bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A versatile method for circulating cell-free DNA methylome profiling by reduced representation bisulfite sequencing

De Koker A.[1,2,6]*, Van Paemel R.[3,4,6], De Wilde B.[3,5,6], De Preter K.[3,4,6], Callewaert N.[1,2,6]*

[1]VIB-UGent Center for Medical Biotechnology, Technologiepark 71, B-9052 Ghent, Belgium

[2]Department of Biochemistry and Microbiology, Ghent University, Ledeganckstraat 35, B-9000 Ghent, Belgium

[3]Center for Medical Genetics, Ghent University Hospital, C. Heymanslaan 10, 9000 Ghent, Belgium

[4]Department of Biomolecular Medicine, Ghent University Hospital, C. Heymanslaan 10, 9000 Ghent, Belgium

[5]Department of Pediatric Hematology Oncology and stem cell transplantation, Ghent University Hospital, C. Heymanslaan 10, 9000 Ghent, Belgium

[6]Cancer Research Institute Ghent (CRIG), C. Heymanslaan 10, 9000 Ghent, Belgium

*Correspondence should be addressed to: nico.callewaert@ugent.vib.be and andries.dekoker@ugent.vib.be

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

The methylation profile of circulating cell-free DNA (cfDNA) in blood can be exploited to detect and diagnose cancer and other tissue pathologies and is therefore of great diagnostic interest. There is an urgent need for a cost-effective genome-wide methylation profiling method that is simple, robust and automatable and that works on highly fragmented cfDNA. We report on a novel sample preparation method for reduced representation bisulfite sequencing (RRBS), rigorously designed and customized for minute amounts of highly fragmented DNA. Our method works in particular on cfDNA from blood plasma. It is a performant and cost-effective methodology (termed cf-RRBS) which enables clinical cfDNA epigenomics studies.

The methylation profile of DNA can be exploited to detect and diagnose cancer and other tissue pathologies and is therefore of great diagnostic interest. To analyze the genome-wide DNA methylation status in high-molecular weight genomic DNA that is typically obtained from tissues, many profiling methods have been developed[1]. Reduced representation bisulfite sequencing (RRBS) strikes a particularly good balance between genome-wide coverage and accurate quantification of the methylation status and an affordable cost. It is thus ideally suited for epigenomics studies involving large numbers of samples[2]. However, none of the RRBS methods reported to date are suited for analyzing the minute quantities of highly fragmented circulating cell-free DNA (cfDNA) that can be purified from blood plasma[3,4]. We report on a novel biotechnological sample preparation method for RRBS, rigorously designed and customized for minute amounts of highly fragmented DNA. This robust and automatable methodology (termed cf-RRBS) is highly cost-effective and enables clinical cfDNA epigenomics studies.

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Altered DNA methylation is an almost universal hallmark of oncogenic transformation and, in a broader sense, of tissue pathology. The ability to detect such altered methylation in "liquid biopsies" cost-effectively and using a simple sample preparation workflow, would be transformative in molecular diagnostics. Whole genome bisulfite sequencing (WGBS) on the highly fragmented circulating cell-free DNA is too costly for routine use, as the sequencing capacity is diluted over the entire genome, requiring very deep sequencing to attain reliable quantitation of methylation status. To overcome this, capturing CpG-rich regions of the genome has been developed, using expensive capture reagents consisting of millions of capture probes. This results in complex sample preparation workflows that are not easy to deploy at high throughput. On the other hand, in reduced representation bisulfite sequencing (RRBS), a genomic subsampling is performed by digesting high molecular weight genomic DNA with a restriction enzyme (typically MspI) that cuts CpG-rich sequences, thus focusing bisulfite sequencing to just a few CpG-rich percentages of the genome, enabling high coverage and hence accurate methylation status calling. However, the method requires high molecular weight gDNA as input, incompatible with the highly fragmented nature of cfDNA in plasma. We set out to develop methodology to overcome this.

Upon MspI digestion, all RRBS methods critically depend on size selection to isolate the short MspI/MspI-digested fragments from the "off-target" higher molecular weight gDNA. However, as cfDNA is already highly fragmented from the start, it is impossible to effectively purify MspI/MspI-fragments from non MspI/MspI-fragments of cfDNA, due to extensive size overlap (Figure 1a). Hence, the purpose of focusing bisulfite sequencing on just the MspI/MspI subsection of the genome is lost. Here, we devised a novel way around this problem, by specifically degrading all "off-target" cfDNA fragments that were not generated by MspI digestion. In a first step, we dephosphorylate the input fragmented cfDNA prior to MspI digestion. Subsequently, MspI digestion is performed, resulting in fragments with phosphorylated 5'-ends at MspI cut sites. Upon dA-tailing and ligation of hairpin-shaped adapters, only MspI/MspI-fragments subsequently yield unnicked DNA molecules without

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

free ends. Those 'circular' molecules are resistant to exonuclease digestion, whereas all other DNA in the sample is degraded using a mixture of exonucleases. After adapter opening and bisulfite conversion, the DNA can then be amplified, producing a classical RRBS sequencing library (Figure 1b). All of these steps are to be performed directly on the minute quantities of cfDNA available from plasma samples. We avoided the problem of excessive sample loss that would occur during tube transfers and purification steps, by carefully designing the sequence of enzymatic manipulations, such that they could all be performed in a single tube by simple reagent additions to the starting tube containing the isolated cfDNA. All consecutive DNA enzymatic steps were optimized to work in a single buffer system. Also, precise tuning of the concentration of reagents was required at each step to avoid interference with subsequent enzymatic reactions (e.g. dATP is used by Klenow (exo-) fragment, but inhibits T4 DNA ligase[7,8]). Similarly, we used heat-denaturable variants of some enzymes, such that they could be inactivated by simple heating. Recombinant Shrimp Alkaline Phosphatase was used for cfDNA dephosphorylation and Antarctic thermolabile Uracil-DNA Glycosylase was used during adapter opening. As a result, all of the steps in the cf-RRBS protocol can be performed on nanogram quantities of cfDNA that are typically obtained from milliliter-sized plasma samples, prior to any amplification. The entire sample preparation occurs in the liquid phase, and no purifications are needed throughout, until bisulfite sequencing is completed. This rigorous design for simplicity enables full automation on basic liquid handling stations, which facilitates the use of cf-RRBS in large-scale discovery and, especially, in routine diagnostics.

The major fraction of cfDNA has a maximal length of ±166bp corresponding to one nucleosomal winding (Supplementary Figure 1). Hence, most MspI/MspI-fragments derived from cfDNA are expected to be shorter than ±166bp. Indeed, for the cf-RRBS method, 90.67% of the uniquely mapping reads mapped to MspI/MspI-fragments of 20-165bp, as opposed to only 13.73% for a classical, size selection-based RRBS method run on the same sample, demonstrating that cf-RRBS

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

efficiently enriches the MspI/MspI-fragments (Figure 1c and Supplementary Figure 2). This was shown to be highly reproducible over multiple samples (Supplementary Figure 1b-d and Supplementary Figure 3).

