DNA co-methylation networks outline the structure and remodeling dynamics of colorectal cancer epigenome

Izaskun Mallona[1,3], Susanna Aussó[2], Anna Díez-Villanueva[2], Víctor Moreno[2] and Miguel A. Peinado[1,3]

1. Health Research Institute Germans Trias i Pujol (IGTP), Program in Predictive and Personalized Medicine of Cancer, Badalona, Spain.
2. Unit of Biomarkers and Susceptibility, Cancer Prevention and Control Program, Catalan Institute of Oncology (ICO), Oncobell Program, Bellvitge Biomedical Research Institute (IDIBELL) and CIBERESP, L'Hospitalet de Llobregat, Barcelona, Spain; Department of Clinical Sciences, Faculty of Medicine and Health Sciences, University of Barcelona, Barcelona, Spain.
3. Correspondence to mpeinado@igtp.cat and izaskun.mallona@gmail.com.

Abstract

Epigenomic plasticity is interconnected with chromatin structure and gene regulation. In tumor progression, orchestrated remodeling of genome organization accompanies the acquisition of malignant properties. DNA methylation, a key epigenetic mark extensively altered in cancer, is also linked to genome architecture and function. Based on this association, we postulate that the dissection of long-range co-methylation structure unveils cancer cell's genome architecture remodeling.

We applied network-modeling of DNA methylation co-variation in two colon cancer cohorts and found abundant and consistent transchromosomal structures in both normal and tumor tissue. Normal-tumor comparison indicated substantial remodeling of the epigenome covariation and revealed novel genomic compartments with a unique signature of DNA methylation rank inversion.

Keywords

DNA methylation, co-methylation, epigenetics, networks, colon, colorectal, cancer

Running title

Long range DNA co-methylation in colon cancer

Introduction

Cancer cell functional reprogramming involves gene expression dysregulation driven by genetic and epigenetic changes. The contribution of epigenetic mechanisms to malignant phenotypes has been thoroughly studied and includes extensive DNA methylation alterations as prominent features of most cancers types (Portela and Esteller, 2010; Feinberg et al., 2016). DNA methylation mainly occurs in the cytosine of the CpG dinucleotide and is usually associated with a repressed chromatin state. Changes in DNA methylation have multiple effects in genome regulation and have been directly associated with gene overexpression and silencing, chromatin remodeling and chromosomal instability (Eden et al., 2003; Rodriguez et al., 2006; Jones, 2012; Schübeler, 2015; Feinberg et al., 2016). Direct comparison of the DNA methylation profiles in the tumor versus the paired normal tissue reveals both losses (hypomethylation) and gains (hypermethylations) of the epigenetic mark. The extent of the change may range from discrete sites and promoters to large regions (Frigola et al., 2006; Portela and Esteller, 2010; Hansen et al., 2011; Jones, 2012; Feinberg et al., 2016).

Neighboring CpGs have a higher chance of being similarly methylated (Eckhardt et al., 2006; Shoemaker et al., 2010; Libertini et al., 2016); nonetheless, the actual extent of this vicinity effect is disputed, with reports of complete to weak or very low decay of co-methylation as the genomic distance increases in different cell types and tissues (Li et al., 2010; Akulenko and Helms, 2013; Fortin and Hansen, 2015; Salhab et al., 2018). Most studies about the functional impact of DNA methylation changes have focused the analysis to local effects on neighboring genes (Jones, 2012; Schübeler, 2015). More recently, taking advantage of the availability of genome-scale DNA methylation data from large cancer datasets, the study of DNA co-methylation profiles has been addressed from different points of view, including the analysis of long range correlations (Akulenko and Helms, 2013; Fortin and Hansen, 2015; Zhang and Huang, 2017) and gene centered approaches (Li et al., 2014; Gao and Teschendorff, 2016; Saghafinia et al., 2018). Numerous methods have been developed to evaluate differential methylation and differential variability of DNA methylation (Jaffe et al., 2011; Jenkinson et al., 2017; Teschendorff and Relton, 2018; Jenkinson et al., 2018; Rulands et al., 2018; Visakh and Nazeer, 2019; Libertini et al., 2018; Jenkinson et al., 2019).

