

1 **Blood-based profiles of DNA methylation predict the underlying distribution of cell**
2 **types: a validation analysis**
3

4 Devin C. Koestler^{1#}, Brock C. Christensen^{1,2,#}, Margaret R. Karagas¹, Carmen J.
5 Marsit^{1,2}, Scott M. Langevin^{3,4}, Karl T. Kelsey^{3,4}, John K. Wiencke⁵, and E. Andres
6 Houseman⁶
7

8 1: Department of Community and Family Medicine, Geisel School of Medicine at
9 Dartmouth College, Lebanon, NH USA

10 2: Department of Pharmacology and Toxicology, Geisel School of Medicine at
11 Dartmouth College, Hanover, NH USA

12 3: Department of Pathology and Laboratory Medicine, Brown University, Providence, RI
13 USA

14 4: Department of Epidemiology, Brown University, Providence, RI USA

15 5: Department of Neurological Surgery, University of California at San Francisco, San
16 Francisco, CA USA

17 6: Department of Public Health, Oregon State University, Corvallis, OR USA

18 #: Authors contributed equally to this work
19

20 **Short Title:** DNA methylation predicts leukocyte subpopulations
21

22 **Keywords:** DNA methylation, whole-blood, cell mixture analysis, mixture deconvolution,
23 leukocytes
24

25 **Corresponding Author:**

26 E. Andres Houseman, Sc.D.
27 College of Public Health and Human Sciences
28 153 Milam Hall
29 Corvallis, OR 97331
30 Phone: 541-737-3177
31 Fax: 541 737-6914
32 Email: Andres.Houseman@oregonstate.edu

33 **Conflict of Interest**

34 The authors have no competing interests to declare.
35
36
37
38
39
40

41 **Abbreviations**

42

43 CBC – Complete blood cell

44 CP – Constrained projection

45 EWAS – Epigenome-wide association study

46 L-DMR – Leukocyte differentially methylated regions

47 MAPE – Median absolute prediction error

48 PBMC – Peripheral blood mononuclear cell

49 WBC – White blood cell

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86 **Abstract**

87 The potential influence of underlying differences in relative leukocyte distributions in
88 studies involving blood-based profiling of DNA methylation is well recognized and has
89 prompted development of a set of statistical methods for inferring changes in the
90 distribution of white blood cells using DNA methylation signatures. However, the extent
91 to which this methodology can accurately predict cell type proportions based on blood-
92 derived DNA methylation data in a large-scale epigenome-wide association study
93 (EWAS) has yet to be examined. We used publicly available data deposited in the Gene
94 Expression Omnibus (GEO) database (accession no. GSE37008), which consisted of
95 both blood-derived epigenome-wide DNA methylation data assayed using the Illumina
96 Infinium HumanMethylation27 BeadArray and complete blood cell (CBC) counts among
97 a community cohort of 94 non-diseased individuals. Constrained projection (CP) was
98 used to obtain predictions of the proportions of lymphocytes, monocytes, and
99 granulocytes for each of the study samples based on their DNA methylation signatures.
100 Our findings demonstrated high consistency between the average CBC-derived and
101 predicted percentage of monocytes and lymphocytes (17.9% and 17.6% for monocytes
102 and 82.1% and 81.4% for lymphocytes), with root mean squared error (rMSE) of 5% and
103 6%, for monocytes and lymphocytes, respectively. Similarly, there was moderate-high
104 correlation between the CP predicted and CBC-derived percentages of monocytes and
105 lymphocytes (0.60 and 0.61, respectively) and these results were robust to the number
106 of leukocyte differentially methylated regions (L-DMRs) used in CP. These results serve
107 as further validation of the CP approach and highlight the promise of this technique for
108 EWAS where DNA methylation is profiled using whole-blood genomic DNA.

109

110

111 Introduction:

112 Methylation of a cytosine residue in the context of a CpG dinucleotide on DNA is a
113 normal epigenetic regulatory mark that contributes to the control of gene expression and
114 genomic stability. Epigenetic processes such as DNA methylation allow a single
115 genome to elicit the multitude of transcriptional programs characteristic of multicellular
116 organisms, whose various cell types have distinct phenotypes and functions. Of course,
117 because epigenetic patterns are linked to cell-specific gene expression patterns, several
118 studies have successfully identified differentially methylated regions (DMRs) among
119 various cell types,¹⁻⁵ i.e. CpG sites whose methylation state is stable and differs among
120 two or more cell types.

121 When studying DNA methylation in human health and disease, DMRs present an
122 important challenge and a unique opportunity. For instance, DNA from peripheral blood
123 is a mixture of genetic substrate from various leukocyte subtypes, and variation in
124 leukocytes proportions could confound true epigenetic associations between methylation
125 and a dependent variable of interest, since there is the potential for associations
126 between phenotype and DNA methylation to be mediated by shifts in leukocyte
127 proportions. Indeed, the potential for shifts in leukocyte composition to confound
128 associations in epigenome-wide association studies (EWAS) has been recognized.⁶⁻¹²
129 The underlying proportion of leukocytes could also confound or bias other leukocyte
130 DNA biomarker relationships, such as that between telomere length, repetitive element
131 DNA methylation,¹³ or mitochondrial copy number¹⁴ and exposures or disease outcomes.

132 Motivated by work from our group and others that identified L-DMRs that
133 distinguish white blood cell types,^{10, 15-17} we recently developed a set of statistical
134 methods that exploit the use of L-DMRs for inferring changes in cell mixture proportions
135 based solely on DNA methylation profiles of peripheral blood.¹⁸ In this approach (Figure
136 1), data obtained from a *target* set (S_T) consisting of DNA methylation profiles from a

137 heterogeneous mixture of cell populations, is assumed to be a high-dimensional
138 multivariate surrogate for the underlying distribution of cell types. Houseman et al.¹⁸
139 proposed a cell mixture deconvolution methodology that involves the projection of DNA
140 methylation profiles from S_1 onto a *reference* data set (S_0), which is comprised of the
141 DNA methylation signatures for isolated leukocyte subtypes. Under certain constraints,
142 which we describe in more detail in the *Statistical Methods* section, the cell mixture
143 deconvolution approach can be used to approximate the underlying distribution of cell
144 proportions within S_1 via constrained projection (CP).