We compared our cf-RRBS method to the main commercial method currently available for sequencing-based cfDNA methylomics of an enriched genomic sub-region, i.e. the SeqCap Epi Enrichment System (Roche) which enriches its genomic target regions using hybridization capture with a huge number of synthetic oligonucleotides. First, we compared the methods from a protocol complexity point of view (Figure 2a). In a step by step comparison, the SeqCap Epi protocol takes much longer than the cf-RRBS protocol. In stark contrast to the single-tube cf-RRBS protocol, the SeqCap Epi protocol consists of many steps, including sample transfers, solid-phase reversible immobilization (SPRI) magnetic bead purifications and vacuum concentrator drying steps. Such drying and purification steps are more complex to automate and difficult to validate in a regulated clinical diagnosis setting. Also, based on catalog pricing, the cost-of-goods of our cf-RRBS sample preparation protocol (± €15/sample) is 10 times lower than that of the SeqCap Epi protocol (± €150/sample). With regard to sequencing, both cf-RRBS (± 50 Mbp) and SeqCap Epi (± 80 Mbp) target a particular subsection of the genome. Our method covers about 3 million CpGs, versus about 5.5 million for the SeqCap Epi (Figure 2b). As the percentage of covered CpGs and the coverage depth over the CpGs correlates with the depth to which the library is sequenced, we factored in the different size of the target genomic regions in a comparison of coverage across the target region for both methods (Figure 2b). To this end, we chose optimal Illumina sequencing lengths for both methods (for details: see online methods) and the cf-RRBS and SeqCap Epi methods were performed in triplicate (Supplementary Figure 1b). After read mapping and CpG methylation extraction, the number of CpGs covered >=1-fold or >=10-fold [9] was calculated, at varying sequence data yield (expressed in Giga base-pairs, Gbp). We observed that, in both methods, the covered CpG count

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

equally rapidly reaches plateau with increasing sequencing depth (Figure 2b). In diagnostics, missing values due to undersampling of the target search space can be a problem, especially as the search space gets bigger. Consequently, at equal sequencing data yield, SeqCap Epi would be expected to result in more missing values across replicates, as the CpG search space is almost twice as big as for cf-RRBS. For example, at 0.6 Gbp there was indeed a larger fraction of CpGs that were not covered >=10-fold in all three of the SeqCap Epi technical replicates, as compared with cf-RRBS, where the overlap was much better. Both cf-RRBS and SeqCap Epi generate highly reproducible methylation data, with correlation coefficients amongst replicates of 0.96 for CpGs covered >= 10 fold (Figure 2b and Supplementary Figure 4).

The correlation between both methods with regard to individual CpGs methylation levels is also high with correlation coefficients of >0.97 for CpGs covered >=10 fold (Figure 2c and Supplementary Figure 5). As expected, when the methylation values are first averaged over the three replicates of both the cf-RRBS and SeqCap Epi method, an even somewhat tighter correlation between the methylation calls is observed (r > 0.98) (Figure 2c). From these technical studies, we conclude that data quality of cf-RRBS is equivalent to the capture-based method, while the method is much simpler, more cost-effective and thus more scalable.

While a very large application spectrum in clinical diagnostics now awaits exploration with cf-RRBS, we conclude with an example of a possible clinical application. In pediatric oncology, tumor type determination to guide treatment decision is often challenging due to the lack of a distinctive histology. To explore whether epigenetic profiling of cfDNA could assist in determining tumor type, WGBS, SeqCap Epi and our cf-RRBS method were performed on two neuroblastoma serum cfDNA samples (Supplementary Figure 6) and the cf-RRBS method was also performed on an additional eight neuroblastoma plasma cfDNA samples (Supplementary Figure 6). As methylation is typically co-regulated in genome regulatory elements, the methylation status was averaged over the CpGs that are localized in each of the already published CancerLocator clusters[10]. From large-scale epigenomics

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

studies on resected/post-mortem tumor tissues, tumor epigenome data is available for many tumor types, which can serve as a reference for cfDNA methylome-based tumor type assignment. Infinium HumanMethylation 450K microarray data of *MYCN* amplification positive (n = 51) or negative (n = 169) neuroblastoma reference tumors and healthy adrenal tissue (n = 2) was retrieved from the public TARGET database or the NCBI Gene Expression Omnibus. In addition HM450K data of Wilms' tumors (n = 131) was retrieved. The close anatomical proximity of adrenal (neuroblastoma) and renal (e.g. Wilms') tumors can make them difficult to differentiate from one another at diagnosis. Because white blood cells are the most prevalent confounding source of DNA in cfDNA samples, we also used the prepubertal white blood cell methylation array data (n = 51) from Almstrup et al.[11] as a reference. The Uniform Manifold Approximation and Projection (UMAP) dimension reduction technique was used to visualize the similarity between the CancerLocator cluster methylation values obtained for both the patient cfDNA and the tumor tissue/white blood cell DNA from the public resources. Three clusters are formed on the UMAP plot, with two overlapping subclusters in the neuroblastoma group, corresponding to the *MYCN* amplification status. The cfDNA methylation data cluster together with the correct tumor group irrespective of the analytical method, in accordance with the tight correlation of methylation calls in all methods (Figure 3). While this experiment is exploratory, tumor tissue of origin assignment based on cf-RRBS hence is feasible and is currently the subject of larger-scale clinical validation in our laboratories.

In conclusion, we provide the first effective RRBS technology for cfDNA that is able to overcome the challenges associated with low abundant and highly fragmented input material. cf-RRBS stands out as striking a particularly good balance between genome methylation coverage, reproducibility, ease of execution and affordability. The method was entirely designed with clinical laboratory validation and automation in mind. Beyond the neuroblastoma tumor type assignment proof of concept presented here, clinical studies to further explore its applications are in process. Building on ongoing

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

rapid developments in tumor tissue epigenome sequencing and epigenome sequencing of a host of non-cancerous pathologies (such as immune/inflammatory[12], degenerative diseases[13], infectious diseases[14] and substance abuse[15]), cf-RRBS is poised for broad clinical utility. cf-RRBS is performed using a series of ordinary enzymatic reagents that can be easily validated for use in molecular diagnostic products. Being species-independent, our method can be used e.g. for veterinary medicine applications as well. This method hence cost-effectively unlocks the power of epigenetic profiling of cfDNA in liquid biopsies, in a methodologically straightforward manner.

Methods

Methods, including statements on data availability and any associated accession codes and references, are available at (insert hyperlink for online methods).