We hypothesize that dynamically methylated elements reveal the functional organization of human cancer cell's genome. To get insights into the structure, functional determinants and underlying mechanisms of DNA methylation dynamics we examined the DNA methylomes of colon cancer patients by evaluating

loci with shared patterns of DNA methylation. The rationale was that samples subjected to complex physiopathological processes (e.g. tumor initiation and progression), albeit being highly heterogeneous, share common driver and passenger events from which biologically relevant phenotypic traits arise.

Methods

We separately run our analysis using two independent colon cancer datasets. Our primary cohort, Colonomics (http://www.colonomics.org), included 90 paired primary tumors (stage IIA and IIB) and their adjacent normal tissue. Of the 90 patients, 67 were males and 23 females, aged 43-86 years (mean: 70.37), and 20 developed metastasis. All tumors were microsatellite-stable. Samples were evaluated for DNA methylation (Illumina Infinium HumanMethylation450 BeadChip Array) and gene expression (Affymetrix Human Genome U219), and somatic mutations (exome sequencing) (Closa et al., 2014; Cordero et al., 2014; Sanz-Pamplona et al., 2015; Sole et al., 2014).

The Cancer Genome Atlas (TCGA) series was composed by 256 primary tumor and 38 adjacent non-tumor samples from the colon adenocarcinoma (COAD) cohort (Zhu et al., 2014). Patients were aged 31 to 90 years at diagnosis (mean 65.61), and included 141 males, 144 females and one unassigned. Pathologic stages included Stage I (40), Stage II (97), Stage III (75), Stage IV (32); 11 were not available or discrepant. Regarding microsatellite instability, 10 were positive, 65 negative and 181 were either not tested or had an unknown status. Samples readouts included DNA methylation by Illumina Infinium HumanMethylation450 BeadChip Array, gene expression by RNA-Seq counts and somatic mutations by exome sequencing.

Data processing and workflow is depicted in Figure 1. Briefly, DNA methylation beta values were subjected to serial pairwise correlation analysis for the Colonomics tumor series (primary dataset) and normal adjacent tissue, as well and both TCGA normal samples and tumors (external datasets). Strong associations (effect size Spearman's $\rho \geq 0.8$) were stored. Co-methylation networks were built upon the correlations data (Cullen and Frey, 1999; Csardi and Nepusz, 2006; Clauset et al., 2009; Gillespie, 2014; Delignette-Muller and Dutang, 2015), from which highly connected modules according to the fast greedy community detection algorithm were isolated (Csardi and Nepusz, 2006; Clauset et al., 2004) were isolated. Purity correction was estimated as in (Zheng et al., 2017; Aran et al., 2015). Next, modules were functionally annotated by comparing them to the literature (Lister et al., 2009; Hansen et al., 2011; Aryee et al., 2014; Fortin and Hansen,

2015) and molecular features databases (Liberzon et al., 2011; Heinz et al., 2010; Smyth, 2005; Quinlan and Hall, 2010; Love et al., 2014; Gel et al., 2015). Modules characterization, including reproducibility assessment, consisted in mutual profiles comparison and differential expression analysis among different cohorts (Hubert and Arabie, 1985; Shannon et al., 2003; Krzywinski et al., 2009; Langfelder et al., 2011; Akdemir and Chin, 2015).

We implemented a Web service at `http://maplab.cat/corre` to facilitate browsing the DNA co-methylation events of any locus or gene. The tool allows candidate queries either by gene symbol or Illumina Infinium probename, providing the annotated co-methylations full list. Apart of downloadable spreadsheets, Corre renders interactive plots to evaluate zonal (chromosome) and functional (chromatin color) enrichments (Ernst et al., 2011; Gesmann and de Castillo, 2011; Zhang et al., 2013; Conway et al., 2016). Source code is available at `https://bitbucket.org/imallona/corre` under the GPL terms. Corre can be accessed freely and without registration.

Extended methods are available as supplementary material.

Results and discussion

Close and distant CpGs co-methylate in colon cancer samples

We retrieved DNA methylation data as measured by Infinium HumanMethylation450 Array β values from 90 tumor and 90 adjacent normal tissues (Closa et al., 2014; Cordero et al., 2014; Sanz-Pamplona et al., 2015; Sole et al., 2014) to feed the co-methylation analysis (Figure 1). Quality check consisted in three steps: First, we excluded probes non uniquely mapping to a genomic location, being polymorphic or located in sex chromosomes (Price et al., 2013). Second, probes with low variability (standard deviation $s_\beta < 0.05$) were filtered out to get rid of correlations led by outliers, while enriching for CpGs with intermediate methylation (Du et al., 2010). Finally, probes with missing data in any sample were removed. Next we sequentially calculated the bulk pairwise Spearman's correlations between any possible pair of probes adjusting for multiple testing (Table S1) .

