SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» DNA methylation

» Paediatric cancer

» Next-generation sequencing

DNA methylation profiling of primary neuroblastoma tumors using methyl-CpG-binding domain sequencing

Anneleen Decock[1,2], Maté Ongenaert[1], Wim Van Criekinge[3,4,5], Frank Speleman[1,2] & Jo Vandesompele[1,2,6]

Received: 23 November 2015

Accepted: 04 January 2016

Published: 02 February 2016

Comprehensive genome-wide DNA methylation studies in neuroblastoma (NB), a childhood tumor that originates from precursor cells of the sympathetic nervous system, are scarce. Recently, we profiled the DNA methylome of 102 well-annotated primary NB tumors by methyl-CpG-binding domain (MBD) sequencing, in order to identify prognostic biomarker candidates. In this data descriptor, we give details on how this data set was generated and which bioinformatics analyses were applied during data processing. Through a series of technical validations, we illustrate that the data are of high quality and that the sequenced fragments represent methylated genomic regions. Furthermore, genes previously described to be methylated in NB are confirmed. As such, these MBD sequencing data are a valuable resource to further study the association of NB risk factors with the NB methylome, and offer the opportunity to integrate methylome data with other -omic data sets on the same tumor samples such as gene copy number and gene expression, also publically available.

Design Type(s)	parallel group design • DNA methylation profiling by high throughput sequencing design
Measurement Type(s)	DNA residue methylation
Technology Type(s)	DNA methylation profiling assay
Factor Type(s)	disease stage • NMYC Gene Amplification • survival assessment • Prognostic Factor
Sample Characteristic(s)	Homo sapiens

[1]Center for Medical Genetics, Ghent University Hospital, De Pintelaan 185, Ghent 9000, Belgium. [2]Cancer Research Institute Ghent (CRIG), De Pintelaan 185, Ghent 9000, Belgium. [3]Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, Ghent 9000, Belgium. [4]MDxHealth, 15279 Alton Parkway, Suite 100, Irvine, California 92618, USA. [5]NXTGNT, Ghent University, Ottergemsesteenweg 460, Ghent 9000, Belgium. [6]Bioinformatics Institute Ghent From Nucleotides to Networks (BIG N2N), De Pintelaan 185, Ghent 9000, Belgium. Correspondence and requests for materials should be addressed to J.V. (email: joke.vandesompele@ugent.be).

Background & Summary

Neuroblastoma (NB), a neuro-ectodermal tumor that originates from precursor cells of the sympathetic nervous system, represents the most common extra-cranial solid tumor of early childhood and is considered a heterogeneous disease driven by genetic aberrations, as during the past decades mainly genetic factors have been described to influence the pathogenesis and disease course (including *MYCN* amplification, *ALK* amplification and mutation, hyperdiploidy, and gains and losses of specific chromosome arms (1p, 3p, 11q and 17q))[1]. Also, recent comprehensive whole-genome sequencing studies of primary NB tumors pinpointed chromothripsis and defects in neuritogenesis genes as important tumor-driving events in a subset of NB[2], and indicated that *MYCN*, *TERT* and *ATRX* alterations define major subgroups of high-risk NB[3,4]. However, also epigenetic mechanisms, such as DNA methylation alterations, seem to contribute to the NB biology and clinical behaviour.

As reviewed in Decock *et al.*[5], multiple DNA methylation alterations have been described in NB, but given the rare occurrence of the disease, the number of comprehensive genome-wide DNA methylation studies analyzing primary tumor samples is limited. Hence, most studies initially make use of NB cell lines and only validate the most obvious methylation alterations in primary NB tumors. For example, a frequently applied methodology to NB cell lines is assessment of gene expression reactivation upon 5'-aza-2'-deoxycytidine (DAC) treatment, a cytosine analogue that cannot be methylated, leading to progressive DNA demethylation upon cell division. However, major drawbacks of these studies are that their discovery phases fall short in covering the NB heterogeneity, as NB cell lines are considered models for aggressive high-risk tumors, and that DNA methylation detection is indirectly assessed, as the influence of the demethylating effect is measured at the transcriptional level[6–8]. To accommodate this, the Illumina 27 and 450 K methylation arrays, directly interrogating the status of approximately 27,000 and 485,000 methylation sites, respectively, recently were applied to primary NB tumors[6,9–12]. Yet, also this technology has important limitations: the design of the arrays is heavily biased to interrogation of CpG sites previously described in literature and covers less than 2% of all CpG sites in the human genome[13].

Therefore, we generated a data set comprising of 102 primary NB tumors in which DNA methylation is assessed by massively parallel sequencing of methylation enriched DNA fragments. The applied method is based on the use of MeCP2, a member of the methyl-CpG-binding domain (MBD) protein family which specifically binds to methylated cytosines and enables precipitation of methylated DNA fragments. This data set is unique in the NB research field, as it is the first sample cohort in which the full tumor heterogeneity is being assessed by genome-wide methylation analysis using next-generation sequencing (NGS); it was originally collected for the identification of prognostic biomarker candidates. Selected candidates were validated in independent cohorts using methylation-specific PCR and we showed that MBD sequencing allowed selection of valuable markers which would not have been identified using the Illumina methylation arrays[14].