Acknowledgments

We thank Hans Van Vlierberghe and Xavier Verhelst for the collection of blood samples from patients with liver disease. We thank Tim Lammens and Geneviève Laureys for providing the neuroblastoma samples. We thank Jo Vandesompele, Anneleen Decock, Wim Van Criekinge and their teams for the reuse of WGBS data on two neuroblastoma patients. A.D.K. and R.V.P were funded by a predoctoral fellowship from the Research Foundation Flanders (FWO), B.D.W is an FWO senior clinical investigator. Research was funded by VIB and Ghent University.

Competing financial interest

De Koker A. and Callewaert N. are listed as inventors in patent application PCT/EP2017/056850 related to the methods disclosed in this manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Materials & Correspondence

Andries De Koker andries.dekoker@vib-ugent.be

Nico Callewaert nico.callewaert@vib-ugent.be

References

1. Yong, W.-S., Hsu, F.-M. & Chen, P.-Y. Profiling genome-wide DNA methylation. *Epigenetics Chromatin* **9**, 26 (2016).
2. Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33**, 5868–5877 (2005).
3. Wen, L. *et al.* Genome-scale detection of hypermethylated CpG islands in circulating cell-free DNA of hepatocellular carcinoma patients. *Cell Res.* **25**, 1250–1264 (2015).
4. Tanić, M. & Beck, S. Epigenome-wide association studies for cancer biomarker discovery in circulating cell-free DNA: technical advances and challenges. *Curr. Opin. Genet. Dev.* **42**, 48–55 (2017).
5. Jiang, P. & Lo, Y. M. D. The Long and Short of Circulating Cell-Free DNA and the Ins and Outs of Molecular Diagnostics. *Trends Genet.* **32**, 360–371 (2016).
6. Breitbach, S. *et al.* Direct Quantification of Cell-Free, Circulating DNA from Unpurified Plasma. *PLoS ONE* **9**, (2014).
7. Setlow, P. DNA polymerase I from Escherichia coli. *Methods Enzymol.* **29**, 3–12 (1974).
8. Cherepanov, A. V. & de Vries, S. Kinetics and thermodynamics of nick sealing by T4 DNA ligase. *Eur. J. Biochem.* **270**, 4315–4325 (2003).
9. Encode protocol for RRBS. *Encode protocol for RRBS* Available at: https://genome.ucsc.edu/encode/protocols/dataStandards/. (Accessed: 22nd October 2018)

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

10. Kang, S. *et al.* CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biol.* **18**, 53 (2017).

11. Almstrup, K. *et al.* Pubertal development in healthy children is mirrored by DNA methylation patterns in peripheral blood. *Sci. Rep.* **6**, 28657 (2016).

12. Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* **31**, 142 (2013).

13. Urdinguio, R. G., Sanchez-Mut, J. V. & Esteller, M. Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies. *Lancet Neurol.* **8**, 1056–1072 (2009).

14. Pacis, A. *et al.* Bacterial infection remodels the DNA methylation landscape of human dendritic cells. *Genome Res.* **25**, 1801–1811 (2015).

15. Cecil, C. A. M. *et al.* DNA methylation and substance-use risk: a prospective, genome-wide study spanning gestation to adolescence. *Transl. Psychiatry* **6**, e976 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig. 1