Co-methylations were frequent at neighboring CpGs but also detectable at long distances (Figure 2A), being this effect independent of local probe density (Figure S1). The extent of the comethylation as mea-

sured from the correlation coefficient ρ (ranging from -1 to 1, in which 0 means full independence) were bell-shape distributed, thus indicating that the majority of correlations lied on the non-significant range, as expected, and with a tendence towards positive association at CpGs located at short distances due to local co-methylation. Opposite methylation trends are detectable in trans (Figure 2). Coordinated variation was independent of the CpG location in open or closed chromatin compartments (Figure S2).

To underpin the biological relevance of the findings and rule out the co-methylation structure arising due to technical noise, we evaluated five possible sources of artifacts: multiple testing, batch effects, leading outliers, tumor purity and chip design and none of them appeared to have a significant effect on the results.

To account for the iterative nature of the analysis, consisting in exhaustively computing any pairwise correlation between the Infinium probes with variable DNA methylation, we set an astringent effect size cut-off of the Spearman's correlation cofficient $\rho \geq 0.8$, which is close to the conservative Bonferroni p-adjustment for the datasets used (optimization against the asymptotic p-values as calculated by the Fisher Z transform, Table S1) (Fisher, 1915; Shakhbazov et al., 2016). The absence of notorious clustering of DNA methylation values (Figure S3) indicates that batch effects are unlikely drivers of co-methylation (Leek et al., 2010). To attenuate the leading effect of DNA methylation outliers we filtered in probes with sufficient variation in DNA methylation (setting a standard deviation threshold). The number of variable probes was about twice in the tumor series as compared to normal tissue (Table S1). To reinforce robustness against outliers, nonparametric Spearman correlations were run, which rely on DNA methylation ranks rather than values (Croux and Dehon, 2010). Regarding tumor purity, we found co-methylation modules to retain tight correlation structure after subtracting purity effects (Zheng et al., 2017) (Figure S4). As for the Infinium array design, neither the probes GC content (Figure S5) nor the dye channel (Figure S6) drove the correlations structure; probes mapping to multiple locations or overlapping to SNPs were filtered out (Price et al., 2013).

Anatomy of the distance-agnostic co-methylation network in colorectal cancers

Given the pairwise nature of co-methylation events, we envisioned a co-methylation network linking all significant co-methylations in a genomic location-agnostic manner, thus representing genome-wide comethylation trends, and potentially linking shared drivers. To do this, we selected the top-scoring correlations and assembled a network in which loci and correlations are represented by nodes (vertex) and links (edges). For the sake of simplicity, we only considered edges with positive correlations.

The tumor network resulted in 63,130 nodes and 26 million co-methylations (Table S2). The distribution of each CpG degree (amount of connectivity, number of co-methylating neighbors) showed a heavy-tail shape with the vast majority of nodes being linked to few counterparts, whereas a few nodes displayed a fairly abundant connectivity. The degree distribution did not resemble power-law, lognormal nor exponential, as there was not linear dependency between the cumulative frequencies and the connectivities (Figures S7 and S8). This structure was unaffected by loci features such as chromatin state (Figure S9) and genomic category (Figure S10); but lost when filtering out trans interactions, as probes placed at any distance in cis showed power-law compatible distributions (Figure S11), as described previously (Zhang et al., 2017).

Interestingly, genomic loci acting as co-methylation hubs (frequently co-methylating in trans) presented homogeneous intermediate DNA methylation levels in normal samples (Figures S12 and S13), which have been suggested to play regulatory roles (Elliott et al., 2015), with an important enrichment of imprinted loci (n=53, 11%). On the other side, they were not enriched for partially methylated domains (PMDs), which have been reported as loci with intermediate DNA methylation values and high variability (Lister et al., 2009): 27% (132 CpGs) of the rich probes overlapped PMDs, similarly to the 33% (21,065 CpGs) of the probes with at least a significant co-methylation and what is expected from the background of the whole set of Infinium probes, with an overlap of the 31% (147,257 CpGs).

