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 subtupes.

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.

Main Text:

INTRODUCTION

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

The most popular tissue methylation profiling approaches are based on the array technology of Illumina (Infinium 450K or EPIC BeadChip) (6,7) or on Reduced Representation Bisulfite

67 Sequencing (RRBS) (8), which fail on highly fragmented low input DNA such as it is found
68 as circulating cell-free DNA (cfDNA) in blood plasma. At present, whole-genome bisulfite
69 sequencing (WGBS) on cfDNA can be used, but remains too costly for routine application,
70 mainly because of its non-targeted nature, requiring very substantial sequencing budget in
71 order to enable robust quantitation of methylation levels (9). Hence, the key to cost-
72 effectively obtaining robust information on the methylation status of regulatory sections of
73 the genome leaked by dying cells as circulating free plasma DNA, is not to dilute sequencing
74 budget over the entire genome, but to focus it on CpG-rich genomic regions (10). Therefore,
75 methods to capture CpG-rich regions have been developed, using expensive custom-designed
76 and often proprietary reagents, such as capture probes in SeqCap Epi (11) and similar
77 products, or methylated DNA-binding antibodies in cfMeDIP-seq (12), involving complex
78 and lengthy protocols that are difficult to automate (Figure 1A).
79
80 Classical reduced representation bisulfite sequencing (RRBS) (8), in contrast, achieves its
81 CpG-rich region focus by digesting high molecular weight genomic DNA (gDNA) extracted
82 from tissues with a restriction enzyme (typically MspI) that cuts CpG-rich sequences (i.e.,
83 'CpG islands'). This is followed by size-based gel electrophoresis separation of the short
84 resulting MspI/MspI-fragments from the genomic DNA input material, empowering high
85 sequencing coverage of the CpG-enriched ~2% of the genome within these short MspI/MspI
86 fragments. The limited required sequencing budget is similar to what is presently routinely
87 used in Non-Invasive Prenatal Testing (NIPT) (13), which serves as a benchmark for clinical
88 implementation, as it is the most widespread genomic sequencing-based molecular diagnostic
89 test in use today. RRBS does not require any customized target-selective reagents, making it
90 a universal method for genome-wide methylome profiling across pathologies. While RRBS is
91 hence very attractive as a concept for clinical liquid biopsy, its current implementations are
92 totally unsuited for use in liquid biopsy applications, as the minute amounts of cell-free
93 plasma DNA input come naturally already (highly) fragmented, making the critical size
94 separation of the MspI/MspI fragments from the fragmented input cfDNA impossible (see
95 Figure 1B for an illustration of this problem).
96
97 To overcome this, we set out to develop cell-free RRBS (cf-RRBS), which incorporates key
98 RRBS innovations to fit the fragmented cfDNA input material. At the same time, we
99 removed low-throughput manual steps such as gel separation, enabling high-throughput
100 automation that is essential for robust use in the diagnostic molecular pathology laboratory.

101 Here we report on the development of cf-RRBS, characterization and benchmarking of the
102 methodology and we illustrate its power in a key application: non-invasive histology
103 subtyping of lung cancer.
104

105 # RESULTS

106 **cf-RRBS method development**

107 Conceptually, rather than affinity-purifying a CpG-dense genomic target subsample as in
108 previous methods, we inverted the problem, and chose to enzymatically destroy all off-target
109 DNA in the sample upon protecting only the targeted MspI/MspI restriction-site flanked
110 CpG-rich genomic subsample. This ~2% of the genome has been amply demonstrated as a
111 very effective measure of genome-wide methylation status ever since RRBS was invented
112 (8,14).

113 In short (Figure 1C), in cf-RRBS, input DNA is first dephosphorylated by recombinant
114 Shrimp Alkaline Phosphatase (rSAP). Fragments are subsequently digested by MspI,
115 resulting in novel phosphorylated 5'-ends, marking MspI cuts. Upon 5' overhang filling, dA-
116 tailing and hairpin-shaped adapter ligation, exclusively MspI/MspI-fragments are typified by
117 fully ligated and hence unnicked ends. These 'circular' (i.e., with hairpins at both ends)
118 MspI/MspI-fragments are uniquely resistant to an optimized mixture of single- and double-
119 stranded 3'>5' and/or 5'>3' exonucleases, whereas all other DNA is degraded, effectively
120 achieving the RRBS target enrichment. Finally, following adapter opening and bisulfite
121 conversion, fragments are amplified, and a sequencing library is obtained. With robustness
122 and ease of automation in mind, we thoroughly optimized the implementation of this method
123 concept, such that only sequential reagent additions and incubations are needed in a single
124 tube, up to the first DNA amplification. This required the reagent concentrations and
125 compositions to be precisely tuned to avoid interference with subsequent enzymatic
126 reactions: e.g., dATP is used by Klenow for dA-tailing, but an excessive amount inhibits T4
127 DNA ligase during hairpin ligation (15). Similarly, heat-labile variants of selected enzymes
128 were introduced, including rSAP and Antarctic thermolabile Uracil-DNA Glycosylase (used
129 for adapter opening), such that enzyme inactivation can simply be achieved with an
130 intermittent heating step.

131 The elimination of all intermittent purification steps and tube transfers up to bisulfite
132 conversion allowed for robustness in working with the very minute DNA samples extractable

133 from patient plasma (Figure 1A). Other design optimization parameters included short
134 runtime and low cost, achieving a total cost including sequencing not surpassing that of the
135 routinely used NIPT test, a benchmark for successful sequencing-based molecular pathology
136 tests. All of this required substantial optimization and we recommend users to adhere tightly
137 to the finalized protocol parameters set out in detail in the Online Methods, including how to
138 automate cf-RRBS on a TECAN EVO200 liquid handling platform. Adaptation to other
139 commonly used liquid handling platforms should be straightforward.

140 **cf-RRBS target enrichment performance**

141 Most of the cfDNA comes in fragments of the length of one nucleosomal winding (i.e.
142 166bp), with a smaller fraction corresponding to the length of DNA wound around two
143 nucleosomes, and vanishingly small amounts of multiple-nucleosome fragments (16). Hence,
144 we defined the target genomic subsample of cf-RRBS as all MspI/MspI-fragments sized 20-
145 200 bp, as longer MspI/MspI genomic fragments would only rarely be present. These
146 targeted MspI/MspI fragments make up ~2% of the human genome, consistent with the 42.4
147 +/- 3.2-fold enrichment of this target observed with the method (n=16) (Figure 2A), resulting
148 in >90% of the reads mapping to the target MspI/MspI regions (Figure 2B). In Figure 2A we
149 also illustrate that the >40-fold target enrichment factor achieved on the highly fragmented
150 cfDNA is in the same range as is achieved with classical RRBS on high-MW gDNA. The
151 data also make it clear that this gel-based size-separation classical RRBS implementation is
152 incapable of achieving high and robust target enrichment when applied to fragmented
153 cfDNA.

