



Deposited via The University of York.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/133612/>

Version: Accepted Version

---

**Article:**

Pellacani, Davide, Droop, Alastair P, Frame, Fiona M et al. (2018) Phenotype-independent DNA methylation changes in prostate cancer. *British journal of cancer*. GG-2018-5107R. ISSN: 1532-1827

<https://doi.org/10.1038/s41416-018-0236-1>

---

**Reuse**

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

# **Phenotype-independent DNA methylation changes in prostate cancer**

## **Running title: Prostate cancer specific DNA methylation**

Davide Pellacani<sup>1,2\*</sup>, Alastair P. Droop<sup>1,3</sup>, Fiona M. Frame<sup>1</sup>, Matthew S. Simms<sup>4</sup>, Vincent M. Mann<sup>1</sup>, Anne T. Collins<sup>1</sup>, Connie J. Eaves<sup>2</sup> and Norman J. Maitland<sup>1,5</sup>

1) Cancer Research Unit, Department of Biology, University of York, Wentworth Way, York, YO10 5DD, UK.

2) Terry Fox Laboratory, BC Cancer Agency, 675 West 10th Avenue, Vancouver, BC V5Z 1L3, Canada

3) Leeds Institute for Data Analytics, Worsley Building, Clarendon Way, Leeds LS2 9NL, UK

4) Department of Urology, Castle Hill Hospital (Hull & East Yorkshire Hospitals NHS Trust), Cottingham HU16 5JQ, UK

5) Hull York Medical School, University of York, Heslington, York YO10 5DD, UK.

### **Correspondence:**

Dr. Davide Pellacani

Terry Fox Laboratory,  
BC Cancer Research Centre  
675 West 10th Avenue  
Vancouver, BC  
V5Z 1L3, Canada

**Email:** [dpellacani@bccrc.ca](mailto:dpellacani@bccrc.ca)

**Phone:** 604-675-8120 Ext 7731

**FAX:** 604-877-0712

1 *Abstract*

2 **Background:** Human prostate cancers display numerous DNA methylation changes  
3 compared to normal tissue samples. However, definitive identification of features related  
4 to the cells' malignant status has been compromised by the predominance of cells with  
5 luminal features in prostate cancers.

6 **Methods:** We generated genome-wide DNA methylation profiles of cell subpopulations  
7 with basal or luminal features isolated from matched prostate cancer and normal tissue  
8 samples.

9 **Results:** Many frequent DNA methylation changes previously attributed to prostate  
10 cancers are here identified as differences between luminal and basal cells in both normal  
11 and cancer samples. We also identified changes unique to each of the two cancer  
12 subpopulations. Those specific to cancer luminal cells were associated with regulation of  
13 metabolic processes, cell proliferation and epithelial development. Within the prostate  
14 cancer TCGA dataset, these changes were able to distinguish not only cancers from  
15 normal samples, but also organ-confined cancers from those with extra-prostatic  
16 extensions. Using changes present in both basal and luminal cancer cells, we derived a  
17 new 17-CpG prostate cancer signature with high predictive power in the TCGA dataset.

18 **Conclusions:** This study demonstrates the importance of comparing phenotypically  
19 matched prostate cell populations from normal and cancer tissues to unmask biologically  
20 and clinically relevant DNA methylation changes.

21

22

23 ***Background***

24 Treatment-naïve prostate cancer (PCa) is characterized by an abnormal accumulation of  
25 proliferative cells with a molecular phenotype similar to the luminal cells present in the  
26 normal prostate<sup>1,2</sup>. However, PCa samples also contain a small population of tumour cells  
27 with basal features. These cells possess “cancer stem cell” features, appear to be  
28 treatment-resistant, and are proposed to serve as a reservoir for tumour recurrence after  
29 castration therapy<sup>3-6</sup>. DNA methylation of bulk PCa samples has been well studied<sup>7</sup> and  
30 aberrant methylation of promoter regions found to be a consistent feature<sup>8</sup>, albeit with  
31 high variability both between patients and within single tumours<sup>9</sup>. Their frequency and  
32 presence in pre-malignant tissues support a strong selective pressure for DNA  
33 methylation changes during cancer development<sup>7</sup>. However, DNA methylation is  
34 dynamically regulated during tissue development and cell differentiation<sup>10</sup>, and distinct  
35 cell types possess specific DNA methylation profiles within the same tissue<sup>11-13</sup>.  
36 Therefore, the luminal molecular features of bulk PCa samples, in contrast to the almost  
37 equal proportion of basal and luminal cells in normal prostate tissues, complicate the  
38 interpretation of datasets generated on whole tissue extracts, where changes associated to  
39 differences in cell types may mask the presence of malignancy-associated signatures.

40       Recent developments in tissue processing and the identification of surface  
41 markers suitable for the prospective isolation of viable basal and luminal cells from  
42 normal prostate tissues have enabled studies of their molecular and biological  
43 characteristics<sup>14-17</sup>. Use of this approach has revealed that many of the genes  
44 downregulated in normal luminal cells compared to basal cells are frequently  
45 hypermethylated in PCa<sup>18</sup>. This data implies a functional link between DNA

46 hypermethylation and the observed expansion of cells with a luminal phenotype in PCa.  
47 However, very little is known about the specific DNA methylation features of PCa cells  
48 with basal and luminal phenotypes in comparison to their normal counterparts. To  
49 address this issue, we generated genome-wide DNA methylation profiles of FACS-  
50 purified populations of cells with basal and luminal features from a series of freshly  
51 isolated patient-matched tumour and normal samples from individuals undergoing radical  
52 prostatectomy. Our results show that many DNA methylation changes frequently seen in  
53 PCa are characteristic differences between luminal and basal cells from both normal and  
54 cancer samples. From these datasets, we were also able to identify two sets of tumour-  
55 specific changes of potential clinical interest. One set consists of changes that are specific  
56 to PCa luminal cells; the other set are changes shared by both basal and luminal tumour  
57 but not normal prostate cells.

58

59

60 **Methods**

61 **Tissue processing:**

62 Prostate tissues were obtained from patients undergoing radical prostatectomy at Castle  
63 Hill Hospital (Cottingham, UK) with informed patient consent and approval from the  
64 NRES Committee Yorkshire & The Humber (LREC Number 07/H1304/121). Tissues  
65 were sampled immediately after surgery. For radical prostatectomies, three core needle  
66 biopsies were taken from four different sites (left base, left apex, right base, right apex)  
67 and were directed by previous pathology, imaging and palpation. Tissues were  
68 transported in RPMI-1640 with 5% FCS and 100U/ml antibiotic/antimitotic solution at  
69 4°C, and processed immediately upon arrival. PCa diagnosis was confirmed by  
70 histological examination of the whole prostate. Tissues were disaggregated as previously  
71 described<sup>19</sup>, and all reagents were supplemented with 10 nM R1881 to better preserve the  
72 viability of luminal cells.

73

74 **Fluorescence activated cell sorting (FACS) and characterization of cell populations:**

75 Single-cell suspensions were labelled with Lineage Cell Depletion Kit (human) and  
76 CD31 MicroBead Kit (Miltenyi Biotec) and Lin<sup>+</sup>/CD31<sup>+</sup> cells depleted twice using  
77 MACS LS Columns (Miltenyi Biotec). Lin<sup>-</sup>/CD31<sup>-</sup> cells were then labelled with EpCAM-  
78 APC, CD49f-FITC and CD24-PE (Miltenyi Biotec) and DAPI and  
79 EpCAM<sup>+</sup>/CD49f<sup>+</sup>/CD24<sup>-</sup> and EpCAM<sup>+</sup>/CD49f<sup>+</sup>/CD24<sup>+</sup> sorted at >95% purity using a  
80 MoFlo (Beckman Coulter) cell sorter. Sorted populations were characterized by  
81 immunofluorescence and qRT-PCR as previously described<sup>18</sup>.

82

83 **Reduced Representation Bisulphite Sequencing (RRBS):**

84 DNA was extracted from FACS-sorted populations using phenol/chloroform extraction  
85 and ethanol precipitation. DNA was quantified using a NanoDrop 1000  
86 Spectrophotometer (Thermo Fisher Scientific), and shipped to Zymo Research for RRBS  
87 analysis. Bisulphite conversion, library preparation, sequencing, and initial  
88 bioinformatics analyses were performed by Zymo Research following the Methyl-  
89 MiniSeq pipeline.

