Title: Development of circulating cell-free DNA reduced representation bisulfite sequencing for clinical methylomics diagnostics

Authors: Andries De Koker[1,3,8]†, Lennart Raman[2,4,5]†, Ruben Van Paemel[4,5,8], Malaïka Van der Linden[6,8], Sofie Van de Velde[4,5,8], Paul Van der Leest[9], Bram Van Gaever[6,8], Anxhela Zaka[1,3,8], Karim Vermaelen[8,10], Ingel Demedts[11], Veerle Surmont[8,10], Ulrike Himpe[11], Franceska Dedeurwaerdere[12], Liesbeth Ferdinande[2], Kathleen Claes[4,5,8], Jilke De Wilde[4,5,8], Wim Timens[9], Vincent D. de Jager[9], T. Jeroen N. Hiltermann[13], Bram De Wilde[4,7,8], Björn Menten[4,5,8], Katleen De Preter[4,5,8], Ed Schuuring[9], Jo Van Dorpe[2,8]* and Nico Callewaert[1,3,8]*

Affiliations:

[1] Center for Medical Biotechnology, VIB; Technologiepark 75, 9052 Zwijnaarde, Belgium.

[2] Department of Pathology, Ghent University, Ghent University Hospital; Corneel Heymanslaan 10, 9000 Ghent, Belgium.

[3] Department of Biochemistry and Microbiology, Ghent University; Technologiepark 75, 9052 Zwijnaarde, Belgium.

[4] Center for Medical Genetics, Ghent University, Ghent University Hospital; Corneel Heymanslaan 10, 9000 Ghent, Belgium.

[5] Department of Biomolecular Medicine, Ghent University, Ghent University Hospital; Corneel Heymanslaan 10, 9000 Ghent, Belgium.

[6] Department of Diagnostics Sciences, Ghent University, Ghent University Hospital; Corneel Heymanslaan 10, 9000 Ghent, Belgium.

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

[8] Cancer Research Institute Ghent (CRIG); Corneel Heymanslaan 10, 9000 Ghent, Belgium.

[9] Department of Pathology, University of Groningen, University Medical Center Groningen, Hanzeplein 1, 9713 GZ Groningen, Netherlands.

[10] Department of Respiratory Medicine, Ghent University, Ghent University Hospital; Corneel Heymanslaan 10, 9000 Ghent, Belgium.

[11] Deparment of Pulmonary Diseases, AZ Delta; Deltalaan 1, 8800 Roeselare, Belgium.

[12] Deparment of Pathology, AZ Delta; Deltalaan 1, 8800 Roeselare, Belgium.

[13] Department of Pulmonary Medicine, University of Groningen, University Medical Center Groningen, Hanzeplein 1, 9713 GZ Groningen, Netherlands.

†A.D.K. and L.R. share first authorship. *N.C. and J.V.D. share corresponding authorship.
Email: Nico.Callewaert@vib-ugent.be and Jo.vandorpe@uzgent.be.

One Sentence Summary: Novel methodology enables facile and cost-effective methylation profiling of fragmented plasma DNA, allowing for routine liquid biopsy clinical diagnostics, exemplified here for differential diagnosis of lung cancer subtypes.

Abstract: The methylation profile of circulating cell-free DNA (cfDNA) is of great interest as a liquid biopsy for the detection and monitoring of cancer and other pathologies. Here we introduce 'circulating cell-free DNA reduced representation bisulfite sequencing' (cf-RRBS), enabling the use of highly effective RRBS on highly fragmented plasma cfDNA as well as on FFPE-archived cancer tissues. The method enriches the CpG-rich RRBS target regions by enzymatic degradation of all off-target DNA rather than by targeted capture, in contrast to previous methods. Critical steps are fully enzymatic in a single-tube, making it rapid, cost-effective, robust and easily implemented on a liquid-handling station for high-throughput sample preparation. We benchmark cf-RRBS results to those obtained by previous more complex methods and exemplify its use for accurate non-invasive subtyping of lung cancer, a frequent onco-pathology task. cf-RRBS enables any molecular pathology lab to tap into the cfDNA methylome, only making use of off-the-shelf reagents and open-source data analysis tools.

67 **Main Text:**

68 **INTRODUCTION**

69

70 Altered DNA methylation is an almost universal hallmark of oncogenic transformation and,
71 in a broader sense, of tissue pathology (1). Recent studies (2–5) have demonstrated that
72 methylation profiles derived from cfDNA not only enable the detection of tumors, but also
73 reveal the tissue of origin for various cancer entities. To fully realize the potential of DNA
74 methylation based diagnostics, obviating tissue biopsy is necessary, instead detecting altered
75 DNA methylation patterns in easily sampled body fluids like blood plasma, i.e. through
76 'liquid biopsy'.

77

78 The most popular tissue methylation profiling approaches are based on the array technology
79 of Illumina (Infinium 450K or EPIC BeadChip) (6,7) or on Reduced Representation Bisulfite
80 Sequencing (RRBS) (8), which fail on highly fragmented low input DNA such as it is found
81 as circulating cell-free DNA (cfDNA) in blood plasma. At present, whole-genome bisulfite
82 sequencing (WGBS) on cfDNA can be used, but remains too costly for routine application,
83 mainly because of its non-targeted nature, requiring very substantial sequencing budget in
84 order to enable robust quantitation of methylation levels (9). Hence, the key to cost-
85 effectively obtaining robust information on the methylation status of regulatory sections of
86 the genome leaked by dying cells as circulating free plasma DNA, is not to dilute sequencing
87 budget over the entire genome, but to focus it on CpG-rich genomic regions (10). Therefore,
88 methods to capture CpG-rich regions have been developed, using expensive custom-designed
89 and often proprietary reagents, such as capture probes in SeqCap Epi (11) and similar
90 products, or antibodies binding to methylated DNA in cfMeDIP-seq (12). These methods all
91 involve highly complex and lengthy protocols that are difficult to automate (Figure 1A).

92

93 Classical reduced representation bisulfite sequencing (RRBS) (8), in contrast, achieves its
94 CpG-rich region focus by digesting high molecular weight genomic DNA (gDNA) extracted
95 from tissues with a restriction enzyme (typically MspI) that cuts CpG-rich sequences (i.e.,
96 'CpG islands'). This is followed by size-based gel electrophoresis separation of the short
97 resulting MspI/MspI-fragments from the genomic DNA input material, empowering high
98 sequencing coverage of the CpG-enriched ~2% of the genome within these short MspI/MspI
99 fragments. The limited required sequencing budget is similar to what is presently routinely

100 used in Non-Invasive Prenatal Testing (NIPT) (13), which serves as a benchmark for clinical
101 implementation, as it is the most widespread genomic sequencing-based molecular diagnostic
102 test in use today. RRBS does not require any customized target-selective reagents, making it
103 a universal method for genome-wide methylome profiling across pathologies. While RRBS is
104 hence very attractive as a concept for clinical liquid biopsy, its current implementations are
105 totally unsuited for use in liquid biopsy applications, as the minute amounts of cell-free
106 plasma DNA input come naturally already (highly) fragmented, making the critical size
107 separation of the MspI/MspI fragments from the fragmented input cfDNA impossible (see
108 Figure 1B for an illustration of this problem).
109
110 To overcome this, we set out to develop cell-free RRBS (cf-RRBS), which incorporates key
111 RRBS innovations to fit the fragmented cfDNA input material. At the same time, we
112 removed low-throughput manual steps such as gel separation, enabling high-throughput
113 automation that is essential for robust use in the diagnostic molecular pathology laboratory.
114 Here we report on the development of cf-RRBS, characterization and benchmarking of the
115 methodology and we illustrate its power in a key application: non-invasive histology
116 subtyping of lung cancer.
117

118 # RESULTS

119 **cf-RRBS method development**

120 Conceptually, rather than affinity-purifying a CpG-dense genomic target subsample as in
121 previous methods, we inverted the problem, and chose to enzymatically destroy all off-target
122 DNA in the sample upon protecting only the targeted MspI/MspI restriction-site flanked
123 CpG-rich genomic subsample. This ~2% of the genome has been amply demonstrated as a
124 very effective measure of genome-wide methylation status ever since RRBS was invented
125 (8,14).