154 Two key steps in the cf-RRBS method enable specific MspI/MspI-fragment enrichment
155 without relying on DNA size: first, marking of natural fragment ends through
156 dephosphorylation by rSAP, enabling bistranded ligation of closed hairpin adapters only to
157 MspI/MspI digested target fragments (which regain 5' phosphates at both ends in the digest);
158 and, second, degrading unprotected off-target fragments by a mixture of exonucleases (Figure
159 1C). We evaluated the effect of omitting one or both of these steps. As illustrated by
160 visualizing the resulting sequencing reads that map to a randomly chosen 2.5 kb genomic
161 stretch on chromosome 1, removing either the initial dephosphorylation or the exonuclease
162 digestion of off-target DNA indeed obviates the precise enrichment of MspI/MspI-fragments
163 (Figure 2C). These observations were confirmed at genome-wide scale (Figure 2D): only
164 with inclusion of both key steps can the >40-fold MspI/MspI target enrichment be achieved.

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

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

184 A selected cfDNA sample was analyzed in triplicate by both cf-RRBS and SeqCap Epi to
185 estimate technical variation. Both methods performed highly consistently (r>0.95; Figure
186 3C). The CpG site call 'overlap' between technical replicates was more than 80% and was
187 largely similar for both methods (Figure 3C, right). In exploratory experiments, and in
188 keeping with the method having been designed to operate on highly fragmented DNA, we
189 have also successfully applied cf-RRBS to DNA extracted from formalin-fixed paraffin-
190 embedded (FFPE) tissues and report on this elsewhere (17). Hence, the same cf-RRBS
191 method can be used both on patient cfDNA and on the most commonly available archived
192 samples of diseased tissues in histopathological collections, which should greatly facilitate
193 making direct correlations between cfDNA methylation alterations and those in the diseased
194 tissues under study (which can be used as reference methylome atlases to deconvolute
195 cfRRBS data obtained from plasma).

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

197 The functional genomic coverage of cf-RRBS was compared to method alternatives by
198 mapping individual CpG calls to CpG islands and publicly available tracks of regulatory
199 elements (i.e., promoters; promoter flanking regions; enhancers; CTCF (CCCTC-Binding
200 factor) binding sites; and remaining unannotated transcription factor binding sites and open
201 chromatin regions). A representative 25 kb chunk of the methylation profiles of the same
202 cfDNA sample, obtained either by cf-RRBS or by SeqCap Epi, visually demonstrates that the
203 spatial distribution of CpG calls is, as expected, highly enriched in CpG islands and
204 regulatory elements (Figure 3D).

205 The number of CpG methylation calls is of course highly related to sequencing depth. At 20
206 million reads (Paired-End 75), cf-RRBS provides methylation information on more than 3
207 million CpG sites (Figure 3E). Since the SeqCap Epi kit was designed by the manufacturer to
208 capture-enrich a larger genomic subsample than cf-RRBS, it naturally covers more CpG sites
209 when this target is sequenced at a depth according to protocol guidelines, requiring about 50
210 million reads (Paired-End 100). However, the fraction of CpGs calls in islands is, overall,
211 greater for cf-RRBS than for SeqCap Epi (Figure 3F). In contrast to WGBS, both of the CpG-
212 rich genomic area targeting methods have a much higher CpG call rate in annotated
213 regulatory elements of the genome (Figure 3G). Notably, 53.4% of CpG calls made by cf-
214 RRBS cover promoters, whilst promoters, as currently annotated in the human genome only
215 embody ~2.5% of the human genome (18). Differential promoter methylation is a key
216 information source as to the cell type of origin of methylated DNA and hence of key
217 diagnostic importance (19).

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

220 While cf-RRBS is very widely applicable and studies are ongoing in our lab and several
221 others in a multitude of clinical diagnostic settings (2, 17, 20, 21), we illustrate its utility by
222 providing a proof-of-principle study on lung cancer. Lung cancer is fundamentally
223 subdivided in small cell lung cancer (SCLC) and non-SCLC. Approximately 85% of
224 diagnoses are non-SCLC of which the most common subtypes are adenocarcinoma (LUAD)
225 or squamous cell carcinoma (LUSC). These distinctions have long been therapeutically
226 relevant, especially for targeted treatment approaches (22) and even for chemotherapy
227 selection (23,24). Most importantly, diagnosing SCLC currently suffices to initiate

228 (chemo)therapy, as currently, no targeted treatments are routinely used, rendering further
229 molecular characterization as yet unnecessary. Also, transformation to SCLC is often seen as
230 a mechanism of resistance upon treatment of non-SCLC (25), an event which might be
231 intercepted using liquid biopsy-based disease monitoring instead of invasive lung biopsy.

232 We analyzed a first cohort, obtained from the University Hospital of Ghent (UZGent), of 56
233 liquid biopsies (lung cancer n=47 and non-tumor pulmonary disease n=9) by cf-RRBS. A
234 second cohort, obtained from the University Medical Center of Groningen (UMCG), of 69
235 liquid biopsies (lung cancer n=39 and healthy controls n=30) was analyzed as well by cf-
236 RRBS for classifier validation purposes. As a healthy control reference, 46 healthy controls
237 cfDNA samples (UZGent) were used. Clinical parameters per cohort are shown in Table 1.

