# Supplementary material

***For manuscript:*** **Comparative Correlation Structure of Colon Cancer Locus Specific Methylation: Characterisation of Patient Profiles and Potential Markers across 3 Array-Based Datasets**

**1. Subset of genes selected for preliminary analysis.**

CHFR, ALX4, BMP3, HIC1, KCNK13, SST, MLH1, SFRP1, TCF7L1, MGMT, CDKN2A, RUNX3, SLIT2, RAB32, NEUROG1, TIMP3, SFRP2, GDF7, SFRP5, CACNA1G, NELL2, SOCS1, PTPRO, MCC, KIT, DKK1, SOX7, DAPK1, PTEN, EYA2, CBS, CRABP1, DPYSL3, PLEKHC1, VIM, ADAMTS19, IGF2, APC, LRRC4, RASSF2, SLC30A3, SFRP4, CLGN, NRG2, HLTF, HIN1, RASSF1A, ID4, APBA2, EVL, APBA3, ST3GAL1.

**2. List of genes retained for clustering from the GEO array-based datasets.**

**D1:**

CHFR, ALX4, BMP3, HIC1, KCNK13, SST, MLH1, SFRP1, TCF7L1, MGMT, CDKN2A, RUNX3, SLIT2, RAB32, NEUROG1, TIMP3, SFRP2, GDF7, SFRP5, CACNA1G, NELL2, SOCS1, PTPRO, MCC, KIT, DKK1, SOX7, DAPK1, PTEN, EYA2, CBS, CRABP1, DPYSL3, PLEKHC1, VIM, ADAMTS19, IGF2, APC, LRRC4, RASSF2, SLC30A3, SFRP4, CLGN, NRG2, HLTF, HIN1, RASSF1A, ID4.

**D2:**

LRRC4, VIM, SOCS1, SST, PTPRO, IGF2, DAPK1, HIN1, NELL2, CHFR, BMP3, RASSF2, APC, SOX7, NEUROG1, MCC, CRABP1, EYA2, SFRP2, HLTF, MGMT, TCF7L1, RUNX3, GDF7, ADAMTS19, CDKN2A, CBS, SFRP1, SLIT2, SFRP5, CACNA1G, ALX4, SFRP4, APBA2.

**D3:**

VIM, GDF7, HLTF, TCF7L1, NELL2, BMP3, APC, CBS, RASSF2, CHFR, PTPRO, SOX7, MGMT, SFRP5, MLH1, ID4, ADAMTS19, IGF2, SOCS1, SFRP2, SLIT2, SFRP4, DKK1, PTEN, PLEKHC1, RAB32, NEUROG1, MCC, CRABP1, KIT, SFRP1, LRRC4, CLGN, SST, RASSF1A, NRG2, RUNX3, ALX4, SLC30A3, CDKN2A, KCNK13, DAPK1, CACNA1G.

**3. Two-way bootstrapping of the D3.**

In order to test the effect of adding or removing tissue samples in D3, a two-way bootstrapping method was applied, using a modified version of the pvclust method [1]. Our modification consisted in replacing hierarchical clustering by Bayesian Clustering (BC), used to cluster the selected datasets.

In this case, (i) genes can be clustered for bootstrap sets of tissues or (ii) the tissues can be clustered for bootstrap sets of genes.

Pvclust permits multiscale bootstrap resampling, where the data size of the bootstrap samples is permitted to have several values [1]. A large number of bootstrap samples (user selected, in our case nboot=100) are generated for each of the following data sizes 0.5*n, 0.6*n, 0.7*n, …, 1.3*n and 1.4*n. pvclust provides two *p*-values: AU (Approximately Unbiased) *p*-value and BP (Bootstrap Probability) *p*-value. AU *p*-value, computed by multiscale bootstrap resampling, is generally less biased than the BP value, computed by normal bootstrap resampling [2].

The bootstrapped gene cluster is given in Fig. S1. The probability that each subcluster is supported by the data rather than being random is given by the AU value (in red). The AU values are moderately high for the different subclusters of genes in Supplementary Fig. 1, suggesting that sampling the tissues does not strongly affect the underlying data structure.