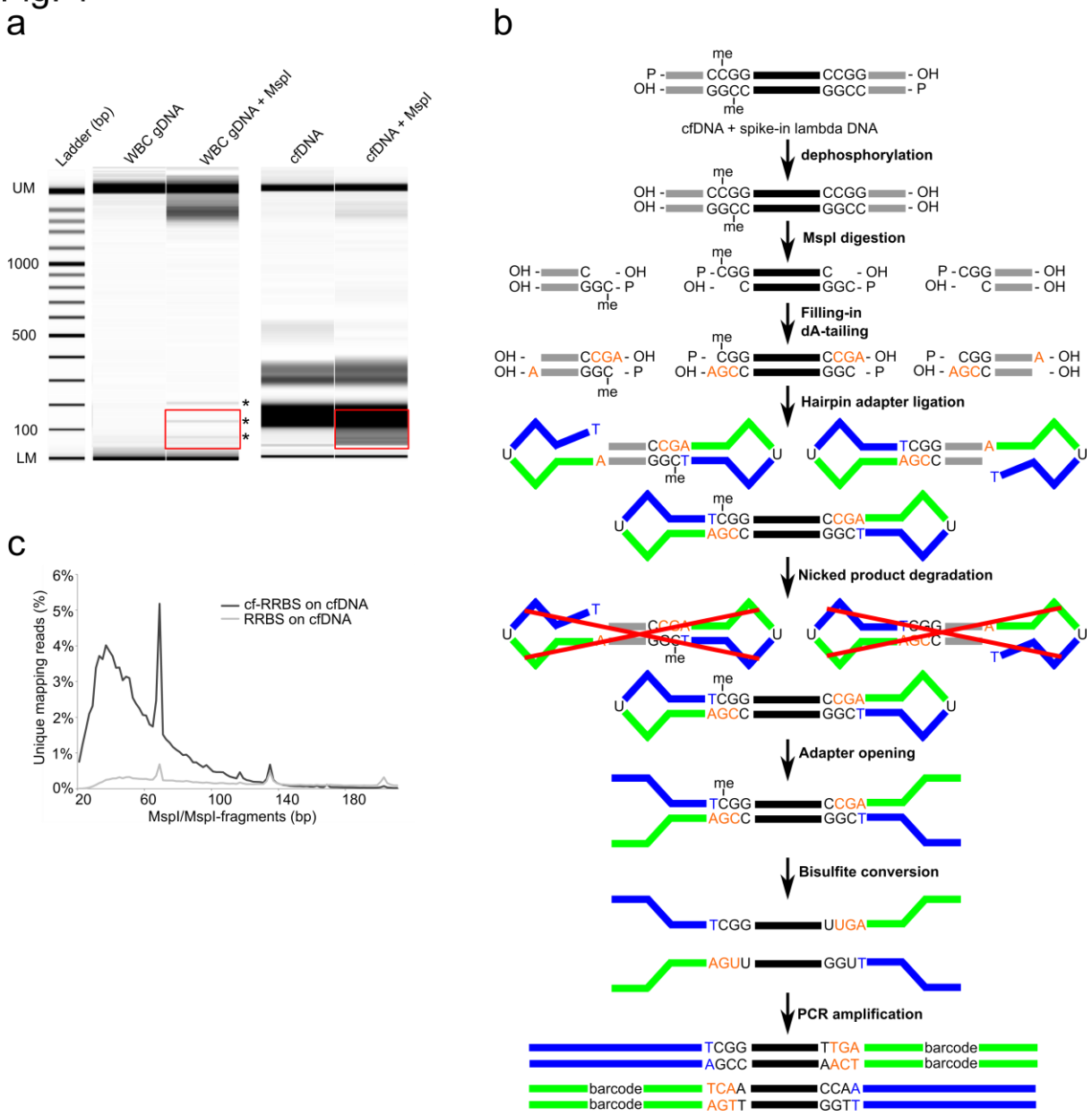


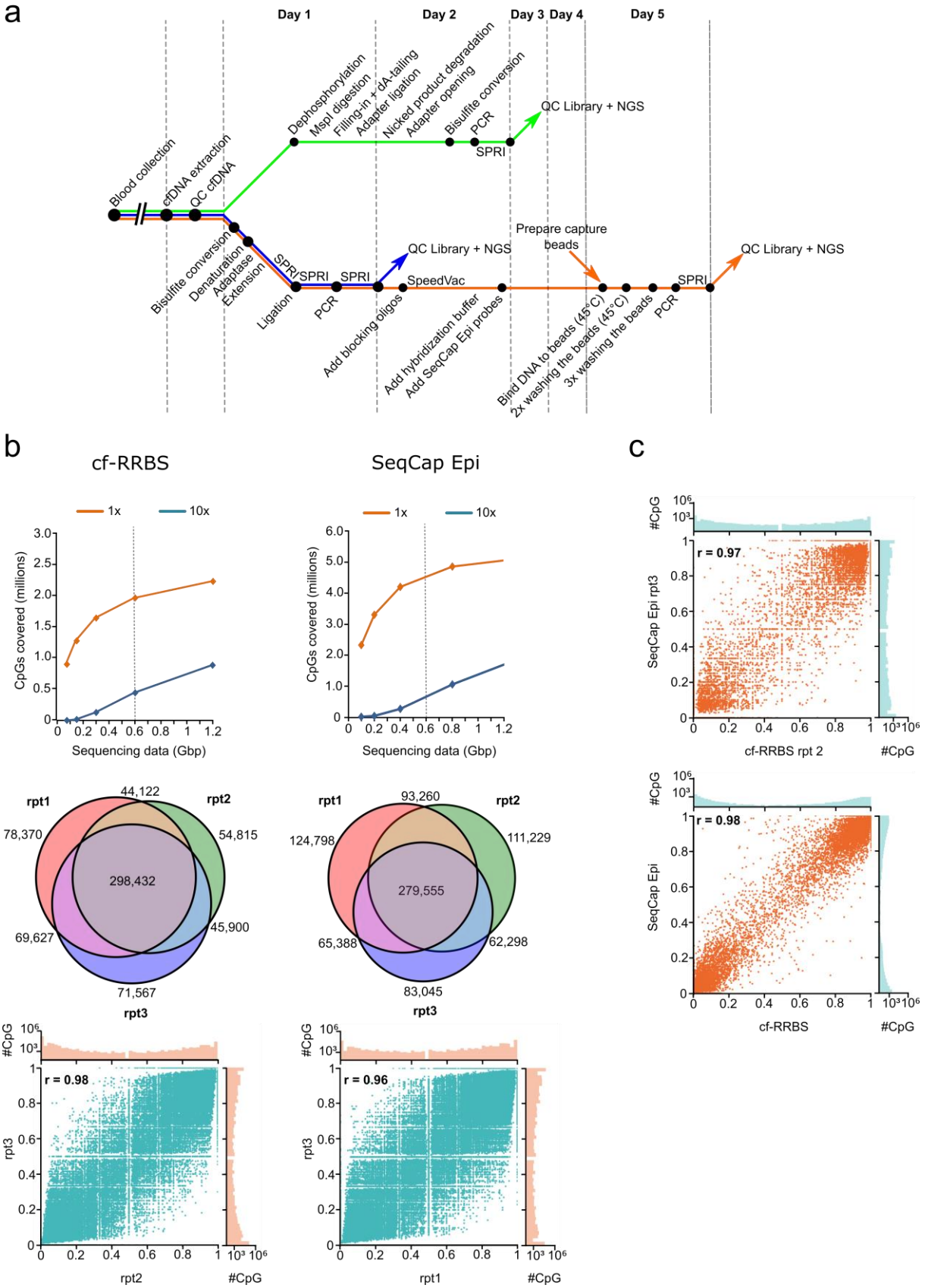
Figure 1: The cf-RRBS method is optimized to obtain a sequencing library enriched in MspI/MspI – fragments when starting from highly fragmented circulating cell-free DNA (cfDNA). (a) Capillary gel electropherograms of white blood cell (WBC) genomic DNA (gDNA) and cfDNA, either undigested or MspI-digested. The undigested WBC gDNA runs above the upper marker (UM) and is not observed on the gel image. After MspI digestion, a smear of shorter DNA appears. Also three characteristic satellite DNA bands can be observed resulting from the digestion of high-abundant DNA repeats in the genome. The undigested cfDNA sample shows the expected apoptotic DNA pattern of a multiple of nucleosome repeats. After MspI digestion, especially the DNA fragment of three nucleosome repeats is not observed anymore. In addition, a smear of shorter DNA fragments is observed beneath the one nucleosome repeat. In classic RRBS, the MspI/MspI-fragments ranging from 30 to 160 bp (indicated with a red rectangle) would be extracted from gel. Starting from cfDNA, extracting these fragments without co-extracting input cfDNA is impossible as these completely overlap in size range.

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(b) In the cf-RRBS workflow the cfDNA sample is first dephosphorylated prior to MspI digestion. Only MspI/MspI-fragments yield unnicked DNA molecules without free ends upon dA-tailing and ligation of hairpin-shaped adapters. Those 'circular' molecules are resistant to exonuclease digestion, whereas all other DNA is degraded using a combination of exonucleases. All these steps are performed on the nanogram quantities of cfDNA typically obtainable from patient's blood samples, necessitating thorough development of these reactions so that they can now be performed without any tube transfers or purifications by subsequent reagent addition and incubation. (c) The graph shows that a large portion of the uniquely mapping reads originate from MspI/MspI-fragments between 20 to 165 bp when performing cf-RRBS on cfDNA, while this is not the case when performing classical RRBS on the same cfDNA (AUC(cf-RRBS) = 90.67% vs AUC(RRBS) = 13.73%).

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig. 2



bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 2: Comparison of the cf-RRBS method to the SeqCap Epi enrichment method. (a) The timelines of the cf-RRBS method in green, WGBS method in blue and SeqCap Epi in orange. The SeqCap Epi protocol takes longer than the cf-RRBS protocol and comprises of more steps that are difficult to automate, such as sample transfers (indicated with black dots), solid-phase reversible immobilization (SPRI) magnetic bead purifications, vacuum concentration (SpeedVac) and temperature-sensitive steps (45°C). (b) Three cf-RRBS and three SeqCap Epi libraries were generated starting from the same cfDNA. The libraries were sequenced by method-optimal Illumina sequencing. For one technical replicate (rpt), the number of CpGs covered by minimum 1 or 10 sequencing reads at different amounts of generated raw sequencing data (in Giga base pairs (Gbp)) for cf-RRBS (left) and SeqCap Epi (right) is shown. The Venn-diagrams show the number of shared CpGs covered 10x between the 3 technical replicates, when 0.6Gbp raw data is generated, meaning 8 million PE75 reads for cf-RRBS and 6 million PE100 reads for SeqCap Epi. Also, a representative scatter plot is shown, comparing the methylation data obtained in two technical cf-RRBS (left) and SeqCap Epi replicates (right). A Pearson correlation coefficient of 0.98 and 0.96, respectively, indicates the technical reproducibility of the two methods. (c) (top) A representative scatter plot is shown comparing the methylation data obtained on a cf-RRBS and a SeqCap Epi analysis of the same cfDNA sample. A Pearson coefficient of 0.97 indicates that cf-RRBS produces highly similar data as the SeqCap Epi method. (bottom) If the methylation level per CpG is averaged over all three replicates for cf-RRBS as well as for SeqCap Epi, an even more narrow correlation is observed, as expected.