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

Patient cohort UZGent		
Variable	**Parameter**	**No. of patients**
Age	Median (range)	66 (44-80)
Sex	Male	31 (66%)
	Female	16 (34%)
Smoking history	Smoker	14 (30%)
	Ex-smoker	29 (62%)
	Non-smoker	3 (6%)
	unknown	1 (2%)
Stage	III	6 (13%)
	IV	41 (87%)
Histology	Adenocarcinoma	20 (42%)
	Squamous cell carcinoma	14 (30%)
	Small cell	13 (28%)

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%)

239

240 A customized data processing and modeling approach for the methylation-based tumor
241 classification was developed. First, we clustered the CpGs of which the methylation level

242 could be determined in >50% of the UZGent healthy control samples, according to a scheme
243 (detailed in the Supplementary Methods) that optimally accommodates that cf-RRBS-
244 detected CpGs are located close together in short MspI/MspI fragments. This results in more
245 than 500,000 such clusters (Table S2), and methylation levels on these clustered CpGs were
246 averaged into a cluster-specific beta value. These clusters were also cross-checked in these
247 healthy control cf-RRBS datasets (n=46) to determine which ones had stable methylation
248 level across healthy controls, in this way minimizing interference of e.g. age-or sex-related
249 methylation alterations (26,27). Illustrating how publicly available tumor methylation data
250 can be reutilized in designing classification models for cf-RRBS-based diagnosis, we
251 extracted these same CpG-cluster methylation values from 719 lung tumor methylomics
252 datasets, so as to be able to determine which ones were most informative for the lung tumor
253 subtype classification task at hand. This procedure resulted in approximately 450 to 500
254 clusters that were propagated in our classification models (Table S3). t-SNE analysis using
255 the selected clusters demonstrated an almost complete separation of the different lung cancer
256 subtypes in the publicly sourced tumor methylation reference dataset (Figure 4A). Hence, the
257 prioritized set of CpG clusters was used to perform deconvolution (5), designed to assign cf-
258 RRBS data of patients to four subfractions: healthy, LUAD, LUSC and SCLC.

259 This model was then used to classify the cf-RRBS-profiled patients and controls according to
260 the model subgroup for which the highest disease-specific fraction was assigned. A very high
261 diagnostic accuracy was observed: the multi-class area under the curve across both cohorts to
262 differentiate the three lung cancer classes and controls was 0.94. A ROC curve is shown
263 differentiating control versus tumor in Fig S2. For the clinical key question of differentiation
264 lung cancer subtypes in patients with a predetermined presence of a tumor, the multi-class
265 area under the curve rises to 0.95. For treatment purposes, the differentiation of NSCLC and
266 SCLC is most important; the receiver operating characteristic curve using the SCLC score
267 with an AUC of 0.93 is shown in Fig 4A. In Supplementary Fig S2, ROC curves are also
268 shown for the differentiation of LUAD versus LUSC using the LUAD test score (AUC =
269 0.92) and LUSC test score (AUC = 0.91), respectively. At a conservative cutoff of 3.5% of
270 the classifier's attribution to one of the cancer subtypes, none of the control plasma was
271 falsely predicted with a lung cancer. The very few false predictions that remained (Figure 4B
272 & 4C– brown-colored lines), had all been annotated by expert pathologists as uncertain
273 diagnosis (Figure 4B & 4C – brown-colored crosses) except for one patient per cohort
274 (Figure 4B & 4C – brown-colored dots). Only considering the cases where pathology was

275 univocal, the confusion matrices across both cohorts were updated (Figure 4D-E). We
276 suspected that the required model fraction cutoff at 3.5% derived from the well-known liquid
277 biopsy problem that some patient's tumors do not shed enough DNA for the differential
278 methylation signal to raise above the healthy tissue-derived cfDNA background. Our data
279 show that ~79% (36/47 for UZGent cohort and 32/39 for UMCG cohort) of patients have
280 sufficient circulating tumor DNA to allow for reliable non-invasive histological subtyping by
281 methylation profiling. The patients for whom this methodology can be used, are easily
282 identified by the indicative subtype fraction based on the cf-RRBS methylome data
283 (Methods). Explicitly the term 'indicative subtype fraction' was used as no comparability is
284 implied with 'Variant Allele Fraction' synonym to 'estimated ctDNA fraction'.
285 Deconvolution methods using methylation data can in an *in silico* experiment estimate
286 subfractions in a mixed sample near-perfect, this differs from a real-world scenario, like in
287 this study, where reference profiles are often incomplete, generated in other laboratories, with
288 other protocols or methods and often on impure samples (28). In the UMCG cohort the VAF
289 on different driver mutations was determined using the AVENIO ctDNA Expanded Kit. After
290 log2 transformation, a correlation of R^2=0.5683 between VAF and indicative subtype fraction
291 by methylation was calculated (Fig 4F). It was observed that low VAF did not necessarily
292 mean wrong subclassification.

293

294 # DISCUSSION

295 cf-RRBS is a RRBS method that is applicable on cfDNA by effectively overcoming the
296 challenges associated with low abundant and highly fragmented input material. MspI
297 digestion and subsequent size-based (20-200 bp) fragment selection results in capturing the
298 majority of promoters and CpG islands (14), uncovering most of the methylation-determined
299 information on cellular (patho)physiology (29). With cf-RRBS, the same representative
300 section of the epigenome is obtained, achieving a similar cost-benefit balance (i.e., ~$200 per
301 sample (30)) as provided by classic RRBS, rendering our method feasible for large-scale
302 studies and routine diagnostics, all the more given its design accounts for automatability.

303 On the moment of writing only two 'whole-genome' approaches have been published which
304 similarly target CpG-rich regions in highly fragmented cfDNA: SeqCap Epi and
305 cfMeDIPSeq. Close inspection of associated protocols indicates longer hands-on and total
306 runtimes in comparison to cf-RRBS (Figure 1A). More importantly, in stark contrast to the
307 cf-RRBS' mostly single-tube workflow, SeqCap Epi and cfMeDIPSeq incorporate a great
308 number of complex sample handling steps, including solid-phase reversible immobilization
309 magnetic bead purifications, vacuum concentrator drying stages, and many sample transfer
310 steps temperature-sensitive steps (Figure 1A). Some of these (like drying steps) are difficult
311 to validate in routine diagnostics workflows and the combination of all of these make
312 automation far from trivial. We anticipate that cf-RRBS, like NIPT testing, will be much
313 easier to implement in hospital diagnostic laboratories provided with validated reagents and
314 software, rather than having to rely on a model in which patient samples necessarily have to
315 be shipped to central highly specialized laboratories, which is not practical nor economical in
316 many parts of the world. cf-RRBS only relies on a few enzymes and bisulfite conversion,
317 which offers supplier versatility and more common paths of diagnostic reagent qualification
318 than is the case e.g. with complex libraries of synthetic capture probes.