145 Currently, leukocyte differential counts and flow cytometry measurements (the
146 gold standard for identifying subsets of cells within heterogeneous mononuclear cell
147 samples), are often not possible because they require fresh samples with intact cells, or
148 are too costly. Thus, as epigenome-wide DNA methylation can be measured using
149 archival peripheral blood with relatively straightforward protocols and commercially
150 available array technology or bisulfite sequencing, the capacity to accurately predict cell
151 type proportions using L-DMRs has important implications for any study of health,
152 disease, or pharmacologic intervention where measurement of leukocyte proportions is
153 of interest. For instance, in EWAS¹⁹ (Langevin et al. *under review*) obtaining reliable
154 estimates of relative leukocyte proportions using DNA-based methods could be used for
155 better understanding the extent to which observed differences in whole blood DNA
156 methylation are due to underlying differences in leukocyte subtypes themselves or
157 reflect direct changes in the methylome. Along these lines, the predicted cell type
158 proportions obtained from constrained projection could be added as additional covariate
159 terms to control for the confounding effects of variable leukocyte distribution when
160 examining the association between DNA methylation and some phenotype/exposure of
161 interest. In fact, the approach described in Houseman et al.¹⁸ has been successfully
162 applied in the context of several EWAS¹⁹ (Langevin et al. *under review*, Koestler et al.

163 *provisionally accepted*) and was shown to reliably estimate leukocyte proportions in a
164 small-scale mixture experiment involving six known mixtures of monocytes and B cells
165 and six known mixtures of granulocytes and T cells¹⁸. However, a comprehensive
166 examination of the potential for constrained projection to accurately predict cell type
167 proportions in large-scale epigenome-wide DNA methylation data sets has not been
168 shown.

169 Lam et al.²⁰ recently investigated the relation of peripheral blood DNA
170 methylation with demographic, socioeconomic, and psychosocial factors among a cohort
171 of 94 healthy individuals using commercially available epigenome-wide methylation array
172 technology. In addition, these authors subjected each blood sample to a detailed
173 differential blood cell count. As further validation of the methods of Houseman et al.¹⁸ for
174 estimating relative leukocyte proportions in peripheral blood using L-DMRs, here we
175 present an analysis of their methylation and differential blood cell count data.
176 Specifically, we focus our attention on the utility of the constrained projection approach¹⁸
177 for accurately predicting relative leukocyte distributions, comparing our predictions to
178 those obtained from a widely accepted method for determining cell type distributions in
179 blood. Since there is interest in balancing the number of L-DMRs and cell-type
180 prediction performance, we also present an examination of the sensitivity of our
181 predictions to varying numbers of L-DMRs used in the constrained projection procedure.

182

183 **Results:**

184 As previously described,²⁰ proportions of lymphocytes, monocytes, basophils,
185 eosinophils, and neutrophils were assessed in whole blood by complete blood count
186 (CBC) with differential, for each of the 99 samples among the 94 study subjects. The
187 percentage of granulocytes in whole blood, which ranged from 36.1% - 77.5% across the
188 study subjects, comprised the vast majority of underlying cell types, constituting on

189 average 61.7% (SD = 8.6%) (Figure 2A). On average, lymphocytes and monocytes
190 constituted 31.6% (SD = 8.3%) and 6.7% (SD = 2.1%) of the underlying cell types, and
191 like granulocytes, exhibited substantial variability across the study subjects (range
192 15.1% – 57.4% and 1.5% - 13.1%, respectively) (Figure 2A).

193 Since DNA methylation was assessed in PBMCs, which are mostly devoid of
194 granulocytes, the percentage of lymphocytes and monocytes in PBMCs were taken to
195 be the percentage of these cell types in the absence of granulocytes. From this, we
196 estimated the average percentage of lymphocytes and monocytes in PBMCs to be
197 82.1% and 17.9%, respectively (Figure 2B).

198 As in Lam et al.²⁰ we first began by implementing a principal components
199 analysis (PCA) to gain an understanding of the extent to which variation in DNA
200 methylation across the array could be explained by differences in the underlying
201 distribution of cell types. PCA represents a feature extraction technique where the data
202 is orthogonally transformed, such that the first principal component has the largest
203 possible variance (accounts for maximal amount of variability in the data), and each
204 succeeding component in turn has the next highest variance possible. As we detected
205 substantial variability in DNA methylation due to BeadChip (Supplementary Figure 1), we
206 first applied the ComBat²¹ methodology to normalize the methylation data based on
207 Beadchip. After adjusting out the effects of BeadChip on variability in DNA methylation,
208 we computed the principal components, or otherwise eigen-probes, based on the
209 adjusted DNA methylation data. Not surprisingly, the CBC-derived proportions of
210 lymphocytes and monocytes were found to be associated with the first and third eigen-
211 probes ($p = 0.07$ and $p = 0.06$, respectively), which accounted for 16.5% and 5.4% of the
212 variation of DNA methylation across the array. The second eigen-probe, which
213 accounted for 9.1% of the variation in DNA methylation, was found to be significantly
214 associated with exercise (minutes per week) ($p = 0.04$), ethnicity (Caucasian versus non-

215 Caucasian) ($p = 0.03$), and marginally significantly associated with age, gender, and
216 smoking status (yes versus no) ($p = 0.07, 0.06, 0.09$, respectively). Thus, even among
217 the study subjects considered here, which were all non-diseased at the time of sample
218 collection, differences in white blood cell distributions are contributing to the observed
219 variation in PMBC DNA methylation. These results provide further support for the
220 adjustment cell type distributions when analyzing blood-derived DNA methylation data,
221 particularly in situations where the phenotype or exposure of interest is responsible for
222 shifts in leukocyte subpopulations.

223 We next examined the extent to which CP is capable of producing reliable and
224 accurate estimates of the underlying relative distribution of leukocytes. To discern L-
225 DMRs, we examined the association between methylation and leukocyte subtype (i.e.,
226 CD4+ T cells, CD8+ T cells, B cells, ect.) for each of the 26,486 autosomal CpG loci.
227 This revealed 10,370 significantly differentially methylated CpGs among the leukocyte
228 subtypes (fdr q-value < 0.05), which we ranked by q-value. Consistent with Liu et al.¹⁹
229 we applied CP using the top 500 L-DMRs, allowing us to obtain predictions for the
230 proportions of CD8+ T-lymphocytes (CD8T), CD4+ T-lymphocytes (CD4T), Natural Killer
231 cell (NK), B-cells (Bcell), Monocytes (Mono), and Granulocytes (Gran) across the 99
232 individual samples. As there were 5 subjects in S_1 with replicate samples (collected at
233 the same time), we had the unique opportunity to assess the similarity in cell type
234 predictions within a replicate pair; which would be expected to be high. The results of
235 this analysis are given in Figure 2C, and show a high-degree of similarity between the
236 predicted cell type proportions among the 5 technical replicates – indicated by colored
237 points. Within a specific cell type, differences between the predicted percentages
238 among technical replicates was minimal, with a mean difference of 2% (SD = 2%). This
239 is also captured in the intra-class correlation coefficients (ICCs), which ranged from 0.85

240 – 0.95, demonstrating a high-degree of similarity in the predicted cell type proportions
 241 among technical replicates.