Here, we provide a detailed description of the methodological approach and bioinformatics analyses, as well as easy access to the (analyzed) MBD sequencing data and analysis tools, allowing other researchers (inexperienced with MBD sequencing) to reuse it. Importantly, the analyzed samples are well annotated; besides overall and event-free survival data, also following NB characteristics are available: age of the patient at diagnosis, tumor stage according to the International Neuroblastoma Staging System (INSS)[15] and *MYCN* amplification status. As such, these data offer the opportunity to further explore the association of these risk factors with the NB methylome. Furthermore, integration of methylome data with other -omic data sets should be examined in order to fully map the NB biology on a genome-wide level. The present MBD sequencing data greatly facilitate these integration analyses, considering that for part of the profiled samples matching expression and array comparative genomic hybridization (aCGH) data are available[16–18] (see Methods for details).

In summary, this data descriptor outlines details on the generation and analysis of MBD sequencing data of 102 primary NB tumors (Fig. 1). As NB is a rare disease and comprehensive DNA methylation studies scarce, these MBD sequencing data are very valuable and permit further unravelling the role of DNA methylation in the NB biology.

Methods

DNA sample collection

Two independent cohorts of 42 and 60 primary tumor DNA samples, respectively annotated as MBD cohort I and II, were sequenced. Samples of fresh frozen tumors were collected at the Ghent University Hospital ($n=49$; Ghent, Belgium), the Hospital Clínico Universitario ($n=42$; Valencia, Spain), the University Children's Hospital Essen ($n=8$; Essen, Germany) and the Our Lady's Children's Hospital Dublin ($n=3$; Dublin, Ireland), according to previously published criteria[7,14], and stage 4S tumors were also included. Detailed clinical characteristics of the patients are given in Table 1 (available online only). For samples 809 and 912, DNA was extracted from different parts of the same primary tumor. Informed consent was obtained from each patient's guardian and the study was approved by the ethical committee of the Ghent University Hospital (approval number B67020109912). Matching expression data[16,17] of 38 tumors are available through the NCBI Gene Expression Omnibus (GEO) database (GSE21713 and GSE32664; sample IDs in Table 1 (available online only)). Matching aCGH data[18] of 38 tumors are available through ViVar[19] (https://www.cmgg.be/vivar/; login: review, password: review, project: Kumps *et al.* 2013; sample IDs in Table 1 (available online only)).

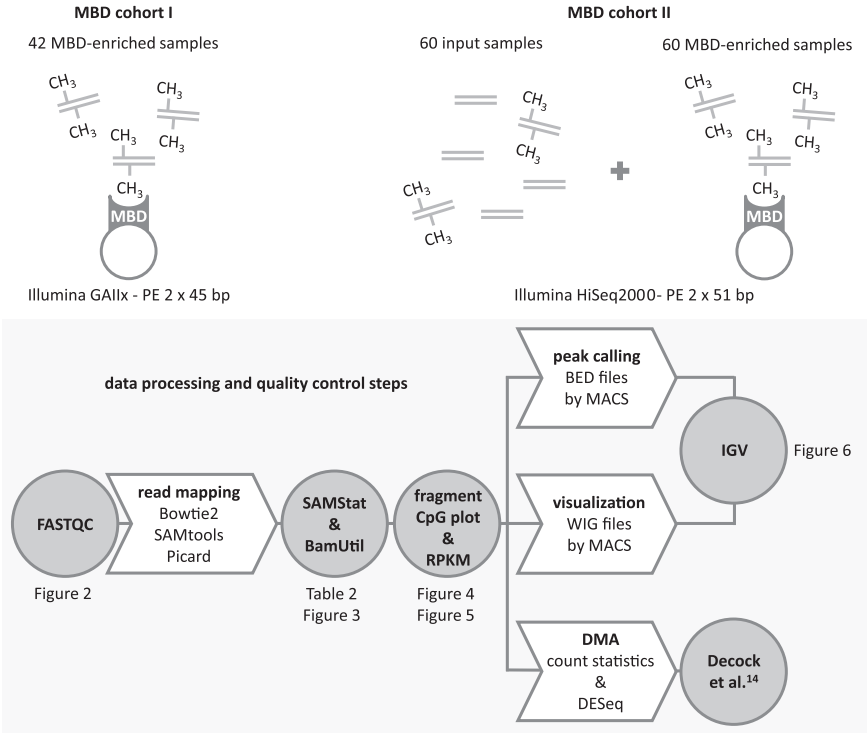

Figure 1. The MBD sequencing data of 102 primary neuroblastoma tumors are processed using different analysis tools. Depicted are the available MBD sequencing data sets and downstream data processing and technical validation steps. These steps are represented as arrows and circles, respectively. For each step, the applied tool or analysis is indicated. For the technical validation steps, also the corresponding data descriptor figures and tables are indicated. DMA, differential methylation analysis; IGV, Integrative Genomics Viewer; PE, paired-end; RPKM, reads per kilobase CpG island per million.

Methyl-CpG-binding domain (MBD) sequencing

DNA fragmentation. For each sample, between 400 to 1000 ng DNA was sheared to obtain DNA fragments with an average length of 200 bp. The DNA was loaded in 120 µl TE buffer (1:5), transferred to a Snap Cap microTUBE (Covaris) and exposed to Covaris S2 Adaptive Focused Acoustics. Fragment distribution and concentration was determined on a High Sensitivity DNA chip (Agilent Technologies).

Methylated DNA capturing. Subsequently, capturing of methylated DNA fragments was done according to the MethylCap kit protocol of Diagenode using 200–500 ng DNA. Elution of the captured fraction was performed in 150 µl High Elution Buffer and DNA was purified using the MinElute PCR purification kit (Qiagen). For MBD cohort II, also input samples (10%) were prepared.