319 cf-RRBS is performed using a series of ordinary enzymatic reagents, which should allow
320 straightforward implementation in clinics without complex laboratory platform changes that
321 often frustrate the translation of 'omics' studies into clinical practice. Moreover, other than
322 human DNA, species with similar characteristics concerning methylation and spatial CpG
323 distribution, such as mouse models, represent excellent cf-RRBS candidates, requiring no
324 protocol adaptations. Also, preliminary results indicate its capability to analyze fragmented
325 formalin-fixed paraffin-embedded DNA, potentially activating a large biobank of archival
326 material for retrospective research.

327 Recently two other RRBS-based protocols which can be applied on cfDNA, cfMethyl-Seq
328 (EarlyDx) (31), and FFPE-derived fragmented DNA, FFPE-RRBS (32), were published. Both
329 methods resemble closely our cf-RRBS protocol and also crucially depend on the pre-
330 treatment of input fragmented DNA with a DNA phosphatase. While the latter protocol does
331 not include a strategy to avoid the use of non-MspI digested 3'-ends, the first protocol is
332 actively blocking non-MspI digested 3'-ends with dideoxynucleotides. In contrast, our cf-
333 RRBS protocol is actively removing DNA pieces that were not digested by MspI. The
334 authors of the cfMethyl-Seq method state that their method could be superior because of two
335 reasons: Firstly, they expect that cfMethyl-Seq retains significantly more cfDNA because
336 exonuclease digestion removes all DNA fragments that contain nicks, a scenario that has
337 been estimated to occur in 30% of cfDNA fragments, leading to the loss of precious cfDNA.
338 And secondly, in cf-RRBS, adapters and ligases are consumed by the majority of cfDNA
339 fragments, including those without the desired digestion sites, leading to lower ligation
340 efficiency for the fragments of interest and lower library yields. However, the authors fail to
341 show real data evidence for this statement. Raw cfMethyl-Seq data access was asked for, but
342 we never received this data. We were able to perform the protocol as published but we are
343 still in doubt if the libraries are of the same quality as those from the original authors.
344
345 The method's application is illustrated on a relevant problem related to lung cancer diagnosis.
346 Here, for patients with inaccessible lesions or substantial comorbidity, tissue examination
347 might be delayed or simply not possible (33). Moreover, SCLC transformation, seen as a
348 mechanism of therapy resistance (25), is difficult to intercept by classic disease monitoring
349 standards (i.e., no serial tumor biopsies possible; mostly imaging-based). Our first cohort
350 included one such case, correctly subtyped as SCLC according to its cfDNA methylation
351 profile. Unfortunately, a plasma sample at initial diagnosis, prior to transformation, was
352 lacking. Overall, we show that accurate histological subtyping of lung cancer is feasible when
353 the indicative subtype fraction is above 3.5%.

354 Beyond an already-published pediatric oncology study (2) and the above-presented proof-of-
355 principle analysis on lung cancer, clinical efforts to further explore cf-RRBS's range of
356 application are in process. Building on ongoing rapid developments in epigenomic
357 sequencing of tumors and non-cancerous pathologies, such as immune/inflammatory diseases
358 (34), degenerative diseases (35), infectious diseases (36), substance abuse (37) and prenatal
359 complications (38), cf-RRBS is poised for broad clinical utility.

360 **MATERIALS AND METHODS**

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

363

364 **List of Supplementary Materials**

365 Supplementary Methods

366 Figure S1 to S2

367 Table S1 to S12 (excel files)

368 References (S1–S13)

369

370 **References and Notes**

371

372 1. Kulis M, Esteller M. DNA Methylation and Cancer. In: Herceg Z, Ushijima T,
373 editors. Advances in Genetics [Internet]. Academic Press; 2010 [cited 2022 May 25]. p. 27–
374 56. (Epigenetics and Cancer, Part A; vol. 70). Available from:
375 https://www.sciencedirect.com/science/article/pii/B9780123808660600022
376 2. Van Paemel R, De Koker A, Vandeputte C, van Zogchel L, Lammens T, Laureys G,
377 et al. Minimally invasive classification of paediatric solid tumours using reduced
378 representation bisulphite sequencing of cell-free DNA: a proof-of-principle study.
379 Epigenetics. 2020 Jul 14;1–13.
380 3. Kang S, Li Q, Chen Q, Zhou Y, Park S, Lee G, et al. CancerLocator: non-invasive
381 cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA.
382 Genome Biol. 2017 Mar 24;18(1):53.
383 4. Li W, Li Q, Kang S, Same M, Zhou Y, Sun C, et al. CancerDetector: ultrasensitive
384 and non-invasive cancer detection at the resolution of individual reads using cell-free DNA
385 methylation sequencing data. Nucleic Acids Res. 2018 Sep 6;46(15):e89.
386 5. Moss J, Magenheim J, Neiman D, Zemmour H, Loyfer N, Korach A, et al.
387 Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free
388 DNA in health and disease. Nat Commun. 2018 Nov 29;9(1):5068.
389 6. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of
390 the Infinium Methylation 450K technology. Epigenomics. 2011 Dec;3(6):771–84.
391 7. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and
392 innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017 Feb
393 28;45(4):e22.
394 8. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced
395 representation bisulfite sequencing for comparative high-resolution DNA methylation
396 analysis. Nucleic Acids Res. 2005;33(18):5868–77.
397 9. Yong WS, Hsu FM, Chen PY. Profiling genome-wide DNA methylation. Epigenetics
398 Chromatin. 2016;9:26.
399 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