The colon cancer co-methylation landscape is reproducible across cohorts and tumor type-specific

To test the reproducibility of the network we repeated the analysis using an independent dataset, the COAD cohort from TCGA, consisting of 256 primary colon adenocarcinomas. In TCGA, the DNA methylation β value calling procedure differs from the original dataset, and therefore reduces the chance of covariation artifacts arising due to the data processing bias.

Near a quarter million probes fulfilled the variability criteria in TCGA colon tumors (Table S1). The overall correlations distribution and the co-methylation decay with distance matched that of the primary dataset (Figures 4 and 2). Next, we evaluated whether the correlation value for each pair of probes was conserved, including the non-significant pairs. To do so, we computed exhaustive pairwise correlations of

CpGs located at the chromosome 10 against itself and plotted the original ρ values of each CpG pair against validation dataset from TCGA. The linearity of both landscapes (Figure 4) indicated a high concordance of the overall co-methylation levels. Similarly, the TCGA colon cancer co-methylome network was comparable in terms of graph properites (Table S2 and Figure S7). Interestingly, the comethylation landscape was very dissimilar to other tumor types' (Figure S14), whilst it was comparing colon datasets from the starting and the validation cohorts, strongly supporting the specificity of the colon cancer co-methylation network.

In a similar vein and to compare the structure of both networks, we checked whether the correlating nodes present in both cohorts displayed the same connectivity to other nodes. The influence score of each node (e.g. based in the number of links arising from it) was estimated using the PageRank score (Page et al., 1999) in both datasets. Both primary and validation tumor co-methylation networks showed a reproducible distribution of nodes' PageRanks mostly composed by lowly influential CpGs (Figure S15).

DNA co-variation in normals pinpoints massive co-methylation rearrangements in tumors

We wondered whether some of the co-methylations found in cancers were already detectable in adjacent normal colonic mucosa, and to which extent the co-methylome structure differed from the tumor. The normal colon co-methylome network was built using 90 non-tumor tissues (Table S1). We found that both the number of probes fulfilling the variance prerequisite (n=99,346) and the number of total correlations (7,430,741) decreased to 39% and 22% of the tumor's ones, respectively. This result was consistent with the higher DNA methylation variability in tumors. As expected, a predominance of positive correlations was observed, being more intense for close probes (Figure 2A). Anti-methylations showed much more abundant in tumors than in normal tissue (Figure 2). In agreement with the associations found in tumors, co-methylations were underrepresented in active promoters (Figure S16) and the co-methylation network' connectivities were not power-law distributed (Figure S7). Probes pairs correlation values showed partial agreement between the normal and tumor datasets (Figure 4). Interestingly, age and other demographic features, common to both the tumor and adjacent normal data, were not primary drivers of the co-methylation (Figure S17).

In order to assess the reproducibility of the network, we repeated the analysis using normal samples from the external dataset from TCGA. It should be noted that this dataset only includes 38 normal samples, and

as the correlation significance depends on the sample size (Fisher, 1915), keeping the same cut-offs is likely to boost the number of false positives. On the other hand, increasing the ρ cut-off to an equivalent detection threshold ($\rho = 0.96$, Table S1) produced a very small network whose properties might be out of scale with the previous analysis. With this cautionary note in mind, keeping the $\rho = 0.8$ cut-off the network confirmed the distinctive distribution of pairwise correlations (Figure 4), whose differences are especially conspicuous at short ranges (Figure 2).

Co-variation in DNA methylation segments the epigenome into sparse compartments

In order to dig into the observed trend of preferential co-methylation of some loci, we explored whether the network had highly connected subnetworks (also known as modules or communities). Modules consist of clusters of nodes heavily interconnected as compared to the rest of the network ers of nodes heavily interconnected (as compared to the rest of the network) constitute a community (Newman and Girvan, 2004; Fortunato, 2010). Modularity is quantified as the fraction of edges connecting nodes of the same type minus what it is expected in a randomly wired network. Scores of 0 indicate no modularity and networks with modular structure typically range from 0.3 to 0.7 (Newman and Girvan, 2004). The tumor co-methylation network was found to be modular (modularity = 0.47) (Table S2), and we partitioned it into 3,270 modules ranging from two to 18,727 nodes using the Clauset, Newman and Moore's fast greedy method (Clauset et al., 2004). Interestingly, the normal tissue network exhibited a higher co-methylome modularity (0.62) (Table S2) and network segmentation resulted in 1,265 modules ranging from two to 17,758 nodes. To rule out the effect of tumor purity on correlation structure, we regressed it out and found that co-methylation modules retained tight correlation structure (Zheng et al., 2017) (Figure S4).