90

91 **Sequence data processing and methylation calls:**

92 Fastq files were trimmed using Trim Galore! v0.4.1  
93 ([http://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)) with the following  
94 parameters: --fastqc --illumina --paired --rrbs --non\_directional. Trimmed sequences were  
95 aligned to the human genome (hg19 downloaded from UCSC, 08-Mar-2009 version)  
96 using bsmmap v2.90<sup>20</sup> and the following parameters: -m 0 -x 1000 -n 1 -p 8 -S 1. The  
97 resulting bam files were sorted and indexed using samtools v0.1.19<sup>21</sup>, and methylation  
98 and coverage calls for each CpG site calculated using the methratio.py script in the bsmmap  
99 software (Supplementary Table 1). Methylation calls were then filtered for low (<3) and  
100 high (>99.95%) read coverage and merged in non-overlapping genomic bins of 100 bp  
101 using the methylKit package v0.99.2<sup>22</sup> within R v3.3.1 to increase comparability between  
102 samples. All subsequent analyses were carried out using only those genomic bins covered  
103 in all samples, with the exception of the results presented in Supplementary Fig. 2 and  
104 Supplementary Table 3 which were generated using single GpG information.  
105

106 **Identification of differentially methylated regions (DMRs):**

107 DMRs were calculated using methylKit<sup>22</sup>; with all pairwise comparisons between the four  
108 cell populations carried out and similar populations from different donors defined as  
109 biological replicates. The patient of origin was used as a categorical covariate to account  
110 for the strong inter-donor variability seen. All p-values were generated using a logistic  
111 regression model and corrected for multiple testing using the SLIM method<sup>23</sup>. DMRs  
112 were defined as those genomic bins with q-values <0.05 and absolute methylation  
113 difference >10% in each pairwise comparison.

114

115 **Characterization of DMRs:**

116 All genomic features were downloaded from the UCSC Table browser  
117 (genome.ucsc.edu) for the hg19 genome. Gene models: “refGene” (RefSeq Genes), CpG  
118 Islands: “cpgIslandExt”, Evolutionary conservation: “phastCons100way”, DNase  
119 hypersensitivity sites (DHSs): “wgEncodeRegDnaseClusteredV3”, transcription factor  
120 binding sites (TFBSs): “wgEncodeRegTfbsClusteredV3”, repeats: “rmsk”  
121 (RepeatMasker). Overlaps and distances of DMRs to other genomic features were  
122 calculated using BEDtools v2.26.0<sup>24</sup>, and significance of enrichments or depletions was  
123 calculated using custom R scripts. All p-values <10<sup>-300</sup> were approximated to 10<sup>-300</sup> to  
124 avoid reaching the minimum value for a floating-point number (2.2\*10<sup>-308</sup>). Average  
125 conservation signals around DMRs were calculated using bwtool v1.0<sup>25</sup>. P-values were  
126 calculated using a bootstrapping approach comparing the average conservation of the  
127 distal DMRs with the average of an equal number of randomly selected, non-overlapping,  
128 distal genomic bins, 1000 times. Gene ontology (GO) analysis was performed using

129 GREAT v3.0<sup>26</sup>, using all covered genomic bins as background and the default “Basal plus  
130 extension” association rules. Results were filtered to include only GO categories, with a  
131 Benjamini–Hochberg corrected (FDR) hypergeometric test p-value <0.05 and  $\geq 3$  genes  
132 with associated regions. K-means clustering of GO categories (biological processes only)  
133 was based on information similarity values calculated using the GOSim package within R  
134 v3.3.1. Promoters frequently altered in PCa were downloaded from the review by Massie  
135 et al., 2017<sup>7</sup>. Only promoters reported by  $\geq 3$  studies were considered frequently altered.  
136 Genome browser plots were generated using the package Sushi within R v3.3.1 and  
137 custom scripts.

138

#### 139 **TCGA data analysis:**

140 Illumina Infinium HumanMethylation450 data generated within the The Cancer Genome  
141 Atlas (TCGA) consortium<sup>27</sup> was downloaded (pre-processed Level 3 data only) from the  
142 NCI Genomic Data Commons website using the provided GDC Data Transfer Tool (data  
143 downloaded on 7<sup>th</sup> Dec 2016). Clinical data was downloaded from firebrowse.org (8<sup>th</sup>  
144 Dec 2016). [The presence of evident batch effects was excluded by visualizing the data on TCGA  
145 Batch Effects \(<http://bioinformatics.mdanderson.org/tcgambatch/>\)](#). A data matrix containing  
146 the beta values for each sample was generated using custom scripts. Probes were mapped  
147 to hg19 using the positions officially reported by Illumina. Overlap of array probes with  
148 DMRs was carried out using BEDtools v2.26.0. Hierarchical clustering was based on  
149 Euclidean distances of unscaled beta-values. Logistic model training using least absolute  
150 shrinkage and selection operator (LASSO) regression was performed using the glmnet  
151 package within R v3.3.1 on a random selection of 70% of the samples. 200 lambda values

152 ranging from  $e^{-7}$  to  $e^{-2}$  were tested and 10-fold cross validation performed. The lambda  
153 with the minimum mean cross-validated error was selected and resulted in 17 probes with  
154 non-zero coefficients. The optimal model was then tested on the remaining 30% of  
155 samples and receiver operator curve and area under the curve (AUC) calculated using the  
156 ROCR package.

157

158

159 **Results**

160 **Phenotypically defined prostate cells from patient-matched normal and PCa**

161 **samples show donor-specific DNA methylation profiles**

162 Matched tumour-directed (cancer) and contralateral (normal) core needle biopsies (1 or 2  
163 per site) were obtained from 4 treatment-naïve prostate cancer patients undergoing radical  
164 prostatectomies. These samples were then enzymatically dissociated and labeled with  
165 antibodies against EpCAM, CD49f and CD24 to enable the prospective isolation of  
166 luminal (EpCAM+CD49f-CD24+) and basal (EpCAM+CD49f+CD24-) cells at >95%  
167 purity (Fig. 1A). EpCAM+CD49f+CD24- cells expressed higher levels of molecular  
168 markers associated with basal cells and lower levels of luminal markers compared to  
169 EpCAM+CD49f-CD24+ cells from the same biopsy, both at the mRNA and protein level  
170 (Supplementary Fig. 1A-B). For convenience, we named the paired subsets as follows:  
171 Cancer Luminal (CL) EpCAM+CD49f-CD24+ cells purified from tumour-directed  
172 biopsies; Cancer Basal (CB) EpCAM+CD49f+CD24- cells purified from tumour-directed  
173 biopsies; Normal Luminal (NL) EpCAM+CD49f-CD24+ cells from contralateral  
174 biopsies; Normal Basal (NB) EpCAM+CD49f+CD24- cells purified from contralateral  
175 biopsies. This yielded 4 CL and CB populations, and 3 matched NL and NB populations,  
176 as in one prostate the palpable tumour was extended to most of the prostate and it was not  
177 possible to obtain a contralateral “normal” tissue biopsy (Supplementary Fig. 1C). DNA  
178 obtained from each of these isolates was then subjected to Reduced Representation  
179 Bisulphite Sequencing (RRBS). On average, this generated information on the DNA  
180 methylation status of  $>8.9 \times 10^6$  cytosines within CpG sites per sample (range  $8 \times 10^6$  –  
181  $9.6 \times 10^6$ , with an average coverage of 7.5 reads, Supplementary Table 1). The data was

182 processed as described in Methods, and binned into 100 bp genomic regions to maximize  
183 the comparability between samples (932,905 bins covering  $4.1 \times 10^6$  CpGs in all samples).  
184 Unsupervised hierarchical clustering of the top 1% most variable regions (bins) across all  
185 samples showed clustering primarily according to the patient of origin, rather than the  
186 subset analyzed (Fig. 1B). This indicates a high donor-determined variation in CpG  
187 methylation, consistent with previous reports of similarly accrued data<sup>28</sup>.

188

### 189 **Distinct DNA methylation profiles in basal and luminal cells**

190 We then calculated DMRs for all pairwise comparisons between the 4 sorted populations  
191 (Fig. 1C, Supplementary Table 2). Among these, the comparison between CB and NB  
192 cells (CB-NB comparison) produced the smallest number of DMRs. In contrast, a large  
193 number of DMRs were seen when either normal or cancer luminal cells were compared  
194 with either source of basal cells (i.e., NL-NB, NL-CB, CL-NB and CL-CB, Fig. 1D). Of  
195 the DMRs revealed in these latter comparisons,  $\sim 2/3$  were hypermethylated in luminal  
196 cells, which correlates with the higher levels of DNMT3a seen in these cells<sup>18</sup>. We also  
197 calculated differential methylation on single CpGs (prior the 100bp binning) with very  
198 similar results (Supplementary Fig. 2 and Supplementary Table 3). Moreover, integration  
199 of the DMRs identified in NL-NB proximal ( $\pm 5$  kb) to annotated transcriptional start sites  
200 (TSSs) with RNA-seq data of similarly purified cells<sup>15</sup> showed the expected inverse  
201 correlation (Supplementary Fig. 3A).

202 We also found an extensive overlap in the DMRs obtained from both the NL-NB and NL-  
203 CB comparisons, and also from the CL-NB and CL-CB comparisons (Supplementary Fig.

204 3B-C). Accordingly, we focussed our subsequent analyses on comparisons of NL-NB and  
205 CL-CB, where cells from the same biopsy could be compared directly.

206 Characterization of the genomic features of the DMRs thus identified showed that  
207 >50% of them fell outside of CpG islands, shores or shelves (Fig. 1E), and >70% were >5  
208 kb away from any annotated TSSs (Fig. 1F-G). These features were particularly  
209 pronounced (highly significant hypergeometric test) for the hypomethylated DMRs  
210 identified in the comparisons of NL-NB, CL-CB and CL-NL. Because hypermethylated  
211 and hypomethylated DMRs might be anticipated to differ in their genomic context, their  
212 impact on the biological properties of basal and luminal cells could also be different.