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig. 3

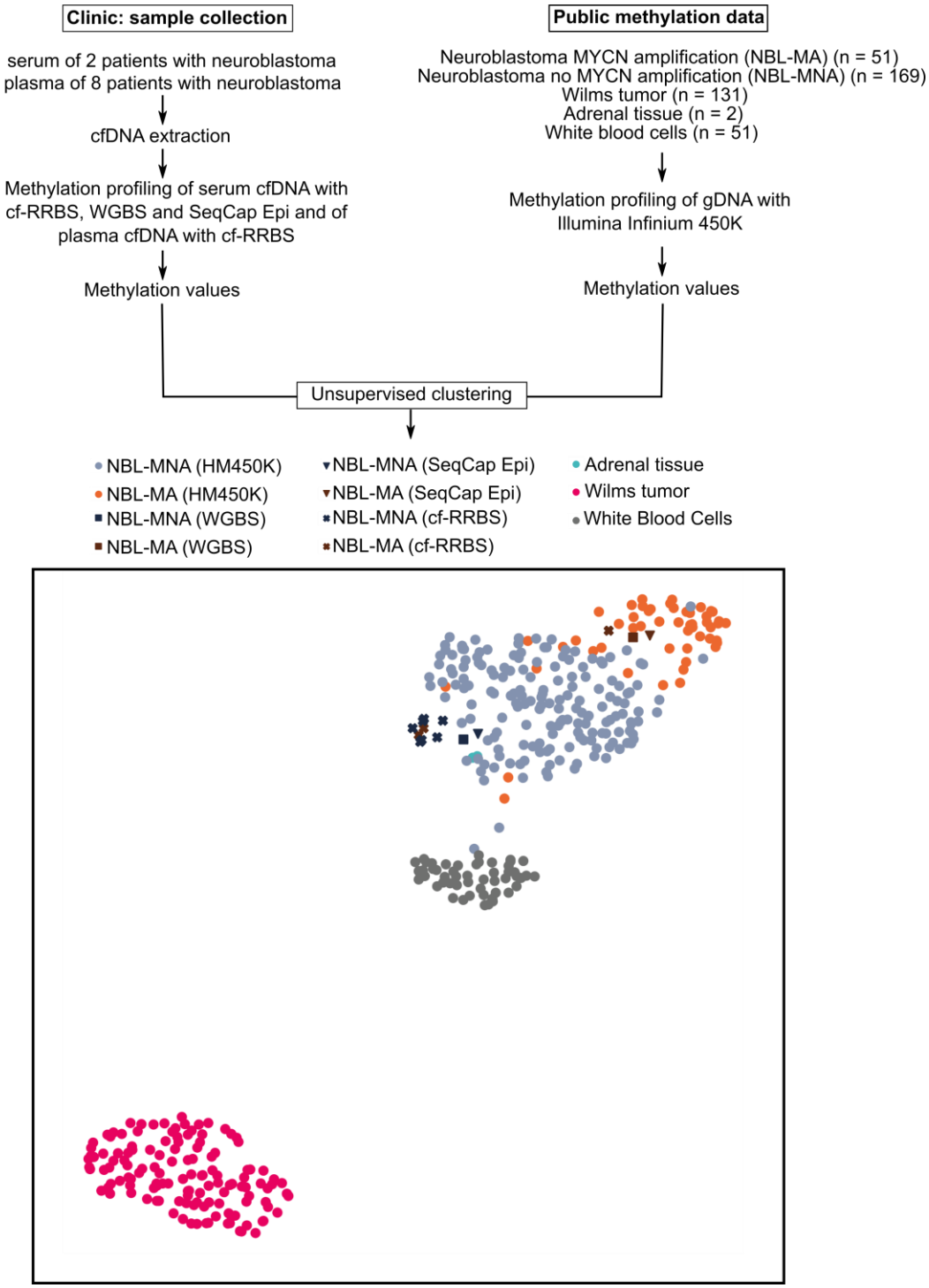


Figure 3: Methylation data obtained by cf-RRBS allows for clinically useful assignment of tumor type in a neuroblastoma case study. Unsupervised clustering on CancerLocator region-averaged methylation status, followed by Uniform Manifold Approximation and Projection (UMAP) dimension reduction. cfDNA methylation data obtained by performing WGBS, SeqCap Epi and cf-RRBS on serum cfDNA samples (n = 2) and cf-RRBS on plasma cfDNA samples (n = 8) of neuroblastoma patients. Illumina Infinium 450K methylome data obtained from the public TARGET database for *MYCN*

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

amplification positive (n = 51) or negative (n = 169) neuroblastoma patients, Wilms' tumor patients (n = 131), adrenal tissue samples (n = 2) and white blood cell samples (n = 51). Three clusters are observed on the UMAP plot, with two neuroblastoma subclusters largely depending on MYCN amplification status. The cfDNA samples clearly cluster to the neuroblastoma group, regardless of the method used for methylation assessment, highlighting the correlation between the three methods and demonstrating that cf-RRBS is fully suitable for this task. In addition, neuroblastoma serum cfDNA samples cluster according to the correct neuroblastoma subgroup. However, neuroblastoma plasma cfDNA samples do not. Clustering of a few of the neuroblastoma samples downloaded from the TARGET database, together with white blood cells (n = 2), is presumably due to contamination of the biopsy in these previous studies.

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Online methods

Sample preparation

Circulating cell-free DNA from blood plasma of hepatocellular carcinoma patients. 10 mL blood was drawn in PAXgene Blood ccfDNA tubes (Qiagen, Cat. No. 768115) from males with hepatocellular carcinoma (Approval by Ethical committee UZGent - Prof. Dr. D Matthys – EC UZG 2016/0532). Within 4 days, plasma was obtained by centrifugation (without brake) of the blood at 1900xg for 15 minutes at room temperature. Circulating cell-free DNA was extracted from 5 mL plasma using the Quick-cfDNA™ Serum & Plasma Kit (ZymoResearch, Cat. No. D4076) following the manufacturer's instructions.