242 As DNA methylation was profiled in PBMCs, we were also interested in
 243 examining the specificity of CP by investigating the predicted proportions of granulocytes
 244 – which would be expected to be approximately zero, allowing for some small residual
 245 contamination in purification. As noted in Figure 2C the predicted percentage of
 246 granulocytes was minimal, ranging from 0% - 7% with a mean value across the study
 247 samples of 1% (SD = 1.3%). Examining the correlation between the predicted
 248 percentage of lymphocytes, obtained by summing the individual predictions among the
 249 lymphoid-derived cells (i.e., CD4T, CD8T, NK, and Bcells), and the percentage of
 250 lymphocytes via CBC (Figure 3A), demonstrated a moderate-high correlation between
 251 predicted lymphocyte proportions and those obtained from CBC ($r = 0.61$; $p < 0.0001$).
 252 Similarly, we also observed a moderate-high correlation between predicted monocyte
 253 proportions and those obtained from CBC ($r = 0.60$; $p < 0.0001$) (Figure 3B).

254 Across the study samples, there was remarkable consistency between the
 255 average percentage of monocytes and lymphocytes via CBC (17.9% and 82.1%,
 256 respectively) and the average predicted percentage of monocytes (17.6%) and
 257 lymphocytes (81.3%). Furthermore, the root mean squared error (rMSE) (i.e.,

258 $\sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{W}_{ik} - W_{ik}^{(CBC)})^2}$, for $k \in \{1, 2, \dots, K\}$) based on comparisons of the predicted and

259 CBC percentages of lymphocytes and monocytes was 6% and 5%, respectively. We
 260 also note that the vast majority of our subject-specific cell type predictions were within
 261 the global 95% bootstrap prediction interval (Figure 3C,D). Examining the bias in our
 262 cell type predictions based on characteristics of the study subjects showed some
 263 evidence of an association between cell type-specific prediction error and stress ($p =$
 264 0.06 and 0.05 for monocytes and lymphocytes, respectively), depression ($p = 0.07$ and

265 0.03 for monocytes and lymphocytes, respectively), and current SES status ($p = 0.02$ for
266 lymphocytes) (Supplementary Tables 1-2). However none of the aforementioned p -
267 values remained statistically significant after controlling for multiple comparisons.

268 Two possibilities to explain this phenomenon include that the accuracy of CBC
269 counts are themselves associated stress, depression, and current SES. Since this is
270 unlikely, the second possibility is that the accuracy of cell type predictions via CP is
271 associated with these covariates. Or in other words, a subject's value for these
272 covariates is adversely influencing the accuracy of our predictions. Since our predictions
273 were based on CP using the top 500 L-DMRs, this would necessarily imply that the
274 methylation status of the top 500 L-DMRs are themselves altered based on the values of
275 these covariates. To this end, we conducted an additional analysis aimed at
276 investigating the association between the methylation status of the top 500 L-DMRs and
277 each of the previously mentioned covariates. For this analysis, we fit a series of
278 generalized estimating equations (GEE) that modeled the methylation M-values for the
279 top 500 L-DMRs, the above covariates as a dependent variable, and incorporated
280 dependency based on replicate samples from the same subject. These models were
281 also adjusted for either the CBC-derived proportion of lymphocytes or the predicted
282 proportion of lymphocyte subtypes (i.e., CD4T, CD8T, ect.) to remove the confounding
283 effect due to interpersonal differences in immune cell subsets. The p -values reported in
284 Supplementary Table 3 reflect the omnibus p -value obtained from a permutation test
285 (further details provided in the Supplementary Material) and demonstrated no
286 association between the top 500 L-DMRs and stress, depression, and current SES ($p =$
287 0.45, 0.56, and 0.12, respectively). While a number of other covariates demonstrated a
288 significant association with the top 500 L-DMRs (i.e., age, gender, ethnicity), none of
289 these covariates were associated with bias in our cell type predictions (Supplementary
290 Tables 1-2). Furthermore, removing the L-DMRs that were significantly associated with

291 age, gender, and ethnicity followed by the subsequent application of CP using the
292 remaining L-DMRs, showed a very high correlation with the previously obtained cell-type
293 estimates (Pearson correlation = 0.99, 0.99, 0.98, 0.96, 0.99, and 0.97 for CD4T, CD8T,
294 Bcell, NK, Mono, and Gran, respectively).

295 We also considered a negative control analysis as a further validation of CP and
296 of the utility of L-DMRs in inferring cell type proportions. While our previous analysis
297 used the top 500 L-DMRs (Supplementary Figure 2A) for predicting cell type proportions
298 in our target data set, as a negative control we used 500 CpGs among the set of non-L-
299 DMRs (i.e., those with $\text{fdr } q\text{-value} > 0.05$). Specifically, the 500 least discriminative
300 CpGs across the leukocyte subtypes were selected for this analysis (Supplementary
301 Figure 2B) and used in the previously described CP procedure to arrive at predictions for
302 cell type proportions. The results of this analysis, showed very little correlation between
303 the between predicted cell type proportions and those obtained from CBC ($r = 0.10$ for
304 both monocytes and lymphocytes, respectively) and an rMSE of 34% between the
305 predicted and CBC-derived cell type percentages for both lymphocytes and monocytes.

306 We next implemented a sensitivity analysis aimed at understanding the
307 sensitivity of the predicted cell type proportions attempted to gain an understanding of
308 the sensitivity of our predictions when m , or the number of L-DMRs used in CP, was
309 varied. For this analysis, m was varied from 20 to 10,000 and as previously, for each
310 selection of m , the correlation and rMSE were used to compare the predicted and CBC-
311 derived proportions of monocytes and lymphocytes. Additionally, for each selection of
312 m , the predicted proportion of granulocytes was recorded and used to assess the
313 specificity of CP across differing selections of m . Our choice of 10,000 L-DMRs as the
314 upper limit in our sensitivity analysis was based on the number of statistically significant
315 CpG loci across the leukocyte subtypes (10,370 with $\text{fdr } q\text{values} < 0.05$). The results of
316 this analysis are given in Figure 4 and show minimal variation in the correlation

317 coefficients between the predicted and CBC-derived proportions of monocytes and
318 lymphocytes across different selections of m (Figure 4A). Specifically, beyond 1000 L-
319 DMRs correlations between the predicted and CBC-derived proportions of monocytes
320 and lymphocytes varied by at most 0.04 on the correlation scale and both appeared to
321 achieve maximum correlation at $m = 6,000$ (i.e., the top 6,000 L-DMRs). Similarly, there
322 was minimal variation in the rMSE between the predicted and CBC-derived proportions
323 of monocytes and lymphocytes across different selections of m , with differences of at
324 most 1.5% in rMSE across the selected numbers of L-DMRs (Figure 4B). While the
325 median percent of granulocytes was minimized at approximately $m = 800$, like the
326 correlation and rMSE between predicted and CBC-derived proportions of monocytes and
327 lymphocytes, there was a fair degree of stability in the median percent of granulocytes
328 as a function of m (Figure 4C).