213

#### 214 **Distal hypermethylated DMRs are enriched in enhancer features**

215 Given that most of the DMRs identified were outside CpG islands and far from TSSs, we  
216 asked whether they might affect distal regulatory elements (enhancers). We therefore  
217 examined three genomic characteristics of such elements: evolutionary conservation<sup>29</sup>,  
218 open chromatin shown by hypersensitivity to DNase I<sup>30</sup>, and presence of TFBSs<sup>31</sup>. Distal  
219 hypermethylated DMRs in each comparison were enriched for evolutionarily conserved  
220 sequences (Fig. 2A, bootstrapped p-value) and overlapped significantly with both DHSs  
221 and ChIP-seq-defined TFBSs (identified within the ENCODE project, Fig. 2B-C,  
222 hypergeometric test). Distal hypomethylated DMRs generally scored lower than the  
223 hypermethylated counterparts for each metric measured. DMRs hypomethylated in the  
224 CL-CB and CL-NL comparisons showed the weakest enrichments. However, all distal  
225 hypomethylated DMRs had high overlaps with genomic repetitive elements (Fig. 2D).

226 Specifically, LINE and LTR elements, but not SINE elements, were significantly  
227 enriched in the distal CL hypomethylated regions.

228 GO enrichment analysis (Fig. 2E, Supplementary Fig. 4) showed that  
229 hypermethylated DMRs in NL-NB were enriched for more than 500 terms, many of  
230 which were linked to prostate development or epithelial stem cell regulation; while  
231 hypomethylated DMRs in the same comparison were enriched for terms related to  
232 androgen receptor signalling and response to cytokines. In the CL-CB comparison,  
233 hypermethylated DMRs were also enriched for more than 500 terms, 311 of which were  
234 also identified in the NL-NB comparison, suggesting a high functional overlap in  
235 hypermethylated regions in luminal cells from both normal and cancer samples. In the  
236 CL-NL comparison, hypermethylated DMRs were enriched in terms related to cell  
237 adhesion, while hypomethylated DMRs were enriched in terms related to epithelial  
238 morphogenesis. These results indicate that several pathways fundamental to the  
239 establishment and maintenance of the normal prostate epithelium are altered in cancer  
240 cells with a luminal phenotype.

241

#### 242 **Phenotype-specific DMRs are shared in normal and cancerous prostate tissues**

243 As suggested by the enriched GO analyses, we found a 28% overlap in all the DMRs  
244 identified from the NL-NB and the CL-CB comparisons (3852/13816, Fisher's exact test  
245  $p$ -value  $< 10^{-300}$ , Fig. 3A). Hierarchical clustering of all samples based on both sets of  
246 DMRs separated them by phenotype (Fig. 3B), reinforcing the presence of a strong  
247 phenotypic signature independent of disease state. These shared DMRs were enriched in  
248 features characteristic of enhancers (Supplementary Fig. 5A-D) and linked to GO terms

249 related to prostate development, regulation of epithelial stem cells and androgen receptor  
250 signalling (Supplementary Fig. 5E-F). Moreover, hypermethylated DMRs were highly  
251 enriched for TFBSs of *TP63*, *TP53* and *NF1*, and hypomethylated DMRs for *FOXA1*,  
252 *p65-NFkB* and *GATA3* (Fig. 3C), all well-known regulators of basal and luminal  
253 epithelial cells, respectively. Interestingly, 26 of the 168 genes described as frequently  
254 differentially methylated in PCa<sup>7</sup>, showed hyper- or hypomethylated DMRs within 5 kb  
255 of their TSSs in both the NL-NB and CL-CB comparisons (Fig. 3D). These included the  
256 frequently hypermethylated genes, *GSTP1* and *CCDC8* (Fig. 3E-F).

257 In summary, these analyses identified a large set of phenotype-specific and  
258 disease-independent DMRs, both of which contained many binding sites for TFs with  
259 known regulatory roles in the normal prostate.

260

### 261 **CL hypermethylate PRC2 target sites and hypomethylate repetitive elements**

262 A second group of genes frequently hypermethylated in PCa were found hypermethylated  
263 in both the CL-CB and CL-NL comparisons (Fig. 4a), but not in the NL-NB comparison.  
264 These might be expected to reflect a PCa-specific methylation signature. DMRs  
265 identified in the CL-CB and CL-NL comparisons showed that many were shared (1472  
266 DMRs, Fisher's exact test p-value < 10<sup>-300</sup>, Fig. 4B) with very few also different between  
267 NL and NB cells (106 DMRs). 65% of these CL-specific hypermethylated DMRs were  
268 distal to TSSs and were again highly enriched for enhancer features, but significantly  
269 depleted in repetitive elements (Supplementary Fig. 6A-E). These regions were  
270 associated with GO terms related to metabolic processes, cell proliferation and epithelial  
271 development (Fig. 4C) and showed a high enrichment of DNA sequences potentially

272 bound by EZH2 and SUZ12, two main members of the PRC2 complex (Supplementary  
273 Fig. 6F). On the other hand, distal hypomethylated DMRs were not enriched for any  
274 feature of putative regulatory regions, but significantly overlapped with LINE and LTR  
275 elements.

276 Since the CL subset represents the majority of the cells in untreated PCa samples,  
277 we hypothesized that aberrant methylation of these DMRs would be measurable even  
278 when whole tissue homogenates are analysed. We therefore interrogated the DNA  
279 methylation array dataset for PCa made available by the TCGA consortium, which  
280 consists of 50 PCa samples with matched normal counterparts, 452 additional PCa  
281 samples without normal counterparts, and 1 metastatic PCa sample<sup>27</sup>. 255 array probes  
282 overlap these 1472 DMRs. Hierarchical clustering of the 50 matched normal and PCa  
283 samples showed an almost perfect subdivision based on the malignancy status of the  
284 samples (TPR = 0.92, TNR = 0.92, Chi-squared test p-value =  $2.4 \times 10^{-16}$ , Fig. 4D). The  
285 same analysis carried out on all 553 samples produced similar results, with one cluster  
286 highly enriched in normal samples (Chi-squared test p-value =  $1.7 \times 10^{-39}$ , Supplementary  
287 Fig. 6G). This clustering also appeared to divide the PCa samples into two main groups,  
288 according to their differences from the normal samples. Exclusive analysis of the cancer  
289 samples confirmed this clustering pattern (Fig. 4E) and showed one cluster to be  
290 significantly enriched for samples with extra-prostatic extensions (pT3 or pT4 in TNM  
291 classification, Chi-squared test p-value < 0.005) in the absence of significant differences  
292 in Gleason score (Chi-squared test p-value > 0.1).

293 Overall, these results indicate that phenotypic luminal PCa cells possess an  
294 aberrant methylation signature characterized by hypermethylation of putative regulatory

295 sequences involved in tissue development, and hypomethylation of LINES and LTRs  
296 repetitive elements. This signature was also able to distinguish cancer samples from  
297 normal, and organ-confined from extraprostatic disease.

298

### 299 **Identification of PCa-specific, phenotype-independent DMRs**

300 Comparisons of the DMRs in the CL-NL and CB-NB pairs showed a small but  
301 significant overlap of both hyper- and hypomethylated DMRs in each (189 DMRs in  
302 total, Fig. 5A). These common DMRs were able to cluster all samples according to their  
303 disease state in a phenotype-independent manner (Supplementary Fig. 7A). Notably, they  
304 included DMRs close to many genes previously implicated in prostate cancer (e.g.,  
305 *NEAT1*, *MTOR*, *RHCG*, *KCNC2*, *WT1*, *HOXC12*, *KMT2B*, Fig. 5B). To determine  
306 whether these DMRs would be altered in an independent dataset, we applied the same  
307 analysis to the TCGA dataset, where 66 array probes overlapped these 189 DMRs.  
308 Hierarchical clustering of the 50 matched normal and PCa samples produced a single  
309 cluster containing 46/50 normal samples and 10/50 PCa samples (TPR = 0.8, TNR =  
310 0.92, Chi-squared test p-value =  $1.8 \times 10^{-12}$ , Fig. 5C). Application of the same analysis to  
311 all samples in the TCGA database produced similar results: one cluster was highly  
312 enriched in normal samples (TPR = 0.87, TNR = 0.74, Chi-squared test p-value =  $8.3 \times 10^{-26}$ ,  
313 Supplementary Fig. 7B), indicating that at least some of these DMRs are frequently  
314 altered in PCa.

315 To select the probes most strongly associated with disease state (i.e., PCa vs  
316 normal), we trained a logistic model using LASSO regression on 70% of the TCGA  
317 samples and selected a 17-probe signature (Fig. 5D). We then tested this model on the

318 remaining 30% of the dataset. This resulted in an AUC of 0.92 (TPR = 0.9, TNR = 0.94,  
319 Fisher's exact test p-value =  $2.82 \times 10^{-12}$  at the selected cut-off of 0.8, Fig. 5E-F,  
320 Supplementary Table 4). The 17-probe signature also included sequences proximal to  
321 several genes with recognized importance in PCa (e.g., *PLAGL1/HYMAI*, *HOXC12*,  
322 *KCNC2*), but was completely non-overlapping with other similar signatures recently  
323 developed for PCa<sup>32-36</sup>.

