagues[19] uncovered recurrent mutation or focal deletion of *ARID1A* and *ARID1B* in 11% of cases using a low-coverage whole-genome and targeted sequencing strategy. Given the genetic heterogeneity described in neuroblastoma, we sought to build on these studies through a focused analysis of a large cohort of high-risk stage 4 neuroblastomas, in which the need for translational advances are most pressing, using several genomic approaches.

We examined 240 matched tumor and normal (blood leukocyte) pairs from individuals older than 18 months of age at diagnosis with metastatic (stage 4) disease by whole-exome sequencing (WES; 221 cases), whole-genome sequencing (WGS; 18 cases, 1 of which was sequenced using two different platforms) or both (1 case; **Supplementary Table 1** and **Supplementary Note**). WES of ~33 Mb of coding sequence yielded an average of 124× coverage, with 87% of bases being suitable for mutation detection (**Supplementary Fig. 1** and **Supplementary Table 2**). We used two different WGS approaches, Illumina[20] (ten cases, 29.7× average coverage) and Complete Genomics[21] (ten cases, 59.9× average coverage), to interrogate structural variation and supplement mutation detection (powered to detect mutations at 86% and 94% of mappable exonic bases, respectively). To assess the expression of mutations and fusion transcripts, we generated over 10 Gb of RNA sequencing (RNA-seq) data for the ten cases sequenced by Illumina WGS.

Across the coding regions of the 240 cases, we detected 5,291 candidate somatic mutations in 3,960 genes (**Supplementary Table 3**). We found a median of 18 candidate exomic mutations (17 substitutions and 1 insertion or deletion (indel)) per tumor (range, 0–218 mutations), of which 14 were nonsilent mutations that are predicted to alter protein sequences (range, 0–158 nonsilent mutations; median, 12 missense, 1 nonsense, 1 indel and 0 splice site; **Supplementary Table 1**). This corresponds to a median frequency of 0.60 mutations per Mb (0.48 nonsilent mutations per Mb), considering only exonic bases with sufficient data for mutation detection (**Fig. 1**). This frequency is consistent with unselected neuroblastomas[17–19], medulloblastoma[22] and hematopoietic malignancies[23,24], is twice that of pediatric rhabdoid cancer[25] and is markedly less than that of adult solid tumors[24,26,27], particularly those with strong environmental contribution[24,28–31]. We verified 241 of the 282 coding candidate somatic substitutions (85%; 525 of 605, including noncoding substitutions) and 26 of the 41 coding indels (63%; 27 of 79, including noncoding substitutions) using mass-spectrometric genotyping or PCR-based resequencing (**Supplementary Note**).

We did not find a correlation between mutation frequency and age at diagnosis ($P = 0.28$, Spearman) or other clinical variables (**Supplementary Table 4**). Consistent with a postulated limited environmental contribution to neuroblastoma development[1], context-specific transition and transversion rates were not elevated compared to other cancers (**Supplementary Fig. 2**), and we did not detect sequencing reads corresponding to pathogenic viruses (**Supplementary Table 5**). Two tumors with markedly increased mutation frequencies (7.27 and 4.29 mutations per Mb) harbored alterations of DNA repair genes (nonsense mutation and deletion of *MLH1* and nonsense mutation of *DB1*).

© 2013 Nature America, Inc. All rights reserved.

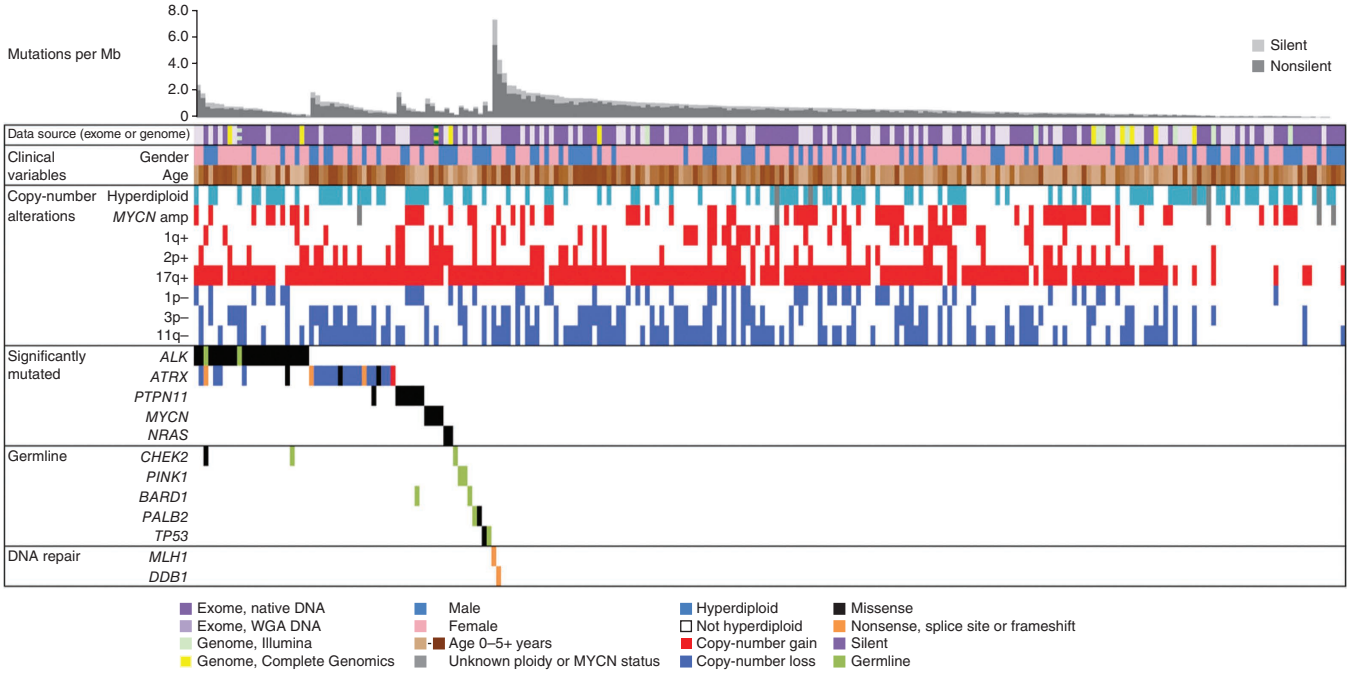

Figure 1 Landscape of genetic variation in neuroblastoma. Data tracks (rows) facilitate the comparison of clinical and genomic data across cases with neuroblastoma (columns). The data sources and sequencing technology used were WES from whole-genome amplification (WGA) (light purple), WES from native DNA (dark purple), Illumina WGS (green) and Complete Genomics WGS (yellow). Striped blocks indicate cases analyzed using two approaches. The clinical variables included were gender (male, blue; female, pink) and age (brown spectrum). Copy number alterations indicates ploidy measured by flow cytometry (with hyperdiploid meaning DNA index >1) and clinically relevant copy number alterations derived from sequence data. Significantly mutated genes are those with statistically significant mutation counts given the background mutation rate, gene size and expression in neuroblastoma. Germline indicates genes with significant numbers of germline ClinVar variants or loss-of-function cancer gene variants in our cohort. DNA repair indicates genes that may be associated with an increased mutation frequency in two apparently hypermutated tumors. Predicted effects of somatic mutations are color coded according to the legend. *MYCN* amp, amplification of *MYCN*.