Circulating cell-free DNA from blood serum or blood plasma of neuroblastoma patients. Two neuroblastoma patients (NBL1 & NBL2) were retrospectively included in the present study (Ethical approval for the use of retrospective samples: EC2008-104/mf). NBL1 was a Caucasian boy of 4 years and 9 months old at diagnosis. The primary site at presentation was the retroperitoneum. Bone marrow puncture and MIBG scan were positive for metastatic disease (INSS stadium 4). Fluorescence in-situ hybridization (FISH) did not show *MYCN* amplification. NBL2 was a Caucasian girl of 3 years and 5 months old at diagnosis. At diagnosis there was extensive disease present, spreading to the retroperitoneum, right kidney and mesenterium. Bone marrow puncture and MIBG scan were positive for metastatic disease spread to the liver, bone and bone marrow (INSS stadium 4). FISH revealed MYCN amplification. Between 4 to 10 mL of blood was drawn in a serum collection tube from these two patients. Serum was obtained by leaving it at room temperature for 20 minutes. The tube was centrifuged for 10 minutes at 1000xg and the serum was separated from the blood clot. Serum was frozen at -80°C before circulating cell-free DNA was extracted using the QIAamp Circulating Nucleic Acid Kit (Qiagen, Cat. No. 55114) following the manufacturer's instructions.

Circulating cell-free DNA was extracted from plasma of eight neuroblastoma patients (NBL3-10).

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

NBL patient	Age at diagnosis (months)	INSS stage	MYCN amplification
3	3	4s	No
4	46	4	No
5	44	4	No
6	21	4	Yes
7	NS	1	No
8	13	1	Yes
9	2	4s	No
10	4	1	No

Therefore, blood was drawn in K_2-EDTA-tubes from boys and girls with neuroblastoma (EC: EC2008-104/mf). Plasma was obtained by centrifugation (acceleration and deceleration of 2) of the blood at 1600xg for 10 minutes at 4°C followed by centrifugation at 16000xg for 10 minutes at 4°C and collection of the supernatant. The plasma was stored at -80°C until processing for cfDNA extraction. Circulating cell-free DNA was extracted from 3.5 mL plasma (the total volume was adjusted to 3.5 mL by adding 1X PBS if the plasma volume was smaller than 3.5 mL) using the Maxwell RSC LV ccfDNA kit (Promega: AS1480) following the manufacturer's instructions. DNA was eluted in 50 µL of AVE buffer (Qiagen).

Unmethylated phage lambda genomic DNA was obtained from Promega (Cat. No. D1521).

NGS library preparation and sequencing

The **RRBS protocol**: 10 ng of cfDNA and 0,1 % w/w of unmethylated lambdaDNA was digested for 30 minutes at 37°C with 10 units of MspI (New England Biolabs (NEB), Cat. No. R0106S) in a 20 µL CutSmart-buffered (NEB, Cat. No. B7204S) reaction. To the mixture, 0.9 µL of 1 mM dNTP mix (Promega, Cat. No. U1511), 2.5 units of Klenow Fragment (3' to 5' exo-) (NEB, Cat. No. M0212S), 0.5 µL 10x CutSmart buffer and 3.1 µL of H_2O (Thermo Fisher Scientific, Cat. No. AM9938) was added and incubated for 20 minutes at 30°C and 20 minutes at 37°C. The reaction was stopped by heat inactivation during 30 minutes at 75°C. Before purification by Nucleospin gel and PCR cleanup (Macherey-Nagel, Cat. No. 740609.250), 400 ng of carrier DNA was added. As carrier DNA we used 4

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

µg of unmethylated lambda DNA which was previously incubated for 1 hour at 37°C with 20 units of MspI and 1 unit of rSAP (NEB, Cat. No. M0371S) in a CutSmart-buffered 20 µL reaction. The reaction was stopped by heat inactivation for 30 minutes at 75°C. A completely methylated adapter was prepared. The adapter was ordered at IDT as two oligos (5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3' and 5'-/Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3'). To allow for adapter hybridization, 10 µL of both 100µM oligo solutions were mixed with 2.5 µL 10X T4 DNA ligase buffer, 2.5 µL H_2O and heated for 1 minute at 95°C, then cooled down to room temperature at 0.1°C/s in a thermocycler. After cleanup of the Klenow (3' to 5' exo-) reaction, the eluate (14 µL) was mixed with 1 µL of 40 µM methylated adapter, 2 µL of 10X T4 DNA ligase buffer (NEB, Cat. No. B0202S) and 2 µL of H_2O. Then, 400 units of T4 DNA ligase (NEB, Cat. No. M0202S) was added. The reaction was incubated overnight at 16°C and stopped by heat inactivation for 20 minutes at 65°C. The ligation mixture was loaded on a 3% agarose gel and ran for 1.5 hours at 105V. The gel was imaged with Ethidium Bromide and the gel section corresponding to 130-280 bp was cut out. DNA was extracted from the gel using the Nucleospin gel and PCR cleanup kit and eluted in 20 µL of H_2O. The DNA was bisulfite converted using the EZ DNA Lightning kit (ZymoResearch, Cat. No. D5030) following the manufacturer's instructions. The bisulfite-treated sample (14µL) was then mixed with 15 µL of KAPA HiFi HotStart Uracil+ ReadyMix PCR Kit (KAPA Biosystems, Cat. No. KK2801) and 0.9 µL of 10 µM of NEBNext universal primer and 0.9 µL of 10µM of NEBNext index primer, both ordered at IDT, and amplified according to the PCR protocol: 5 minutes 95°C, 20x(20 seconds 98°C, 15 seconds 65°C, 45 seconds 72°C), 5 minutes 72°C, terminal cool to 4°C.

The **cf-RRBS protocol** was performed in a BioRad thermocycler with heated lid (105°C). 10 ng (5 µL) of circulating cell-free DNA (the sample was quantified using the FEMTO Pulse Automated Pulsed-Field CE Instrument from Advanced Analytical Technologies) was dephosphorylated in a 10 µL reaction by adding a 5 µL mixture containing 1 µL 0.01 ng (i.e. 0,1 % w/w) of unmethylated