324

### 325 ***Discussion***

326 PCa is characterized by frequent aberrant DNA methylation of many genomic sites that  
327 may contain clinically relevant signatures<sup>7,37</sup>. The early establishment (presence in pre-  
328 neoplastic tissues) and high prevalence of these aberrant patterns is also suggestive of  
329 their direct involvement in PCA tumorigenesis. However, the normal prostate epithelium  
330 is composed of similar numbers of luminal and basal cells, whereas most treatment-naïve  
331 prostate cancers are largely composed of cells with many luminal features. This shift in  
332 favor of a transcriptional and epigenomic program of normal luminal cells might mask or  
333 complicate the identification of cancer-specific features in prostate cancer when bulk  
334 analyses are performed on this type of tumour.

335 We now report a detailed comparison of genome wide methylation profiles  
336 obtained separately from epithelial cells with luminal and basal phenotypes, isolated with  
337 a high purity from patient-matched normal and cancer biopsy samples. From comparative  
338 analyses of these profiles, we found a major proportion of the methylation differences  
339 between normal basal and luminal cells were conserved in their malignant counterparts.  
340 These affected many promoters frequently described as aberrantly methylated in bulk

341 PCa compared to normal tissues, consistent with the increased representation of cells with  
342 a luminal phenotype in PCa, in which a higher proportion of cells carrying a methylation  
343 signature of normal luminal cells might then be expected.

344         However, our study made it possible to identify, for the first time, regions  
345 specifically altered in the luminal fraction of PCa. The hypermethylated DMRs in this  
346 group were genes associated to genes involved in metabolic processes, cell proliferation  
347 and epithelial development, all functions clearly deregulated in prostate cancer, therefore  
348 potentially containing major cancer driver events. Furthermore, hypomethylated DMRs  
349 were highly enriched in repetitive elements, a feature also previously reported in many  
350 cancer types, where they have been thought to contribute to genomic instability and  
351 aberrant gene expression<sup>38-40</sup>.

352         Importantly, this set of DMRs was able to discriminate not only normal and PCa  
353 samples in the TCGA dataset, but also PCa samples with or without extra-prostatic  
354 extensions, the former being indicative of highly aggressive, invasive cancers. Since this  
355 distinction was not evident from the Gleason grades of these tumours, the epigenetic data  
356 may reflect a an acquisition of specific aberrant epigenomic changes that herald disease  
357 progression<sup>7,41-43</sup>. Genomic regions consistently altered in both tumour phenotypes in the  
358 PCa samples analyzed also have potential clinical importance. Indeed, the new logistic  
359 model constructed from these regions makes use of only 17 probes to distinguish normal  
360 and PCa samples with similar specificity and sensitivity to previously developed, non-  
361 overlapping models<sup>35,36</sup>, and may be useful in the context of the low mutagenic burdens  
362 seen in most hormone-naïve prostate cancers.

363 The results reported here show that many DNA methylation changes commonly  
364 associated with PCa cells are explained by a predominant luminal phenotype of the  
365 treatment-naïve PCa population, and are not cancer-specific nor are likely to contain  
366 driver events. Importantly however, we were able to identify two separate classes of PCa-  
367 specific DNA methylation changes: those specific to cancer luminal cells that can  
368 distinguish both normal from cancer samples and organ-confined cancers from those with  
369 extra-prostatic extensions; those changes common to basal and luminal cancer cells that  
370 are able to distinguish PCa efficiently from normal samples. These two novel sets of  
371 cancer-specific changes clearly demonstrate the potential of profiling normal and cancer  
372 cell subpopulations in identifying signatures that may contain previously unrecognized  
373 driver events in the development and progression of PCa.

374

#### 375 *Additional Information*

##### 376 **Ethics approval and consent to participate**

377 Prostate tissues were obtained from patients undergoing radical prostatectomy at Castle  
378 Hill Hospital (Cottingham, UK) with informed patient consent and approval from the  
379 NRES Committee Yorkshire & The Humber (LREC Number 07/H1304/121).

380

##### 381 **Availability of data and materials**

382 The methylation and coverage calls for all RRBS libraries generated are available from  
383 GEO [GSE107596]. For patients' privacy reasons, raw data (FASTQ and BAM files) for  
384 the RRBS libraries are not publicly available, but can be available from the  
385 corresponding author on request.

386

387 **Competing Interests**

388 The authors declare no competing financial and non-financial interests.

389

390 **Funding**

391 This work was supported by The Freemasons' Grand Charity (DP and NJM), Yorkshire  
392 Cancer Research program grant Y257PA (DP, NJM, FMF, APD, and ATC), and British  
393 Columbia Cancer Agency (Strategic Priorities Fund, DP and CJE).

394

395 **Authors' Contributions**

396 DP and NJM designed the project. MSS and VMM procured the tissue samples. DP  
397 processed and sorted the tissue samples, and performed all other experiments. DP, FMF  
398 and ATC developed the tissue processing and sorting protocol. DP and APD conducted  
399 all bioinformatic analyses. DP, CJE and NJM wrote the manuscript. All authors  
400 contributed to the interpretation of the results and read and approved the manuscript.

401

402 **Acknowledgements**

403 We thank the urology surgeons and patients from Castle Hill Hospital for kind donations  
404 of clinical prostate samples. We thank Artem Babaian, Rod Docking, Dr. Kieran O Neill,  
405 Hye-Jung E. Chun, Dr. Misha Bilenky, Dr. Alireza Heravi-Moussavi and Dr. Martin  
406 Hirst for the useful discussions regarding the bioinformatic analyses conducted in this  
407 study.

408

409

**References:**

- 410 1. Ananthanarayanan V, Deaton RJ, Yang XJ, Pins MR, Gann PH. Alteration of  
411 proliferation and apoptotic markers in normal and premalignant tissue associated  
412 with prostate cancer. *BMC Cancer*. 2006;6:73.
- 413 2. De Marzo AM, Meeker AK, Epstein JI, Coffey DS. Prostate stem cell  
414 compartments: expression of the cell cycle inhibitor p27Kip1 in normal,  
415 hyperplastic, and neoplastic cells. *Am J Pathol*. 1998 Sep;153(3):911–9.
- 416 3. Polson E, Lewis JL, Celik H, Mann VM, Stower MJ, Simms MS, et al.  
417 Monoallelic expression of TMPRSS2/ERG in prostate cancer stem cells. *Nat*  
418 *Commun*. 2013 Mar 27;4:1623.
- 419 4. Frame FM, Pellacani D, Collins AT, Simms MS, Mann VM, Jones G, et al. HDAC  
420 inhibitor confers radiosensitivity to prostate stem-like cells. *Br J Cancer*. 2013 Dec  
421 10;109(12):3023–33.
- 422 5. Birnie R, Bryce SD, Roome C, Dussupt V, Droop A, Lang SH, et al. Gene  
423 expression profiling of human prostate cancer stem cells reveals a pro-  
424 inflammatory phenotype and the importance of extracellular matrix interactions.  
425 *Genome Biol*. 2008;9(5):R83.
- 426 6. Collins AT, Berry PA, Hyde C, Stower MJ, Maitland NJ. Prospective  
427 identification of tumorigenic prostate cancer stem cells. *Cancer Res*. 2005 Dec  
428 1;65(23):10946–51.
- 429 7. Massie CE, Mills IG, Lynch AG. The importance of DNA methylation in prostate  
430 cancer development. *J Steroid Biochem Mol Biol*. 2017 Feb;166:1–15.
- 431 8. Goering W, Kloth M, Schulz WA. DNA methylation changes in prostate cancer.  
432 *Methods Mol Biol*. 2012;863:47–66.
- 433 9. Aryee MJ, Liu W, Engelmann JC, Nuhn P, Gurel M, Haffner MC, et al. DNA  
434 methylation alterations exhibit intraindividual stability and interindividual  
435 heterogeneity in prostate cancer metastases. *Sci Transl Med*. 2013 Jan  
436 23;5(169):169ra10.
- 437 10. Gifford CA, Ziller MJ, Gu H, Trapnell C, Donaghey J, Tsankov A, et al.  
438 Transcriptional and Epigenetic Dynamics during Specification of Human  
439 Embryonic Stem Cells. 2013 May 23;153(5):1149–63.
- 440 11. Farlik M, Halbritter F, Müller F, Choudry FA, Ebert P, Klughammer J, et al. DNA  
441 Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. *Cell*  
442 *Stem Cell*. 2016 Dec 1;19(6):808–22.
- 443 12. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky  
444 M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. *Nature*.  
445 2015 Feb 19;518(7539):317–30.