Table 1 Genes with a significant frequency of somatic mutation across 240 neuroblastomas

Gene	Description	Mutations	Cases	Unique sites	Missense	Loss of function[a]	q	Expressed in neuroblastoma[b]
ALK	Anaplastic lymphoma receptor tyrosine kinase	22	22	7	22	0	$<1.8 \times 10^{-7}$	Yes
PTPN11	Protein tyrosine phosphatase, nonreceptor type 11	7	7	6	7	0	1.8×10^{-5}	Yes
ATRX	Alpha thalassemia/mental retardation syndrome X-linked	6	6	6	3	3	0.031	Yes
OR5T1	Olfactory receptor, family 5, subfamily T, member 1	3	2	3	3	0	0.040	No
PDE6G	Phosphodiesterase 6G, cGMP-specific, rod, γ	2	2	2	2	0	0.052	No
MYCN	V-myc myelocytomatosis viral related oncogene, neuroblastoma	4	4	1	0	0	0.093	Yes
NRAS	Neuroblastoma RAS viral (v-ras) oncogene homolog	2	2	2	2	0	0.017 (COSMIC only)	Yes

[a]Nonsense, splice site or frameshift. [b]'No' indicates very low or absent mRNA expression in the RNA-seq or microarray datasets.

Using the MutSig algorithm[32], we identified six genes mutated at a significant frequency in the 240 tumors ($q < 0.1$; **Table 1** and **Supplementary Table 6**). A seventh gene, *NRAS*, was implicated by restricting the analysis to genes listed in the Catalogue of Somatic Mutations in Cancer (COSMIC, v48)[33]. Using neuroblastoma data from our RNA-seq cohort (10 cases), the TARGET RNA microarray project (250 cases) and a publicly available RNA microarray project (416 cases)[34], we determined that *OR5T1* and *PDE6G* have very low or absent mRNA expression in neuroblastoma (**Supplementary Fig. 3**). Therefore, we focused our analysis on five genes with statistical and biological rationale for neuroblastoma involvement: *ALK*, *PTPN11*, *ATRX*, *MYCN* and *NRAS*.

ALK and *PTPN11* were previously reported to be mutated in up to 10% and 3.4% of neuroblastoma cases, respectively[8,9,35–37], which is consistent with our screen here. All 22 somatic *ALK* mutations (9.2%) were restricted to the kinase domain, and all 7 *PTPN11* mutations (2.9%) have been previously reported[33,37–40]. Although we found no pathogenic germline *PTPN11* variants, two cases had germline *ALK* variants encoding two amino acid alterations: the pathogenic, activating p.Arg1275Gln alteration and the probable benign, kinase-dead p.Ile1250Thr alteration. Activating *ALK* variants were not associated with *MYCN* amplification ($P = 0.28$). Contrary to a previous report[41], we did not find Phe1174 alterations in a higher proportion of cases with amplification of *MYCN* than of cases with wild-type *ALK* ($P = 0.53$). Notably, *ALK* was the only significantly mutated gene to have an association with clinical outcome, as mutation-positive cases had a decreased overall survival probability ($P = 0.0103$; **Supplementary Fig. 4**).

Loss-of-function mutations or deletions of RNA-helicase *ATRX* have recently been described in neuroblastoma[17,18]. We found putative loss-of-function *ATRX* alterations in 9.6% of cases (6 mutations and 17 multiexon deletions; **Supplementary Fig. 5**). We confirmed previous observations[18] that alterations of *ATRX* and *MYCN* were mutually exclusive and that *ATRX* alterations were enriched in older children ($P = 0.0021$; **Supplementary Fig. 6**). One case had an apparent gain of exons 18–26 with an unclear functional effect.

High-level *MYCN* amplification has long been known as a negative prognostic indicator in neuroblastoma[42], but activating mutations have not been described. In our cohort, four cases without *MYCN* amplification had an identical p.Phe44Leu alteration. Tumors from all four of these cases had regional single-copy gain of chromosome 2p, three of which had a gain of the mutant allele. In a tumor with matched RNA-seq data, the mutant allele was expressed at a level twice that of wild type. This alteration has been documented in single cases of glioblastoma, medulloblastoma and pancreatic adenocarcinoma[33] and

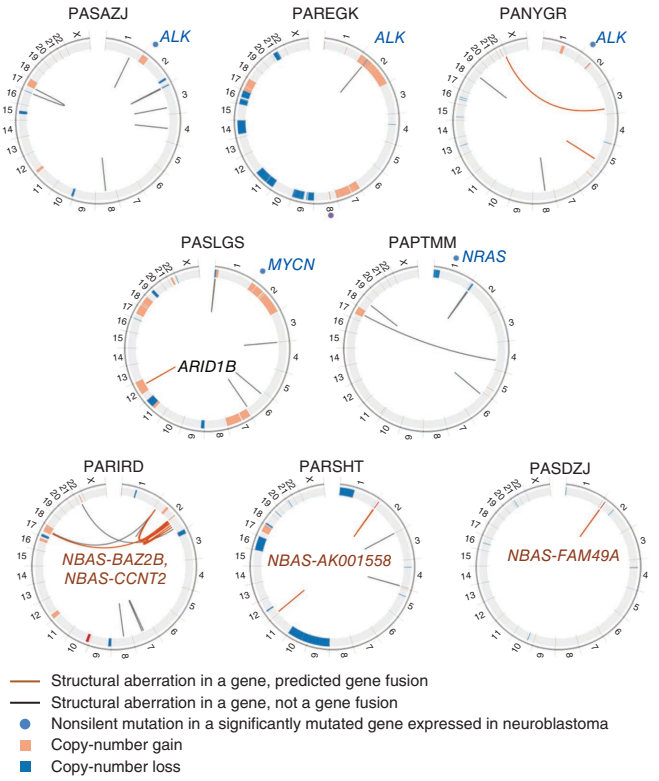

Figure 2 Structural variation in neuroblastoma genomes. CIRCOS[68] plots of cases with recurrent somatic alterations labeled using TARGET identifiers. Chromosomes are arranged end to end in the outer-most ring. Mutations in significantly mutated genes are depicted in light blue outside of each diagram. The inside ring shows somatic copy number gains and losses (high-level gains are red, low-level gains are orange, and losses are blue) detected by WGS. The innermost arcs depict genic structural aberrations (gene fusions are orange and all others are black) detected by RNA-seq and confirmed by local reassembly of WGS reads. Nongenic rearrangements are not shown. The top five cases have mutations in the significantly mutated genes *ALK*, *MYCN* and *NRAS*. The bottom three cases each have several rearrangements of *NBAS*, with expressed fusion transcripts as annotated.