Library preparation. As MBD cohort I and II were profiled in a different time frame and NGS methodologies evolve at rapid pace, a different library preparation protocol and sequencing technology was applied to each of them. For MBD cohort I, DNA library preparation was performed using the NEBNext DNA Library Prep Master Mix Set for Illumina (New England Biolabs) in combination with the Multiplexing Sample Preparation Oligonucleotide Kit (Illumina) for paired-end adapter ligation. Size selection of the library is done on a 2% agarose gel (Bio-Rad). Fragments between 250 and 350 bp were excised and purified using a Qiagen Gel Extraction Kit. For MBD cohort II, library preparation was automated on an Apollo 324 Next Generation Sequencing Library Preparation System (IntegenX), making use of the PrepX ILM DNA Library Kit (IntegenX). For paired-end adapter ligation the Multiplexing Sample Preparation Oligonucleotide Kit was used. Size selection was done with 1X AMPure XP beads (Agencourt) and PEG-Bead Solution.

Library amplification. PCR library amplification with appropriate Index Primers for each sample was performed using the Multiplexing Sample Preparation Oligonucleotide Kit and following PCR conditions: 30 s at 98 °C, 21 amplification cycles (10 s at 98 °C, 30 s at 65 °C and 30 s at 72 °C), 5 min at 72 °C, and held at 4 °C. PCR product purification was done using the High Pure PCR Purification Kit (Roche). QC was performed on a DNA 1000 chip (Agilent) and concentration was determined by qPCR according to the qPCR Quantification Protocol Guide of Illumina. Samples were pooled and profiled on

Statistic	MBD cohort I—enriched samples			MBD cohort II—enriched samples			MBD cohort II—input samples		
	range	*mean*	*median*	*range*	*mean*	*median*	*range*	*mean*	*median*
total read number (e6)	4.65–18.20	13.38	14.17	29.74–66.59	45.09	44.41	20.86–59.51	36.00	33.19
duplicate reads (%)	0.70–72.00	6.46	3.39	2.55–79.69	31.04	19.89	2.24–10.47	4.17	3.68
properly paired reads (%)	48.29–94.51	85.64	89.29	86.86–97.57	95.33	95.72	94.78–97.55	96.50	96.59

Table 2. Using BamUtil, basic sequencing statistics of MBD cohort I and II are computed. Total read number: the total number of reads in the two paired FASTQ files of a sample; duplicate reads as a percentage of the total read number; properly paired reads as a percentage of the total read number.

an Illumina GAIIx (PE 2 × 45 bp) for MBD cohort I and on an Illumina HiSeq2000 (PE 2 × 51 bp) for MBD cohort II.

Data processing and analysis

Sequencing data. All crucial steps in the processing and analysis of the MBD sequencing data are summarized in Fig. 1. Raw sequencing data were demultiplexed and converted to FASTQ files (with sequencing reads and quality scores). Quality control on the raw data was performed by FASTQC (version 0.9.2; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Read mapping. Next, the sequencing reads were mapped/aligned to the human reference genome (hg19), using the Bowtie2 (ref. 20) mapper (version 2.0.0 beta7) and FASTQ files as input. For each sample, two paired FASTQ files are available (as we performed paired-end sequencing), in which the data lines correspond to each other. To improve the mapping quality, reads were only taken into account if the sequences in both files could be mapped to the reference genome (maximum 500 bp between both paired ends). Also sequencing quality scores were used in the mapping process. The BAM format was used as output file type. PCR duplicates were marked with Picard (version 1.79; http://broadinstitute.github.io/picard/) and the BAM files were sorted and indexed using SAMtools[21] (version 0.1.18) and index commands. These files have been deposited as raw data files in the NCBI Gene Expression Omnibus (GEO) database (Data Citation 1 for MBD cohort I; Data Citation 2 and Data Citation 3 for MBD cohort II). FASTQ records can be extracted from the sequence alignments in the BAM files using the BEDTools bamtofastq conversion utility[22]. Starting from the SRA files, the NCBI SRA Toolkit (fastq-dump) can be used to generate the FASTQ files. Mapping quality was evaluated using SAMStat[23] (version 1.08) and BamUtil (version 1.0.2; http://genome.sph.umich.edu/wiki/BamUtil). Technical validation of MBD enrichment is performed by fragment CpG plot analysis[24] and by plotting the densities of the median numbers of mapped reads per kilobase per million (RPKM[25]) in all CpG islands ($n = 28{,}691$) across the different subcohorts.

Peak calling. The process of converting mapped sequencing reads to coverage vectors and the detection of enriched regions (peaks) is referred to as peak detection or peak calling. Here, peak calling was done using the MACS[26] software tool (version 1.4.0 beta) and BAM files as input. BED files were generated (Data Citation 1 for MBD cohort I; Data Citation 2 and Data Citation 3 for MBD cohort II), indicating the location and score (linked to the *P*-value) of the identified peaks.

Visualization. MACS is also used to output WIG files (Data Citation 1 for MBD cohort I; Data Citation 2 and Data Citation 3 for MBD cohort II), which are transformed to a binary format (TDF file; Data Citation 1 for MBD cohort I; Data Citation 2 and Data Citation 3 for MBD cohort II) by igvtools (https://www.broadinstitute.org/igv/igvtools) for visualization in the Integrative Genomics Viewer (IGV)[27]. An example IGV XML-session file for MBD cohort II and instructions on how to make use of this file are included in the GitHub repository (see Code availability).