- 446 13. Pellacani D, Bilenky M, Kannan N, Heravi-Moussavi A, Knapp DJHF, Gakkhar S,  
447 et al. Analysis of Normal Human Mammary Epigenomes Reveals Cell-Specific  
448 Active Enhancer States and Associated Transcription Factor Networks. *Cell Rep.*  
449 2016 Nov 15;17(8):2060–74.
- 450 14. Goldstein AS, Huang J, Guo C, Garraway IP, Witte ON. Identification of a cell of  
451 origin for human prostate cancer. *Science.* 2010 Jul 30;329(5991):568–71.
- 452 15. Zhang D, Park D, Zhong Y, Lu Y, Rycak K, Gong S, et al. Stem cell and  
453 neurogenic gene-expression profiles link prostate basal cells to aggressive prostate  
454 cancer. *Nat Commun.* 2016 Feb 29;7:10798.
- 455 16. Drost J, Karthaus WR, Gao D, Driehuis E, Sawyers CL, Chen Y, et al. Organoid  
456 culture systems for prostate epithelial and cancer tissue. *Nat Protoc.* 2016  
457 Feb;11(2):347–58.
- 458 17. Karthaus WR, Iaquina PJ, Drost J, Gracanin A, van Boxtel R, Wongvipat J, et al.  
459 Identification of multipotent luminal progenitor cells in human prostate organoid  
460 cultures. 2014 Sep 25;159(1):163–75.
- 461 18. Pellacani D, Kestoras D, Droop A, Frame FM, Berry PA, Lawrence MG, et al.  
462 DNA hypermethylation in prostate cancer is a consequence of aberrant epithelial  
463 differentiation and hyperproliferation. *Cell Death Differ.* 2014 May;21(5):761–73.
- 464 19. Frame FM, Pellacani D, Collins AT, Maitland NJ. Harvesting Human Prostate  
465 Tissue Material and Culturing Primary Prostate Epithelial Cells. *Methods Mol*  
466 *Biol.* 2016;1443:181–201.
- 467 20. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPPING program. *BMC*  
468 *Bioinformatics.* 2009;10:232.
- 469 21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence  
470 Alignment/Map format and SAMtools. *Bioinformatics.* 2009 Aug 15;25(16):2078–  
471 9.
- 472 22. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A,  
473 et al. methylKit: a comprehensive R package for the analysis of genome-wide  
474 DNA methylation profiles. *Genome Biol.* 2012 Oct 3;13(10):R87.
- 475 23. Wang H-Q, Tuominen LK, Tsai C-J. SLIM: a sliding linear model for estimating  
476 the proportion of true null hypotheses in datasets with dependence structures.  
477 *Bioinformatics.* 2011 Jan 15;27(2):225–31.
- 478 24. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing  
479 genomic features. *Bioinformatics.* 2010 Mar 15;26(6):841–2.
- 480 25. Pohl A, Beato M. bwtool: a tool for bigWig files. *Bioinformatics.* 2014 Jun  
481 1;30(11):1618–9.

- 482 26. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT  
483 improves functional interpretation of cis-regulatory regions. *Nat Biotechnol.* 2010  
484 May;28(5):495–501.
- 485 27. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary  
486 Prostate Cancer. 2015 Nov 5;163(4):1011–25.
- 487 28. Yu YP, Ding Y, Chen R, Liao SG, Ren B-G, Michalopoulos A, et al. Whole-  
488 genome methylation sequencing reveals distinct impact of differential  
489 methylations on gene transcription in prostate cancer. *Am J Pathol.* 2013  
490 Dec;183(6):1960–70.
- 491 29. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, et al.  
492 Histone modifications at human enhancers reflect global cell-type-specific gene  
493 expression. *Nature.* 2009 May 7;459(7243):108–12.
- 494 30. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The  
495 accessible chromatin landscape of the human genome. *Nature.* 2012 Aug  
496 29;489(7414):75–82.
- 497 31. Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell  
498 type-specific enhancers. *Nat Rev Mol Cell Biol.* 2015 Mar;16(3):144–54.
- 499 32. Zhao S, Geybels MS, Leonardson A, Rubicz R, Kolb S, Yan Q, et al. Epigenome-  
500 Wide Tumor DNA Methylation Profiling Identifies Novel Prognostic Biomarkers  
501 of Metastatic-Lethal Progression in Men Diagnosed with Clinically Localized  
502 Prostate Cancer. *Clin Cancer Res. American Association for Cancer Research;*  
503 2017 Jan 1;23(1):311–9.
- 504 33. Geybels MS, Zhao S, Wong C-J, Bibikova M, Klotzle B, Wu M, et al. Epigenomic  
505 profiling of DNA methylation in paired prostate cancer versus adjacent benign  
506 tissue. *Prostate.* 2015 Dec;75(16):1941–50.
- 507 34. Geybels MS, Wright JL, Bibikova M, Klotzle B, Fan J-B, Zhao S, et al. Epigenetic  
508 signature of Gleason score and prostate cancer recurrence after radical  
509 prostatectomy. *Clin Epigenetics. BioMed Central;* 2016;8(1):97.
- 510 35. Mundbjerg K, Chopra S, Alemozaffar M, Duymich C, Lakshminarasimhan R,  
511 Nichols PW, et al. Identifying aggressive prostate cancer foci using a DNA  
512 methylation classifier. *Genome Biol. BioMed Central;* 2017 Jan 12;18(1):3.
- 513 36. Tang Y, Jiang S, Gu Y, Li W, Mo Z, Huang Y, et al. Promoter DNA methylation  
514 analysis reveals a combined diagnosis of CpG-based biomarker for prostate cancer.  
515 *Oncotarget. Impact Journals;* 2017 Aug 29;8(35):58199–209.
- 516 37. Strand SH, Ørntoft TF, Sørensen KD. Prognostic DNA methylation markers for  
517 prostate cancer. *Int J Mol Sci. Multidisciplinary Digital Publishing Institute;* 2014  
518 Sep 18;15(9):16544–76.

- 519 38. Chen RZ, Pettersson U, Beard C, Jackson-Grusby L, Jaenisch R. DNA  
520 hypomethylation leads to elevated mutation rates. *Nature*. 1998 Sep  
521 3;395(6697):89–93.
- 522 39. Eden A, Gaudet F, Waghmare A, Jaenisch R. Chromosomal instability and tumors  
523 promoted by DNA hypomethylation. *Science*. American Association for the  
524 Advancement of Science; 2003 Apr 18;300(5618):455–5.
- 525 40. Babaian A, Mager DL. Endogenous retroviral promoter exaptation in human  
526 cancer. *Mob DNA*. BioMed Central; 2016;7(1):24.
- 527 41. Ellinger J, Kahl P, Gathen Von Der J, Rogenhofer S, Heukamp LC, Gütgemann I,  
528 et al. Global levels of histone modifications predict prostate cancer recurrence.  
529 *Prostate*. 2010 Jan 1;70(1):61–9.
- 530 42. Angulo JC, Andrés G, Ashour N, Sánchez-Chapado M, López JI, Roper S.  
531 Development of Castration Resistant Prostate Cancer can be Predicted by a DNA  
532 Hypermethylation Profile. *J Urol*. 2016 Mar;195(3):619–26.
- 533 43. McDonald OG, Li X, Saunders T, Tryggvadottir R, Mentch SJ, Warmoes MO, et  
534 al. Epigenomic reprogramming during pancreatic cancer progression links anabolic  
535 glucose metabolism to distant metastasis. *Nat Genet*. 2017 Mar;49(3):367–76.

536

537 **Figure Legends**

538 **Fig. 1: Identification of DMRs between prostate cancer cell populations. (A)**

539 Representative FACS profiles of a cell suspension prepared from core needle biopsies of  
540 a radical prostatectomy sample. **(B)** Heatmap showing scaled methylation values of the  
541 top 1% most variable regions (100 bp bins) in the samples analyzed. Hierarchical  
542 clustering is based on Euclidean distance of the unscaled values and complete linkage.  
543 **(C)** Diagram showing all pairwise comparisons carried out. **(D)** Number of DMRs found  
544 in each comparison. **(E)** Overlap of DMRs with CpG islands, shores (2 kb flanking  
545 islands) or shelves (2 kb flanking shores). P-values from hypergeometric test against all  
546 regions. E = enriched, D = depleted. **(F)** Distribution of distances of DMRs to the closest  
547 TSS. Grey box indicates  $\pm 5$  kb from a TSS. Purple lines: hypermethylated DMRs, orange

548 lines: hypomethylated DMRs, gray line: all regions. **(G)** Proportion of DMRs proximal or  
549 distal to TSSs. P-values from hypergeometric test against all regions. E = enriched, D =  
550 depleted.