329 In a manner similar to the sensitivity analysis described above, we also examined
330 the correlation and rMSE based on predicted and CBC-derived proportions of
331 monocytes and lymphocytes as a function of the number of non L-DMRs used in CP..
332 Contrary, to the relative stability in both correlation and rMSE as a function of the
333 number of L-DMRs, the correlation between predicted and CBC-derived proportions of
334 monocytes and lymphocytes based on non L-DMRs varied considerably as a function of
335 m (-0.18 – 0.25 across the range of m) and tended to be centered at 0 (Figure 4D).
336 Similarly, the rMSE for monocytes and lymphocytes was as large as 47% and 48%,
337 respectively, but stabilized at around $m = 6,000$ with an rMSE of approximately 15% for
338 both monocytes and lymphocytes (Supplementary Figure 3).

339

340 **Discussion:**

341 Using a publicly available data set consisting of PBMC-derived DNA methylation and
342 CBC counts for 99 samples across 94 healthy non-diseased subjects, we have

343 investigated the extent to which the constrained projection approach of Houseman et
344 al.¹⁸ provides reliable and accurate estimates of the underlying relative leukocyte
345 distribution in blood. Owing to the fact that blood is a readily accessible tissue and
346 because peripheral blood leukocytes have been suggested to directly or indirectly
347 participate in the pathophysiology of a vast array of disease states,²²⁻²⁴ DNA methylation
348 analyses using blood-derived genomic DNA have been conducted across a variety of
349 different human diseases^{8, 9, 12, 25-28} and also in the context of exposures.²⁹⁻³¹ The
350 validation analysis considered here is motivated both by (i) the increasing number of
351 EWAS using blood-based assessment of DNA methylation and (ii) the recognized
352 potential for confounding based on underlying interpersonal differences in circulating
353 immune profiles characteristic of such studies.¹⁵ While several recent works have
354 adjusted for CBC counts^{20, 32} for identifying differential patterns of methylation in blood
355 that are independent of the underlying distribution of leukocytes, in many cases CBC
356 counts may not be readily available (or even feasible), and in any case can not provide
357 complete information on immune variability due to their inability to discriminate
358 lymphocyte subtypes. While FACS can be used to identify lymphocyte subtypes, this
359 method bears a relatively high cost per sample and generally requires fresh samples.
360 The limitations of current approaches underscore the vast potential utility of DNA
361 methylation-based methods and accompanying statistical techniques that are capable of
362 accurately and reliably estimating the distribution of cell types.

363 Our initial analyses, which used the top 500 L-DMRs in CP, first focused on the
364 investigating the specificity of this approach. As DNA methylation was profiled in
365 PBMCs (devoid of multinucleate granulocytes) for the samples in our *target* methylation
366 data set, the percentage of granulocytes would be expected to be approximately zero –
367 potentially subject to some granulocyte contamination in the isolation process.³³ Our
368 findings, which demonstrated a minimal predicted percentage of granulocytes across the

369 *target* study samples (1%; mean across study samples) illustrate the specificity of CP
370 and are even more noteworthy when considering that granulocytes typically comprise
371 the vast majority of white blood cell types (50% - 70%) in the whole blood of non-
372 diseased individuals. As our *target* methylation data set consisted of technical replicate
373 samples for five subjects, we were also interested in investigating the reproducibility of
374 CP by comparing the predicted cell type proportions within a replicate pair. Overall, our
375 results demonstrated a high-degree of similarity in the predicted cell type proportions
376 within replicate pairs. The minor differences between the predicted cell type proportions
377 between replicate samples from the same subject is not unexpected given the less than
378 perfect nature of cell sorting (>97% purity) to obtain S_0 , the role of measurement error in
379 array-based DNA methylation assessment,³⁴ and the well established issue of technical
380 variability in DNA microarrays arising from plate/BeadChip effects.³⁵⁻³⁷

381 Our findings also demonstrated high consistency between the average CBC-
382 derived and predicted percentage of monocytes and lymphocytes (17.9% and 17.6% for
383 monocytes and 82.1% and 81.4% for lymphocytes), with rMSE of 5% and 6%, for
384 monocytes and lymphocytes, respectively. Moreover, bias in our estimates of the
385 proportion of monocytes appeared to be independent of most potential confounders in
386 DNA methylation array analyses. Of those that showed some evidence of an
387 association with cell type-specific prediction error (i.e., stress, depression, and current
388 SES status), none were significantly associated with the methylation status of the top
389 500 L-DMRs. While there were some covariates that exhibited significant associations
390 with the top 500 L-DMRs (i.e., age, gender, and ethnicity), these results are likely to be
391 conservative as the models we fit controlled for only the CBC-derived proportion of
392 lymphocytes and not individual lymphocyte subtypes, which were not available in the
393 *target* data set. Moreover, removal of those specific L-DMRs followed by the
394 subsequent estimate of cell type proportions based on the remaining L-DMRs showed a

395 very high correlation with the previously obtained estimates. Based on these findings, L-
396 DMRs that exhibited covariation with subject-specific characteristics do not seem to be
397 substantially influencing our estimates of cell type proportions.

398 Reinforcing the potential of CP for producing accurate cell type predictions, we
399 also observed a moderate-high correlation between predicted monocyte and lymphocyte
400 proportions and those obtained from CBC counts. These results, taken together with the
401 findings of our negative control analysis indicating poor prediction performance when CP
402 is based on the least discriminative L-DMRs, stand as testament to the value of L-DMRs
403 in deconvoluting cellular mixtures based on blood-derived DNA methylation profiles in a
404 *target* methylation data set. While our *reference* data set allowed us to predict the
405 proportion of specific lymphocyte subtypes (i.e., CD4T, CD8T, ect.), such a detailed
406 speciation of lymphocytes was not available for the *target* data set considered here. As
407 a result, this limited our ability to assess the predictive accuracy of CP for these cell
408 types. We do note however that future work involving measurements of individual
409 lymphocyte subtypes in a *target* data set is currently underway.

410 As the capacity to accurately predict the underlying relative leukocyte distribution
411 in blood is principally driven by DMRs across leukocyte subtypes, Illumina's most recent
412 BeadArray, the HumanMethylation450 BeadArray, which simultaneously profiles the
413 methylation status for >485,000 CpGs, is likely to reveal additional L-DMRs. In doing so,
414 these additional L-DMRs could be added to our existing set of L-DMRs, which might
415 further improve the accuracy and precision of cell type predictions. As a cautionary
416 note, attention should be given toward selecting L-DMRs containing SNPs at/near the
417 targeted probe, which might affect the measurement of DNA methylation. Furthermore,
418 while the top 500 L-DMRs used here comprised only autosomal CpG loci (X and Y
419 linked loci were removed) due to the potential for gender associated biases, the
420 application of the methods described here to gender-specific data sets (i.e., ovarian

421 cancer, prostate cancer, ect.) could be augmented by including both autosomal and non-
422 autosomal L-DMRs. However, we expect only marginal differences in cell type
423 estimates, as only a small fraction of the top L-DMRs were associated with non-
424 autosomal CpG loci (i.e. 6 out of 500). Similarly, future work involving methylation
425 profiling of additional sorted cell types, such as nucleated red blood cells present at birth
426 and in cord blood, M1 and M2 macrophages, and myeloid derived suppressor cells,
427 have the potential to further refine studies of infant cord blood methylation profiles.