Differential methylation analyses. Differential methylation analyses between sample groups are described in detail in Decock *et al.*[14]. Briefly, for each subcohort, two count data sets were constructed, in which for each sample the numbers of mapped reads in the promoter region of the different Ensembl Transcripts or 5 kb genomic windows are indicated. Here, we provide access to these count data sets (Supplementary Tables 3–8), which can directly be used for differential methylation analyses in DESeq[14,28].

Code availability. All tools and code that are necessary to generate the described file types are provided in a Docker container (Docker Hub; https://hub.docker.com/r/mateongenaert/mbdtoolbox/). More advanced analysis scripts can be found in the GitHub repository (https://github.com/mateongenaert/MBDToolBox).

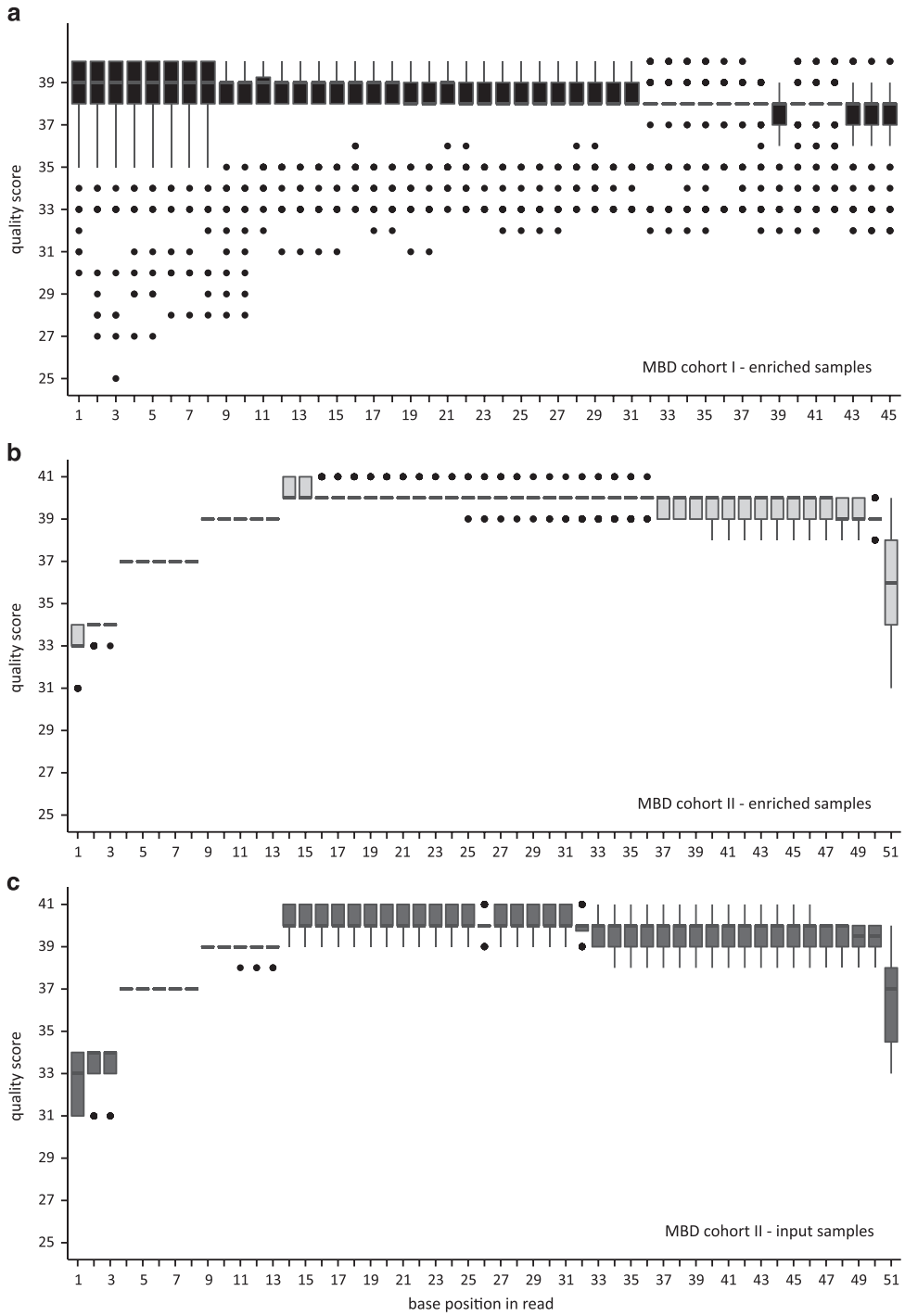

Figure 2. The per base sequence quality scores indicate that the raw sequencing data are of good quality. Shown are the distributions of the median per base quality score (determined by FASTQC) of the enriched samples of MBD cohort I (**a**), and of the enriched (**b**) and input (**c**) samples of MBD cohort II. In the boxplots, the lower and upper hinge of the boxes represents the 25th and 75th percentile, respectively. The whiskers extend to the lowest and highest value that is within 1.5 times the interquartile range. Data beyond the end of the whiskers are outliers and plotted as dots.