450 Dec 10;14(10):3156.
451 27. Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, et al. Distinct DNA
452 methylomes of newborns and centenarians. Proc Natl Acad Sci. 2012 Jun 26;109(26):10522–
453 7.
454 28. De Ridder K, Che H, Leroy K, Thienpont B. Benchmarking of methods for DNA
455 methylome deconvolution. Nat Commun. 2024 May 16;15(1):4134.
456 29. Bock C. Epigenetic biomarker development. Epigenomics. 2009 Oct;1(1):99–110.
457 30. Boyle P, Clement K, Gu H, Smith ZD, Ziller M, Fostel JL, et al. Gel-free multiplexed
458 reduced representation bisulfite sequencing for large-scale DNA methylation profiling.
459 Genome Biol. 2012 Oct 3;13(10):R92.
460 31. Stackpole ML, Zeng W, Li S, Liu CC, Zhou Y, He S, et al. Cost-effective methylome
461 sequencing of cell-free DNA for accurately detecting and locating cancer. Nat Commun.
462 2022 Sep 29;13(1):5566.
463 32. Zhang S, He S, Zhu X, Wang Y, Xie Q, Song X, et al. DNA methylation profiling to
464 determine the primary sites of metastatic cancers using formalin-fixed paraffin-embedded
465 tissues. Nat Commun. 2023 Sep 14;14:5686.
466 33. Stokstad T, Sørhaug S, Amundsen T, Grønberg BH. Medical complexity and time to
467 lung cancer treatment – a three-year retrospective chart review. BMC Health Serv Res. 2017
468 Jan 17;17(45).
469 34. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al.
470 Epigenome-wide association data implicate DNA methylation as an intermediary of genetic
471 risk in rheumatoid arthritis. Nat Biotechnol. 2013 Feb;31(2):142–7.
472 35. Urdinguio RG, Sanchez-Mut JV, Esteller M. Epigenetic mechanisms in neurological
473 diseases: genes, syndromes, and therapies. Lancet Neurol. 2009 Nov;8(11):1056–72.
474 36. Pacis A, Tailleux L, Morin AM, Lambourne J, MacIsaac JL, Yotova V, et al.
475 Bacterial infection remodels the DNA methylation landscape of human dendritic cells.
476 Genome Res. 2015 Dec;25(12):1801–11.
477 37. Cecil C a. M, Walton E, Smith RG, Viding E, McCrory EJ, Relton CL, et al. DNA
478 methylation and substance-use risk: a prospective, genome-wide study spanning gestation to
479 adolescence. Transl Psychiatry. 2016 Dec 6;6(12):e976.
480 38. Yuen RK, Peñaherrera MS, von Dadelszen P, McFadden DE, Robinson WP. DNA
481 methylation profiling of human placentas reveals promoter hypomethylation of multiple
482 genes in early-onset preeclampsia. Eur J Hum Genet. 2010 Sep;18(9):1006–12.

483

484 **Acknowledgments:** We thank Hans Van Vlierberghe and Xavier Verhelst for providing
485 blood samples from patients with liver disease. We thank Tim Lammens and Geneviève
486 Laureys for providing the neuroblastoma samples. We thank Jo Vandesompele, Anneleen
487 Decock, Wim Van Criekinge and their teams for the reuse data on neuroblastoma patients.

488

489 **Funding:**
490 Research Foundation Flanders 11R8616N (ADK)
491 Research Foundation Flanders 11B3718N (RVP)
492 Research Foundation Flanders 1S90621N (JDW)

Research Foundation Flanders 18B1712N (BDW)
Research Foundation Flanders (AZ)
Bijzonder Onderzoeksfonds BOF.STA.2017.0002.01 (LR)
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
Software: ADK, LR, RVP, SVdV, BVG
Validation: ADK, LR, RVP, MVdL, JDW
Formal analysis: ADK, LR, RVP, MVdL
Investigation: ADK, LR, RVP, MVdL
Resources: KV, ID, VS, UH, FD, LF, KC
Data curation: ADK, LR, RVP, SVdV, MVdL, JDW
Writing – original draft: ADK, LR, NC
Writing – review & editing: ADK, LR, RVP, SVdV, AZ, KDR, JVD, NC
Visualization: ADK, LR, RVP, SVdV
Supervision: BDW, BM, KDR, JVD, NC
Project administration: ADK, LR, RVP, MVdL
Funding acquisition: BDW, BM, KDR, JVD, NC

Competing interests:

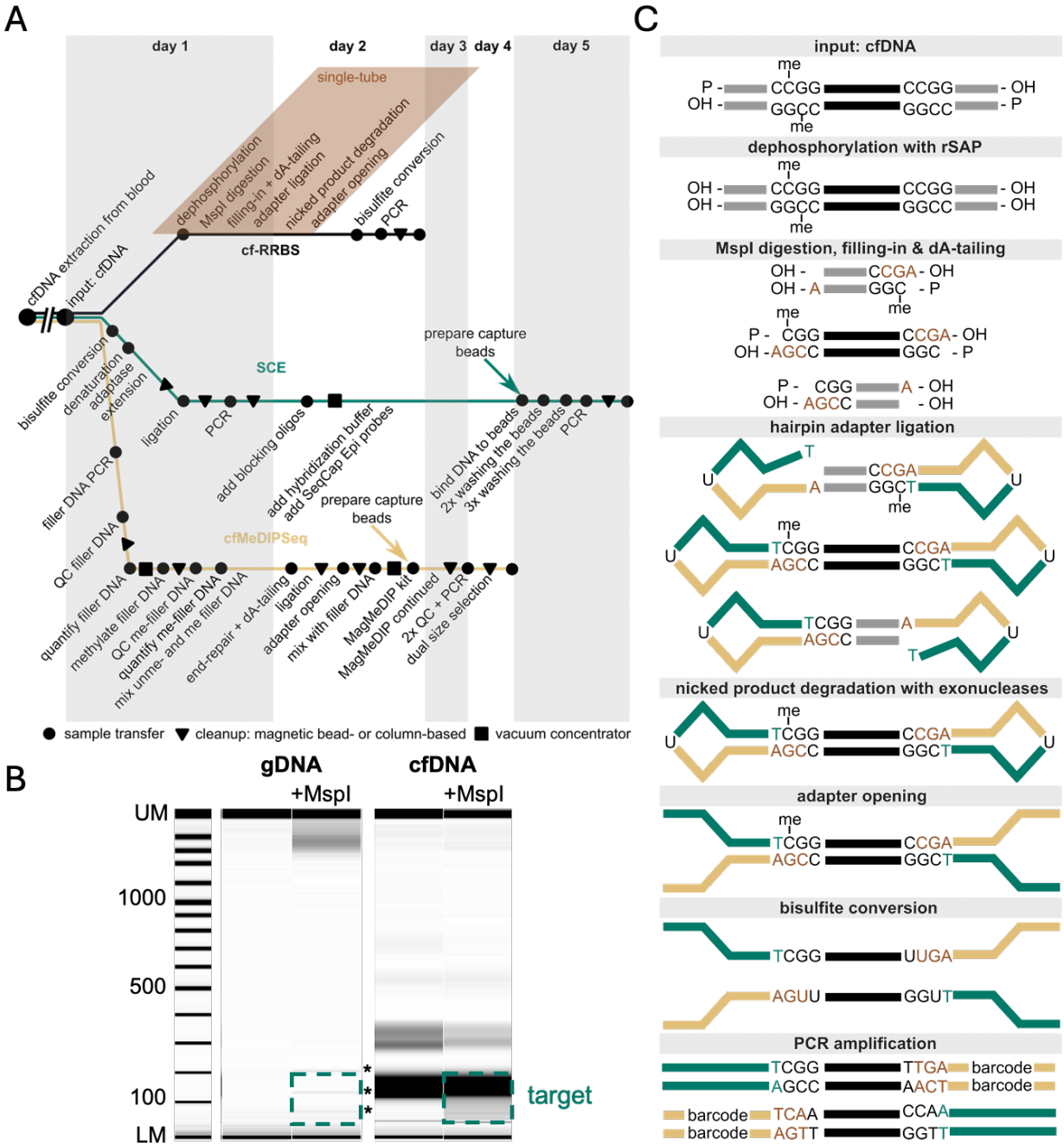
A.D.K. and N.C. are listed as inventors in patent application PCT/EP2017/056850: MEANS AND METHODS FOR AMPLIFYING NUCLEOTIDE SEQUENCES related to the methods disclosed in this manuscript.