428 It should also be noted that while confounding in blood-based assessment of
429 DNA methylation by variation in circulating immune cells motivated the methods
430 described in Houseman et al.¹⁸, the underlying proportion of leukocytes could also
431 confound other leukocyte DNA biomarker relationships, including the relationship
432 between telomere length, repetitive element DNA methylation,¹³ or mitochondrial copy
433 number¹⁴ and exposures or disease phenotypes. Thus, future applications might involve
434 an extension of the methods of Houseman et al.¹⁸ for deconvoluting cell mixtures using
435 DNA methylation data and controlling for this confounding in studies of these and other
436 leukocyte-based biomarkers.

437 Our sensitivity analysis of cell type predictions as a function of the number of L-
438 DMRs, m , used in CP, demonstrated that both the rMSE and the correlation between
439 predicted and CBC-derived cell type proportions were relatively stable as a function of
440 m . While we and others¹⁹ have found that using the top 500 L-DMRs in CP works well,
441 we recommend that investigators interested in implementing this methodology do so
442 using a range of L-DMRs, checking ensure that cell type predictions remain relatively
443 stable as a function of m . We also note that other algorithms, i.e., ones other than using
444 omnibus F-statistic for the one-way ANOVA problem for identifying L-DMRs, might result
445 in different optimal number of L-DMRs. For example, t-statistics for pairwise
446 comparisons of CpG-specific DNA methylation across leukocyte subtypes may actually

447 result in a fewer number of total L-DMRs while maintaining or exceeding the prediction
448 performance.

449 In summary this work serves as further validation of the CP approach of
450 Houseman et al.¹⁸ using an independent data set based on a large-scale EWAS focused
451 on healthy non-diseased adults. The increasing numbers of EWAS that involve DNA
452 methylation profiling in unfractionated whole blood coupled with the well-established role
453 of confounding due to cell type distributions, highlight the promise and future
454 applications of this technique.

455

456 **Materials and Methods:**

457 *Target samples S_1 , DNA methylation from heterogeneous mixture of cell types*

458 To investigate the extent to which patterns of blood-based DNA methylation can be used
459 for inferring the underlying distribution of cell types, we used publicly available data
460 deposited in the Gene Expression Omnibus (GEO) database (accession no.
461 GSE37008). This study, which has been previously described,²⁰ consisted of
462 epigenome-wide assessment of DNA methylation based on genomic DNA derived from
463 purified peripheral blood mononuclear cells (PBMCs) from a community cohort of 94
464 non-diseased individuals in the Vancouver, BC lower mainland area.³⁸ Individuals in this
465 study ranged in age from 24 to 45 years (median = 33, SD = 5.08), were predominantly
466 female (63%; $n = 59$), and non-smokers (87%; $n = 82$).

467 The Illumina Infinium HumanMethylation27 array platform, which enables the
468 quantitative assessment of the DNA methylation status of 27,578 CpG loci at single-
469 nucleotide resolution, was used to measure DNA methylation in genomic DNA derived
470 from PBMCs. The methylation status for each individual CpG locus was calculated as
471 the ratio of fluorescent signals ($\beta = \text{Max}(M,0)/[\text{Max}(M,0)+\text{Max}(U,0) + 100]$), ranging from
472 0 (no methylation) to 1 (complete methylation), using the average probe intensity for the

473 methylated (M) and unmethylated (U) alleles. CpG loci associated with X and Y
474 chromosomes were removed from our analyses, due to gender-associated biases. The
475 DNA methylation status was assessed in replicate for 5/94 of the individuals in this study
476 (samples collected at the same time point), giving rise to a total of 99 samples that
477 comprised the *target* set (S_1) used in our validation analysis.

478

479 *Assessment of cell type proportions in target samples, S_1*

480 As previously described by,²⁰ blood draw samples were processed immediately with
481 density-gradient centrifugation for isolation of peripheral blood mononuclear cells
482 (PBMCs). At the time of blood draw, samples were subjected to complete blood count
483 (CBC) with differential using an Advia 70 Hematology System (Siemens Medical) to
484 estimate the proportions of lymphocytes, monocytes, basophils, eosinophils, and
485 neutrophils. In addition, in a subset of PBMC samples, subpopulations of lymphocytes
486 were captured by immunomagnetic selection for CD14+ lymphocytes as well as CD3+
487 monocytes.

488

489 *Reference samples S_0 , DNA methylation from isolated cells*

490 As previously described,^{10, 18} our *reference* set (S_0) consisted of sorted, normal, human,
491 peripheral blood leukocyte subtypes, purchased from AllCells. Leukocytes were isolated
492 from different, anonymous, nondiseased individuals' whole blood by magnetic-activated
493 cell sorting (MACS) using a combination of negative and positive selection with highly
494 specific cell surface antibodies conjugated to magnetic beads. The purity of separated
495 cells was confirmed with flow cytometry to be >97% and included 46 white blood cell
496 samples, comprising lymphoid (B cells, Natural Killer (NK) cells, and Pan-T-cells) and
497 myeloid (Monocytes and Granulocytes) derived cells (Table 1). Genomic DNA was
498 extracted and purified from cell pellets using a commercially available method (Qiagen),

499 treated with sodium bisulfite (Zymo Research) and subjected to methylation profiling
 500 using the Infinium HumanMethylation27 BeadArray (Illumina); the same platform used for
 501 the DNA methylation analysis of the *target* samples described above.

502

503 *Statistical methods*

504 While a complete description of the constrained projection (CP) approach for predicting
 505 cell type proportions based on DNA methylation signatures from a heterogeneous
 506 mixture of cells has been described previously Houseman et al.¹⁸, below we summarize
 507 the salient aspects of this approach with specific attention given towards those that
 508 relate to this validation analysis. As described above, let S_0 denote the reference
 509 sample of DNA methylation profiles from isolated cells and let S_1 denote the
 510 corresponding set of target DNA methylation profiles, which are assumed to arise from
 511 mixtures of the cell types isolated in S_0 (Figure 1B). Here, S_0 is comprised of the DNA
 512 methylation profiles for n_0 specimens (i.e., $n_0 = 46$ based on our *reference* data set), \mathbf{Y}_{0i} , i
 513 $= 1, 2, \dots, n_0$, an $m \times 1$ vector of DNA methylation measurements. Similarly, S_1 consists
 514 of the DNA methylation profiles for n_1 samples (i.e., $n_1 = 99$ based on our *target* data
 515 set), \mathbf{Y}_{1i} , $i = 1, 2, \dots, n_1$, for the same m CpG sites in \mathbf{Y}_{0i} (and in the same order). Each
 516 element in \mathbf{Y}_{hi} , $h \in \{0, 1\}$ corresponds to a specific, pre-selected L-DMR chosen to
 517 distinguish one or more of the cellular subtypes assayed in S_0 and contributing to the
 518 mixtures measured in S_1 . As previously described,^{10, 18} L-DMRs were identified by rank
 519 ordering CpGs based on the F-statistics for distinguishing cell types, obtained from a
 520 series of linear mixed effects models fit to each CpG independently among the
 521 specimens in S_0 . Assuming that S_0 is comprised of K different cell types (i.e., $K = 6$
 522 based on our *reference* data set), each of which has mean m_k , we have that
 523 $E(\mathbf{Y}_{0i} | \mathbf{c}_i = k) = m_k$, where \mathbf{c}_i denotes cell type and $\mathbf{c}_i = \{1, 2, \dots, K\}$. Therefore,