Data Records

An overview of the sample annotation and data outputs is given in Table 1 (available online only). The outputs of each step in the data processing (read mapping: BAM files, peak calling: BED files, and visualization: WIG and TDF files) have been deposited in the GEO database. For MBD cohort I, the accession number is GSE69224 (Data Citation 1), for MBD cohort II, GSE69243 (Data Citation 2) and

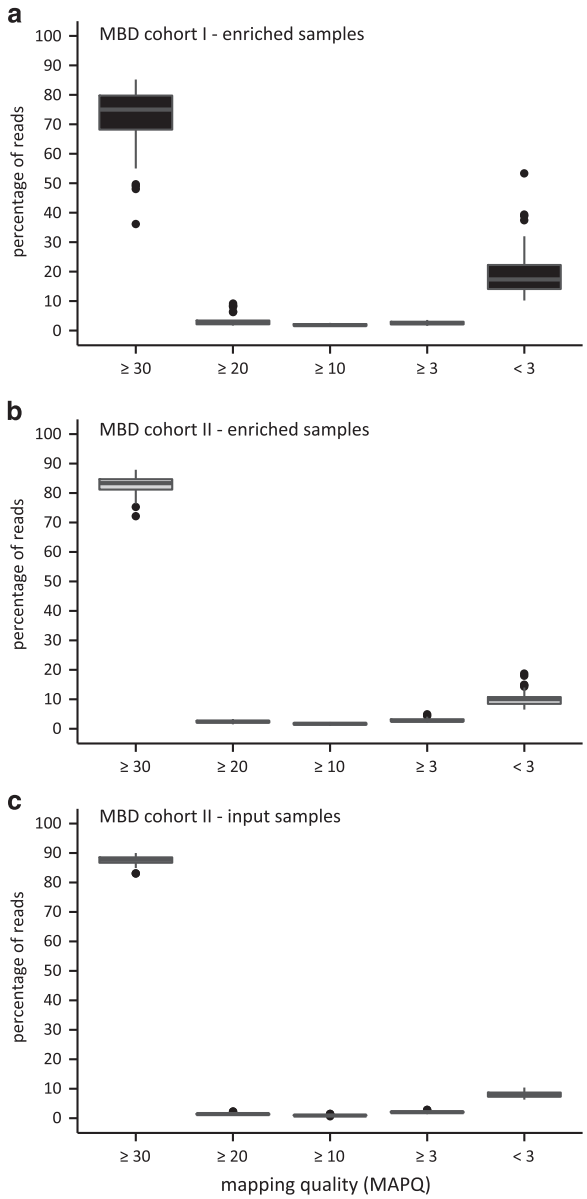

Figure 3. The mapping quality scores illustrate high mapping accuracy. Shown are the distributions of the percentages of mapped reads across the different mapping quality ranges, as determined by SAMStat ((**a**) enriched samples of MBD cohort I, (**b**) enriched samples of MBD cohort II and (**c**) input samples of MBD cohort II). In the boxplots, the lower and upper hinge of the boxes represents the 25th and 75th percentile, respectively. The whiskers extend to the lowest and highest value that is within 1.5 times the interquartile range. Data beyond the end of the whiskers are outliers and plotted as dots.

GSE69268 (Data Citation 3). In GEO, these data sets were submitted as SubSeries of the SuperSeries GSE69279 (Data Citation 4). We also provide a Docker container, made available through Docker Hub, that embeds all necessary tools to generate the data files and illustrates the analysis pipeline. More advanced analysis scripts are given in the GitHub repository (see Code availability).

Technical Validation

Validation of raw and mapped sequencing data

The total read number and percentage of duplicate and properly paired reads in each sample are given in Supplementary Table 1, and a summary of these sequencing statistics across the different sample cohorts can be found in Table 2.

To ensure raw data quality, FASTQC analyses were performed to determine the per base sequence quality which reflects the probability that a base has been called incorrectly[29]. Quality scores between 41 and 28, 28 and 20, and below 20 are considered base calls of very good quality, calls of reasonable quality and calls of poor quality, respectively. In order to obtain a general overview of the range of quality values

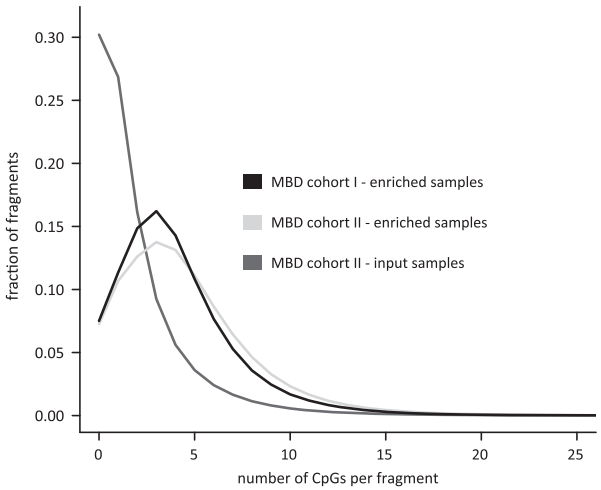

Figure 4. Fragment CpG plots demonstrate that the MBD-enriched samples have a high fraction of CpG dense sequencing fragments. Shown are the fractions of mapped MBD sequencing fragments with different CpG counts. Per cohort, 100,000 randomly selected fragments of each sample were used to construct the plots.