126 In short (Figure 1C), in cf-RRBS, input DNA is first dephosphorylated by recombinant
127 Shrimp Alkaline Phosphatase (rSAP). Fragments are subsequently digested by MspI,
128 resulting in novel phosphorylated 5'-ends, marking MspI cuts. Upon 5' overhang filling, dA-
129 tailing and hairpin-shaped adapter ligation, exclusively MspI/MspI-fragments are typified by
130 fully ligated and hence unnicked ends. These 'circular' (i.e., with hairpins at both ends)
131 MspI/MspI-fragments are uniquely resistant to an optimized mixture of single- and double-
132 strand-specific 3'>5' and/or 5'>3' exonucleases, whereas all other DNA is degraded,

133 effectively achieving the RRBS target enrichment. Finally, following adapter opening and
134 bisulfite conversion, fragments are amplified, and a sequencing library is obtained. With
135 robustness and ease of automation in mind, we thoroughly optimized the implementation of
136 this method concept, such that only sequential reagent additions and incubations are needed
137 in a single tube, up to the first DNA amplification. This required the reagent concentrations
138 and compositions to be precisely tuned to avoid interference with subsequent enzymatic
139 reactions: e.g., dATP is used by Klenow for dA-tailing, but an excessive amount inhibits T4
140 DNA ligase during hairpin ligation (15). Similarly, heat-labile variants of selected enzymes
141 were introduced, including rSAP and Antarctic thermolabile Uracil-DNA Glycosylase (used
142 for adapter opening), such that enzyme inactivation can simply be achieved with an
143 intermittent heating step.

144 The elimination of all intermittent purification steps and tube transfers up to bisulfite
145 conversion allowed for robustness in working with the very minute DNA samples extractable
146 from patient plasma (Figure 1A). Other design optimization parameters included short
147 runtime and low cost, achieving a total cost, including sequencing, not surpassing that of the
148 routinely used NIPT test, a benchmark for successful sequencing-based molecular pathology
149 tests. All of this required substantial optimization and we recommend users to adhere tightly
150 to the finalized protocol parameters set out in detail in the Online Methods, which also
151 includes how to automate cf-RRBS on a TECAN EVO200 liquid handling platform.
152 Adaptation to other commonly used liquid handling platforms should be straightforward.

153 **cf-RRBS target enrichment performance**

154 Most of the cfDNA comes in fragments of the length of one nucleosomal winding (i.e.
155 166bp), with a smaller fraction corresponding to the length of DNA wound around two
156 nucleosomes, and vanishingly small amounts of multiple-nucleosome fragments (16). Hence,
157 we defined the target genomic subsample of cf-RRBS as all MspI/MspI-fragments sized 20-
158 200 bp, as longer MspI/MspI genomic fragments would only rarely be present. These
159 targeted MspI/MspI fragments make up ~2% of the human genome, consistent with the 42.4
160 +/- 3.2-fold enrichment of this target observed with the cfRRBS method (n=16) (Figure 2A),
161 resulting in >90% of the reads mapping to the target MspI/MspI regions (Figure 2B). In
162 Figure 2A we also illustrate that the >40-fold target enrichment factor achieved on the highly
163 fragmented cfDNA is in the same range as is achieved with classical gel-separation based
164 RRBS on high-MW gDNA. The data also make it clear that this gel-based size-separation

classical RRBS implementation is indeed incapable of achieving high and robust target enrichment when applied to fragmented cfDNA.

Two key steps in the cf-RRBS method enable specific MspI/MspI-fragment enrichment without relying on DNA size: first, marking of natural fragment ends through dephosphorylation by rSAP, enabling bistranded ligation of closed hairpin adapters only to MspI/MspI digested target fragments (which regain 5' phosphates at both ends in the digest); and, second, degrading unprotected off-target fragments by a mixture of exonucleases (Figure 1C). We evaluated the effect of omitting one or both of these steps. As illustrated by visualizing the resulting sequencing reads that map to a randomly chosen 2.5 kb genomic stretch on chromosome 1, removing either the initial dephosphorylation or the exonuclease digestion of off-target DNA indeed obviates the precise enrichment of MspI/MspI-fragments (Figure 2C). These observations were confirmed at genome-wide scale (Figure 2D): only with inclusion of both key steps can the >40-fold MspI/MspI target enrichment be achieved.

cf-RRBS quantifies methylation status highly reproducibly and yields equivalent results with established more complex and/or more costly cfDNA methylation profiling methods.

The methylation profiling accuracy of cf-RRBS was benchmarked by pairwise concordance analyses using cfDNA samples that were also analyzed by SeqCap Epi (genome-wide targeted capture method from Roche-Nimblegen). Both CpG site- and island-level 'beta values' (i.e., the ratio of the number of reads that indicate methylation over the total number of reads mapping to a CpG site or CpG island) were considered, measuring agreement by Pearson correlation and mean absolute error (~the average 'disagreement' between both methods). The genomic region target of SeqCap Epi and cf-RRBS are only partially overlapping, so these analyses are limited to the genomic regions that are sampled by both technologies. Results from a representative sample are shown (Figure 3A). Correlations were very high between both methods: average r = 0.96 for sites and r = 0.98 for islands, and mean absolute errors were consistently below 6% (Figure 3B). Of note, both target-enrichment methods compared equally well to results from WGBS analysis: concordance with WGBS was substantially better at the CpG island level (which averages the methylation status of many CpG sites) than at individual CpG level, which is likely caused by inaccurate single CpG methylation status sequencing by the low per-site coverage that is obtained in WGBS (Figure S1).

197 A selected cfDNA sample was analyzed in triplicate by both cf-RRBS and SeqCap Epi to
198 estimate technical variation. Both methods performed highly consistently (r>0.95; Figure
199 3C). The CpG site call 'overlap' between technical replicates was more than 80% and was
200 largely similar for both methods (Figure 3C, right). In exploratory experiments, and in
201 keeping with the method having been designed to operate on highly fragmented DNA, we
202 have also successfully applied cf-RRBS to DNA extracted from formalin-fixed paraffin-
203 embedded (FFPE) tissues and report on this elsewhere (17). Hence, the same cf-RRBS
204 method can be used both on patient cfDNA and on the most commonly available archived
205 samples of diseased tissues in histopathological collections, which should greatly facilitate
206 making direct correlations between cfDNA methylation alterations and those in the diseased
207 tissues under study (which can be used as reference methylome atlases to deconvolute
208 cfRRBS data obtained from plasma).

209 **cf-RRBS predominantly targets gene promoters**

210 The functional genomic coverage of cf-RRBS was compared to method alternatives by
211 mapping individual CpG calls to CpG islands and publicly available tracks of regulatory
212 elements (i.e., promoters; promoter flanking regions; enhancers; CTCF (CCCTC-Binding
213 factor) binding sites; and remaining unannotated ChIP-Seq transcription factor binding sites
214 and DNase I hypersensitivity open chromatin regions). A representative 25 kb chunk of the
215 methylation profiles of the same cfDNA sample, obtained either by cf-RRBS or by SeqCap
216 Epi, visually demonstrates that the spatial distribution of CpG calls is, as expected, highly
217 enriched in CpG islands and regulatory elements (Figure 3D).