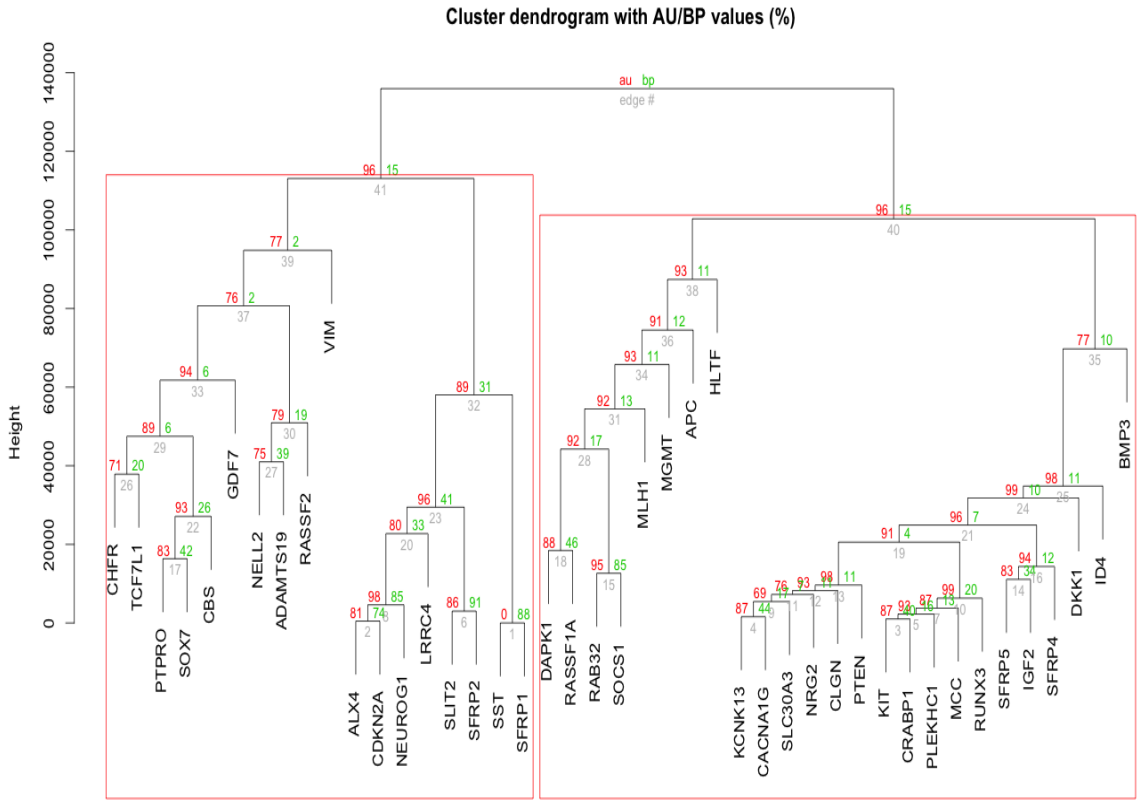
Tissue samples have also been clustered using BC for different bootstrap gene sets. The resulting cluster of tissues with AU values, is given in Fig. S2. Clusters which are the most stable and well-defined (rectangular outline) are found within adjacent healthy tissues. The outlined clusters occur 90% of the time, making a strong argument for their reliability, with resampling from random start points, (not biased to previous run results). This contrasts with the occurrence of non-designated clusters, predominantly found within the cancer tissues, which are variable in content and correspondingly less reliably defined. This is not surprising given that the cancer data are quite highly dispersed from the locus-specific methylation viewpoint. Inclusion of many or few adjacent values does not materially alter the picture since healthy tissue clusters are predominantly compact and well defined.

**References:**


1. Suzuki R and Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. Applications Note. 2006; 22(12), 1540-1542.
2. http://www.is.titech.ac.jp/~shimo/prog/pvclust/