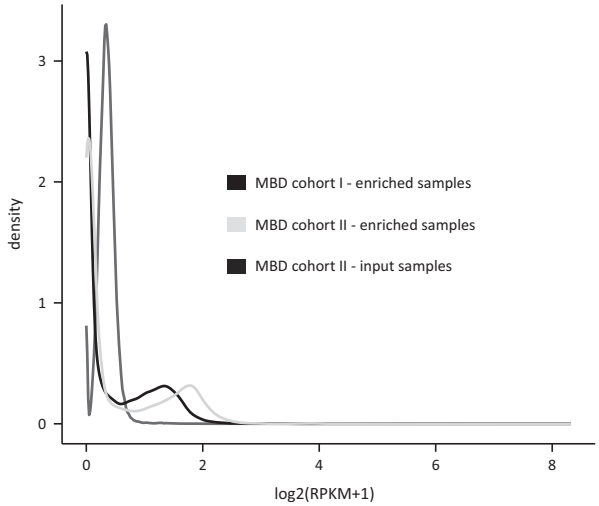

Figure 5. CpG island RPKM values confirm enrichment towards methylated DNA fragments upon MBD capture. Shown are the densities of the median RPKM values per subcohort. RPKM: reads per kilobase CpG island per million.

across all bases at each position, the median quality score for each position in each FASTQ file was determined. Fig. 2 shows the distribution of these median per base quality scores across the different sample cohorts. In general, the quality scores of both MBD cohort I and II are of reasonable to very good quality. Given the different sequencing technologies that were used for MBD cohort I (Illumina GAIIx) and II (Illumina HiSeq2000), it is expected that the read quality of MBD cohort II is higher than that of MBD cohort I. The steadily increase and subsequent decrease in quality along the read is also expected for Illumina-based experiments[29,30].

Mapping quality is ensured by analyzing the mapping quality scores of the alignments in each sample (Supplementary Table 2). In Fig. 3, the distributions of the percentages of mapped reads across the different mapping quality ranges are shown. For all subcohorts, the reads are clearly mapped with high accuracy, as almost for every sample, more than half of the mapped reads has a MAPQ $\geq$ 30 (ref. 23).

Validation of MBD-based enrichment

Over the past years several companies developed commercial kits for MBD-based capturing of methylated fragments. Although all of them claim to be of high quality, differences in performance exist. Careful kit selection is thus of utmost importance[24]. Here, sheared tumor DNA was enriched towards methylated fragments using the MethylCap kit of Diagenode, that makes use of the methylCap protein, consisting of the MBD of human MeCP2 fused with gluthatione-S-transferase (GST) containing an

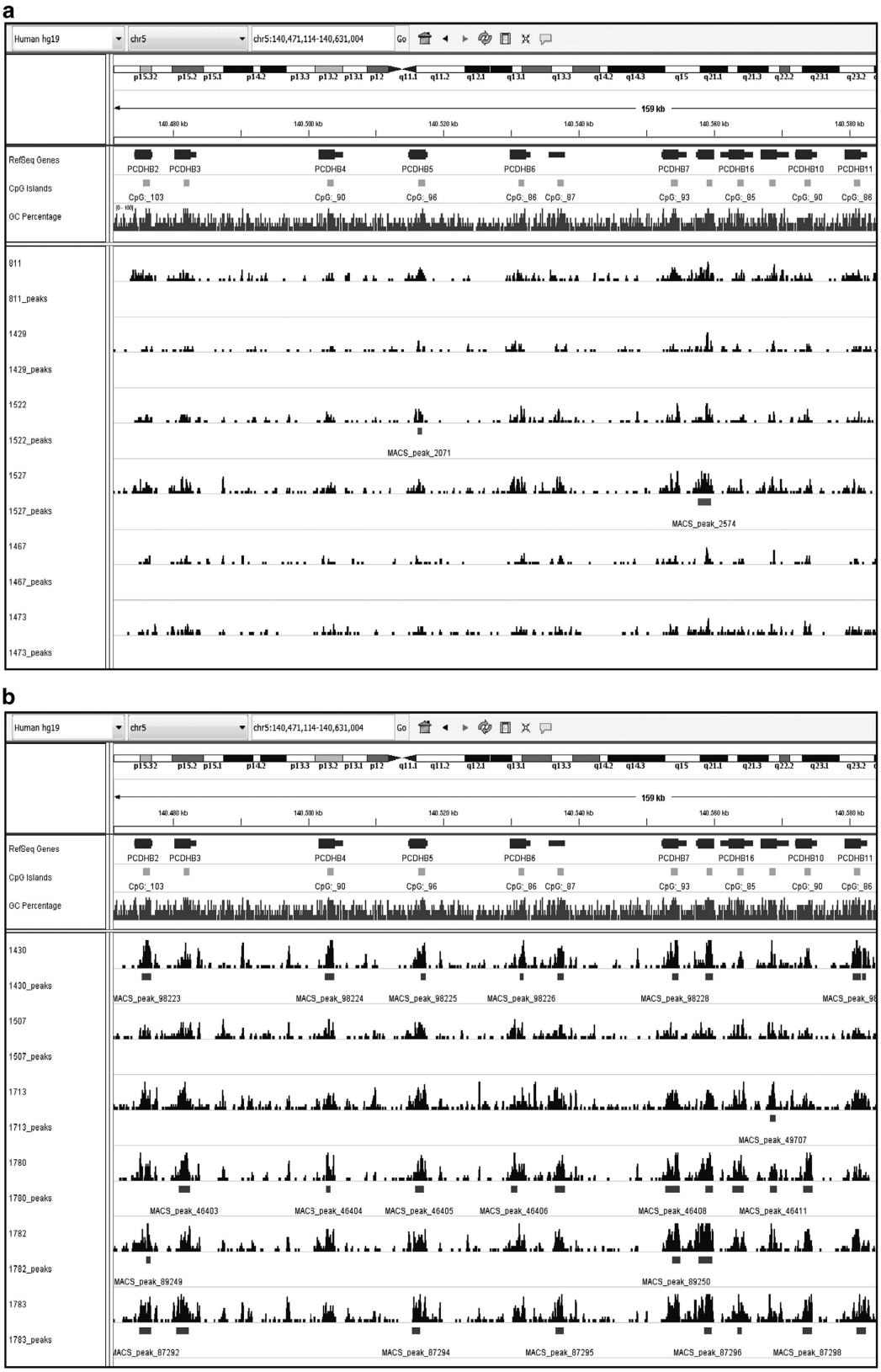

Figure 6. Visualization of the MBD sequencing data in IGV confirms methylation of the *PCDHB* gene cluster. In (**a**) the data of MBD cohort I is shown, in (**b**) the data of MBD cohort II. The upper panels show the genes in the cluster, the location of CpG islands and the GC percentage. In the lower panels, sequence coverage of 6 high-risk patient samples is shown (peak pattern), as well as the location of identified peaks (horizontal bars).