© 2013 Nature America, Inc. All rights reserved.

Table 2 Candidate pathogenic germline variants in 222 cases with neuroblastoma sequenced by WES

Gene	Subject identifier	Genome position (build hg19)	cDNA change	Protein change
ALK	PARVLK	Chr2:29,432,664	c.3824G>A	p.Arg1275Gln
CHEK2	PAKXDZ	Chr22:29,121,242	c.433C>T	p.Arg145Trp
CHEK2	PAPTFZ	Chr22:29,121,015	c.542G>A	p.Arg181His
CHEK2	PARJMX	Chr22:29,121,018	c.539G>A	p.Arg180His
PINK1	PANYBL	Chr1:20,972,133	c.1040T>C	p.Leu437Pro
PINK1	PATINJ	Chr1:20,971,042	c.836G>A	p.Arg279His
BARD1	PAHYWC	Chr2:215,657,051	c.334C>T	p.Arg112*
BARD1	PATGWT	Chr2:215,595,215	c.1921C>T	p.Arg641*
TP53	PAICGF	Chr17:7,578,194	c.655C>T	p.Pro219Ser
PALB2	PAPZYZ	Chr16:23,646,182	c.1684+1C>A	Splice at Gly562

cDNA, complementary DNA; the asterisks indicate the introduction of a stop codon after the amino acid listed.

is scored as functional by PolyPhen2 (ref. 43) (score, 0.971), SIFT[44] (score, 0), MutationTaster[45] (score, 1.0) and AlignGVGD[46] (score, C65). The residue is highly conserved across the MYC superfamily (pfam01056; **Supplementary Fig. 7**)[47], and an additional tumor had a mutation in the homologous domain of *MYC* (encoding p.Thr58Ile), a common alteration in Burkitt's lymphoma[48]. Despite the relative infrequency of *MYCN* mutations in neuroblastoma, these mutations may be clinically relevant if they confer MYC dependency similar to high-level amplification.

We next searched, as previously described[22,27], for enrichment of somatic mutations in components of canonical pathways[49], chromatin modifiers or splice factors[50–52]. Of 857 gene sets, 12 were enriched for somatic mutation ($q < 0.1$; **Supplementary Table 7**), 8 of which implicated RAS-MAPK signaling components. Contrary to a previous report[17], we did not find any mutations in *TIAM1* or any enrichment of mutations in genes regulating neuritogenesis and GTPase activity through the RAC-RHO pathway ($q > 0.275$; **Supplementary Note**). However, an analysis of the mutation list from the previous report[17] using our methods recapitulated their finding of significant mutation frequencies in guanine nucleotide exchange factors ($q = 6.26 \times 10^{-3}$) and GTPase activating proteins ($q = 3.15 \times 10^{-5}$) but not in any of the 12 pathways identified by analysis of our cohort ($q \geq 0.848$). This comparison suggests limitations to the current gene set and pathway analysis methods, especially when mutation frequencies are low.

Somatic mutations in *FANCM* and *FAN1* in two cases with chromothripsis have been previously reported[17]. Although we found three cases with *FANCM* mutations, Fanconi anemia genes were not enriched for somatic mutation ($q = 0.764$; **Supplementary Table 7**), and we did not detect any exonic breakpoints in cases with *FANCM* mutations. This is perhaps not surprising given the relatively small portion of the genome queried by exome sequencing, so we cannot rule out an association of *FANCM* mutations and chromothripsis at this time.

Of the five recurrently mutated genes reported previously[19], we found mutations in *ALK*, *ARID1A* (encoding p.Gly1139Val and p.Gly1942Asp) and *VANGL1* (encoding p.Gly308Trp). Two cases had focal deletions of *ARID1B*: individual PASLGS had an exon 2 deletion (**Fig. 2** and **Supplementary Table 8**) and individual PARGKK had loss of exons 1–3. Of the 113 genes with apparent hemizygous mutations on chromosomes 1p, 3p and 11q (arms frequently lost in neuroblastoma), only *PBRM1* had loss-of-function mutations in two cases, whereas all other variations were singletons (**Supplementary Note**).

To identify rare germline variants predisposing to neuroblastoma, we searched for enrichment of clinically annotated variants from the ClinVar database and loss-of-function variants in cancer genes[53–55] in the blood-derived DNA samples from our WES cohort compared to normal DNAs from 1,974 European American individuals sequenced by the Exome Sequencing Project (ESP)[56] (Online Methods, **Supplementary Tables 9** and **10** and **Supplementary Fig. 8**). This approach nominated five genes with candidate germline pathogenic variants: *ALK*, *CHEK2*, *PINK1*, *TP53* and *BARD1* (**Table 2**). The *ALK* variant encoding p.Arg1275Gln has been reported as the most common pathogenic variant in familial neuroblastoma[8,9]. Three *CHEK2* germline variants destabilize the protein[57,58] and are reported cancer predisposition alleles[58,59] that have not been previously described in neuroblastoma. The *TP53* variant encoding p.Pro219Ser has been associated with Li-Fraumeni syndrome[60], consistent with previous reports of neuroblastomas occurring in these families[61]. Two *PINK1* variants are associated with Parkinson's disease[62–64], and this gene is known to be transcriptionally regulated by MYC proteins[65]. In addition, two loss-of-function variants in *BARD1*, a recently discovered neuroblastoma susceptibility locus[14], support the concept that rare variants may exist at loci where common polymorphisms affect disease occurrence. Another member of the BRCA complex, *PALB2*, had a germline variant that was predicted to ablate a splice site in one case (**Table 2**) and a somatic missense mutation in another (**Fig. 1** and **Supplementary Table 3**). Taken together, our conservative approach to identifying putatively pathogenic germline variants suggests that these events may have a larger role in neuroblastoma initiation than was previously suspected.

Structural analysis of the 19 cases sequenced by WGS identified a median of 41.5 breakpoints per case (range, 29–143; **Fig. 2** and **Supplementary Fig. 9**). Overall, 83 rearrangements affected 97 genes, 22 of which had evidence in the RNA-seq data (**Supplementary Table 8**). Whereas we found 11q;17q translocations in 3 of the 19 cases (1 case with two events), we did not find any recurrent fusion transcripts in our cohort. *NBAS*, located near *MYCN* on the short arm of chromosome 2, was the most commonly rearranged gene and harbored 11 distinct events in three cases with amplified *MYCN* (**Fig. 2**). We found substantial local rearrangement in three cases, all affecting the vicinity of *MYCN* and *NBAS* loci, but the numerous complex copy number states and retention of heterozygosity in lower–copy number regions are more consistent with an episomal model[66] than chromothripsis[67] in the 19 cases evaluated here (**Supplementary Fig. 10**). We found no other areas of clustered chromosomal breakpoints suggestive of chromothripsis in the cases sequenced by WGS, and there were no clusters evident within coding regions of 142 cases sequenced using WES of native DNA.