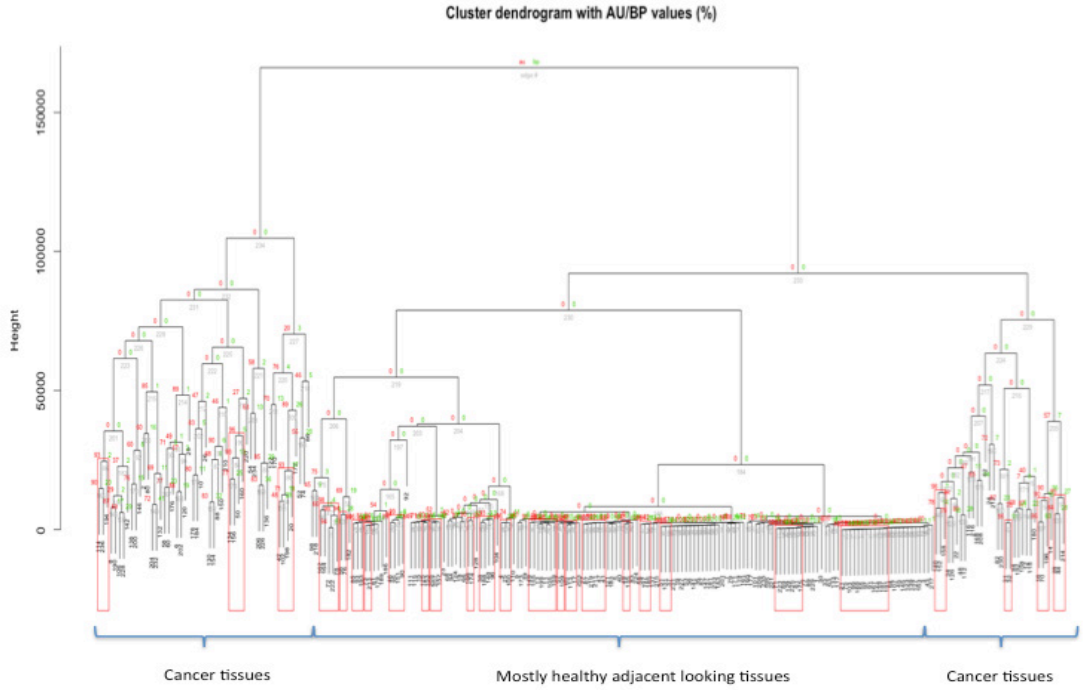
## Supplementary Figures

## Supplementary Figure S1.




**Bootstrapping D3 gene cluster, based on resampling the tissues and clustering the genes.** Clustering of the initial selection of genes from D3, obtained from Bayesian Clustering with bootstrapping by multiscale resampling of the tissues. To assess the effect of using a subset of tissues in this dataset, bootstrapping was carried out using functions from pvclust R package with modifications (to replace hierarchical by BC clustering). pvclust provides two *p*-value measures: AU (Approximately Unbiased) p-value and BP (Bootstrap Probability) p-value. AU *p*-value, computed by multiscale bootstrap resampling, is less biased than BP *p*-value, computed by normal bootstrap resampling. The *p*-value of a sub-cluster indicates probability of occurrence of this sub-cluster. For example, a sub-cluster with AU=0.95 or 95% would imply that it occurs with high reliability. Gene sub-clusters have moderately high AU values, suggesting that resampling the tissues within this same dataset may affect the underlying data structure only slightly. AU and BP values are given at the base at each subcluster, AU to the left and BP to the right.

**Supplementary Figure S2.**




**Bootstrapping D3 tissue cluster: resampling the genes and clustering the tissues.** Multiscale bootstrap resampling of genes for AU values for tissue sub-clusters. The red rectangles highlight clusters with AU values > 0.9.

## Supplementary Tables:

**Supplementary Table S3.** Genes, classified in the same gene clusters by Discriminant Analysis based on PCA, common for D1 and D3 (the two large datasets) and common for all three datasets, respectively, are given. Italicised entries are genes from the initial gene list. Others are genes differentially methylated in cancer tissues that have not been considered for the initial BC and PCA analysis. Tissue clusters are labelled from *c1* to *c4* for the tumour tissues and *adj* for the adjacent tissues.