218 The number of CpG methylation calls is of course highly related to sequencing depth. At 20
219 million reads (Paired-End 75), cf-RRBS provides methylation information on more than 3
220 million CpG sites (Figure 3E). Since the SeqCap Epi kit was designed by the manufacturer to
221 capture-enrich a larger genomic subsample than cf-RRBS, it naturally covers more CpG sites
222 when this target is sequenced at a depth according to protocol guidelines, requiring about 50
223 million reads (Paired-End 100). However, the fraction of CpGs calls in islands is, overall,
224 greater for cf-RRBS than for SeqCap Epi (Figure 3F). In contrast to WGBS, both of the CpG-
225 rich genomic area targeting methods have a much higher CpG call rate in annotated
226 regulatory elements of the genome (Figure 3G). Notably, 53.4% of CpG calls made by cf-
227 RRBS cover promoters, whilst these, as currently annotated in the human genome only
228 embody ~2.5% of the human genome (18). Differential promoter methylation is a key

229 information source as to the cell type of origin of methylated DNA and hence of key
230 diagnostic importance (19).

231 **Non-invasive cfDNA methylation profiling by cf-RRBS allows specific histological lung**
232 **cancer subtyping**

233 While cf-RRBS is very widely applicable and studies are ongoing in our lab and several
234 others in a multitude of clinical diagnostic settings (2, 17, 20, 21), we illustrate its utility in
235 the present original method invention paper by providing a proof-of-principle study on lung
236 cancer. Lung cancer is fundamentally subdivided in small cell lung cancer (SCLC) and non-
237 SCLC. Approximately 85% of diagnoses are non-SCLC, of which the most common
238 subtypes are adenocarcinoma (LUAD) or squamous cell carcinoma (LUSC). These
239 distinctions have long been therapeutically relevant, especially for targeted treatment
240 approaches (22) and even for chemotherapy selection (23,24). Most importantly, diagnosing
241 SCLC currently suffices to initiate (chemo)therapy, as currently no targeted treatments are
242 routinely used, rendering further molecular characterization as yet unnecessary. Also,
243 transformation to SCLC is often seen as a mechanism of treatment resistance and relapse
244 upon treatment of non-SCLC (25), an event which might be intercepted using liquid biopsy-
245 based disease monitoring instead of invasive lung biopsy.

246 We analyzed a first cohort, obtained from the University Hospital of Ghent (UZGent), of 51
247 liquid biopsies (lung cancer n=42 and non-tumor pulmonary disease n=9) by cf-RRBS. A
248 second cohort, obtained from the University Medical Center of Groningen (UMCG), of 63
249 liquid biopsies (lung cancer n=33 and healthy controls n=30) was analyzed as well by cf-
250 RRBS, for external classifier validation purposes. Clinical parameters per cohort are shown in
251 Table 1.

252 A customized data processing and modeling approach for the methylation-based tumor
253 classification was developed. First, we clustered the CpGs of which the methylation level
254 could be determined in >50% of the UZGent healthy control cfDNA samples (n = 46, 19
255 males and 27 females, median age 31 (23-61)), according to a clustering scheme (detailed in
256 the Supplementary Methods) that optimally accommodates that cf-RRBS-detected CpGs are
257 located close together in short MspI/MspI fragments. This results in more than 500,000 such
258 clusters (Table S3A), and methylation levels on these clustered CpGs were averaged into a
259 cluster-specific beta value (i.e. methylation fraction). Unstable clusters were removed
260 (standard deviation beta value >1%) within this set of healthy control cf-RRBS datasets (n =

261 46), in this way minimizing interference of e.g. age-or sex-related methylation alterations
262 (26,27). This resulted in a set of 28,128 of such CpG clusters with stable methylation in
263 healthy controls. Illustrating how publicly available tumor methylation data can be reutilized
264 in designing classification models for cf-RRBS-based diagnosis, we extracted these same
265 CpG-cluster methylation values from 682 lung tumor methylomics datasets, so as to be able
266 to determine which ones were most informative for the lung tumor subtype classification task
267 at hand. This procedure resulted in 473 clusters that were propagated in our classification
268 models (Table S3B). t-SNE analysis using the selected clusters demonstrated an almost
269 complete separation of the different lung cancer subtypes in the publicly sourced tumor
270 methylation reference dataset (Figure 4A). Hence, the prioritized set of CpG clusters was
271 used to perform deconvolution (5), designed to assign cf-RRBS data of patients to four
272 subfractions: healthy, LUAD, LUSC and SCLC.

273 This model was then used to classify the cf-RRBS-profiled patients and controls according to
274 the model subgroup for which the highest disease-specific fraction was assigned. A very high
275 diagnostic accuracy was observed: the multi-class area under the curve across both cohorts to
276 differentiate the three lung cancer classes and controls was 0.96. A ROC curve is shown
277 differentiating control versus tumor in Fig S2. For the clinical key question of differentiation
278 lung cancer subtypes in patients with a predetermined presence of a tumor, the multi-class
279 area under the curve rises to 0.96. For treatment purposes, the differentiation of NSCLC and
280 SCLC is most important; the receiver operating characteristic curve using the SCLC score
281 with an AUC of 0.93 is shown in Fig 4A. In Supplementary Fig S2, ROC curves are also
282 shown for the differentiation of LUAD versus LUSC using the LUAD test score (AUC =
283 0.93) and LUSC test score (AUC = 0.88), respectively. At a conservative cutoff of 3.5% of
284 the classifier's attribution to one of the cancer subtypes, none of the control plasma was
285 falsely predicted with a lung cancer. Confusion matrices across both cohorts are shown
286 (Figure 4B-C). We suspected that the required model fraction cutoff at 3.5% derived from the
287 well-known liquid biopsy problem that some patient's tumors do not shed enough DNA for
288 the differential methylation signal to raise above the healthy tissue-derived cfDNA
289 background. Our data show that ~72% (29/42 for UZGent cohort and 25/33 for UMCG
290 cohort) of patients have sufficient circulating tumor DNA to allow for reliable non-invasive
291 histological subtyping by methylation profiling. The patients for whom this methodology can
292 be used, are easily identified by the indicative subtype fraction based on the cf-RRBS
293 methylome data (Methods). Explicitly the term 'indicative subtype fraction' was used as no

294 comparability is implied with 'Variant Allele Fraction' synonym to 'estimated ctDNA
295 fraction'. Deconvolution methods using methylation data can in an *in silico* experiment
296 estimate subfractions in a mixed sample near-perfect, this differs from a real-world scenario,
297 like in this study, where reference profiles are often incomplete, generated in other
298 laboratories, with other protocols or methods and often on impure samples (28). In the
299 UMCG cohort the VAF on different driver mutations was determined using the AVENIO
300 ctDNA Expanded Kit. After log2 transformation, a correlation of R^2=0.5712 between VAF
301 and indicative subtype fraction by methylation was calculated (Fig 4D). It was observed that
302 low VAF did not necessarily mean wrong subclassification.

303

304 DISCUSSION

305 cf-RRBS is a RRBS method that is applicable on cfDNA by effectively overcoming the
306 challenges associated with low abundant and highly fragmented input material. MspI
307 digestion and subsequent size-based (20-200 bp) fragment selection results in capturing the
308 majority of promoters and CpG islands (14), uncovering most of the methylation-determined
309 information on cellular (patho)physiology (29). With cf-RRBS, the same representative
310 section of the epigenome is obtained, achieving a similar cost-benefit balance (i.e., ~$100 per
311 sample (30)) as provided by classic RRBS, rendering our method feasible for large-scale
312 studies and routine diagnostics, especially as we have designed and implemented it for liquid
313 handler automation.

314 Until now, two other 'whole-genome' method concepts have been published which similarly
315 target CpG-rich regions in highly fragmented cfDNA: oligonucleotide capture-based products
316 (SeqCap Epi (Roche NimbleGen) and NGS methylation detection system (Twist Bioscience))
317 and antibody-based enrichment (cfMeDIPSeq). These methods, in stark contrast to cf-RRBS'
318 mostly single-tube workflow, are time consuming, and require a great number of complex
319 sample handling steps, including solid-phase reversible immobilization magnetic bead
320 purifications, vacuum concentrator drying stages, and many sample transfer steps and
321 temperature-sensitive steps. Some of these (like drying steps) are difficult to validate in
322 routine diagnostics workflows and the combination of all of these make automation far from
323 trivial. We anticipate that cf-RRBS, like NIPT testing, will be much easier to implement in
324 most molecular pathology diagnostic laboratories and cf-RRBS, as we have already observed
325 in the laboratories with which we have collaborated for pilot clinical studies (2,17,20). cf-

326 RRBS only relies on a few standard molecular biology enzymes and bisulfite conversion,
327 which offers supplier versatility instead of making important diagnostic workflows dependent
328 on a single commercial supplier of the proprietary reagents that are needed in these other
329 methods.