524 $\mathbf{M} = (m_1, m_2, \dots, m_K)$ represents an $m \times K$ matrix of mean methylation for the m selected L-
 525 DMRs across the K different leukocyte subtypes. Here, we used a series of mixed
 526 effects models (i.e., treating chip as a random effect) to obtain $\hat{\mathbf{M}}$.

527 Assuming that subject i assayed in S_1 is a mixture of the K leukocyte subtypes
 528 assayed in the reference set S_0 , with mixing coefficients represented by a $K \times 1$ vector

529 W_i , $\sum_{k=1}^K W_{ik} \leq 1$, where $W_{ik} \geq 0$, then $E(\mathbf{Y}_{1i} | W_i = W) = \mathbf{M}W$. That is, the methylation

530 profile of subject i in the *target* data is assumed to arise as the weighted methylation
 531 profile across the K leukocyte subjects, such that the contribution of each subtype, or
 532 otherwise the proportion of each leukocyte subtype, is reflected by W . Thus, interest
 533 here is focused on the estimation of W_i . Houseman et al.¹⁸ demonstrate that W_i can be

534 estimated using constrained projection, i.e., by setting \hat{W}_i to the value of W that

535 minimizes $\|\mathbf{Y}_{1i} - \hat{\mathbf{M}}W\|$ with the constraint $W_k \geq 0, k \in \{1, 2, \dots, K\}$ and the additional

536 constraint that $\sum_{k=1}^K W_k \leq 1$. The former constraint ensures non-negativity among for

537 estimated proportion of particular cell type and the later ensures that the coefficients
 538 have the “multinomial” interpretation of additive proportions.

539 Bootstrap resampling was used to quantify uncertainty in the estimation of W_i .

540 Since there are several sources of variability; including variability in the observed
 541 methylation values for the samples in S_1 and in the estimate of \mathbf{M} , a parametric bootstrap
 542 procedure was used to obtain resampled estimates of the cell type proportions,

543 $\hat{W}^{(r)}, r = 1, 2, \dots, 1000$ for each sample in S_1 . The standard deviation of the resampled

544 estimates of the cell type proportions were computed and used to construct 95%

545 prediction intervals for \hat{W}_i . Further details regarding the parametric bootstrap procedure
546 are provided elsewhere¹⁸.

547 In our examination, we focused first on obtaining the estimate \hat{W}_i , followed by the
548 subsequent comparison of \hat{W}_i and $W_i^{(CBC)}$, where $W_i^{(CBC)}$ represents the proportions of
549 the K leukocyte subjects obtained using complete blood cell count measurements.
550 Additionally m , or the number of L-DMRs used in the constrained projection, is a tuning
551 parameter. Thus, we also examined the sensitivity of \hat{W}_i as value of m was varied from
552 20 to 10,000.

553 We note a few considerations that arise in the comparison of \hat{W}_i and $W_i^{(CBC)}$. As
554 previously mentioned, our *target* data set consisted of whole-blood, CBC counts of
555 lymphocytes, monocytes, basophils, eosinophils, and neutrophils (whereas DNA
556 methylation was profiled in PBMCs). The percentage of these cells in whole blood was
557 taken to be the count of the various cell types per 10^{-9} liter of whole blood divided by the
558 sum of the counts over all cell types. The percentage of granulocytes was computed as:
559 granulocyte(%) = basophil(%) + eosinophils(%) + neutrophil(%). Since DNA methylation
560 was assessed in PBMCs, which contain a negligible proportion of granulocytes, the
561 percentage of lymphocytes and monocytes in PBMCs were taken to be the percentage
562 of these cell types in the absence of granulocytes, i.e., the count of these cell types per
563 10^{-9} liter of whole-blood by the sum of the counts of only lymphocytes and monocytes per
564 10^{-9} liter of whole-blood. Thus, $W_i^{(CBC)}$ is a 1×3 vector, representing the proportions of
565 lymphocytes, monocytes, and granulocytes in PBMCs.

566 In combination with the methylation data available for our *target* data, our
567 *reference* data on isolated leukocyte subtypes, allowed us to obtain estimates of the
568 proportions of each of the cell types given in Table 1. Since such a detailed speciation

569 of leukocytes was not available from the CBC measurements in the *target* data -
570 particularly for the lymphoid derived cell types - we took our estimate of the proportion of
571 lymphocytes to be the sum of the individual estimates of the lymphoid derived cells, i.e.,
572 $\text{Lymphocyte}(\%) = \text{CD4+Tcell}(\%) + \text{CD8+Tcell}(\%) + \text{NK cell}(\%) + \text{Bcell}(\%)$. Hence, \hat{W}_i
573 represents a 1×3 vector, indicating the estimated proportions of lymphocytes,
574 monocytes, and granulocytes for sample i within the *target* data.

575 Given the potential for confounding in the analysis of DNA methylation data
576 based on factors such as age, gender, race and smoking status etc., we conducted a
577 series of analyses aimed at examining the association between the prediction error and
578 absolute prediction error ($(\hat{W}_i - W_i^{(CBC)})$ and $|\hat{W}_i - W_i^{(CBC)}|$) and potential confounders.
579 Specifically, we examined the extent to which bias in our predictions are associated with:
580 age (yrs), gender, smoking status (yes/no), childhood socio-economic status (high/low),
581 current socio-economic status (high/low), alcohol consumption (drinks per week), BMI,
582 exercise (min. per week), stress (perceived stress scale questionnaire), depression
583 (center for epidemiologic studies depression scale), and ethnicity (Caucasian/non-
584 Caucasian). For each of the above factors, a linear mixed effects model was fit that
585 modeled prediction error or absolute prediction error as the response, the potential
586 confounder as the independent variable, and a included random effect term for subject
587 to account for correlated errors among replicate samples collected from the same
588 subject. Unadjusted and false discovery rate adjusted P-values were computed for each
589 of aforementioned factors. Along these lines, we also examined the association
590 between the top 500 L-DMRs and each of the covariates described above. Further
591 details regarding the methods used in the analysis are given in the Supplementary
592 Material.