551

552 **Fig. 2: Hypermethylated distal DMRs have features of enhancers.** **(A)** Average plots  
553 of evolutionary conservation scores of the distal DMRs in each set. Purple lines:  
554 hypermethylated DMRs; orange lines: hypomethylated DMRs, gray line: all regions. P-  
555 values are from bootstrapping analysis. **(B)** Proportion of distal DMRs overlapping with  
556 DHSs (identified by ENCODE). P-values from hypergeometric test against all regions. E  
557 = enriched, D = depleted. **(C)** Overlap of distal DMRs with ChIP-seq derived TFBSs  
558 (identified by ENCODE). P-values are from hypergeometric tests against all regions. E =  
559 enriched, D = depleted. **(D)** Overlap of each set of distal DMRs with repetitive elements  
560 (UCSC repeatMask), SINEs, LINEs and LTRs. P-values from hypergeometric tests  
561 against all regions. E = enriched, D = depleted. **(E)** Number of GO terms enriched by  
562 each set of DMRs. GO terms identified using GREAT (FDR<0.05 and at least 3 genes in  
563 the set).

564

565 **Fig. 3: Shared phenotype-specific DMRs.** **(A)** Overlap between the DMRs identified in  
566 the NL-NB and CL-CB comparisons. P-values derived from Fisher's exact test. **(B)**  
567 Heatmap showing scaled methylation values of the DMRs identified in the NL-NB (left)  
568 or CL-CB (right) comparisons. Hierarchical clustering is based on Euclidean distances of  
569 the unscaled values and complete linkage. **(C)** TFBSs enriched in the hypermethylated  
570 (purple) or hypomethylated (orange) DMRs common between the NL-NB and CL-CB

571 comparisons. Left panel: analysis performed using HOMER findMotifs, p-values from  
572 binomial test. Right panel: enrichment of ENCODE defined TFBSs, p-values from  
573 hypergeometric test against all regions. **(D)** Frequently hyper- or hypomethylated genes  
574 in PCa<sup>7</sup> that were also hypermethylated (purple) or hypomethylated (orange) in the NL-  
575 NB and CL-CB comparisons. **(E-F)** Genome browser plots of the promoter regions of  
576 GSTP1 **(E)** and CCDC8 **(F)**. Grey squares are the bins analyzed. Lines and shaded areas  
577 represent mean  $\pm$ SEM of each category (NB=light blue, NL=light red, CB=dark blue,  
578 CL=dark red). DMRs are shown on top: hypermethylated=purple,  
579 hypomethylated=orange.

580

581 **Fig. 4: Aberrant methylation in CL.** (A) Frequently hyper- or hypomethylated genes in  
582 PCa<sup>7</sup> that are also hypermethylated (purple) or hypomethylated (orange) in the CL-CB  
583 and CL-NL comparisons. (B) Overlap between the DMRs identified in the CL-CB and  
584 CL-NL comparisons. P-values derived from Fisher's exact test. (C) Clustering of the  
585 gene ontologies (biological process) enriched in DMRs common between the CL-CB and  
586 CL-NL comparisons based on information similarity. Each circle shows an individual GO  
587 term enriched in regions hypermethylated (purple), hypomethylated (orange) or both  
588 (green), the size of the circles is proportional to the enrichment p-value. The 2 main  
589 clusters of GO terms determined by k-means are highlighted (light blue and pink), and  
590 named after the most frequent terms. (D) Heatmap showing scaled methylation values ( $\beta$ -  
591 values) of probes overlapping the DMRs common to the CL-CB and CL-NL comparisons  
592 in the PCa samples (magenta) and matched normal samples (green) within the TCGA  
593 dataset. Hierarchical clustering based on Euclidean distances of the unscaled values and

594 complete linkage. The dark green and gray clusters were generated by cutting the tree at  
595 the first bifurcation. (E) Heatmap showing scaled methylation values ( $\beta$ -values) of probes  
596 overlapping the DMRs common to the CL-CB and CL-NL comparisons in the PCa  
597 samples (matched normal samples not included) of the TCGA dataset. Hierarchical  
598 clustering based on Euclidean distance of the unscaled values and complete linkage. The  
599 dark green and gray clusters are generated by cutting the tree at the first bifurcation.

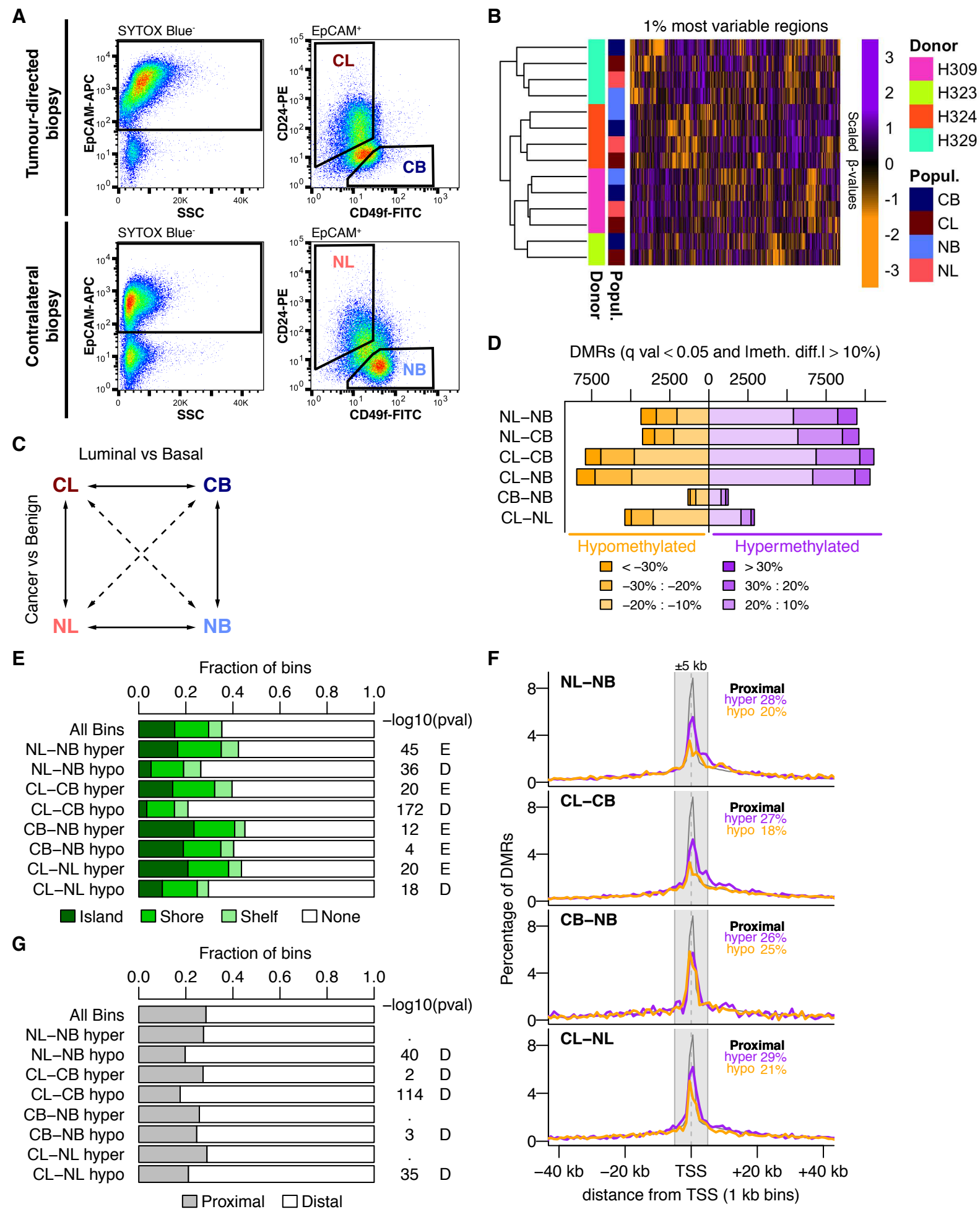
600

601 **Fig. 5: PCa-specific DMRs shared between CB and CL.** (A) Overlap between the  
602 DMRs identified in the CL-NL and CB-NB comparisons. P-values derived from Fisher's  
603 exact test. (B) Genome browser views of KCNC2 promoter (top) and RHCG exon 2  
604 (bottom). Grey squares are the bins analyzed. Lines and shaded areas represent mean  
605  $\pm$ SEM of each category (NB=light blue, NL=light red, CB=dark blue, CL=dark red).  
606 DMRs are shown on top: hypermethylated=purple, hypomethylated=orange. (C)  
607 Heatmap showing scaled methylation values of probes overlapping the DMRs common  
608 between CL-CB and CB-NB in the matched normal and cancer samples within the TCGA  
609 dataset. Hierarchical clustering based on Euclidean distances of the unscaled values and  
610 complete linkage. The dark green and gray clusters were generated by cutting the tree at  
611 the first 2 bifurcations. (D) Selection of a 17-probe signature distinguishing normal and  
612 PCa samples applying LASSO regression on a logistic model of the training dataset (70%  
613 of the TCGA samples). Lines show the changes in coefficients in relation to different  
614 lambdas. The vertical dashed line shows the optimal lambda identified using cross-  
615 validation. (E) Receiver-operating characteristic curve generated by applying the optimal  
616 logistic model to the test dataset (30% of the TCGA samples). (F) Heatmap showing

617 scaled methylation values of the 17-probe signature in the test dataset (30% of the TCGA  
618 samples). The bar plot on the left side shows the final coefficients for each probe in the  
619 model, and the bar plot on top shows the logistic probability generated by for each  
620 sample (Green: normal samples, magenta: cancer samples).