330 It also deserves note that the methodology can directly be applied to any species' DNA and
331 hence applications can be anticipated in veterinary medicine, plant epigenetics etc., where
332 cost pressures are even higher than in human diagnostics. As cf-RRBS was designed to
333 operate on highly fragmented DNA, we also already found that that the method works well to
334 analyze the fragmented DNA extracted from formalin-fixed paraffin-embedded DNA,
335 enabling the production of cf-RRBS methylation reference atlases directly from the rich
336 collections of pathological archival materials. A study on this is forthcoming.

337 The method's application is illustrated on a relevant problem related to lung cancer diagnosis.
338 Here, for patients with inaccessible lesions or substantial comorbidity, biopsy tissue
339 examination might be delayed or simply not possible (31). Moreover, SCLC transformation,
340 seen as a mechanism of therapy resistance (25), is difficult to intercept by classic disease
341 monitoring standards (i.e., no serial tumor biopsies possible; mostly imaging-based).

342 Beyond an already-published pediatric oncology study (2) and the above-presented proof-of-
343 principle analysis on lung cancer, clinical efforts to further explore cf-RRBS's range of
344 application are in process. Since the preprint of our method, other groups are also producing
345 further derived methods that use the same concepts (32). Building on ongoing rapid
346 developments in epigenomic sequencing of tumors and non-cancerous pathologies, such as
347 immune/inflammatory diseases (33), degenerative diseases (34), infectious diseases (35),
348 substance abuse (36) and prenatal complications (37), cf-RRBS is poised for broad clinical
349 utility.

350

351

352

353

354

355

356

MATERIALS AND METHODS

All wet-lab and computational methods, as well as patient characteristics are included in the Online Methods.

List of Supplementary Materials

Supplementary Methods

Figure S1 to S2

Table S1 to S5 (excel files)

References (S1–S13)

References and Notes

1. Kulis M, Esteller M. DNA Methylation and Cancer. In: Herceg Z, Ushijima T, editors. Advances in Genetics [Internet]. Academic Press; 2010 [cited 2022 May 25]. p. 27–56. (Epigenetics and Cancer, Part A; vol. 70). Available from: https://www.sciencedirect.com/science/article/pii/B9780123808660600022
2. Van Paemel R, De Koker A, Vandeputte C, van Zogchel L, Lammens T, Laureys G, et al. Minimally invasive classification of paediatric solid tumours using reduced representation bisulphite sequencing of cell-free DNA: a proof-of-principle study. Epigenetics. 2020 Jul 14;1–13.
3. Kang S, Li Q, Chen Q, Zhou Y, Park S, Lee G, et al. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. Genome Biol. 2017 Mar 24;18(1):53.
4. Li W, Li Q, Kang S, Same M, Zhou Y, Sun C, et al. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. Nucleic Acids Res. 2018 Sep 6;46(15):e89.
5. Moss J, Magenheim J, Neiman D, Zemmour H, Loyfer N, Korach A, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun. 2018 Nov 29;9(1):5068.
6. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. Epigenomics. 2011 Dec;3(6):771–84.
7. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017 Feb 28;45(4):e22.
8. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 2005;33(18):5868–77.
9. Yong WS, Hsu FM, Chen PY. Profiling genome-wide DNA methylation. Epigenetics Chromatin. 2016;9:26.
10. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011 May 15;25(10):1010–22.
11. Wendt J, Rosenbaum H, Richmond TA, Jeddeloh JA, Burgess DL. Targeted Bisulfite Sequencing Using the SeqCap Epi Enrichment System. In: Tost J, editor. DNA Methylation Protocols [Internet]. New York, NY: Springer; 2018 [cited 2022 May 25]. p. 383–405. (Methods in Molecular Biology). Available from: https://doi.org/10.1007/978-1-4939-7481-8_20
12. Shen SY, Burgener JM, Bratman SV, De Carvalho DD. Preparation of cfMeDIP-seq libraries for methylome profiling of plasma cell-free DNA. Nat Protoc. 2019 Oct;14(10):2749–80.
13. Kostenko E, Chantraine F, Vandeweyer K, Schmid M, Lefevre A, Hertz D, et al. Clinical and Economic Impact of Adopting Noninvasive Prenatal Testing as a Primary Screening Method for Fetal Aneuploidies in the General Pregnancy Population. Fetal Diagn Ther. 2019;45(6):413–23.
14. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc. 2011 Apr;6(4):468–81.
15. Rossi R, Montecucco A, Ciarrocchi G, Biamonti G. Functional characterization of the T4 DNA ligase: a new insight into the mechanism of action. Nucleic Acids Res. 1997 Jun 1;25(11):2106–13.

16. Jiang P, Lo YMD. The Long and Short of Circulating Cell-Free DNA and the Ins and Outs of Molecular Diagnostics. Trends Genet. 2016 Jun 1;32(6):360–71.
17. De Wilde J, Van Paemel R, De Koker A, Roelandt S, Van de Velde S, Callewaert N, et al. A Fast, Affordable, and Minimally Invasive Diagnostic Test for Cancer of Unknown Primary Using DNA Methylation Profiling. Lab Investig J Tech Methods Pathol. 2024 Aug;104(8):102091.
18. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The Ensembl Regulatory Build. Genome Biol. 2015 Mar 24;16(1):56.
19. Taryma-Leśniak O, Sokolowska KE, Wojdacz TK. Current status of development of methylation biomarkers for in vitro diagnostic IVD applications. Clin Epigenetics. 2020 Jul 6;12(1):100.
20. Cornelli L, Van Paemel R, Ferro dos Santos MR, Roelandt S, Willems L, Vandersteene J, et al. Diagnosis of pediatric central nervous system tumors using methylation profiling of cfDNA from cerebrospinal fluid. Clin Epigenetics. 2024 Jul 5;16(1):87.
21. Baetens M, Van Gaever B, Deblaere S, De Koker A, Meuris L, Callewaert N, et al. Advancing diagnosis and early risk assessment of preeclampsia through noninvasive cell-free DNA methylation profiling. Clin Epigenetics. 2024 Dec 18;16:182.
22. Chan BA, Hughes BGM. Targeted therapy for non-small cell lung cancer: current standards and the promise of the future. Transl Lung Cancer Res [Internet]. 2015 Feb [cited 2022 May 25];4(1). Available from: https://tlcr.amegroups.com/article/view/2699
23. Scagliotti GV, Parikh P, von Pawel J, Biesma B, Vansteenkiste J, Manegold C, et al. Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. J Clin Oncol Off J Am Soc Clin Oncol. 2008 Jul 20;26(21):3543–51.
24. Huang CY, Ju DT, Chang CF, Muralidhar Reddy P, Velmurugan BK. A review on the effects of current chemotherapy drugs and natural agents in treating non–small cell lung cancer. BioMedicine. 7(4):23.
25. Oser MG, Niederst MJ, Sequist LV, Engelman JA. Transformation from non-small-cell lung cancer to small-cell lung cancer: molecular drivers and cells of origin. Lancet Oncol. 2015 Apr;16(4):e165–72.
26. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013 Dec 10;14(10):3156.
27. Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, et al. Distinct DNA methylomes of newborns and centenarians. Proc Natl Acad Sci. 2012 Jun 26;109(26):10522–7.
28. De Ridder K, Che H, Leroy K, Thienpont B. Benchmarking of methods for DNA methylome deconvolution. Nat Commun. 2024 May 16;15(1):4134.
29. Bock C. Epigenetic biomarker development. Epigenomics. 2009 Oct;1(1):99–110.
30. Boyle P, Clement K, Gu H, Smith ZD, Ziller M, Fostel JL, et al. Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. Genome Biol. 2012 Oct 3;13(10):R92.
31. Stokstad T, Sørhaug S, Amundsen T, Grønberg BH. Medical complexity and time to lung cancer treatment – a three-year retrospective chart review. BMC Health Serv Res. 2017 Jan 17;17(45).
32. Stackpole ML, Zeng W, Li S, Liu CC, Zhou Y, He S, et al. Cost-effective methylome sequencing of cell-free DNA for accurately detecting and locating cancer. Nat Commun. 2022 Sep 29;13(1):5566.
33. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013 Feb;31(2):142–7.