N-terminal His6-tag. A previous evaluation assessed the quality of this kit for combination with NGS by comparison with four other commercially available kits[24]. This study also compared the MBD sequencing data with reduced representation bisulfite sequencing (RRBS) and Illumina 27 K methylation array data of the same samples. Together, these analyses showed that the MethylCap kit outperforms the others, due to a consistent combination of high yield, sensitivity and specificity[24]. In order to demonstrate that the samples of MBD cohort I and II were enriched for methylated DNA fragments after MBD-based capturing, we made use of the fragment CpG plot[24]. As this plot depicts the CpG content of the mapped fragments and the MethylCap kit theoretically only captures methylated cytosines in a CpG dinucleotide context, the fragment CpG plot can be used to evaluate the MBD-based enrichment. An overview of the CpG content of the mapped fragments per sample cohort is depicted in Fig. 4. This fragment CpG plot clearly illustrates that the MBD-enriched samples of MBD cohort I and II have a high fraction of CpG dense fragments, while the input (non-MBD-enriched) samples of MBD cohort II are not enriched in CpG content. Additionally, using the number of mapped reads per kilobase CpG island per million (RPKM) values[25], the methylation level of each CpG island across the different subcohorts was determined. The density plot in Fig. 5 indicates that the MBD-enriched samples have a higher fraction of CpG islands with an RPKM>1 compared to the input samples of MBD cohort II. Based on these analyses, it can be concluded that the MBD-based capture successfully led to the enrichment of methylated DNA fragments.

Validation of methylated genes in neuroblastoma

Finally, TDF and BED files, containing sequence coverage and peak locations respectively, were loaded into IGV to visually inspect genes previously described to be methylated in NB. As an example, the MBD sequencing data of the *PCDHB* gene cluster is shown in Fig. 6. This gene cluster is frequently methylated in NB[5,31], which is confirmed by the MBD sequencing data of both MBD cohort I and II. Additionally, 78 regions identified in the MBD sequencing data as being methylated, were validated in two independent patient cohorts using methylation-specific PCR (MSP)[14]. These data confirm the validity of MBD sequencing in identifying methylated regions in NB.

Usage Notes

The MBD sequencing data can be downloaded from the GEO database via accession numbers GSE69224 (for MBD cohort I; Data Citation 1), GSE69243 and GSE69268 (for MBD cohort II; Data Citation 2 and Data Citation 3; SuperSeries GSE69279 (Data Citation 4)). The unique GEO sample accession IDs and clinical annotation can be found in Table 1 (available online only). This table also contains the accession IDs of the matching expression and aCGH data, which allows easy data access and facilitates integration analyses.

All output files from the different steps in the MBD sequencing data processing are provided through GEO. Analysis tools and scripts have been embedded in a Docker container, to deliver an environment that runs on any supported host platform (Windows, MAC, Linux). This Docker container, and all instructions on how it is made and how analyses can be run on the data, are made available through Docker Hub and GitHub (see Code availability). This allows researchers to try out the analysis pipeline that was used to generate the publically available data, without the need of additional infrastructure or software versions. The Docker container guarantees that the provided commands work and allows researchers to start exploring the data at the level they are experienced with.

Alternative processing tools can be tested for read mapping (e.g., BWA[32]) or identification of enriched regions (e.g., PeakRanger[33] or BALM[34]), or absolute methylation scores can be calculated (MEDIPS[35]; see Code availability). Researchers inexperienced with MBD sequencing can easily visualize their genes of interest by downloading the BED and TDF files (see Code availability). Downstream differential methylation analyses can be done with DESeq[28] (as described in Decock *et al.*[14]) using count data sets provided in Supplementary Tables 3–8, or other software can be used, such as DiffBind[36] and edgeR[37]. Differences in absolute methylation scores can be used for RankProd[38] analyses.

References

1. Brodeur, G. M. Neuroblastoma: biological insights into a clinical enigma. *Nat. Rev. Cancer* **3,** 203–216 (2003).
2. Molenaar, J. J. *et al.* Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* **483,** 589–593 (2012).
3. Valentijn, L. J. *et al.* TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors. *Nat. Genet.* **47,** 1411–1414 (2015).
4. Peifer, M. *et al.* Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature* **526,** 700–704 (2015).
5. Decock, A., Ongenaert, M., Vandesompele, J. & Speleman, F. Neuroblastoma epigenetics: from candidate gene approaches to genome-wide screenings. *Epigenetics* **6,** 962–970 (2011).
6. Carén, H. *et al.* Identification of epigenetically regulated genes that predict patient outcome in neuroblastoma. *BMC Cancer* **11,** 66 (2011).
7. Decock, A. *et al.* Genome-wide promoter methylation analysis in neuroblastoma identifies prognostic methylation biomarkers. *Genome Biol.* **13,** R95 (2012).
8. Duijkers, F. A. *et al.* Epigenetic drug combination induces genome-wide demethylation and altered gene expression in neuro-ectodermal tumor-derived cell lines. *Cell. Oncol.* **36,** 351–362 (2013).
9. Mayol, G. *et al.* DNA hypomethylation affects cancer-related biological functions and genes relevant in neuroblastoma pathogenesis. *PLoS ONE* **7,** e48401 (2012).
10. Yáñez, Y. *et al.* Two independent epigenetic biomarkers predict survival in neuroblastoma. *Clin. Epigenetics* **7,** 1–14 (2015).