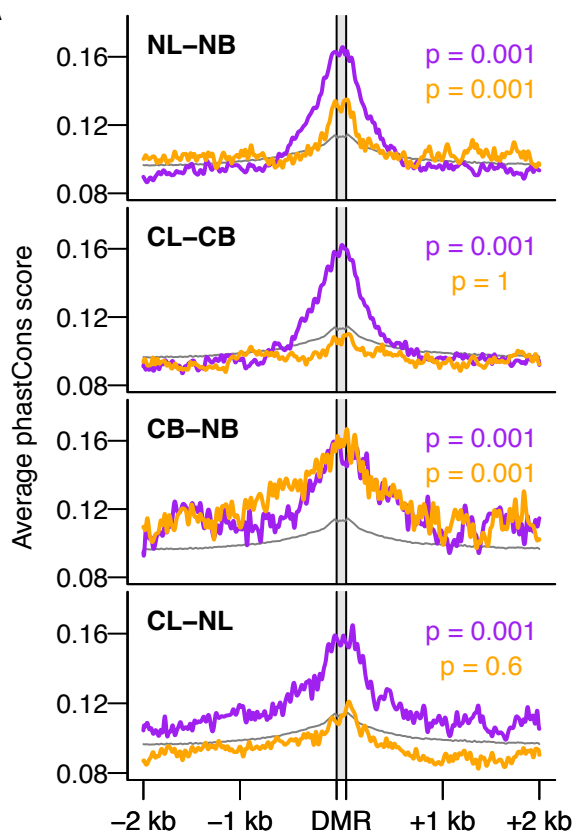
621

622

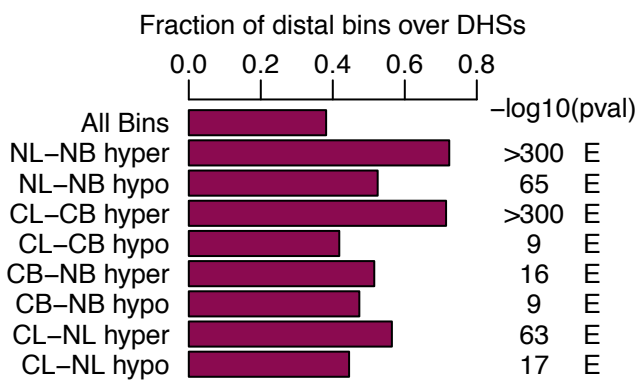
**Figure 1**

**Figure 2**

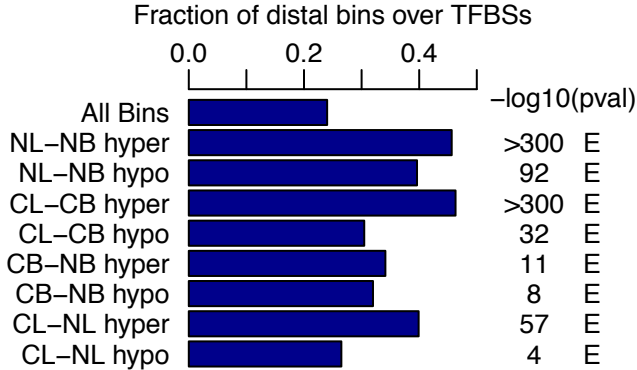
**A**



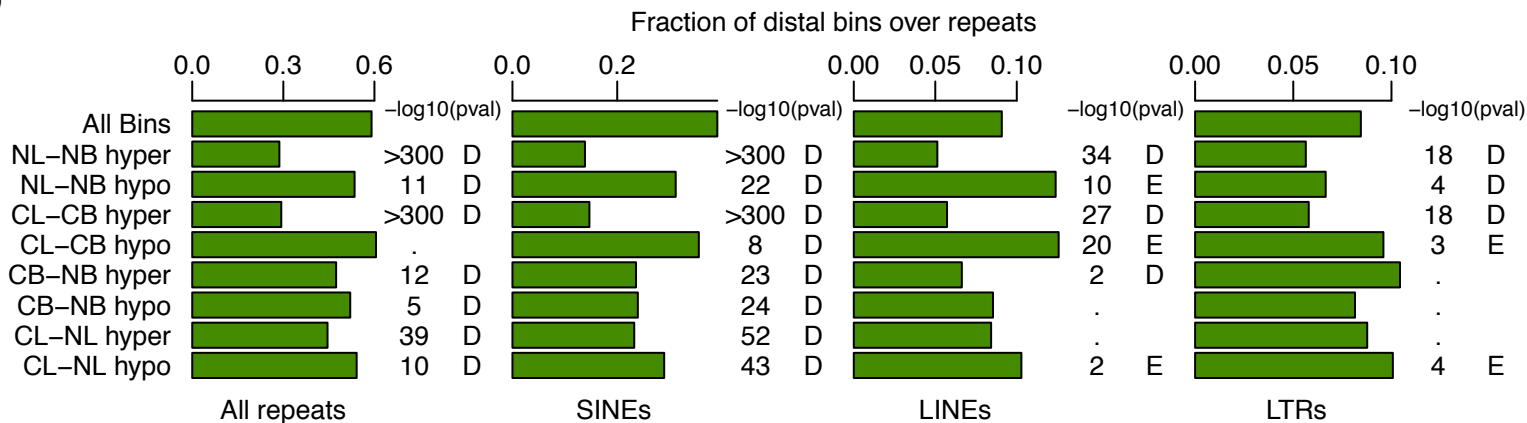
**B**



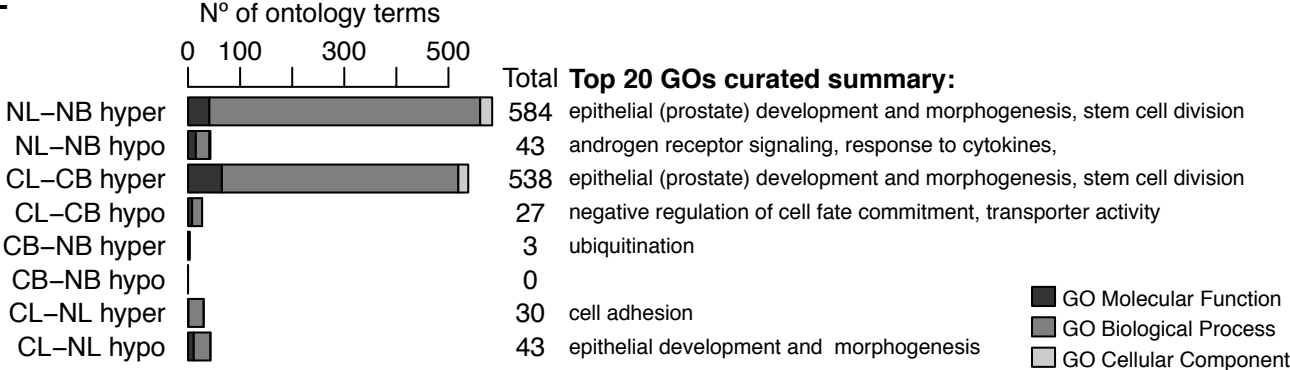
**C**



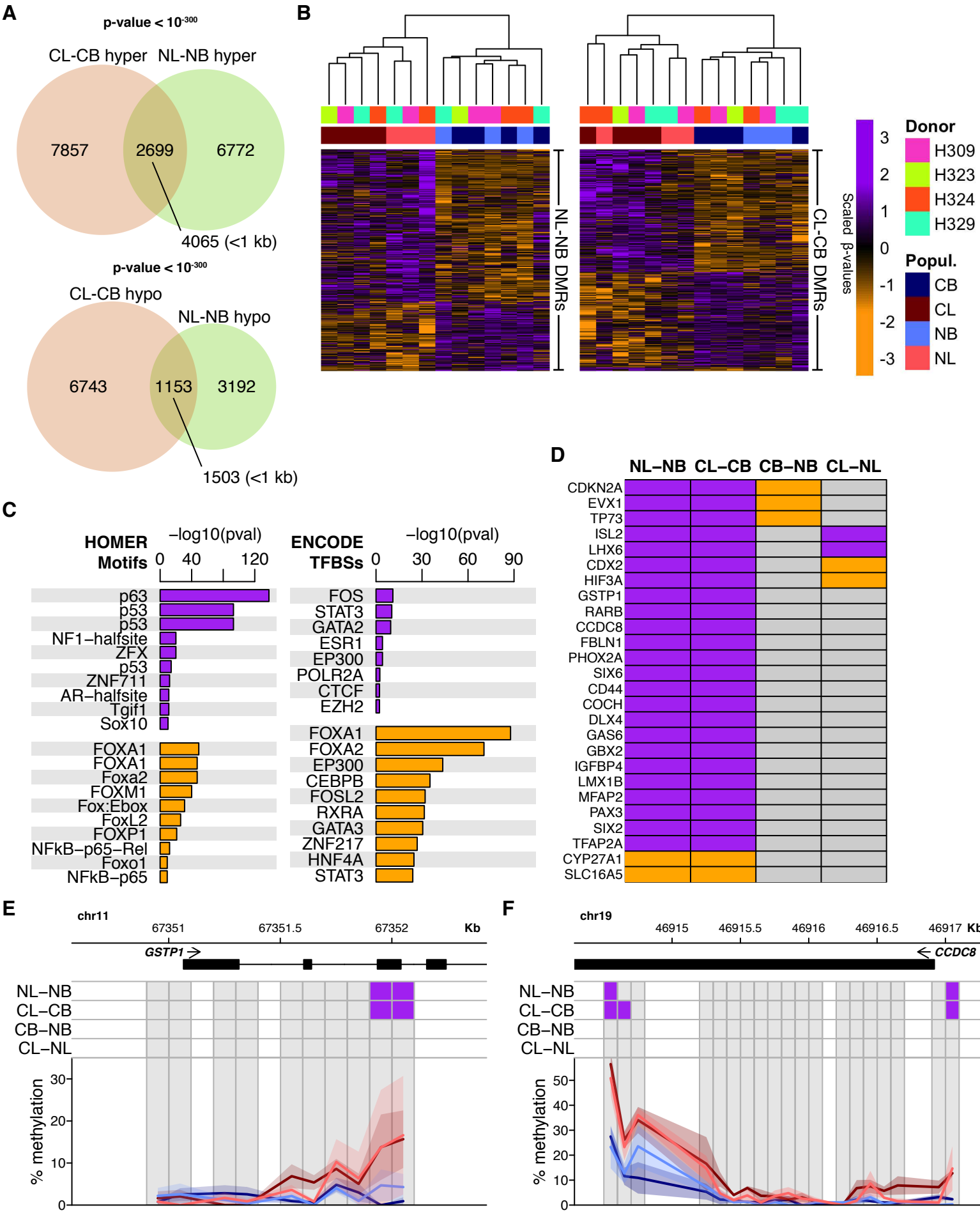