34. Urdinguio RG, Sanchez-Mut JV, Esteller M. Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies. Lancet Neurol. 2009 Nov;8(11):1056–72.
35. Pacis A, Tailleux L, Morin AM, Lambourne J, MacIsaac JL, Yotova V, et al. Bacterial infection remodels the DNA methylation landscape of human dendritic cells. Genome Res. 2015 Dec;25(12):1801–11.
36. Cecil C a. M, Walton E, Smith RG, Viding E, McCrory EJ, Relton CL, et al. DNA methylation and substance-use risk: a prospective, genome-wide study spanning gestation to adolescence. Transl Psychiatry. 2016 Dec 6;6(12):e976.
37. Yuen RK, Peñaherrera MS, von Dadelszen P, McFadden DE, Robinson WP. DNA methylation profiling of human placentas reveals promoter hypomethylation of multiple genes in early-onset preeclampsia. Eur J Hum Genet. 2010 Sep;18(9):1006–12.

Acknowledgments: We thank Hans Van Vlierberghe and Xavier Verhelst for providing 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 data on neuroblastoma patients.

Funding:

Research Foundation Flanders 11R8616N (ADK)

Research Foundation Flanders 11B3718N (RVP)

Research Foundation Flanders 1S90621N (JDW)

Research Foundation Flanders 18B1712N (BDW)

Research Foundation Flanders (AZ)

Bijzonder Onderzoeksfonds BOF.STA.2017.0002.01 (LR)

Bijzonder Onderzoeksfonds BOF24-GOA-033-Callewaert (ADK)

Kom Op Tegen Kanker KotK_UG/2020/12444/1 (ADK)

VIB (ADK, NC)

VIB Grand Challenges project VR 2021 1712 DOC. 1492/4 (ADK, NC)

Ghent University (ADK, LR, RVP, MVdL, JDW, BDW, BM, KDR, JVD, NC)

UZ Innovation Fund (BVG)

Author contributions:

Conceptualization: ADK, LR, RVP, JVD, NC

Methodology: ADK, LR, RVP, JVD, NC

Software: ADK, LR, RVP, SVdV, BVG

Validation: ADK, LR, RVP, MVdL, JDW, PVdL, WT, VDJ, TJNH, ES
Formal analysis: ADK, LR, RVP, MVdL
Investigation: ADK, LR, RVP, MVdL
Resources: KV, ID, VS, UH, FD, LF, KC, PVdL, WT, VDJ, TJNH, ES
Data curation: ADK, LR, RVP, SVdV, MVdL, JDW, PVdL, WT, VDJ, TJNH, ES
Writing – original draft: ADK, LR, NC
Writing – review & editing: ADK, LR, RVP, MVdL, SVdV, PVdL, BVG, AZ, KV, ID, VS, UH, FD, LF, KC, JDW, WT, VDJ, TJNH, BDW, BM, KDR, ES, JVD, NC
Visualization: ADK, LR, RVP, SVdV, NC
Supervision: BDW, BM, KDR, ES, JVD, NC
Project administration: ADK, LR, RVP, MVdL
Funding acquisition: BDW, BM, KDR, ES, JVD, NC

Competing interests:

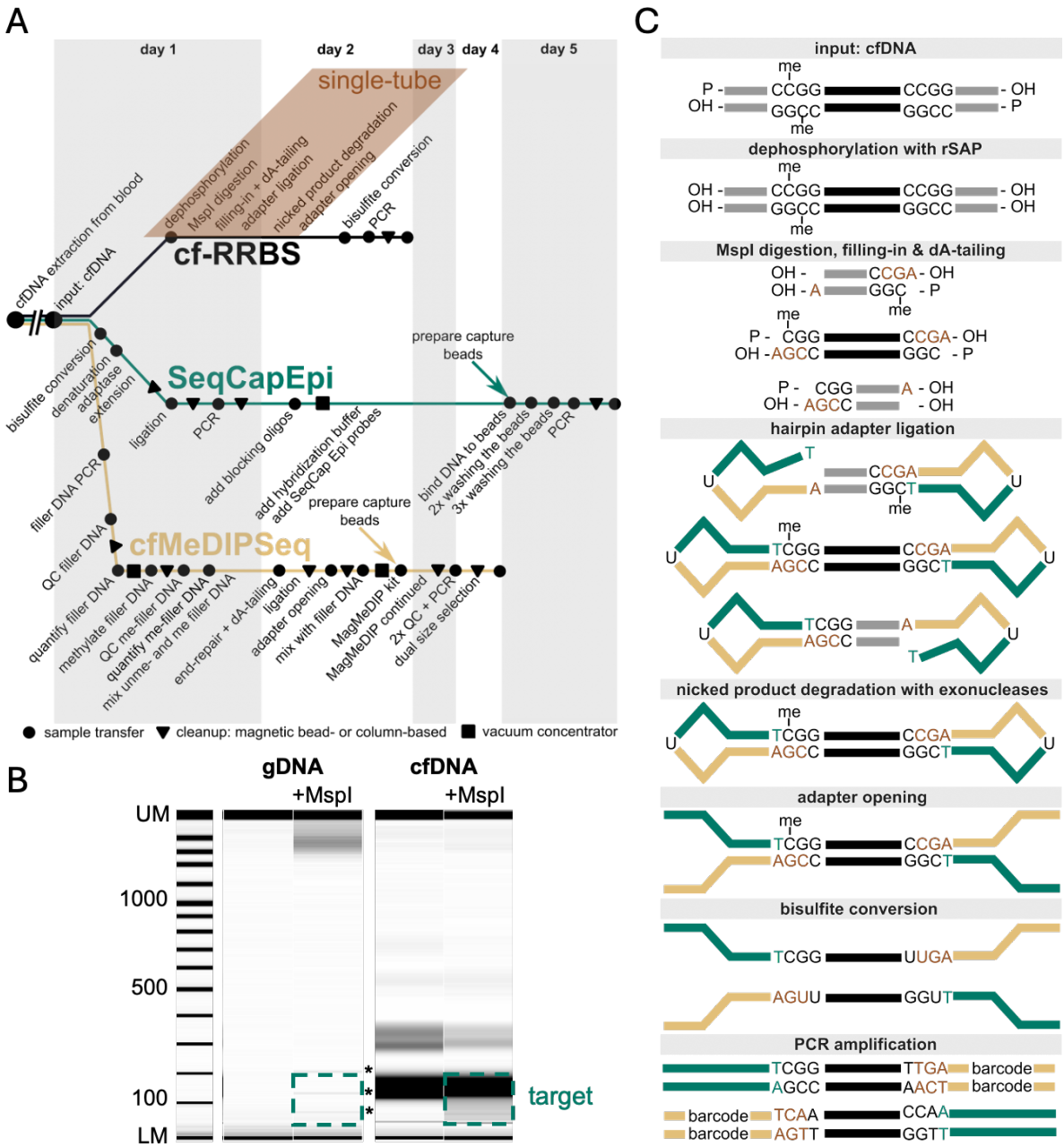
A.D.K. and N.C. are listed as inventors in patent application PCT/EP2017/056850: MEANS AND METHODS FOR AMPLIFYING NUCLEOTIDE SEQUENCES related to the methods disclosed in this manuscript.