High-risk neuroblastomas harbor a very low frequency of recurrent somatic mutations. We do not expect that substantial numbers of mutations went undetected here, as the tumor purities were high and identical methods have identified recurrent mutations in other tumor types[22,23,26,27,30]. The relative paucity of recurrent mutations challenges the concept that druggable targets can be defined in each patient by DNA sequencing alone. Our data suggest that the majority of high-risk neuroblastomas may be driven by rare germline variants and/or copy number alterations and epigenetic modifications during tumor evolution. The notable lack of precisely defined genomic causes of this highly aggressive pediatric neoplasm reinforces the need to understand the interplay of host genetic factors, somatic mutations, chromosomal abnormalities and epigenetic alterations in the context of nervous system development.

URLs. Therapeutically Applicable Research to Generate Effective Treatments (TARGET), http://target.cancer.gov/; Broad Institute

© 2013 Nature America, Inc. All rights reserved.

Cancer Genome Analysis tools, http://www.broadinstitute.org/cancer/cga/; ClinVar, http://www.ncbi.nlm.nih.gov/clinvar/; Familial Cancer Database, http://www.familialcancerdatabase.nl/; Human DNA repair genes, http://sciencepark.mdanderson.org/labs/wood/DNA_Repair_Genes.html; National Heart, Lung, and Blood Institute Grand Opportunity Exome Sequencing Project, https://esp.gs.washington.edu/drupal/; International Agency for Research on Cancer TP53 Database, http://p53.iarc.fr/; Complete Genomics Analysis Tools, http://www.completegenomics.com/analysis-tools/cgatools/; ALEXA RNA-seq analysis tools, http://www.alexaplatform.org/; Picard analysis tools, http://picard.sourceforge.net/; The R Project for Statistical Computing, http://www.r-project.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequence data used for this analysis are available in dbGaP under accession number phs000467.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the Children's Oncology Group for the collection and annotation of samples for this study, and all TARGET co-investigators for scientific support of this project. Funding was provided by US National Institutes of Health grants CA98543 and CA98413 to the Children's Oncology Group, RC1MD004418 to the TARGET consortium, CA124709 (J.M.M.) and CA060104 (R.C.S.) and National Human Genome Research Institute grant U54HG003067 (E.S.L., D.A., S.B.G., G.G. and M.M.), as well as a contract from the National Cancer Institute, US National Institutes of Health (HHSN261200800001E). Additional support included a Canadian Institutes of Health Research Fellowship (T.J.P.), a Roman M. Babicki Fellowship in Medical Research at the University of British Columbia (O.M.), the Canada Research Chair in Genome Science (M.A.M.), the Giulio D'Angio Endowed Chair (J.M.M.), the Alex's Lemonade Stand Foundation (J.M.M.), the Arms Wide Open Foundation (J.M.M.) and the Cookies for Kids Foundation (J.M.M.). We thank E. Nickerson, S. Channer, K. Novik, C. Suragh and R. Roscoe for project management support. We also thank the staff of the Genome Sciences Centre Biospecimen Core, Library Construction, Sequencing and Bioinformatics teams, and the staff of the Broad Institute Biological Samples, Genome Sequencing and Genetic Analysis Platforms for their expertise in genomic processing of samples, and generating the sequencing data used in this analysis.

AUTHOR CONTRIBUTIONS

J.M.M., J. Khan, R.C.S., D.S.G. and M.A.S. conceived and led the project. M.A.M. and M.M. conceived of and supervised all aspects of the sequencing work at the British Columbia (BC) Cancer Agency Genome Sciences Centre and Broad Institute, respectively. T.J.P. and O.M. performed the analyses and interpreted the results. E.F.A., S.A., J.S.W., K.A.C., M.D., S.J.D., A.C.W., Y.P.M., L.J., T.B., Y.M., J.M.G.-F. and M.D.H. selected and characterized samples, provided disease-specific expertise in data analysis and edited the manuscript. R.S. and W.B.L. provided statistical support and analyses of clinical covariates. D.A., E.S., C. Sougnez, M.D. and J.M.G.A. provided overall project management and quality control support. S.L.C., K.C., M. Hanna, A.K., J. Kim, M.S.L., L.L., A.M., A.H.R., A.S. and C. Stewart supported analysis of somatic and germline alterations in the exome sequencing data. C.S.P. performed the pathogen discovery analysis. I.B., K.L.M., R.C., S.D.J. and J.Q. performed *de novo* assembly of Illumina sequencing data. Y.Z. led the library construction effort for the Illumina libraries. A.T. and Y.Z. planned the sequencing verification, and A.A. and B.K. performed the experiments. R.D.C. performed copy number analysis of genome sequencing data. M.K. performed verification of candidate rearrangements. N.T. performed gene- and exon-level quantification analysis of RNA-seq data. A.L. and A.H.K. helped interpret data provided by Complete Genomics. R.A.M. and M. Hirst led the sequencing effort for the Illumina genome and transcriptome libraries. S.B.G. and E.S.L. led the sequencing effort for the exome sequencing libraries. G.G. and S.J.M.J. supervised the bioinformatics group at the Broad Institute and BC Cancer Agency Genome Sciences Centre, respectively. T.J.P., O.M., D.S.G., M.A.M., M.M. and J.M.M. cowrote the manuscript with input from all coauthors.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/doifinder/10.1038/ng.2529.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Maris, J.M. Recent advances in neuroblastoma. *N. Engl. J. Med.* **362**, 2202–2211 (2010).
2. Smith, M.A. *et al.* Outcomes for children and adolescents with cancer: challenges for the twenty-first century. *J. Clin. Oncol.* **28**, 2625–2634 (2010).
3. 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**, 1165–1173 (1999).
4. Yu, A.L. *et al.* Anti-GD2 antibody with GM-CSF, interleukin-2, and isotretinoin for neuroblastoma. *N. Engl. J. Med.* **363**, 1324–1334 (2010).
5. Oeffinger, K.C. *et al.* Chronic health conditions in adult survivors of childhood cancer. *N. Engl. J. Med.* **355**, 1572–1582 (2006).
6. Trochet, D. *et al.* Germline mutations of the paired-like homeobox 2B (*PHOX2B*) gene in neuroblastoma. *Am. J. Hum. Genet.* **74**, 761–764 (2004).
7. Mosse, Y.P. *et al.* Germline *PHOX2B* mutation in hereditary neuroblastoma. *Am. J. Hum. Genet.* **75**, 727–730 (2004).
8. Mosse, Y.P. *et al.* Identification of *ALK* as a major familial neuroblastoma predisposition gene. *Nature* **455**, 930–935 (2008).
9. Janoueix-Lerosey, I. *et al.* Somatic and germline activating mutations of the ALK kinase receptor in neuroblastoma. *Nature* **455**, 967–970 (2008).
10. Diskin, S.J. *et al.* Common variation at 6q16 within *HACE1* and *LIN28B* influences susceptibility to neuroblastoma. *Nat. Genet.* **44**, 1126–1130 (2012).
11. Wang, K. *et al.* Integrative genomics identifies *LMO1* as a neuroblastoma oncogene. *Nature* **469**, 216–220 (2011).
12. Nguyen, L.B. *et al.* Phenotype restricted genome-wide association study using a gene-centric approach identifies three low-risk neuroblastoma susceptibility loci. *PLoS Genet.* **7**, e1002026 (2011).
13. Diskin, S.J. *et al.* Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* **459**, 987–991 (2009).
14. Capasso, M. *et al.* Common variations in *BARD1* influence susceptibility to high-risk neuroblastoma. *Nat. Genet.* **41**, 718–723 (2009).
15. Maris, J.M. *et al.* Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N. Engl. J. Med.* **358**, 2585–2593 (2008).
16. Deyell, R.J. & Attiyeh, E.F. Advances in the understanding of constitutional and somatic genomic alterations in neuroblastoma. *Cancer Genet.* **204**, 113–121 (2011).
17. Molenaar, J.J. *et al.* Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* **483**, 589–593 (2012).
18. Cheung, N.-K.V. Association of age at diagnosis and genetic mutations in patients with neuroblastoma. *J. Am. Med. Assoc.* **307**, 1062–1071 (2012).
19. Sausen, M. *et al.* Integrated genomic analyses identify *ARID1A* and *ARID1B* alterations in the childhood cancer neuroblastoma. *Nat. Genet.* **45**, 12–17 (2012).
20. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
21. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
22. Pugh, T.J. *et al.* Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature* **488**, 106–110 (2012).
23. Wang, L. *et al.* *SF3B1* and other novel cancer genes in chronic lymphocytic leukemia. *N. Engl. J. Med.* **365**, 2497–2506 (2011).
24. Cancer Genome Atlas Research Network. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012); erratum **491**, 288 (2012).
25. Lee, R.S. *et al.* A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. *J. Clin. Invest.* **122**, 2983–2988 (2012).
26. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409 (2012).
27. Chapman, M.A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467–472 (2011).
28. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).
29. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
30. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157–1160 (2011).
31. Nikolaev, S.I. *et al.* Exome sequencing identifies recurrent somatic *MAP2K1* and *MAP2K2* mutations in melanoma. *Nat. Genet.* **44**, 133–139 (2012).
32. Getz, G. *et al.* Comment on 'The Consensus Coding Sequences of Human Breast and Colorectal Cancers'. *Science* **317**, 1500 (2007).
33. Forbes, S.A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
34. Shi, L. *et al.* The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* **28**, 827–838 (2010).
35. George, R.E. *et al.* Activating mutations in *ALK* provide a therapeutic target in neuroblastoma. *Nature* **455**, 975–978 (2008).
36. Chen, Y. *et al.* Oncogenic mutations of *ALK* kinase in neuroblastoma. *Nature* **455**, 971–974 (2008).