593

594 **Acknowledgements:** This work was supported by the US National Institutes of Health
595 grants: R01 CA078609 to K.T.K., R25 CA134286, ES018175 and RD83459901 to
596 M.R.K, RO1 CA126831 to J.K.W., and R01 MH094609 to C.M.J.

597

598 **References:**

599

- 600 1. Rakyan VK, Down TA, Thorne NP, Flicek P, Kulesha E, Graf S, et al. An
601 integrated resource for genome-wide identification and analysis of human tissue-
602 specific differentially methylated regions (tDMRs). *Genome Res* 2008; 18:1518-29.
- 603 2. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels
604 JL, et al. Aging and environmental exposures alter tissue-specific DNA
605 methylation dependent upon CpG island context. *PLoS Genet* 2009; 5:e1000602.
- 606 3. Baron U, Turbachova I, Hellwag A, Eckhardt F, Berlin K, Hoffmuller U, et al.
607 DNA methylation analysis as a tool for cell typing. *Epigenetics* 2006; 1:55-60.
- 608 4. Bocker MT, Hellwig I, Breiling A, Eckstein V, Ho AD, Lyko F. Genome-wide
609 promoter DNA methylation dynamics of human hematopoietic progenitor cells
610 during differentiation and aging. *Blood* 2011; 117:e182-9.
- 611 5. Ji H, Ehrlich LI, Seita J, Murakami P, Doi A, Lindau P, et al. Comprehensive
612 methylome map of lineage commitment from haematopoietic progenitors. *Nature*
613 2010; 467:338-42.
- 614 6. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Gayther SA,
615 Apostolidou S, et al. An epigenetic signature in peripheral blood predicts active
616 ovarian cancer. *PLoS One* 2009; 4:e8274.
- 617 7. Wang L, Aakre JA, Jiang R, Marks RS, Wu Y, Chen J, et al. Methylation
618 markers for small cell lung cancer in peripheral blood leukocyte DNA. *J Thorac*
619 *Oncol* 2010; 5:778-85.
- 620 8. Marsit CJ, Koestler DC, Christensen BC, Karagas MR, Houseman EA,
621 Kelsey KT. DNA methylation array analysis identifies profiles of blood-derived
622 DNA methylation associated with bladder cancer. *J Clin Oncol* 2011; 29:1133-9.
- 623 9. Pedersen KS, Bamlet WR, Oberg AL, de Andrade M, Matsumoto ME, Tang
624 H, et al. Leukocyte DNA methylation signature differentiates pancreatic cancer
625 patients from healthy controls. *PLoS One* 2011; 6:e18223.
- 626 10. Koestler DC, Marsit CJ, Christensen BC, Accomando W, Langevin SM,
627 Houseman EA, et al. Peripheral blood immune cell methylation profiles are
628 associated with nonhematopoietic cancers. *Cancer Epidemiol Biomarkers Prev*
629 2012; 21:1293-302.

- 630 11. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlen SE, Greco D, et al.
631 Differential DNA methylation in purified human blood cells: implications for cell
632 lineage and studies on disease susceptibility. *PLoS One* 2012; 7:e41361.
- 633 12. Langevin SM, Koestler DC, Christensen BC, Butler RA, Wiencke JK, Nelson
634 HH, et al. Peripheral blood DNA methylation profiles are indicative of head and
635 neck squamous cell carcinoma: an epigenome-wide association study.
636 *Epigenetics* 2012; 7:291-9.
- 637 13. Marsit C, Christensen B. Blood-derived DNA methylation markers of cancer
638 risk. *Advances in experimental medicine and biology* 2013; 754:233-52.
- 639 14. Hou L, Zhu ZZ, Zhang X, Nordio F, Bonzini M, Schwartz J, et al. Airborne
640 particulate matter and mitochondrial damage: a cross-sectional study.
641 *Environmental health : a global access science source* 2010; 9:48.
- 642 15. Adalsteinsson BT, Gudnason H, Aspelund T, Harris TB, Launer LJ,
643 Eiriksdottir G, et al. Heterogeneity in white blood cells has potential to confound
644 DNA methylation measurements. *PLoS One* 2012; 7:e46705.
- 645 16. Wieczorek G, Asemissen A, Model F, Turbachova I, Floess S, Liebenberg V,
646 et al. Quantitative DNA methylation analysis of FOXP3 as a new method for
647 counting regulatory T cells in peripheral blood and solid tissue. *Cancer research*
648 2009; 69:599-608.
- 649 17. Sehouli J, Loddenkemper C, Cornu T, Schwachula T, Hoffmuller U,
650 Grutzkau A, et al. Epigenetic quantification of tumor-infiltrating T-lymphocytes.
651 *Epigenetics* 2011; 6:236-46.
- 652 18. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ,
653 Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture
654 distribution. *BMC Bioinformatics* 2012; 13:86.
- 655 19. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al.
656 Epigenome-wide association data implicate DNA methylation as an intermediary
657 of genetic risk in rheumatoid arthritis. *Nat Biotechnol* 2013; 31:142-7.
- 658 20. Lam LL, Emberly E, Fraser HB, Neumann SM, Chen E, Miller GE, et al.
659 Factors underlying variable DNA methylation in a human community cohort.
660 *Proceedings of the National Academy of Sciences of the United States of America*
661 2012; 109 Suppl 2:17253-60.
- 662 21. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray
663 expression data using empirical Bayes methods. *Biostatistics* 2007; 8:118-27.
- 664 22. Chung FM, Tsai JC, Chang DM, Shin SJ, Lee YJ. Peripheral total and
665 differential leukocyte count in diabetic nephropathy: the relationship of plasma
666 leptin to leukocytosis. *Diabetes care* 2005; 28:1710-7.
- 667 23. Lewis SA, Pavord ID, Stringer JR, Knox AJ, Weiss ST, Britton JR. The
668 relation between peripheral blood leukocyte counts and respiratory symptoms,