Data and materials availability:

The data, code and materials underlying this article are available within the main text and its supplementary materials. cov-files containing the methylation data are available in <<<database>>> under accession number #######.

541 **Figures**

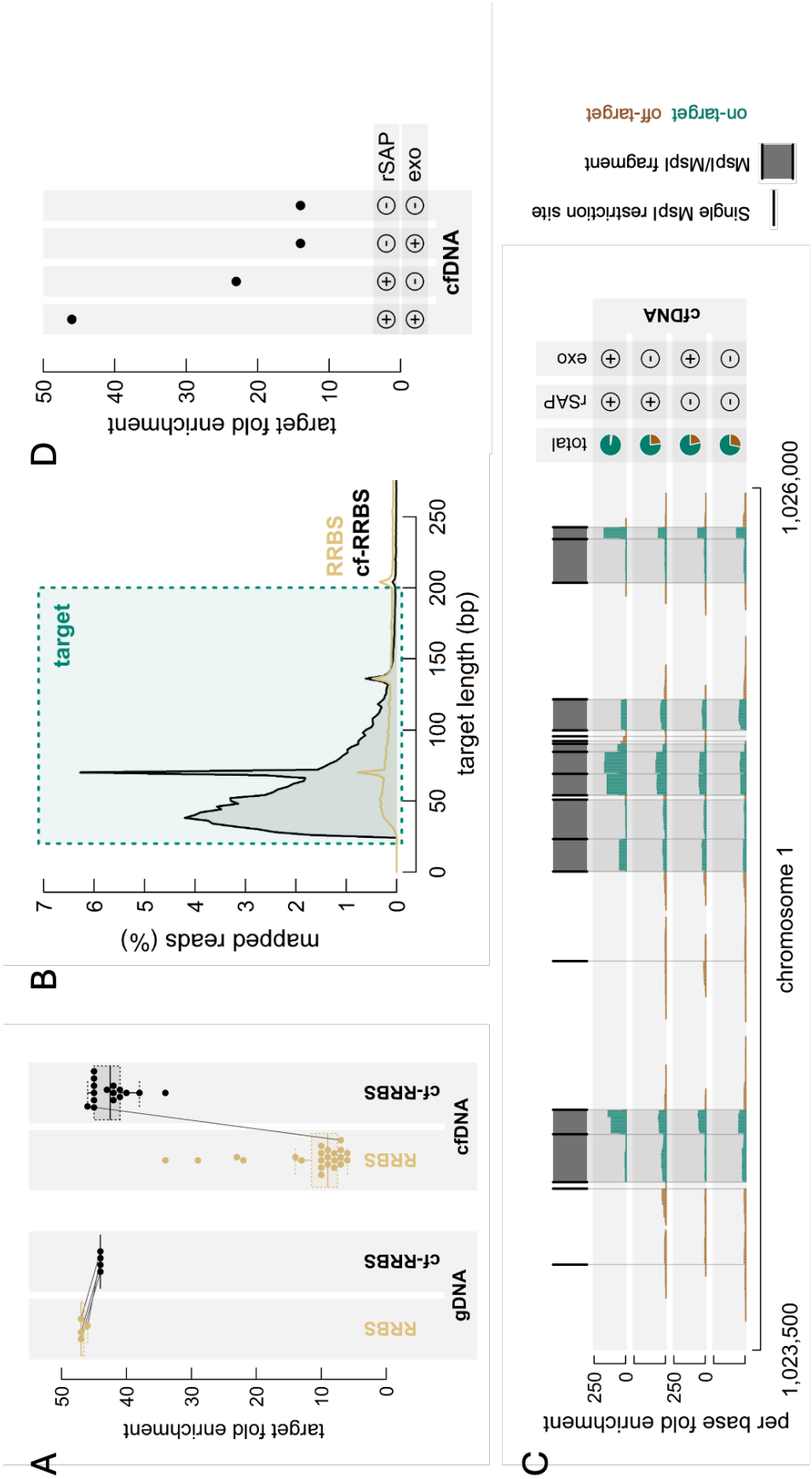
542 **Fig. 1**



543
544
545
546
547
548
549
550
551

552 **Figure 1. cf-RRBS enables reduced representation bisulfite sequencing on highly fragmented**
553 **DNA**. **(A)** cf-RRBS is technically much less complex than target capture-based and antobody capture-
554 based methods. A simplified protocol comparison between cf-RRBS, SeqCap Epi and cfMeDIPSeq in
555 function of time. Both SeqCap Epi and cfMeDIPSeq require a moch more complex sequence of
556 sample processing steps than cf-RRBS, which also takes much less time. Additional steps include
557 sample transfers (dots), solid-phase reversible immobilization magnetic bead purifications (triangles),
558 vacuum concentration steps (rectangles), many of which are not easily amenable to automation. (B) A
559 digital gel image of gDNA and cfDNA, either undigested or MspI-digested. In classic RRBS,
560 MspI/MspI-fragments ranging from 20 to 200 bp (indicated with a green rectangle; 'the target') are to
561 be extracted from the gel. The undigested gDNA runs above the upper marker (UM) and is not
562 observed. After MspI digestion, three (indicated with asterisks) characteristic satellite DNA bands
563 appear near the lower maker (LM) and this region contains the target MspI/MspI short fragments that
564 are then gel-extracted to achieve the enrichment that results in the reduced representation of the
565 genome. With cfDNA, such reduced representation target extraction is impossible, since cell-free
566 fragments are naturally highly fragmented and the size range completely overlaps with the target
567 MspI/MspI fragment size range. (C) A simplified cf-RRBS workflow, where thick lines indicate DNA
568 fragment strands. Fragments are subject to dephosphorylation, MspI digestion, dA-tailing, and
569 ligation of hairpin-shaped adapters. Only MspI/MspI-fragments yield unnicked DNA molecules
570 without free ends, providing protection against exonucleases, which degrade all off-target DNA
571 molecules, hence achieving reduced representation target enrichment even if the starting DNA is
572 highly fragmented, such as in plasma cfDNA and in FFPE-tissue extracted DNA.
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588