© 2013 Nature America, Inc. All rights reserved.

37. Bentires-Alj, M. *et al.* Activating mutations of the Noonan syndrome-associated *SHP2/PTPN11* gene in human solid tumors and adult acute myelogenous leukemia. *Cancer Res.* **64**, 8816–8820 (2004).
38. Tartaglia, M. *et al.* Mutations in *PTPN11*, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. *Nat. Genet.* **29**, 465–468 (2001).
39. Tartaglia, M. *et al.* Somatic mutations in *PTPN11* in juvenile myelomonocytic leukemia, myelodysplastic syndromes and acute myeloid leukemia. *Nat. Genet.* **34**, 148–150 (2003).
40. Sarkozy, A. *et al.* A novel *PTPN11* gene mutation bridges Noonan syndrome, multiple lentigines/LEOPARD syndrome and Noonan-like/multiple giant cell lesion syndrome. *Eur. J. Hum. Genet.* **12**, 1069–1072 (2004).
41. De Brouwer, S. *et al.* Meta-analysis of neuroblastomas reveals a skewed *ALK* mutation spectrum in tumors with *MYCN* amplification. *Clin. Cancer Res.* **16**, 4353–4362 (2010).
42. 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* **224**, 1121–1124 (1984).
43. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
44. Ng, P.C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
45. Schwarz, J.M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7**, 575–576 (2010).
46. Tavtigian, S.V. *et al.* Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J. Med. Genet.* **43**, 295–305 (2006).
47. Marchler-Bauer, A. *et al.* CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, D225–D229 (2011).
48. Love, C. *et al.* The genetic landscape of mutations in Burkitt lymphoma. *Nat. Genet.* **44**, 1321–1325 (2012).
49. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
50. Barbosa-Morais, N.L., Carmo-Fonseca, M. & Aparício, S. Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Res.* **16**, 66–77 (2006).
51. Chen, M. & Manley, J.L. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat. Rev. Mol. Cell Biol.* **10**, 741–754 (2009).
52. Gabut, M., Chaudhry, S. & Blencowe, B.J. SnapShot: The splicing regulatory machinery. *Cell* **133**, 192 e1 (2008).
53. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
54. Sijmons, R.H. Encyclopaedia of tumour-associated familial disorders. Part I: from AIMAH to CHIME syndrome. *Hered. Cancer Clin. Pract.* **6**, 22–57 (2008).
55. Wood, R.D., Mitchell, M. & Lindahl, T. Human DNA repair genes, 2005. *Mutat. Res.* **577**, 275–283 (2005).
56. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
57. Sodha, N., Mantoni, T.S., Tavtigian, S.V., Eeles, R. & Garrett, M.D. Rare germ line *CHEK2* variants identified in breast cancer families encode proteins that show impaired activation. *Cancer Res.* **66**, 8966–8970 (2006).
58. Lee, S.B. *et al.* Destabilization of CHK2 by a missense mutation associated with Li-Fraumeni syndrome. *Cancer Res.* **61**, 8062–8067 (2001).
59. Dong, X. *et al.* Mutations in *CHEK2* associated with prostate cancer risk. *Am. J. Hum. Genet.* **72**, 270–280 (2003).
60. Petitjean, A. *et al.* Impact of mutant p53 functional properties on *TP53* mutation patterns and tumor phenotype: lessons from recent developments in the IARC *TP53* database. *Hum. Mutat.* **28**, 622–629 (2007).
61. Birch, J.M. *et al.* Relative frequency and morphology of cancers in carriers of germline *TP53* mutations. *Oncogene* **20**, 4621–4628 (2001).
62. Choi, J.M. *et al.* Analysis of PARK genes in a Korean cohort of early-onset Parkinson disease. *Neurogenetics* **9**, 263–269 (2008).
63. Marongiu, R. *et al.* *PINK1* heterozygous rare variants: prevalence, significance and phenotypic spectrum. *Hum. Mutat.* **29**, 565 (2008).
64. Klein, C. *et al.* PINK1, Parkin, and DJ-1 mutations in Italian patients with early-onset parkinsonism. *Eur. J. Hum. Genet.* **13**, 1086–1093 (2005).
65. Fredlund, E., Ringnér, M., Maris, J.M. & Påhlman, S. High Myc pathway activity and low stage of neuronal differentiation associate with poor outcome in neuroblastoma. *Proc. Natl. Acad. Sci. USA* **105**, 14094–14099 (2008).
66. Storlazzi, C.T. *et al.* Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. *Genome Res.* **20**, 1198–1206 (2010).
67. Stephens, P.J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
68. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

© 2013 Nature America, Inc. All rights reserved.