Expectedly, we found the vast majority of the small modules to be composed by sets of probes located at close distance from each other (e.g. at CpG islands). As these could be mined by standard block variation or differential methylation analysis, we focused in trans-chromosomal modules, which were more numerous in the tumor co-methylome (1% of the total modules) than in the normal's (1.4%). Trans modules are diverse in physical properties, such as the chromosome, and in functional properties, such as A/B compartment or chromatin state (Figure S20). In agreement with the original dataset results, TCGA tumor co-methylation network was also modular (modularity score 0.41) (Table S2) and segmentation produced 3,421 modules

ranging from two to 8,981 nodes. The application of size and distance filters reduced the number of trans-chromosomal modules to 35.

The validation series TCGA similarly depicted a modular tumor network (modularity score 0.41) (Table S2) and its clustering into modules resulted in a segmentation in 3,421 modules ranging from two to 8,981 nodes. By application of the same size and co-location filters applied to the original series (at least 10 probes, 1 Mb apart or in different chromosomes) only 35 modules remained. Next, we faced both cohorts module by module in order to detect assess their preservation.

We next evaluated the extent of conservation of transchromosomal co-methylation territories between normals and tumors (biological effect) as well as between cohorts (replicability). Networks clustering on adjusted Rand's distance (Hubert and Arabie, 1985) indicated that modules memberships separate tumor's from normal's networks in both datasets (Figure S21A), in line with the similarities in nodes population (Figure S21B), and their spatial co-methylation patterns (Figure S22). It is worth noting that the use of a correlation threshold (i.e. effect size $\rho \geq 0.8$) may underestimate module co-methylation maintenance when the correlations distribution gets displaced towards values below, but close to, the statistical significance cut-off (Figure S23).

Module preservation across tissue types and cohorts was also evaluated by cross-tabulation of the number of shared CpGs. Twelve tumor modules had one or more counterparts in the normal tissue network, and a similar number in the validation TCGA tumor cohort (Fisher's exact test) (Figures S24 and S25). The five-top sized Colonomics tumor modules partially matched to multiple TCGA's modules. This result might be technically due to the resolution limit of modularity-optimizing algorithms to detect modules, which tend to aggregate modules into few giant components, disregarding their inner complexity (Fortunato and Barthelemy, 2007).

A novel and sparse genomic compartment shows inverted DNA methylation ranks in colon tumors

As we detected a noticeable gain of anti-methylations in the tumors as compared to the normals (Figure 2), and that the co-methylation tumor network depicted two separate giant components (Figure 3) that were not present in the normal tissue, we suspected a major restructuring of co-methylation architecture. The external series from TCGA replicated this finding in both tumor and normal epigenomes (Figure 3).

The pervasive nature of anticorrelations overcame age, gender, tumor stage, tumor purity and anatomical site potential effects on modules' DNA methylation variation (Figures S17, S18, S19). Loci with copy number alterations also conveyed the module-specific DNA methylation ranks mimicking the profiles along balanced regions (Figure S26).

This observation was caused by the inversion of the co-variation pattern of sparse co-methylation compartments, whose DNA methylation status gets a quantitative DNA methylation rank inversion during malignant transformation (Figure 5). To describe this phenomenon, which displays properties beyond regional differential methylation nor increased regional DNA methylation variability, we next evaluated the anatomical and functional features of the co-methylation modules.

Co-methylation loci are enriched for functional signatures

To test the hypothesis that co-methylation structures are directly related with functional properties we investigated genomic and functional features of co-methylated CpGs. We subtracted the overrepresentation of promoters and promoter-related features in the Infinium array (Bibikova et al., 2011; Sandoval et al., 2011; Silva-Martínez et al., 2017), as well as their presumable effects driven by spatial clustering (Kim et al., 2008; Barrera and Peinado, 2012; Gaidatzis et al., 2014; MacDonald et al., 2015; Libertini et al., 2016; Wang et al., 2016). The tumor modules displayed important differences in feature enrichment, including chromatin states, genomic categories, CpG islands and association with known motifs (Figure 3). In line with the weighted gene co-expression network analysis, in which functional signatures can be told apart by mining gene co-expression (Langfelder and Horvath, 2008), the module-specific co-regulation patterns denoted by distinctive abundance of genomic and functional states. Among the multiple features analyzed, a striking global enrichment of inactive promoters was observed in a large number of modules (Figure S27), pointing

out potential clusters of co-regulated genes.