**D**



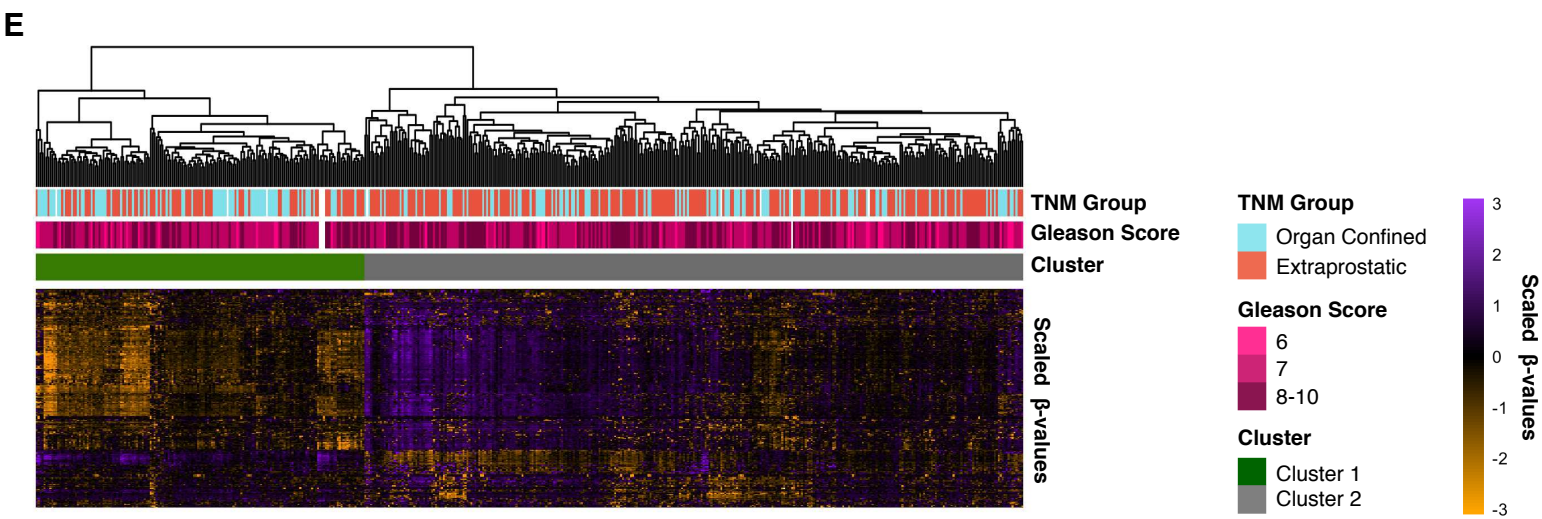
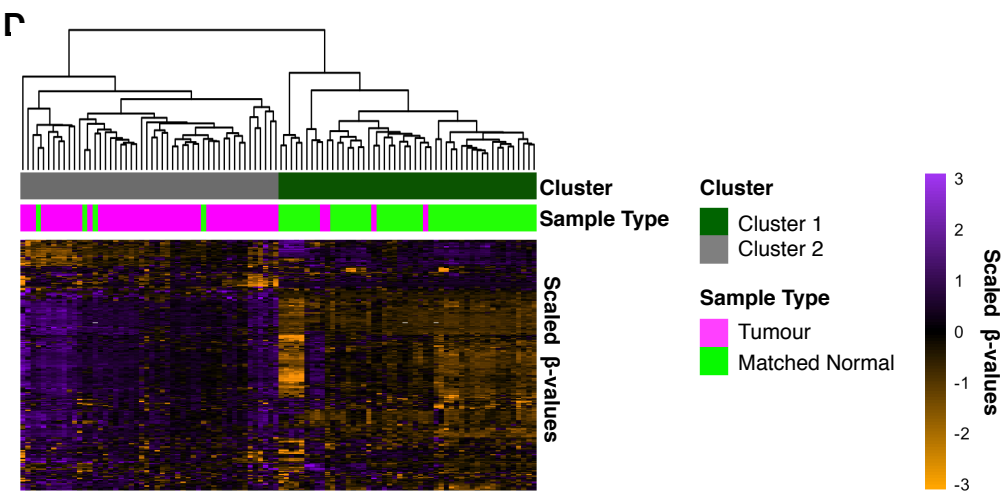
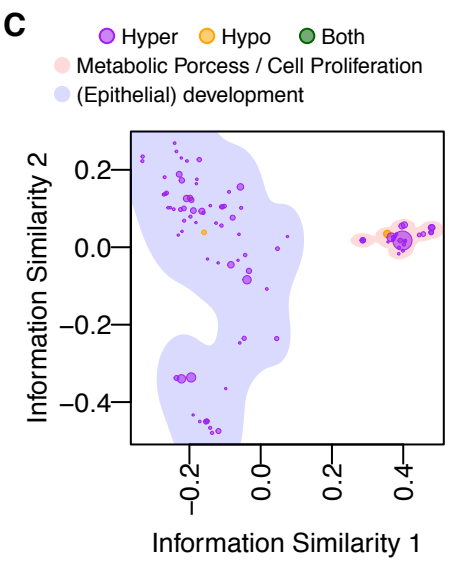
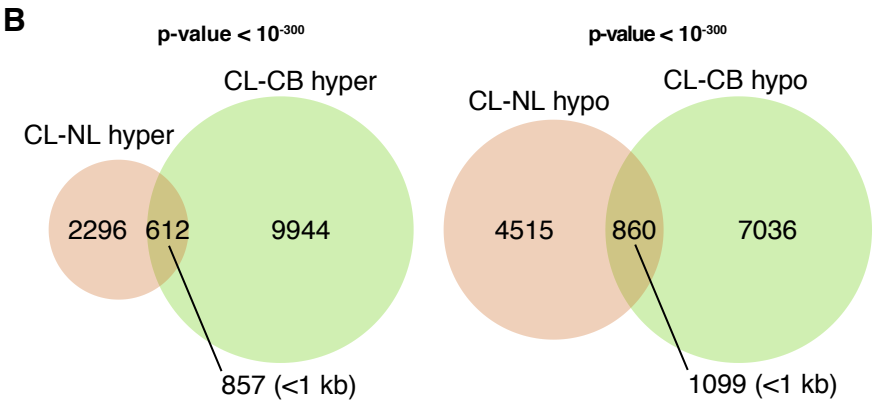
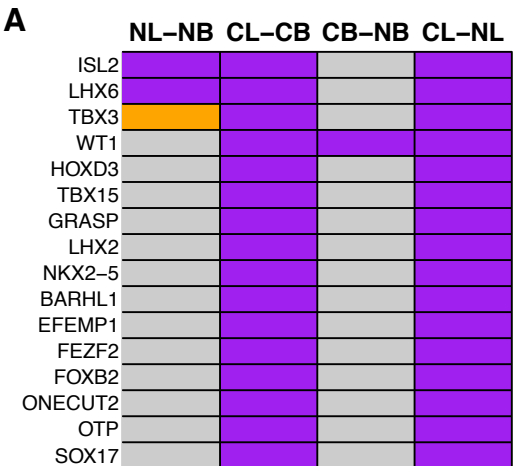
**E**



**Figure 3**



**Figure 4**



**Figure 5**