ONLINE METHODS

Summary. Paired tumor and normal DNA from 240 cases with high-risk neuroblastoma were identified from the Children Oncology Group biobank on the basis of subjects having metastatic disease and, preferably, being between 18 months and 5.5 years of age at diagnosis. Whole-genome sequences were generated for 19 of these pairs using two technologies: nine cases were sequenced by Illumina sequencing by synthesis[20], nine by Complete Genomics probe-anchor ligation[21] and one by both methods. RNA-seq data were generated for the ten cases sequenced by Illumina WGS. Whole-exome sequences of 222 pairs were generated using in-solution hybrid capture[69] followed by Illumina sequencing. Phi29-based WGA was used to generate sufficient tumor and matched normal DNA templates from 80 cases. Reads were aligned to build hg19/GRCh37 of the human genome reference sequence using Burrows-Wheeler Aligner (BWA)[70], and somatic mutations were detected using SNVmix[71] (Illumina genomes and RNA-seq), muTect[27] (exomes) and version 2 of the Complete Genomics' custom caller[21,72] (Complete Genomics genomes). Mutations were annotated using Oncotator, and MutSig[32] was used to identify genes mutated at significant frequencies. PathSeq[73] was used to query exome datasets for reads supporting viral infection. These and other tools used for exome sequence analyses are described on the Broad Institute Cancer Genome Analysis website (see URLs). Rearrangements were detected from whole-genome data using trans-ABySS *de novo* assembly[74] and Complete Genomics' custom software[75]. Somatic mutations and structural alterations were confirmed by mass-spectrometric genotyping (Sequenom) or PCR followed by Sanger or Illumina sequencing.

Sample selection and preparation. This study focused on high-risk neuroblastoma, and we attempted to reduce heterogeneity by restricting eligibility to subjects with stage 4 (metastatic) disease and, preferably, those between 1.5 and 5.5 years of age at diagnosis (median, 3.4 years of age; range, 1.5–16.5 years of age) (**Supplementary Table 1**). All specimens were obtained at original diagnosis after informed consent at Children's Oncology Group member institutions. Males outnumbered females 149 to 91 (62%). Amplification of the *MYCN* oncogene was found in 77 tumors (32%) by fluorescence *in situ* hybridization, and 131 (55%) tumors had a diploid DNA index by flow cytometry. Flash-frozen tumor samples were analyzed for percent tumor content by histopathology, and samples with <75% tumor content were excluded. The Children's Hospital of Philadelphia Institutional Review Board was responsible for oversight of this study.

Genome sequencing and analysis. *Illumina sequencing technology (BC Cancer Agency).* Whole-genome and transcriptome libraries of ten cases sequenced using Illumina technology at the BC Cancer Agency were constructed from input amounts of 2–4 μg DNA and 3–10 μg DNaseI-treated total RNA, respectively, following previously described protocols[76,77]. The sequencing was carried out using Illumina GAIIx instruments as per the manufacturer's instructions. Paired-end reads generated from genome and transcriptome sequencing were aligned to the hg19/GRCh37 reference human genome assembly[78] using BWA[70] version 0.5.7. RNA-seq reads were processed as previously described[79,80].

Single nucleotide variant (SNV) detection in the Illumina tumor genome and transcriptome data was performed using SNVMix2 with filtering to include SNVs such that the combined probability of either heterozygous or homozygous SNV was greater than 0.99 (ref. 71). Reads flagged as poor quality according to the Illumina chastity filter, duplicate reads and reads aligned with a mapping quality <40 were excluded from SNV calling. The somatic status of SNV calls was determined using read evidence from the SAMtools version 0.1.13 pileup[81] constructed at the variant positions in the matched normal genome. Positions with normal genome coverage by least five unique reads supporting the reference allele were considered somatic. The candidate somatic SNV calls were inspected using IGV[82], and only those calls confirmed by visual inspection were used in the analysis.

Short insertions and deletions were detected in the tumor and normal Illumina WGS bam files using two software programs, Pindel[83] and SAMtools[81]. The mean of read-pair insert sizes were calculated for all samples to be ~400 bp, and this value was used in each Pindel run. The Pindel short insertion output was filtered to select events that mapped to annotated genes (Ensembl59). Candidate somatic short insertion events that recurred in at least two cases were manually reviewed using the Integrative Genomics Viewer[82]. The output from SAMtools pileup and varFilter functionality[81] run separately on normal and tumor libraries were filtered to identify somatic events. In the normal samples, any event with a total coverage of less than eight was discarded. In the tumor libraries, only indels supported by at least 16% of reads at a locus were considered. After the filtering, any indel present in one or more normal libraries was flagged as germline. None of the candidate somatic coding indels from the Pindel or SAMtools analysis was confirmed by manual inspection using IGV[82], which is consistent with the low frequency of somatic indels in the rest of the cohort (median of one indel across all other WGS and WES cases, and 86 cases had no indels).

Copy number analysis of the Illumina WGS data was conducted using a previously described hidden Markov model (HMM) method[84]. Briefly, 50 million reads with mapping qualities >10 were randomly selected from matched tumor and normal data. Reads were divided into bins of 200 adjacent alignments, and the ratio of tumor to normal reads was calculated for each bin. These ratios were then normalized by subtracting the median of these ratios across the whole genome. This resulted in a metric of relative read density from the tumors and matched normal samples in bins of variable length along the genome, where bin width was inversely proportional to the number of mapped reads in the normal genome. GC bias correction was applied, and an HMM was used to classify and segment the tumor genome into continuous regions of somatic copy number loss (HMM state 1), neutrality (HMM state 2), slight gain (HMM state 3), gain (HMM state 4) or high gain (HMM state 5).

To identify candidate transcript rearrangements, we used ABySS[85] to perform *de novo* transcriptome assembly of ten RNA-seq datasets. To identify known and new transcript structures, the assembled contigs were aligned to the hg19 (GRCh37) human reference genome assembly and compared to annotated transcript models using the trans-ABySS pipeline[74]. This approach identified all contigs with two discrete genomic BLAT alignments. The top five scoring alignments were manually inspected to remove probable false-positive events caused primarily by few supporting reads. Local rearrangements were identified from contigs with single, gapped BLAT alignments and supporting read evidence from manual review. Targeted assembly of the candidate rearrangement regions was performed to validate the events in the genomic data.