Next, we explored the co-methylation modules overlap with regions of DNA methylation variability previously reported in colon cancer, including differentially methylated regions (Hansen et al., 2011). Interestingly, seven out of 32 tumor modules significantly overlapped tumor hypermethylated blocks (Figure S28). Regarding other types of DNA methylation variability reported by Hansen et al. (2011) , modules showed distinctive profiles, with frequent enrichment in boundary shifts as well in loss of regulation; novel hypomethylation blocks were enriched in three modules only. This complexity reinforces the individuality of co-methylation modules.

A comprehensive summary of structural and functional feature enrichment for each co-methylation module is shown in Tables S4 and S5). To name a few examples, multiple co-methylation modules were significantly enriched for Polycomb-related marks (i.e. H3K27me3, or SUZ12, EED and PRC2 targets; e.g. tumor modules 1, 3, 5 and 598); for frequently mutated at COSMIC molecular signatures (i.e. tumor modules 2, 4 and 8); and for gene expression (i.e. tumor module 8). Finally, we could also confirm that co-methylation network associated features found in TCGA matched Colonomics enrichment signatures, e.g., the underrepresentation of co-methylations within active promoters (Figure S27).

To shed light into causal factors driving dynamic methylome modularity we searched for enriched motifs (i.e. transcription factor binding sites) at the co-methylating loci (Supplementary methods). We found that six out of the 32 Colonomics tumor modules presented one or more significantly enriched motifs (Table S6). Enriched motifs included ETS and RUNX families in modules 1, 3, 327, FOS family members (FRA1, ATF3, BATF, FOSL2, JUN-AP1) in modules 2 and 4), FOXA1-related factors (FOXA1, HNF4a, FOXMA) in module 2, GC box (KLF5 and KLF4) in module 2, C/EBP in module 3, PAX7 and MYF5 in module 5, homeobox in modules 5 and 83, MADS in module 152, and ASCL1 in module 598.

In summary, the modularity of long range DNA co-methylation events in colorectal cancer point to novel properties of epigenetic remodeling. We postulate that coordinated DNA methylation changes at interspersed sites (here identified as belonging to the same module) might be linked to common players and effectors, hypothesis in line with the DNA methylation-mediated binding of transcription factors to regulatory sites (Kribelbauer et al., 2017; Yin et al., 2017).

A colon cancer co-methylome browser

We developed a web tool available at http://maplab.cat/corre to ease browsing of raw and processed data. To illustrate the usage of this tool we queried the INHBB gene encoding activin B, a member of the TGF-beta family, with different biological activities, including a role in cell proliferation and inflammation. Epigenetic silencing of INHBB is frequent in colorectal cancer (Frigola et al., 2006) and has been proposed as indicator of poor outcome (Mayor et al., 2009). The co-methylation landscape of INHBB exposed by the Corre tool showed a large number of positive correlations 20 kb upstream and downstream of the gene in both normal and tumor samples (Figure S29). Negative correlations were only present in tumors and were enriched in poised promoters, indicating the potential remodeling of bivalent states and hypermethylation (McGarvey et al., 2008; Ohm et al., 2007; Rodriguez et al., 2008). Compared with the normal samples, the tumors displayed an increase in the number of links for most sites, although some chromosomes, especially 8 (Figure S29E), but also 13, 14, 17 and 21, showed an opposite trend with a depletion of co-methylations in the tumors as compared with the normal samples (Figure S30). Another interesting result was in regard to cg03699182 probe (Figure S29D, arrowhead) located in the CpG island of the INHBB promoter presented 42 co-methylations in the normal samples against only three in the tumors. Most of the cg03699182 co-methylations affected were located in poised promoters of polycomb regulated genes (Figure S30). The dynamics of the connections and the properties of the affected sites are consistent with the participation of instructive mechanisms resulting in the DNA hypermethylation and long range epigenetic silencing of multiple genes in colorectal cancer (Frigola et al., 2006; Keshet et al., 2006; Michieletto et al., 2017).

Conclusions

Here we describe a network modeling analysis of DNA methylation co-variation. When applied to colorectal cancer and normal colonic mucosa it is demonstrated that the architecture of DNA methylation variability is modular and embodies structural and functional associations with a potential impact in cancer biology. Normal and tumor networks display characteristic and differential structures that are consistently replicated in independent cohorts. Interestingly a quantitative inversion of DNA methylation ranks results in a novel and sparse genomic compartment pinpointing a long-range functional reorganization of cancer cell's epigenome.