- 669 atopy, lung function, and airway responsiveness in adults. *Chest* 2001; 119:105-
670 14.
- 671 24. Nadif R, Siroux V, Oryszczyn MP, Ravault C, Pison C, Pin I, et al.
672 Heterogeneity of asthma according to blood inflammatory patterns. *Thorax* 2009;
673 64:374-80.
- 674 25. Baccarelli A, Wright R, Bollati V, Litonjua A, Zanobetti A, Tarantini L, et al.
675 Ischemic heart disease and stroke in relation to blood DNA methylation.
676 *Epidemiology* 2010; 21:819-28.
- 677 26. Kim M, Long TI, Arakawa K, Wang R, Yu MC, Laird PW. DNA methylation as
678 a biomarker for cardiovascular disease risk. *PLoS One* 2010; 5:e9692.
- 679 27. Mill J, Tang T, Kaminsky Z, Khare T, Yazdanpanah S, Bouchard L, et al.
680 Epigenomic profiling reveals DNA-methylation changes associated with major
681 psychosis. *American journal of human genetics* 2008; 82:696-711.
- 682 28. Zhu X, Liang J, Li F, Yang Y, Xiang L, Xu J. Analysis of associations
683 between the patterns of global DNA hypomethylation and expression of DNA
684 methyltransferase in patients with systemic lupus erythematosus. *International*
685 *journal of dermatology* 2011; 50:697-704.
- 686 29. Alegria-Torres JA, Barretta F, Batres-Esquivel LE, Carrizales-Yanez L,
687 Perez-Maldonado IN, Baccarelli A, et al. Epigenetic markers of exposure to
688 polycyclic aromatic hydrocarbons in Mexican brickmakers: A pilot study.
689 *Chemosphere* 2013; 91:475-80.
- 690 30. Joubert BR, Haberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, et al.
691 450K epigenome-wide scan identifies differential DNA methylation in newborns
692 related to maternal smoking during pregnancy. *Environmental health perspectives*
693 2012; 120:1425-31.
- 694 31. Thapar M, Covault J, Hesselbrock V, Bonkovsky HL. DNA methylation
695 patterns in alcoholics and family controls. *World journal of gastrointestinal*
696 *oncology* 2012; 4:138-44.
- 697 32. Byun HM, Nordio F, Coull BA, Tarantini L, Hou L, Bonzini M, et al. Temporal
698 stability of epigenetic markers: sequence characteristics and predictors of short-
699 term DNA methylation variations. *PLoS One* 2012; 7:e39220.
- 700 33. Bryk JA, Popovic PJ, Zenati MS, Munera V, Pribis JP, Ochoa JB. Nature of
701 myeloid cells expressing arginase 1 in peripheral blood after trauma. *The Journal*
702 *of trauma* 2010; 68:843-52.
- 703 34. Laird PW. Principles and challenges of genomewide DNA methylation
704 analysis. *Nature reviews Genetics* 2010; 11:191-203.
- 705 35. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by
706 surrogate variable analysis. *PLoS Genet* 2007; 3:1724-35.

707 **36. Sun Z, Chai HS, Wu Y, White WM, Donkena KV, Klein CJ, et al. Batch effect**
 708 **correction for genome-wide methylation data with Illumina Infinium platform. BMC**
 709 **medical genomics 2011; 4:84.**

710 **37. Teschendorff AE, Zhuang J, Widschwendter M. Independent surrogate**
 711 **variable analysis to deconvolve confounding factors in large-scale microarray**
 712 **profiling studies. Bioinformatics 2011; 27:1496-505.**

713 **38. Miller GE, Chen E, Fok AK, Walker H, Lim A, Nicholls EF, et al. Low early-**
 714 **life social class leaves a biological residue manifested by decreased**
 715 **glucocorticoid and increased proinflammatory signaling. Proceedings of the**
 716 **National Academy of Sciences of the United States of America 2009; 106:14716-21.**
 717
 718
 719

720 **Figure Legends:**

721
 722 **Figure 1: Illustration of the blood cell mixture deconvolution approach.** This
 723 approach involves, (A) constrained projection of DNA methylation profiles from a *target*
 724 methylation data set (S_1) onto a *reference* data set (S_0), which is comprised of the DNA
 725 methylation signatures for isolated white blood cell types (i.e., shapes reflect different
 726 white blood cell types). The result is an estimate of the underlying distribution of cell
 727 proportions (i.e., circle, triangle, and hexagon) for each sample within S_1 . (B) This
 728 approach assumes that the methylation signature for samples within S_1 are the weighted
 729 sum of the methylation signatures from individual white blood cell types, where the
 730 weights are proportional to the cell type frequencies.

731
 732 **Figure 2: Complete blood cell (CBC) and predicted proportions of white blood cell**
 733 **types in the target methylation data set.** CBC derived proportions (i.e., $W^{(CBC)}$) of
 734 white blood cell types in (A) whole blood and (B) peripheral blood mononuclear cell
 735 (PBMCs) (i.e., devoid of granulocytes) for the samples in the target methylation data set.
 736 (C) Predicted proportions (i.e., \hat{W}) of CD8+ T-lymphocytes (CD8T), CD4+ T-
 737 lymphocytes (CD4T), Natural killer cells (NK), B cells (Bcell), Monocytes (Mono), and
 738 Granulocytes (Gran) for the target samples using constrained projection (CP). Black
 739 bars denote the median and the red dashed bars denote the 75th and 25th percentiles for
 740 the predicted cell type proportions. Colored points indicate subjects with replicate
 741 samples, where two points of the same color denote replicate samples for the same
 742 subject.

743
 744 **Figure 3: Comparison of the predicted and CBC derived proportions of monocytes**
 745 **and lymphocytes among the target samples.** Scatter-plot of the predicted and CBC-
 746 derived proportions of (A) monocytes and (B) lymphocytes. Solid red lines represent the
 747 unity lines (i.e., $y = x$). Bland-Altman plots for (C) monocyte and (D) lymphocyte
 748 proportions. Y-axes represent the difference in the predicted and CBC-derived cell type
 749 proportions and X-axes represent the mean cell type proportions based on CP prediction
 750 and CBC-based proportions. Red-dotted lines indicate the global bootstrap-based 95%
 751 prediction intervals for the difference in predicted and CBC-derived cell type proportions.
 752

753 **Figure 4: Prediction performance as a function of the number of L-DMRs used in**
 754 **CP.** (A) Pearson correlation between the predicted and CBC-derived proportions of

755 monocytes (blue line) and lymphocytes (red line) as a function of the numbers of L-
 756 DMRs used in CP. (B) root mean squared error (rMSE) for monocytes and lymphocytes
 757 and (C) median (%) granulocytes as a function of the numbers of L-DMRs used in CP.
 758 (D) Pearson correlation between the predicted and CBC-derived proportions of
 759 monocytes and lymphocytes as a function of the numbers of non L-DMRs (negative
 760 controls) used in CP.

761

762

763 **Tables:**

764

765 **Table1: Sorted white blood cell types in reference set, S_0 .**

766

Cell lineage	Cell type	Description	Sample size
Lymphoid	B cells	CD19+ B-lymphocytes	6
	NK cells	CD56+ Natural Killer (NK) cells	11
	CD4+ T cells ^{1,2}	CD3+CD4+ T-lymphocytes	8
	CD8+ T cells ^{1,3}	CD3+CD8+ T-lymphocytes	2
	NKT T cells ¹	CD3+CD56+ T-lymphocytes	1
	T cells {other} ¹	CD3+ T-lymphocytes	5
Myeloid	Granulocytes	CD15+ granulocytes	8
	Monocytes	CD14+ monocytes	5
Total	-	-	46

767

1: Considered as a member of the "pan-T-cell" group

768

2: Pan T-cell further refined as also belonging to the "CD4+" group

769

3: Pan T-cell further refined as also belonging to the "CD8+" group

770