Complete Genomics sequencing technology. Whole-genome sequencing libraries of ten cases were constructed from 3.5 μg of DNA and sequenced using Complete Genomics Inc. (CGI) technology[21]. Sequencing and alignment of reads to the hg19/GRCh37 reference human genome assembly was performed by the CGI Cancer Sequencing service, analytic pipeline version 1 (see URLs). Mutation call files provided by this service were used to extract somatic mutations using the criteria in **Supplementary Table 11**. CGI also provided flat files containing candidate rearrangements and segmental relative copy number ratios derived from normalized read counts from matched tumor and normal samples. Copy number calls were converted to the five HMM states described above using the criteria listed in **Supplementary Table 12**.

Exome sequencing and analysis. The generation, sequencing and analysis of 222 pairs of exome libraries at the Broad Institute was performed using a previously described protocol[27]. Because of the small quantities of DNA available, 80 DNA samples were amplified using Phi29-based multiple-strand displacement whole-genome amplification (Repli-g service, QIAgen). Exonic regions were captured by in-solution hybridization using RNA baits similar to those previously described[27] but supplemented with additional probes capturing additional genes listed in ReqSeq[78] in addition to the original Consensus Coding Sequence (CCDS)[78] set. In total, ~33 Mb of genomic sequence was targeted, comprising 193,094 exons from 18,863 genes annotated by the CCDS[86] and RefSeq[86] databases as coding for a protein or microRNA (accessed November 2010). Sequencing of 76-bp paired-end reads was performed using Illumina Genome Analyzer IIx and HiSeq 2000 instruments. Reads were aligned to the hg19/GRCh37 build of the reference human genome sequence[78] using BWA[70]. PCR duplicates were flagged in the bam files for exclusion from further analysis using the Picard MarkDuplicates tool. To confirm sample identity, copy number profiles derived from sequence data were compared with those derived from microarray data when available. Candidate somatic base substitutions

© 2013 Nature America, Inc. All rights reserved.

were detected using muTect (previously referred to as muTector[27]), and indels were detected using IndelGenotyper[27]. Segmental copy number ratios were calculated as the ratio of tumor fraction read depth to the average fractional read depth in normal samples for that region.

Removal of oxidation of guanine bases (oxoG) library preparation artifacts. Cases sequenced using WGA and native DNA were sequenced more than 8 months apart using the Sequencing Platform at the Broad Institute. Initial comparison of candidate mutation calls from these two datasets identified a preponderance of apparent G>T or C>A substitutions of low allele fraction (<0.15) and within specific sequence contexts (**Supplementary Fig. 2a**). We subsequently characterized this artifact and developed a method to detect and remove these events. In brief, these artifacts are introduced at the DNA-shearing step of the library construction process and arise from oxoG by high-energy sonication. During downstream PCR, oxoG bases preferentially pair with thymine rather than cytosine, resulting in apparent G>T or C>A substitutions of low allele fraction that are enriched within specific sequence contexts (**Supplementary Fig. 2b**). Consistent with this mechanism, the intensity of the sonication process was increased with the introduction of a new 150-bp shearing protocol between preparation of the WGA and native DNA samples.

The number of artifacts in a library was apparently sample dependent (**Supplementary Fig. 2c**), and these events were found in unmatched tumor and normal libraries. In some cases, thousands of candidate mutations were called in cases with a heavily affected tumor sample and an unaffected normal sample. However, nearly every sample had at least one such artifact, and we have observed similar events in publicly available datasets from other centers, suggesting a common artifact mode that was exacerbated in some of our samples. To address this problem, we devised a method to differentiate oxoG artifacts from *bona fide* mutations.

Because of the modification of only one strand of a GC base-pair (that is, only the G base), reads supporting the artifact have a characteristic read orientation conferred by adaptor ligation. Therefore, all reads supporting an artifact were almost exclusively derived from the first or second read of the Illumina HiSeq instrument. *Bona fide* variants are supported by near-equal numbers of first and second reads. We made use of the skewed read-orientation combinations and low allele fractions characteristic of this artifact to identify and remove oxoG artifacts from mutation calls in our cohort (meaning removal of all variants with allele fraction <0.1 or exclusively supported by a single read orientation). This method restored the mutation pattern and frequency seen in earlier sequencing of WGA cases (**Supplementary Fig. 2d**).

Verification of somatic mutations and rearrangements. We used a combination of genotyping and sequencing technologies to verify random candidate mutations (PCR/Sanger and PCR/HiSeq sequencing of candidates from the Complete Genomics and BC Cancer Agency Illumina WGS and RNA-seq data), as well as mutations supportive of our significance analyses (Sequenom and PCR/MiSeq of WES and WGS data; **Supplementary Tables 13** and **14**). Combining all of the validation experiments resulted in overall validation rates of 87% for substitutions (525/605 candidates, 241/282 coding) and 34% for indels (27/79 candidates, 26/41 coding). Some mutations were verified using multiple technologies, and therefore the total number of candidate mutations verified is lower than the sum total of mutations described in the **Supplementary Note**. See the **Supplementary Note** for details and cross-platform comparisons.

Integrated analysis of somatic variation from exome and genome datasets. Somatic mutations detected in the WGS, WES and RNA-seq datasets were annotated using Oncotator (see URLs). Genes mutated at a statistically significant frequency were identified using MutSig[32], a method that identifies genes with mutation frequencies greater than those expected by chance given the detected background mutation rates, gene length and callable sequence in each tumor and normal pair. The relationship between mutation frequency and age of diagnosis was tested using the Spearman rank test. The implementation of the Kolmogorov-Smirnov test in R version 2.11.1 (ks.test) was used to test differences in mutation frequency distributions of several clinical variables (**Supplementary Table 4**).

To identify frequently mutated groups of genes, we applied the MutSig algorithm to sets of genes rather than individual genes. These gene sets comprised 853 'canonical pathways' curated by Gene Set Enrichment Analysis[49], as well as lists of chromatin modifiers and splice factors curated from the literature[50–52] (**Supplementary Table 6**; 'CHROMATIN_MODIFIERS', 'EPIGENETIC_COMPLEXES', 'SPLICE_FACTORS' and 'DNA_METHYLATION'). Significance analysis of mutations and pathways reported by Molenaar *et al.*[17] are provided in the **Supplementary Note**.

Expression analysis of significantly mutated genes. Alignments of RNA-seq data were used to estimate gene expression. Gene coverage analysis was based on Ensembl gene annotations (homo_sapiens_core_59_37d). These annotations were collapsed into a single gene model containing the union of exonic bases from all annotated transcripts of the gene. The analysis used SAMtools pileup to get the per-base coverage depths and excluded reads with mapping quality <10 and reads flagged as poor quality according to the Illumina chastity filter. Duplicate reads were kept in this analysis. The reads per kilobase of exon model per million mapped reads (RPKM) metric was used to estimate gene expression[87]. RPKM was calculated using the formula (number of reads mapped to all exons in a gene $\times 10^9$)/(NORM_TOTAL $\times$ sum of the lengths of all exons in the gene), where NORM_TOTAL is the total number of reads that are mapped to nonmitochondrial exons.