Data and materials availability:

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

528 **Figures**

529 **Fig. 1**



530

531

532

533

534

535

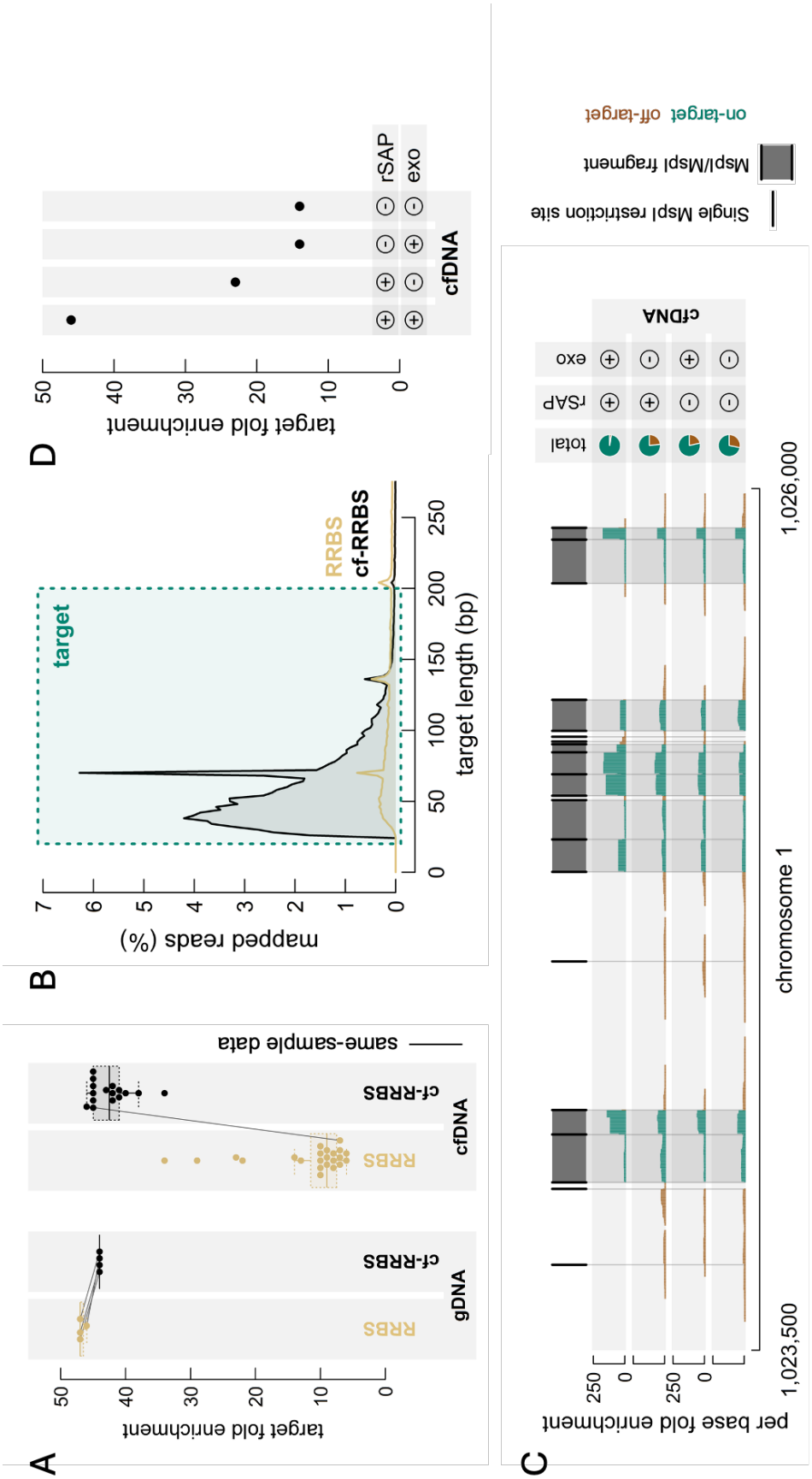
536

537

538

Figure 1. cf-RRBS is technically less complex than SeqCap Epi and cfMeDIPSeq. **(A)** A simplified protocol comparison between (mostly single-tube) cf-RRBS, SeqCap Epi and cfMeDIPSeq in function of time. Both SeqCap Epi and cfMeDIPSeq take longer than cf-RRBS: additional steps include sample transfers (dots), solid-phase reversible immobilization magnetic bead purifications (triangles), vacuum concentration steps (rectangles). (B) A digital gel image of gDNA and cfDNA, either undigested or MspI-digested. In classic RRBS, MspI/MspI-fragments ranging from 20 to 200 bp (indicated with a green rectangle; 'the target') are to be extracted from the gel. The undigested gDNA runs above the upper marker (UM) and is not observed. After MspI digestion, three (indicated with asterisks) characteristic satellite DNA bands appear near the lower maker (LM). With cfDNA, target extraction includes undigested DNA, since cell-free fragments are naturally highly fragmented. (C) A simplified cf-RRBS workflow, where thick lines indicate DNA fragment strands. Fragments are subject to dephosphorylation, MspI digestion, dA-tailing, and ligation of hairpin-shaped adapters. Only MspI/MspI-fragments yield unnicked DNA molecules without free ends, providing protection against exonucleases.