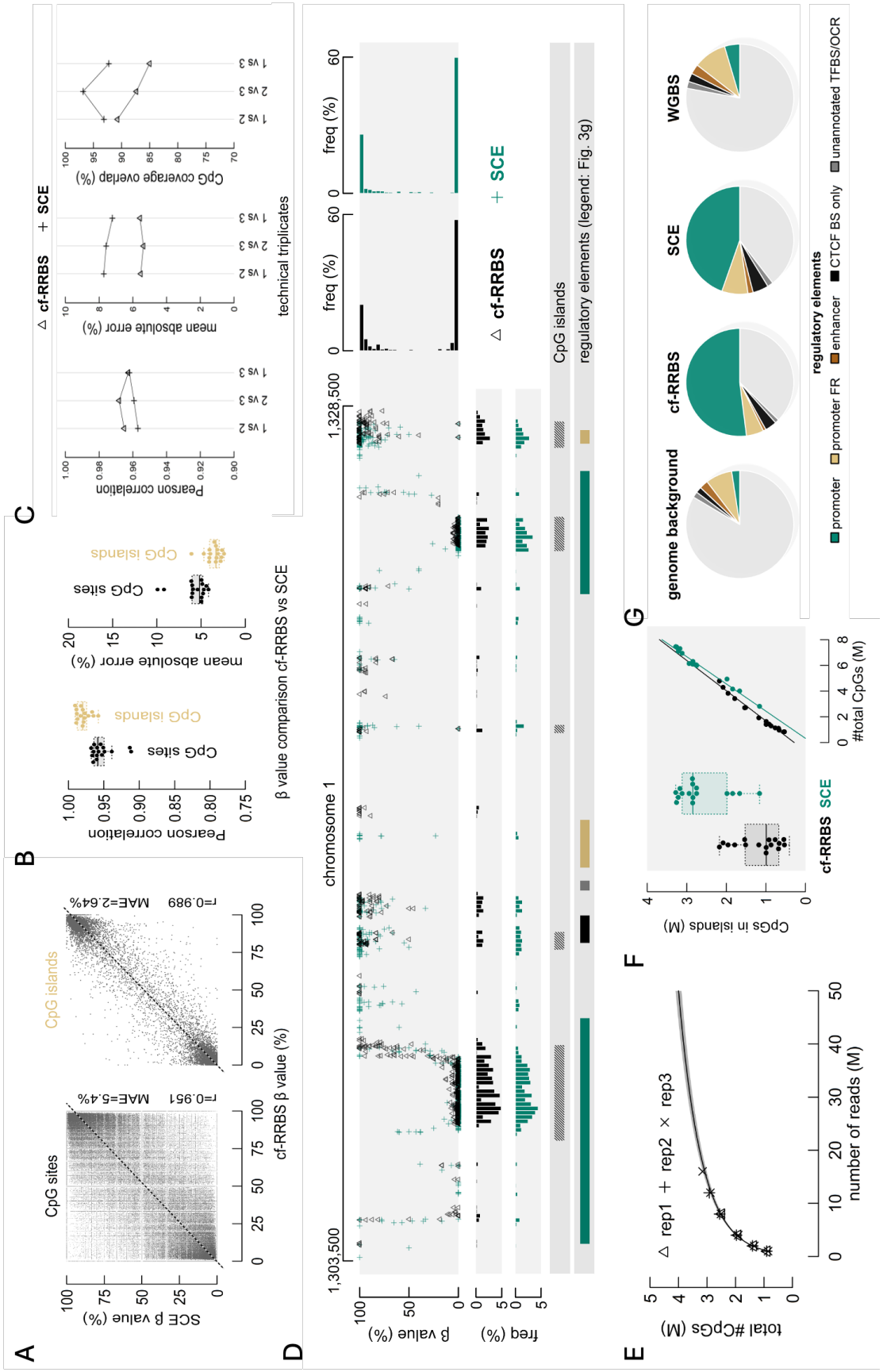
589 **Fig. 2**


A
target fold enrichment
50
40
30
20
10
0
gDNA
cfDNA
RRBS
cf-RRBS
RRBS
cf-RRBS
B
mapped reads (%)
7
6
5
4
3
2
1
0
target
RRBS
cf-RRBS
0
50
100
150
200
250
target length (bp)
D
target fold enrichment
50
40
30
20
10
0
cfDNA
rSAP
exo
C
per base fold enrichment
250
0
250
0
chromosome 1
1,023,500
1,026,000
total
rSAP
exo
cfDNA
Single MspI restriction site
MspI/MspI fragment
on-target off-target

590

591 **Figure 2. cf-RRBS solves the target enrichment problem of classic RRBS on cfDNA. (A)** While a
592 similar target fold enrichment is observed between classic RRBS and cf-RRBS on gDNA (n=4),
593 RRBS demonstrates poor enrichment when executed on cfDNA (n=23), which however can be
594 achieved effectively by using the cf-RRBS method (n=16). Same-sample dots (paired samples) are
595 connected. (B) A histogram of the mapped reads in function of the target length shows that cf-RRBS
596 reads map the target at a much higher frequency than classic RRBS. (C) The per base fold enrichment
597 of a representative chromosomal stretch of 2.5 kb is visualized across four samples, annotated by cf-
598 RRBS protocol control deviations at the right-hand side, including intentionally missing steps
599 concerning rSAP dephosphorylation, exonuclease (exo) degradation, both or none illustrating that all
600 steps in the cfRRBS method contribute to its efficiency at enriching the MspI/MspI fragments and
601 depleting the DNA that is not contained in such fragments. The top track presents *in silico* predicted
602 MspI digestion sites and MspI/MspI-fragments (sized 20-200 bp). Pie charts clarify on- and off-target
603 fractions of the visualized genomic stretch. (D) Similar to C but here the overall genome-wide target
604 fold enrichment is shown instead of for the randomly selected genomic stretch of panel C.
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627

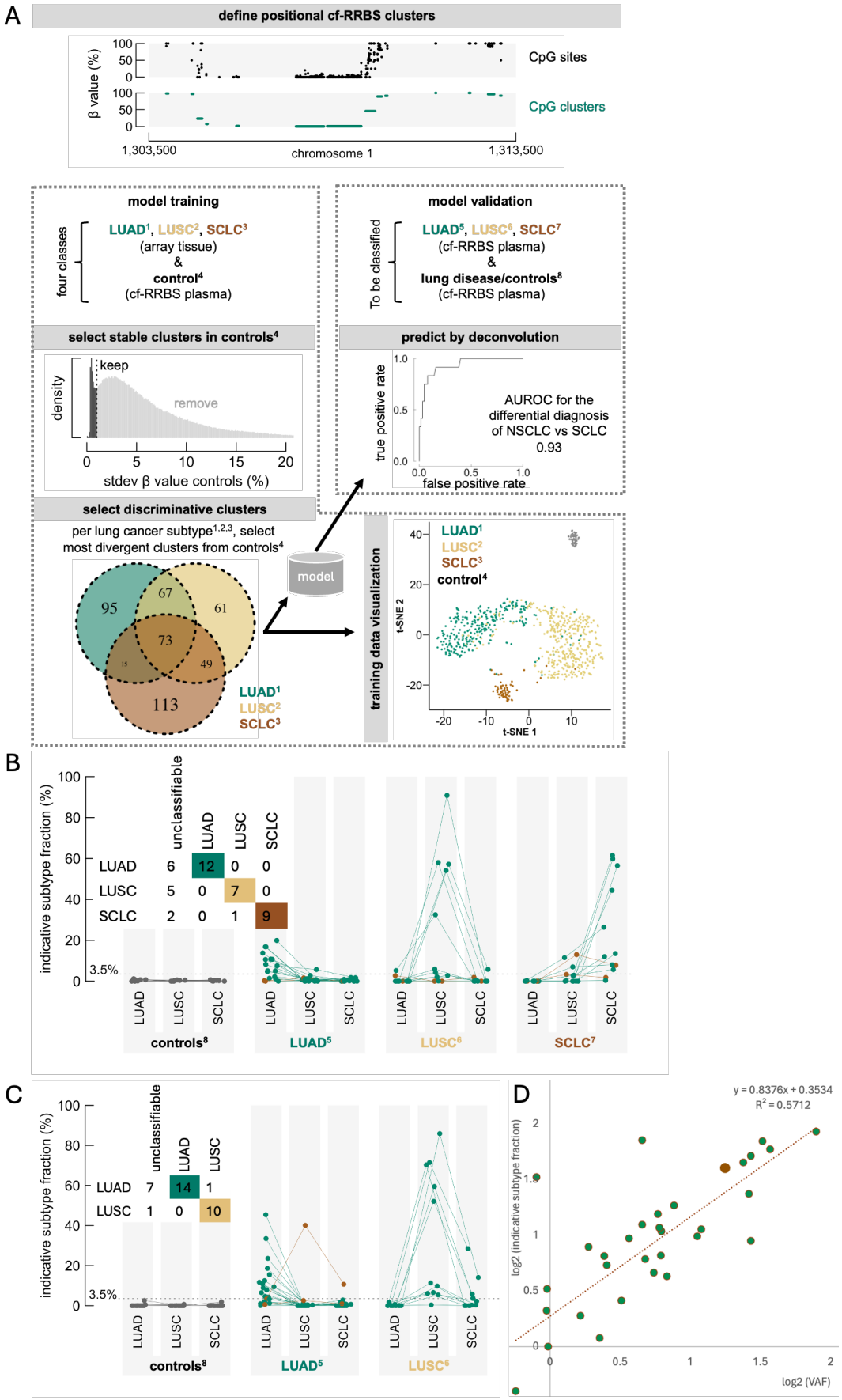
628 **Fig. 3**


A
CpG sites
CpG islands
SCE β value (%)
cf-RRBS β value (%)
r=0.951 MAE=5.4%
r=0.989 MAE=2.64%
B
Pearson correlation
mean absolute error (%)
CpG sites
CpG islands
β value comparison cf-RRBS vs SCE
C
△ cf-RRBS + SCE
Pearson correlation
mean absolute error (%)
CpG coverage overlap (%)
1 vs 2
2 vs 3
1 vs 3
technical triplicates
D
chromosome 1
1,303,500
1,328,500
β value (%)
freq (%)
△ cf-RRBS + SCE
CpG islands
regulatory elements (legend: Fig. 3g)
E
△ rep1 + rep2 × rep3
total #CpGs (M)
number of reads (M)
F
CpGs in islands (M)
cf-RRBS SCE
#total CpGs (M)
G
genome background
cf-RRBS
SCE
WGBS
regulatory elements
promoter
promoter FR
enhancer
CTCF BS only
unannotated TFBS/OCR