The expression threshold for each RNA-seq library was determined as the 95th percentile of the distribution of the expression of silent intergenic regions computed and defined as described on the ALEXA platform website[88]. Using this threshold, we determined that *ALK*, *PTPN11*, *ATRX, MYCN* and *NRAS* were expressed above background in each of the ten cases with available RNA-seq data. In contrast, *OR5T1* and *PDE6G* were not expressed above background in at least nine out of ten cases in our cohort.

The TARGET neuroblastoma Affymetrix Human Exon Array data (S.A., L.J., R.S., Tu, Y., Hadjidaniel, M.D., E.F.A., M.D.H., W.B.L., J.M.G.-F., J.S.W., Guo, X., D.S.G., M.A.S., Khan, J. J.M.M. & R.C.S., unpublished data) of 250 primary diagnostic tumor specimens was normalized by quantile normalization and summarized using robust multichip average (Affymetrix Power Tools software package version 1.12). This dataset includes samples from 220 cases with high-risk disease and 30 cases with low-risk disease. The transcript-level data of the core probe sets for each sample were averaged on the basis of gene symbol annotations provided by the manufacturer (17,422 unique genes). To identify the relative expression of genes in neuroblastomas, the percentile values of *ALK*, *PTPN11*, *ATRX*, *MYCN*, *NRAS*, *OR5T1* and *PDE6G* were computed from the cumulative distribution function calculated for each sample's gene profile. The same analysis was conducted on the Agilent 44K microarray data (19,528 unique genes) of 416 neuroblastoma tumors from the MicroArray Quality Control (MAQC)-II study (Gene Expression Omnibus GSE16716; ArrayExpress E-MATB-179)[88]. This dataset includes tumors from patients diagnosed with high-risk (n = 135), intermediate-risk (n = 34) or low-risk (n = 247) neuroblastoma. Su *et al.*[89] demonstrated that individual tissues express 30–40% of all genes by comparing microarray expressions across panels of human and mouse tissues. The median percentiles for *ALK*, *PTPN11*, *ATRX*, *MYCN* and *NRAS* in both datasets are well within the percentile range of genes that are probably expressed in a tissue. The low median percentiles for *OR5T1* and *PDE6G* (less than the 40th and 25th percentiles in the TARGET and MAQC-II data) suggest low expression in neuroblastoma tumors (**Supplementary Fig. 3**).

Germline variant analysis. Detection of pathogenic germline variation at a basepair resolution in a cohort of patients with cancer is complicated by the selection of an appropriately matched and sized control population, relatively high carrier frequencies for unrelated disorders and complex genetics underlying cancer predisposition. To nominate germline variants predisposing to neuroblastoma, we searched for enrichment of putative functional variants in the blood-derived DNA samples from our cohort sequenced by WES compared to normal DNAs from 1,974 European American individuals sequenced by the National Heart, Lung, and Blood Institute Grand Opportunity ESP[56]. As indel calls from the ESP cohort were not publicly available at the time of our study, we did not include them in our analysis.

To ensure consistency and accuracy of germline variant detection, all cases with neuroblastoma sequenced by WES were called simultaneously with 800 cases sequenced by WES from the 1000 Genomes Project using the UnifiedGenotyper from the Genome Analysis Toolkit. A principal component

© 2013 Nature America, Inc. All rights reserved.

doi:10.1038/ng.2529

analysis of the genotype calls was performed to determine the ethnic background of our cases (**Supplementary Fig. 7**) with respect to three 1000 Genomes populations. As over 80% of our cohort was of European or admixed European ancestry, we downloaded genotyping calls and coverage information from 1,974 European American individuals available on the ESP website to serve as a control population. To focus our analysis on rare variation consistent with the low prevalence of neuroblastoma, we removed from both datasets all variants present in individuals sequenced as part of the 1000 Genomes Project. We then generated two lists of rare variants: overlaps with clinically reported variants recorded in ClinVar (downloaded 4/27/2012; 284 variants in neuroblastoma and 2,947 in ESP) and loss-of-function variants in any of 924 genes in the Cancer Gene Census[53], the Familial Cancer database[54] or a list of DNA repair genes[55] (86 neuroblastoma and 1,068 ESP). We then tested each gene for significant enrichment of variants in the neuroblastoma cohort compared to the ESP cohort (one-tailed Fisher's exact test; **Supplementary Tables 7** and **9**).

The germline ClinVar analysis uncovered four genes of significance driven by single variants present at greater frequency in neuroblastoma compared to ESP: *CYP2D6*, *NOD2*, *SLC34A3* and *HPD*. All of these variants are present at low frequency in an expanded European American ESP cohort (rs5030865 in 1/8,524 chromosomes, rs104895438 in 5/8,600, rs121918239 in 14/8,514 and rs137852868 in 11/8,600), suggesting they are benign polymorphisms. Note that although candidates detected by this approach were not significant after correction for multiple testing, we believe there is sufficient biological rationale and supporting evidence for validation in larger cohorts. We also looked for overlap with sites recorded in COSMIC[33]. This analysis identified a *TP53* variant associated with Li-Fraumeni syndrome[60].

69. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
70. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
71. Goya, R. *et al.* SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* **26**, 730–736 (2010).
72. Carnevali, P. *et al.* Computational techniques for human genome resequencing using mated gapped reads. *J. Comput. Biol.* **19**, 279–292 (2012).
73. Kostic, A.D. *et al.* PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat. Biotechnol.* **29**, 393–396 (2011).
74. Robertson, G. *et al. De novo* assembly and analysis of RNA-seq data. *Nat. Methods* **7**, 909–912 (2010).
75. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–477 (2010).
76. Morin, R.D. *et al.* Somatic mutation of *EZH2* (Y641) in follicular and diffuse large B-cell lymphomas of germinal center origin. *Nat. Genet.* **42**, 181–185 (2010).
77. Morin, R.D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**, 298–303 (2011).
78. Pruitt, K.D., Tatusova, T., Klimke, W. & Maglott, D.R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.* **37**, D32–D36 (2009).
79. Morozova, O. *et al.* System-level analysis of neuroblastoma tumor–initiating cells implicates AURKB as a novel drug target for neuroblastoma. *Clin. Cancer Res.* **16**, 4572–4582 (2010).
80. Morin, R. *et al.* Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45**, 81–94 (2008).
81. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
82. Robinson, J.T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
83. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
84. Shah, S.P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
85. Simpson, J.T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
86. Pruitt, K.D. *et al.* The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* **19**, 1316–1323 (2009).
87. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
88. Griffith, M. *et al.* Alternative expression analysis by RNA sequencing. *Nat. Methods* **7**, 843–847 (2010).
89. Su, A.I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. USA* **99**, 4465–4470 (2002).

© 2013 Nature America, Inc. All rights reserved.