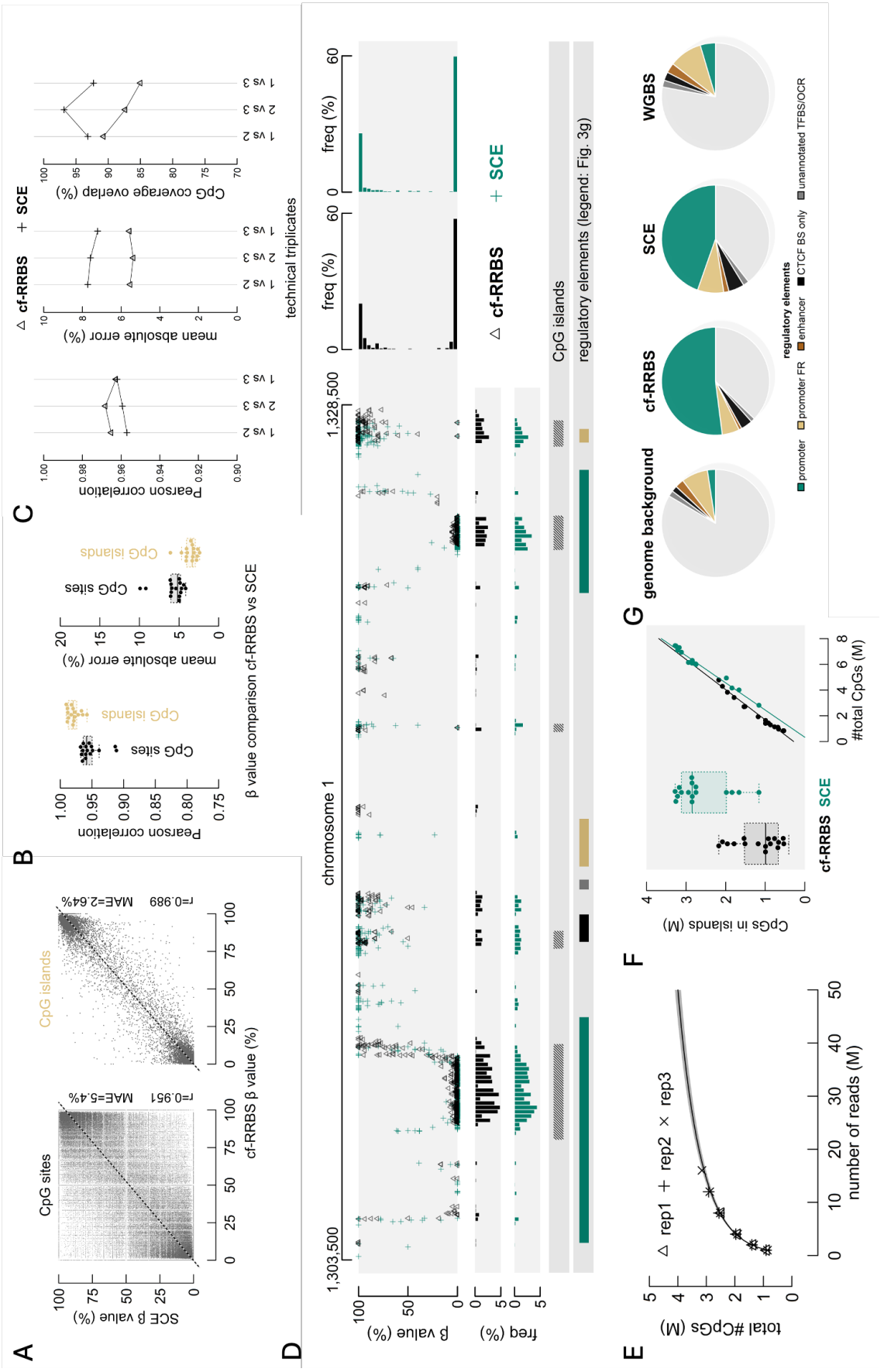
576 **Fig. 2**


A
target fold enrichment
50
40
30
20
10
0
gDNA
cfDNA
RRBS
cf-RRBS
RRBS
cf-RRBS
same-sample data
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
cfDNA
total
rSAP
exo
1,023,500
chromosome 1
1,026,000
Single MspI restriction site
MspI/MspI fragment
on-target off-target

577

578 **Figure 2. cf-RRBS solves the target enrichment problem of classic RRBS on cfDNA. (A)** While a
579 similar target fold enrichment is observed between classic RRBS and cf-RRBS on gDNA (n=4),
580 RRBS demonstrates poor enrichment when executed on cfDNA (n=23), which can be restored by
581 using cf-RRBS (n=16). Same-sample dots are connected. (B) A histogram of the mapped reads in
582 function of the target length shows that cf-RRBS reads map the target at a higher frequency than
583 classic RRBS. (C) The per base fold enrichment of a representative chromosomal stretch of 2.5 kb is
584 visualized across four samples, annotated by cf-RRBS protocol deviations at the right-hand side,
585 including missing steps concerning rSAP dephosphorylation, exonuclease (exo) degradation, both or
586 none. The top track presents *in silico* predicted MspI digestion sites and MspI/MspI-fragments (sized
587 20-200 bp). Pie charts clarify on- and off-target fractions of the visualized genomic stretch. (D)
588 Similar to C but the overall genome-wide target fold enrichment is shown.
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614