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

lambdaDNA, 1 µL recombinant Shrimp Alkaline Phosphatase (rSAP) (1U/µL), 1 µL of 10X CutSmart buffer and 2 µL of H_2O. The mixture was pipetted in 0.2 mL thin-walled PCR tubes and incubated for 1 hour at 37°C. The phosphatase was heat killed for 30 minutes at 75°C. Next, the DNA was digested by adding 5 µL of a mixture containing 0.5 µL of MspI (20U/µL), 0.5 µL of 10X CutSmart and 4 µL of H_2O and incubated for 30 minutes at 37°C. End-repair and A-tailing was performed by adding a 10 µL-mixture of 0.5 µL of Klenow Fragment (3'-->5' exo-) (5U/µL), 0.25 µL of dNTP mix (40µM dATP, 4µM dCTP and 4µM dGTP) (Promega, Cat. No. U1330), 1 µL of 10X CutSmart and 8.25 µL of H_2O. This 25 µL-mixture was incubated for 20 minutes at 30°C, for 20 minutes at 37°C and heat killed for 20 minutes at 75°C. Next, adapter was ligated to the DNA-fragments. A 10 µL-mixture containing 4 µL of 10 mM ATP (NEB, Cat. No. P0756S), 1 µL of 10 µM adapter, 1 µL of 10X CutSmart and 4 µL of H_2O and a 5 µL-mixture containing 0.5 µL of T4 DNA ligase (2000 U/µL) (NEB, Cat. No. M0202M), 0.5 µL of 10X CutSmart and 4 µL of H_2O was added. The adapter is identical to the NEBNext-adapter (seq: 5´-/Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTC/ideoxyU/ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3´) and was ordered at IDT, but without phosphorothioate bond at the 3'-end. Ligation was done overnight (14 hours) at 16°C and the ligase was heat killed at 65°C for 10 minutes. The next day, a mixture of 0.5 µL of exonuclease I (20 U/µL) (NEB, Cat. No. M0293S), 0.5 µL of exonuclease III (50 U/µL) (NEB, Cat. No. M0206S), 0.5 µL of exonuclease VII (10U/µL) (NEB, Cat. No. M0379S), 0.5 µL of 10X CutSmart and 3 µL of H_2O was pipetted to the 40µL-reaction. This mixture was incubated for 2 hours at 37°C and heat killed for 10 minutes at 80°C and 10 minutes at 95°C. As a last step before bisulfite conversion, 1 µL of Antarctic Thermolabile UDG (1U/µL) (NEB, Cat. No. M0372S), 0.5 µL of endonuclease VIII (10U/µL) (NEB, Cat. No. M0299S), 0.5 µL of 10X CutSmart and 3 µL of H_2O was added and incubated for 1 hour at 37°C and heat killed for 10 minutes at 75°C.

The MspI/MspI-digested adapter-ligated DNA fragments were bisulfite converted using the EZ DNA Methylation-Lightning™ Kit according to manufacturer's instructions. Bisulfite-converted DNA was eluted in 14 µL of H_2O. A final amplification step was performed using the KAPA HiFi HotStart Uracil+

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ReadyMix PCR Kit. 15 µL of 2X polymerase mix was combined with 13.2 µL of DNA, 0.9 µL of 10 µM NEBNext universal primer (IDT) and 0.9 µL of 10 µM NEBNext index primer (IDT). The PCR protocol was: 5 minutes 95°C, 17-19x(20 seconds 98°C, 15 seconds 65°C, 45 seconds 72°C), 5 minutes 72°C, forever 12°C.

Libraries prepared using the cf-RRBS and RRBS protocol were cleaned up by magnetic bead cleanup (CleanNA PCR) (GC biotech, Cat. No. CPCR-0050) and eluted in 0.1X TE buffer. The libraries were visualized with the Fragment Analyzer (Advanced Analytical Technologies, Inc.) and quantified using the Qubit dsDNA HS Assay Kit (Thermo Fisher scientific, Cat. No. Q32851). Based on the concentration value obtained, the libraries were pooled and were sequenced on a NextSeq500 instrument (Illumina), performing a PE75 run using 5% phiX. It was chosen to perform PE75 sequencing on the cf-RRBS libraries, because the MspI/MspI-inserts are between 20 to 165 bp long. By sequencing too short, a substantial amount of the insert would not be read, while sequencing too long, a lot of duplicate data would be read and even, when read length is longer than the insert, irrelevant sequencing data that originates from the adapter would be read.

The **WGBS protocol** was performed by the genome analysis service facility, NXTGNT (Ghent, Belgium), on 50 ng of cfDNA extracted from blood serum of neuroblastoma patients. The Accel-NGS methyl-seq DNA library kit for Illumina with associated indexing kit A (Swift Biosciences, Cat. No. 30024 and 36024) was used according to manufacturer's instructions. The EZ DNA Methylation-Lightning™ Kit was used for bisulfite conversion. Sequencing was performed on a HiSeq2500 (Illumina) at BGI (China) using one lane per sample, performing a PE125 run, V4 reagent.

The Accel-NGS Methyl-Seq DNA library kit with associated indexing kit A was combined with the **SeqCap Epi enrichment system** (Roche Nimblegen, Cat. No. 07138881001). As an internal standard, unmethylated lambda DNA was sheared to ± 150bp with a S202 Focused-Ultrasonicator (Covaris) and spiked-in at 0.1% w/w before library preparation. Library preparation was performed on 10 ng of

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

input cfDNA following manufacturer's instructions, taking into account their proposed adaptations when working with cfDNA and combining with the NimbleGen SeqCap Epi Hybridization capture probes. To cut costs, before hybridization capture, six libraries were pooled (166 ng/sample) and also 1 nmol of all appropriate SeqCap HE Index Blocking Oligo's were added. Sequencing was performed on a HiSeq4000 (Illumina) using two lanes, performing a PE150 run, V4 reagent. For SeqCap Epi libraries, it was reasoned PE100 sequencing is optimal, because inserts are the full-length cfDNA fragments of ± 165 bp long. In practice, we sequenced PE150, but to allow for proper comparison with the cf-RRBS method we trimmed the raw reads to 100 bp before further analysis.

Data analysis.

Raw reads were demultiplexed based on sample ID and data from different flow cells were concatenated, generating two files containing the first and the second reads per sample. Depending on the type of analysis, the seqtk-tool (v1.2-r101) was used to crop and/or down-sample the raw sequencing reads. Then, the raw reads were trimmed (default: Phred score cut off: 20, maximal trimming error rate: 0.1, minimal required detected adapter: 1 bp and min required read length: 20 bp) using Trim Galore (v0.4.3), depending on Cutadapt (v1.13) and Python (v2.7.5), with method-specific options enabled, generating two trimmed read files and two trimming reports. Subsequently, the human reference genome (hg19.fa) and lambda genome (lambda.fa) were prepared for mapping bisulfite-converted reads. This entailed converting the genome into a fully bisulfite-converted C>T and G>A version and indexing these by bowtie2, through use of the bismark_genome_preparation package, available in the bismark-tool[16] (v0.17.0). Next, the bismark package was used with the 'un' 'ambiguous' and 'ambig_bam' options enabled. This package bisulfite-converts the forward reads C>T and the reverse reads G>A and makes four parallel alignments whereof the best is kept. Seven files are generated: the bam-file of unique mapping reads, the amig.bam-file of reads mapping equally well to one locus as to another, where after one locus is randomly chosen, two fastq-files listing the ambiguous mapping reads, two fastq-files listing the unmapped reads and a mapping

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

report. Last, a methylation call was performed by use of the bismark_methylation_extractor package with the 'single' (only for cf-RRBS libraries on NBL1 & NBL2) or 'paired', 'comprehensive', 'merge_non_CpG', 'bedGraph', 'cytosine_report' and method-specific options enabled. The 'paired' option will extract the methylation status only once, based on the forward read, in the overlapping part of a read pair. The 'comprehensive' option combined with the 'merge_non_CpG' option will pool the 4 strand-specific methylation calls of 3 C contexts (CpG, CHG and CHH) to two context-dependent files (CpG and non_CpG). The 'bedGraph' combined with the cytosine_report option generates a .bedGraph-file and a .cov-file. The .bedGraph-file gives the CpG coordinate and the methylation percentage, the .cov-file also adds the read coverage per CpG. Two additional files are obtained, a splitting report and a M-bias file. The splitting report is a summary report (this report, generated from reads mapping to the lambda genome, was used to calculate the bisulfite conversion efficiency) and the M-bias file gives the methylation proportion across every position in the read revealing possible technically introduced artefacts. See "supplementary_code_to_raw_data.txt" for exact code and details.