| | **Overlap** |
|---|---|
| **Gene cluster 1:** average methylation values in the tissue clusters are: $c1 \geq 0.5$, $c2 \leq 0.4$, $c3 \leq 0.2$, $c4 \approx 0$, $adj \approx 0$ | |
| *Intersection between D1 and D3:* | ***PLEKHC1, SLC30A3, KIT***, THSD3, ***NRG2, CLGN, RAB32, CACNA1G, MCC,*** MGC62100, ***DKK1, RASSF1A, KCNK13, CRABP1***. |
| *Intersection between D1, D2 and D3:* | ***CACNA1G, MCC.*** |
| **Gene cluster 2:** average methylation values in the tissue clusters are: $c1 \geq 0.5$, $c2 \leq 0.5$, $c3 \leq 0.4$, $c4 \leq 0.1$, $adj \approx 0$ | |
| *Intersection between D1 and D3:* | ARHGEF7, CSPG2, NDRG4, JAKMIP1, PHF21B, RFX4, BHLHB5, NKX2-5, DFNA5, SNCB, ADARB2, EHD3, ***GDF7*****,** ***PTPRO***, CBSLMX1A, LMX1A, EPHA7, STK33, QKI, ***CBS***, RSPO3, SOX5, JPH3, EFNB3, LOC51334, ***VIM***, ELOVL4, C1QL1, MME, ADAM23, KCNK10, ***RASSF2***, RPRM, GFPT2, PPM1E, DMRTA1, ST8SIA1, PCDH9, DUSP26, DGKG, MMP25, KCNK12, OLFM1. |
| *Intersection between D1, D2 and D3:* | ***PTPRO***, ELOVL4. |
| **Gene cluster 3:** average methylation values in the tissue clusters are: $c1\text{-}c2 \geq 0.4$, $c3\text{-}c4 \geq 0.15$, adj $\approx 0$ | |
| *Intersection between D1 and D3:* | ISYNA1, INPP5B, ***RUNX3*****,** ***SFRP4***, SART2, RASSF5, DOCK3, ***IGF2*****,** ***ID4*****.** |
| *Intersection between D1, D2 and D3:* | ISYNA1, INPP5B |
| **Gene cluster 4:** average methylation values in the tissue clusters are: $c1 - c4 \geq 0.35$, $adj \geq 0.1$. | |
| *Intersection between D1 and D3:* | *N*ot given here, overlap is very large. |
| *Intersection between D1, D2 and D3:* | EMILIN3, UNQ739, SCUBE3, DCC, ***ALX4***, OTOP3, GALR1, PPFIA2, NAP1L3, FLJ45983, CDH13, CHL1, ABCA3, NID2, EYA4, AKAP12, PLD5, PTPRM, FLJ32447, FOXG1B, ***CDKN2A***, ADAMTS18, PTGFR, PCDHB1, SLC32A1, ***SLIT2***, MGC33530, HTR1E, ***NEUROG1***, ZNF354C, NAALAD2, FIGN, MAL, CSMD3, TMEFF2, AGC1, TRPC6, C2orf32, PCDH11X, HS3ST3A1, PAX7, ADAMTS5, FLJ37478, UCHL1, HTR1B, ZNF134, PHOX2A, SLC16A12, GRIK1, RAB11FIP4, ZNF415, UNC5C, COL23A1, ADAM12, EFCAB1, ASTN2, TUSC3, MMP2, IGF2AS, PAK7, DMRT3, CHRDL1, NEFL, ITGA8, NME5, PCDH17, GAS7, FGF10, NMBR, DLK1, CASR, ADRB3, WIT-1, CIDEA, FAM43B, BVES, CD40, SNAP91, CBLN2, SPSB4, ZNF660, SALL1, COL14A1, LRAT, GRP, INA, FZD2, COL4A1 |
| **Gene cluster 5:** average methylation values in the tissue clusters are: $c1 - c4 \geq 0.5$, $adj \geq 0.2$. | |
| *Intersection between D1 and D3:* | ADAMTSL1, HCN4, VILL, ZIC1, SULT4A1, FAM107A, ***DAPK1***, A4GALT, COL11A1, TSGA14, CTPS, TMEM35, FRZB, ISL1, SLC6A4, C9orf78, ***MLH1*** |
| *Intersection between D1, D2 and D3:* | SULT4A1 |
| **Gene cluster 6:** average methylation values in the tissue clusters are: $c1 - c4 \sim 0.6$, $adj \geq 0.35$. | |
| *Intersection between D1 and D3:* | *N*ot given here, overlap is very large. |
| *Intersection between D1, D2 and D3:* | AQP1, GNAS, LPA, CDH8, TCL1A, ADCY8, GPR75, USP54, ZNF135, ATP8A2, KLHDC7B, SEC31L2. |