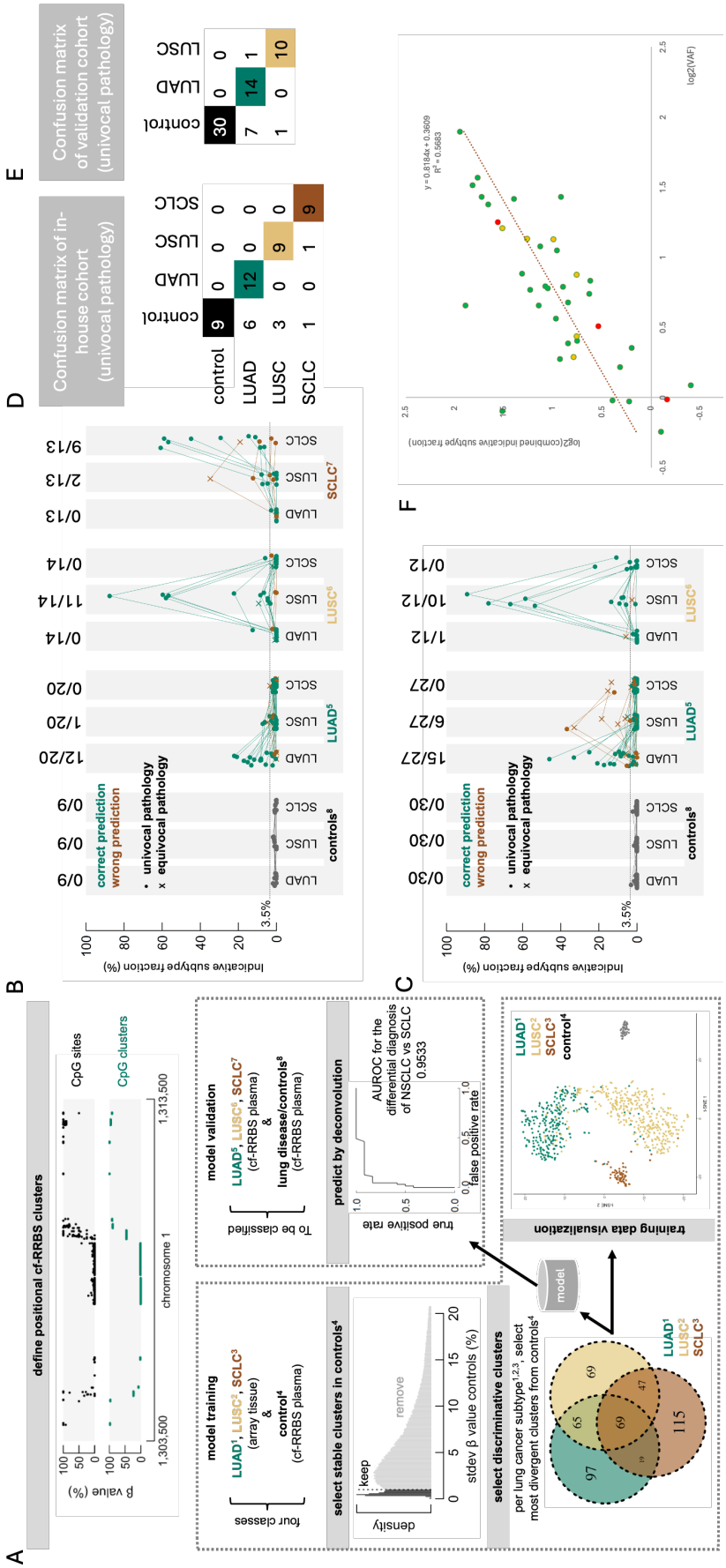
615 **Fig. 3**


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
Pearson correlation
mean absolute error (%)
CpG coverage overlap (%)
△ cf-RRBS + SCE
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
total #CpGs (M)
△ rep1 + rep2 × rep3
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

616

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, of which the legend is provided by 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 sampled 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.

Fig. 4


A
define positional cf-RRBS clusters
CpG sites
CpG clusters
β value (%)
chromosome 1
1,303,500
1,313,500
model training
four classes
LUAD1, LUSC2, SCLC3
(array tissue)
&
control4
(cf-RRBS plasma)
select stable clusters in controls4
keep
remove
density
stdev β value controls (%)
select discriminative clusters
per lung cancer subtype1,2,3, select most divergent clusters from controls4
model
LUAD1
LUSC2
SCLC3
97
65
69
69
47
19
115
training data visualization
control4
model validation
To be classified
LUAD5, LUSC6, SCLC7
(cf-RRBS plasma)
&
lung disease/controls8
(cf-RRBS plasma)
predict by deconvolution
AUROC for the differential diagnosis of NSCLC vs SCLC
0.9533
true positive rate
false positive rate
B
Indicative subtype fraction (%)
correct prediction
wrong prediction
univocal pathology
equivocal pathology
LUAD
LUSC
SCLC
controls8
LUAD5
LUSC6
SCLC7
0/9
0/9
0/9
12/20
1/20
0/20
0/14
11/14
0/14
0/13
2/13
9/13
3.5%
C
Indicative subtype fraction (%)
0/30
0/30
0/30
15/27
6/27
0/27
1/12
10/12
0/12
D
Confusion matrix of in-house cohort (univocal pathology)
control
LUAD
LUSC
SCLC
E
Confusion matrix of validation cohort (univocal pathology)
control
LUAD
LUSC
F
log2(combined indicative subtype fraction)
log2(VAF)
y = 0.818x + 0.3609
R² = 0.5683

657 **Figure 4. Methylation profiles obtained by cf-RRBS enable non-invasive lung cancer subtyping**
658 **at high specificity**. **(A)** Predictive modeling workflow. Prior to modeling, CpG clusters according to
659 spatial distribution optimized for cf-RRBS were defined in order to reduce data complexity with
660 minimal data loss. Control liquid biopsies (n=46); and LUAD, LUSC and SCLC tissue samples
661 (n=719), served as a reference set. Based on these, CpG clusters for modeling were selected using a
662 two-step procedure: first, variable clusters in controls were removed based on the standard deviation
663 (stdev); second, per lung cancer subtype, the most divergent clusters in comparison to controls were
664 kept (cluster overlap between subtypes is indicated by Venn diagram). The reference set is visualized
665 by t-SNE analysis, based on previously selected clusters. Finally, following modeling by
666 deconvolution, validation was executed using two cohorts, patient (n=47 or n=39) and control liquid
667 biopsies (n=9 or n=30). Differential diagnosis of NSCLC versus SCLC across both cohorts is
668 indicated by receiver operating characteristic analysis, summarized by AUROC. (B-C) Indicative
669 subtype fractions of each lung cancer subtype for in-house cohort (B) and UMCG cohort (C). Samples
670 are grouped per type. For controls and patients, every fraction is shown, where same-sample dots are
671 connected. The dotted line at 3.5% represent a 'noise' cutoff: control liquid biopsies are consistently
672 below this limit. (D-E) Confusion table of both cohorts combined at the 3.5% 'noise' cutoff and with
673 the equivocal pathology samples excluded. (F) Correlation log2(VAF) versus log2(combined
674 indicative subtype fraction) on UMCG cohort. Correct prediction + univocal pathology (green),
675 equivocal pathology (orange), wrong prediction + univocal pathology (red)