The CpG-track (chr 1:22, X, Y, M) was obtained using the BSgenome.Hsapiens.UCSC.hg19(v1.4.0)- and the Biostrings (v2.47.0)-package, installed with BiocInstaller(v1.28.0) from Bioconductor(v3.6). The MspI-track (chr 1:22, X, Y, M) was obtained using the BSgenome.Hsapiens.UCSC.hg19(v1.4.0)- and the Biostrings(v2.47.0)-package, but also the rtracklayer(v1.38.2)[17]- and the HiTC(v1.22.0)[18]-package, all available from Bioconductor. A MspI/MspI-fragment was defined as 5'-CGGNN…NNCCG-3'. Making an 'expected cf-RRBS'-subset, i.e. MspI/MspI-fragments of length 20 to 165 bp, was done by the BEDTools suite[19](v2.25.0) and AWK (v4.0.2) programming. Therefore, the MspI/MspI-fragments with length 22 to 167 bp were extracted, because the fragments in the MspI-track are defined with 2 bp extra (e.g. a 20 bp long MspI/MspI-fragment is biologically a double-stranded molecule 5'-CGGN$_{16}$C-3' and 3'-CN$_{16}$GGC-5', but is defined in the MspI-track as 5'-CGGN$_{16}$CCG-3', thus 22 bp). The

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

CancerLocator CpG cluster-track was downloaded from Kang et al.[10]. See "Supplementary_code_to_generate_genome_track.txt" for exact code and details.

The BEDTools suite and AWK programming were used to overlay the sequencing data and the MspI track and to check the performance of the cf-RRBS versus the RRBS protocol. It was checked how many of all mapping reads hit MspI/MspI-fragments of different length intervals. Hereto, for every MspI/MspI-fragment it was counted how many reads were 100% overlapping. Also the MspI/MspI-fragment itself had to be covered by the read for a minimum of 11.976%. This percentage was obtained by dividing the shortest read length (=20 bp) by the longest "cf-RRBS-expected" MspI/MspI-fragment (=167 bp). All MspI/MspI-fragments that were covered more than one time were binned per 2 lengths (e.g. 20&21 bp, 22&23 bp, …) to obtain a smoothed distribution (Figure 1c) and the number of reads hitting each MspI/MspI-fragment belonging to the bin was summed and divided by the total unique mapping reads to obtain the percentage unique mapping reads per bin of MspI/MspI-fragments. See "Supplementary_code_MspI_enrichment.txt" for exact code and details.

The BEDtools suite and AWK programming was used to extract all individual CpGs covered and their methylation percentage, and this for three technical cf-RRBS and SeqCap Epi replicates . For the replicate with the highest number of raw reads, CpGs covered >=1x or >=10x were counted at different downsampled number of raw reads, being 1, 2, 4, 8 and 16 million. At 0.6Gbps, CpGs that are covered >=10x in all three cf-RRBS or SeqCap Epi replicates can be counted and visualized in a Venn-diagram. Python (v3.6) programming was used to pairwise visualize the methylation level of the CpGs that overlap between all three cf-RRBS or SeqCap Epi replicates in scatter plots and to calculate the Pearson correlation coefficient. Also, the methylation level of the CpGs that overlap between all three cf-RRBS and SeqCap Epi replicates were visualized in scatter plots in a pairwise fashion. Finally, the methylation levels of the CpGs that overlap between all three cf-RRBS and SeqCap Epi replicates were averaged before visualization in a scatter plot (Bottom plot Figure 2c). See "Supplementary_code_CpG_coverage.txt" for exact code and details.

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To evaluate whether the methylation signature derived from the cfDNA with cf-RRBS, WGBS and SeqCap Epi is similar to that obtained from gDNA of the primary tumor type, we used the Uniform Manifold Approximation and Projection dimension reduction technique to visualize the similarity between the three techniques. UMAP is similar to t-Distributed Stochastic Neighbor Embedding (t-SNE) but is better at preserving the global structure of the underlying data[20]. For neuroblastoma primary tumors, Wilms tumors and adrenal tissue, Infinium HumanMethylation 450K microarray data were obtained from the TARGET initiative (n = 353) or the NCBI Gene Expression Omnibus (n = 2). Neuroblastoma samples from TARGET with unspecified *MYCN* status were removed from the sample list. As it was possible that a large proportion from the cfDNA methylation signature can variably originate from DNA from white blood cells[21], we also plotted the prepubertal white blood cell methylation profiles from HM450K from Almstrup et al.[11] (n=51) in order to assess such contamination of the cfDNA samples and the biopsies used for the TARGET data. As CpG methylation is typically co-regulated within a particular genome-regulatory region, we used CancerLocator regions as defined from Kang et al.[10] to group the CpGs into clusters in order to obtain a region-averaged methylation value supported by more reads. After merging all samples in a single matrix (Python 3.6.3 & pandas 0.20.3), neuroblastoma samples were annotated with either *MYCN* amplification (NBL-MA, n = 51) or no MYCN amplification (NBL-MNA, n = 169). CpG clusters with missing values were removed (17251, 59.29%). These clusters are those for which not all of the cfDNA and TARGET datasets had a sufficient support for reliable methylation calling (<30 NGS reads per cluster and/or >50% of array-values are NA). The umap module (0.2.4) with default parameters was used for dimensionality reduction and matplotlib 2.1.0 was used for plotting UMAP component 1 and 2. See “Supplementary_code_UMAP.txt” for exact code and details.

bioRxiv preprint doi: https://doi.org/10.1101/663195; this version posted June 7, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

16. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571 (2011).

17. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).

18. Servant, N. *et al.* HiTC: exploration of high-throughput 'C' experiments. *Bioinformatics* **28**, 2843–2844 (2012).

19. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

20. Becht, E. *et al.* Evaluation of UMAP as an alternative to t-SNE for single-cell data. *bioRxiv* 298430 (2018). doi:10.1101/298430

21. Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl. Acad. Sci.* **112**, E5503–E5512 (2015).