629

Figure 3. Methylation profiling with cf-RRBS is accurate. (A) Methylation profile scatter plot comparison between cf-RRBS and SeqCap Epi of representative patient LP10's cfDNA, at CpG site-level (left) and at CpG island-level (right). A least squares fit (dotted line), corresponding Pearson correlation (r), and the mean absolute error (MAE) are given for both comparisons. (B) Swarm/boxplot combinations of the Pearson correlation (left) and mean absolute error (right) between cfDNA samples analyzed by cf-RRBS and SeqCap Epi. (C) Pairwise comparisons between cfDNA samples analyzed in triplicate by cf-RRBS and SeqCap Epi. Paired relations are connected. Analyses include the Pearson correlation between CpG site beta values (left); the mean absolute error between CpG site beta values (middle); and the positional overlap in CpG calls (right), calculated as the number of recurrent CpGs over the total number of CpGs in the sample with the lowest sequencing depth. (D) Methylation profiles of representative patient LP10's cfDNA derived by cf-RRBS (black) and SeqCap Epi (green) of a genomic stretch sized 25 kb. Frequency histograms of CpG beta values are shown, one for each technique (right). Similarly, the spatial distribution of CpG calls is clarified by positional frequency histograms below. The bottom two bars contain tracks of CpG islands and regulatory elements, see legend of panel G. (E) the number of raw cfDNA sequencing reads is shown in relation to the total number of CpG calls for three technical cf-RRBS replicates, randomly subsampled to 1, 2, 4, 8, 12 and 16 million reads. A log regression fit, with 95% confidence interval (given by shading), indicates three million CpG calls are obtained with less than 20 million reads. (F) Swarm/boxplot combinations of the number of CpG calls located in CpG islands compared between cfDNA samples analyzed by cf-RRBS and SeqCap Epi. The scatter plot at the right-hand side provides the relation with the total number of CpG calls for each technique, given by least squares fits. cf-RRBS provides relatively more calls located in CpG islands. (G) Pie charts indicate the average fractions of CpGs that cover regulatory elements per cfDNA methylation profiling method. The far-left pie demonstrates the genomic background. Included regulatory elements are unannotated transcription factor binding sites (TFBSs) and open chromatin regions (OCRs); (parts of) CTCF binding sites (BSs) that do not travel across other elements; enhancers; promoter flanking regions (FRs); and promoters.

668 **Fig. 4**


A
define positional cf-RRBS clusters
β value (%)
CpG sites
CpG clusters
1,303,500
chromosome 1
1,313,500
model training
four classes
LUAD¹, LUSC², SCLC³
(array tissue)
&
control⁴
(cf-RRBS plasma)
model validation
To be classified
LUAD⁵, LUSC⁶, SCLC⁷
(cf-RRBS plasma)
&
lung disease/controls⁸
(cf-RRBS plasma)
select stable clusters in controls⁴
density
keep
remove
stdev β value controls (%)
predict by deconvolution
true positive rate
false positive rate
AUROC for the differential diagnosis of NSCLC vs SCLC
0.93
select discriminative clusters
per lung cancer subtype¹,²,³, select most divergent clusters from controls⁴
model
training data visualization
t-SNE 1
t-SNE 2
LUAD¹
LUSC²
SCLC³
control⁴
B
indicative subtype fraction (%)
unclassifiable
LUAD
LUSC
SCLC
3.5%
controls⁸
LUAD⁵
LUSC⁶
SCLC⁷
C
D
y = 0.8376x + 0.3534
R² = 0.5712
log2 (indicative subtype fraction)
log2 (VAF)

669

670 **Figure 4. Methylation profiles obtained by cf-RRBS enable non-invasive lung cancer subtyping**
671 **at high specificity**. **(A)** Predictive modeling workflow. Prior to modeling, CpG clusters according to
672 spatial distribution optimized for cf-RRBS were defined in order to reduce data complexity with
673 minimal data loss. Control liquid biopsies (n=46); and public methylation datasets of LUAD, LUSC
674 and SCLC tissue samples (n=682), served as a reference set. Based on these, CpG clusters for
675 modeling were selected using a two-step procedure: first, variable clusters in controls were removed
676 based on the standard deviation (stdev); second, per lung cancer subtype, the most divergent clusters
677 in comparison to healthy donor plasma cfDNA controls were kept (cluster overlap between subtypes
678 is indicated by Venn diagram). The reference set is of lung cancer tissue methylation datasets and
679 healthy donor's cfDNA cfRRBS visualized by t-SNE analysis, based on these selected clusters.
680 Finally, a linear deconvolution model for the cfDNA methylation ß-values in these clusters was
681 trained, attributing fractions of the cfDNA to the four deconvoluted entities such that the linear
682 combination of each reference ß-values multiplied by the attributed fractions results in the observed
683 cfDNA ß-value. Following modeling by deconvolution, validation was executed using two cohorts,
684 patient (n=42 or n=33) and control liquid biopsies (n=9 or n=30). Differential diagnosis of NSCLC
685 versus SCLC across both cohorts is indicated by receiver operating characteristic analysis,
686 summarized by AUROC. (B-C) Indicative subtype fractions and confusion table at the 3.5% 'noise'
687 cutoff of each lung cancer subtype for in-house cohort and (B) and UMCG cohort (C). Samples are
688 grouped per type. For controls and patients, every fraction is shown, where same-sample dots are
689 connected. The dotted line at 3.5% represent a 'noise' cutoff: control liquid biopsies are consistently
690 below this limit. (D) Correlation log2(VAF) versus log2(indicative subtype fraction) on UMCG
691 cohort. Correct prediction (green) and wrong prediction (brown).
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706

707 **Tables**

708 **Table 1: Clinical parameters per cohort**

Patient cohort UZGent		
Variable	**Parameter**	**No. of patients**
Age	Median (range)	66 (44-80)
Sex	Male	30 (71%)
	Female	12 (29%)
Smoking history	Smoker	14 (33%)
	Ex-smoker	25 (60%)
	Non-smoker	2 (5%)
	unknown	1 (2%)
Stage	III	6 (14%)
	IV	36 (86%)
Histology	Adenocarcinoma	20 (43%)
	Squamous cell carcinoma	12 (29%)
	Small cell	12 (29%)

Patient cohort UMCG		
Variable	**Parameter**	**No. of patients**
Age	Median (range)	65 (40-83)
Sex	Male	18 (47%)
	Female	20 (53%)
Smoking history	Smoker	28 (74%)
	Ex-smoker	9 (24%)
	Non-smoker	1 (3%)
	unknown	0 (0%)
Stage	III	2 (5%)
	IV	36 (95%)
Histology	Adenocarcinoma	26 (68%)
	Squamous cell carcinoma	12 (32%)
	Small cell	0 (0%)

Control cohort UZGent		
Variable	**Parameter**	**No. of patients**
Age	Median (range)	59 (41-71)
Sex	Male	7 (78%)
	Female	2 (22%)

Control cohort UMCG		
Variable	**Parameter**	**No. of patients**
Age	Median (range)	63 (48-77)
Sex	Male	14 (47%)
	Female	16 (53%)

709