11. Gómez, S. *et al.* DNA methylation fingerprint of neuroblastoma reveals new biological and clinical insights. *Epigenomics* **7,** 1137–1153 (2015).
12. Schramm, A. *et al.* Mutational dynamics between primary and relapse neuroblastomas. *Nat. Genet.* **47,** 872–877 (2015).
13. Stirzaker, C., Taberlay, P. C., Statham, A. L. & Clark, S. J. Mining cancer methylomes: prospects and challenges. *Trends Genet.* **30,** 75–84 (2014).
14. Decock, A. *et al.* Methyl-CpG-binding domain sequencing reveals a prognostic methylation signature in neuroblastoma. *Oncotarget* **7,** 1960–1972 (2015).
15. Brodeur, G. M. *et al.* Revisions of the international criteria for neuroblastoma diagnosis, staging, and response to treatment. *J. Clin. Oncol.* **11,** 1466–1477 (1993).
16. Mestdagh, P. *et al.* The miR-17-92 microRNA cluster regulates multiple components of the TGF-β pathway in neuroblastoma. *Mol. Cell* **40,** 762–773 (2010).
17. Eschenburg, G., Eggert, A., Schramm, A., Lode, H. N. & Hundsdoerfer, P. Smac mimetic LBW242 sensitizes XIAP-overexpressing neuroblastoma cells for TNF-α-independent apoptosis. *Cancer Res.* **72,** 2645–2656 (2012).
18. Kumps, C. *et al.* Focal DNA copy number changes in neuroblastoma target MYCN regulated genes. *PLoS ONE* **8,** e52321 (2013).
19. Sante, T. *et al.* ViVar: a comprehensive platform for the analysis and visualization of structural genomic variation. *PLoS ONE* **9,** e113800 (2014).
20. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
21. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
22. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
23. Lassmann, T., Hayashizaki, Y. & Daub, C. O. SAMStat: monitoring biases in next generation sequencing data. *Bioinformatics* **27,** 130–131 (2011).
24. De Meyer, T. *et al.* Quality evaluation of methyl binding domain based kits for enrichment DNA-methylation sequencing. *PLoS ONE* **8,** e59068 (2013).
25. 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).
26. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
27. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–192 (2013).
28. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
29. Minoche, A. E., Dohm, J. C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.* **12,** R112 (2011).
30. Yu, Y. *et al.* Comprehensive RNA-Seq transcriptomic profiling across 11 organs, 4 ages, and 2 sexes of Fischer 344 rats. *Sci. Data* **1,** 140013 (2014).
31. Banelli, B. *et al.* A pyrosequencing assay for the quantitative methylation analysis of the PCDHB gene cluster, the major factor in neuroblastoma methylator phenotype. *Lab. Invest.* **92,** 458–465 (2012).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
33. Feng, X., Grossman, R. & Stein, L. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics* **12,** 139 (2011).
34. Lan, X. *et al.* High resolution detection and analysis of CpG dinucleotides methylation using MBD-seq technology. *PLoS ONE* **6,** e22226 (2011).
35. Lienhard, M., Grimm, C., Morkel, M., Herwig, R. & Chavez, L. MEDIPS: Genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics* **30,** 284–286 (2014).
36. Stark, R. & Brown, G. D. DiffBind: differential binding analysis of ChIP-seq peak data. http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf (2011).
37. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
38. Hong, F. *et al.* RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* **22,** 2825–2827 (2006).

Data Citations

1. Decock, A., Ongenaert, M., Speleman, F. & Vandesompele, J. *Gene Expression Omnibus* GSE69224 (2015).
2. Decock, A., Ongenaert, M., Speleman, F. & Vandesompele, J. *Gene Expression Omnibus* GSE69243 (2015).
3. Decock, A., Ongenaert, M., Speleman, F. & Vandesompele, J. *Gene Expression Omnibus* GSE69268 (2015).
4. Decock, A., Ongenaert, M., Speleman, F. & Vandesompele, J. *Gene Expression Omnibus* GSE69279 (2015).

Acknowledgements

The authors thank Rosa Noguera, Johannes H. Schulte and Raymond L. Stallings for sharing primary tumor samples, and NXTGNT for the sequencing services. The work is further supported by the Emmanuel Van der Schueren Foundation (scientific partner of the Flemish League Against Cancer (VLK)), the Fournier-Majoie Foundation (FFM) and the Belgian National Lottery. A.D. was a recipient of a grant of the Research Foundation Flanders (FWO) and an Emmanuel Van der Schueren research grant (VLK).

Author Contributions

Drafting the article: A.D. Data generation: W.V. Data processing: M.O. Technical validation analyses: A.D. and M.O. Overall supervision of study: F.S. and J.V. All authors contributed to preparation of the manuscript and approved the final version.

Additional Information

Table 1 is only available in the online version of this paper.

Supplementary Information accompanies this paper at http://www.nature.com/sdata

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Decock, A. *et al.* DNA methylation profiling of primary neuroblastoma tumors using methyl-CpG-binding domain sequencing. *Sci. Data* 3:160004 doi: 10.1038/sdata.2016.4 (2016).