Ethics approval and consent to participate

The study protocol was approved by the Clinical Research Ethics Committee (CEIC) of the Bellvitge Hospital (no. PR178/11).

Availability of data and materials

We provide free access to the co-methylation network data at http://maplab.cat/corre without registration. As for reproducibility, source code is released under the GPL free software license at http://bitbucket.org/imallona/correlations (analysis) and http://bitbucket.org/imallona/corre (shiny app).

Methylation data from the Colonomics project can be downloaded from GEO with accession GSE131013; TCGA data are available at https://portal.gdc.cancer.gov.

Competing interests

MAP is cofounder and equity holder of Aniling, a biotech company with no interests in this paper. MAP lab has received research funding from Celgene. The rest of the authors declare no conflict of interest.

Funding

This work was funded by the Spanish Ministry of Science, Innovation and Universities (FEDER, SAF2015-64521-R, RTI2018-094009-B-100 to MAP), the Agency for Management of University and Research Grants (AGAUR) of the Catalan Government grants 2017SGR723 (to VM) and 2017SGR529 (to MAP). The funding agency had no role in the design of the study, collection, analysis, interpretation of data nor manuscript writing.

Authors' contributions

IM, VM and MAP conceived the study. IM developed the method and wrote the software. IM, SA, ADV, VM and MAP analyzed data. IM and MAP wrote the manuscript. All authors read and approved the manuscript.

Acknowledgements

We thank Iñaki Martinez de Ilarduya and Judith Flo for their excellent technical support and Francisco Chen for contributing to the sequential correlations parallelization. Results presented here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Authors' information

Izaskun Mallona is currently a SIB Swiss Institute of Bioinformatics and University of Zurich member (Department of Molecular Mechanisms of Disease DMMD, and Institute of Molecular Life Sciences DMLS).

References

Akdemir, K. C. and Chin, L. (2015) Hicplotter integrates genomic data with interaction matrices. *Genome Biology*, **16**, 198.

Akulenko, R. and Helms, V. (2013) Dna co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. *Human Molecular Genetics*, ddt158.

Aran, D., Sirota, M. and Butte, A. J. (2015) Systematic pan-cancer analysis of tumour purity. *Nature Communications*, **6**.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D. and Irizarry, R. A. (2014) Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, **30**, 1363–1369.

Barrera, V. and Peinado, M. A. (2012) Evaluation of single cpg sites as proxies of cpg island methylation states at the genome scale. *Nucleic Acids Research*, **40**, 11490–11498.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., Fan, J. B. and Shen, R. (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**, 288–295.

Clauset, A., Newman, M. E. and Moore, C. (2004) Finding community structure in very large networks. *Physical review E*, **70**, 066111.

Clauset, A., Shalizi, C. R. and Newman, M. E. (2009) Power-law distributions in empirical data. *SIAM review*, **51**, 661–703.

Closa, A., Cordero, D., Sanz-Pamplona, R., Sole, X., Crous-Bou, M., Pare-Brunet, L., Berenguer, A., Guino, E., Lopez-Doriga, A., Guardiola, J., Biondo, S., Salazar, R. and Moreno, V. (2014) Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis. *Carcinogenesis*, **35**, 2039–2046.

Conway, J., Eddelbuettel, D., Nishiyama, T., Prayaga, S. K. and Tiffin, N. (2016) *RPostgreSQL: R interface to the PostgreSQL database system.* R package version 0.4-1.

Cordero, D., Sole, X., Crous-Bou, M., Sanz-Pamplona, R., Pare-Brunet, L., Guino, E., Olivares, D., Berenguer, A., Santos, C., Salazar, R., Biondo, S. and Moreno, V. (2014) Large differences in global transcriptional regulatory programs of normal and tumor colon cells. *BMC cancer*, **14**, 1.

Croux, C. and Dehon, C. (2010) Influence functions of the spearman and kendall correlation measures. *Statistical methods & applications*, **19**, 497–515.

Csardi, G. and Nepusz, T. (2006) The igraph software package for complex network research. *InterJournal, Complex Systems*, **1695**, 1–9.

Cullen, A. C. and Frey, H. C. (1999) *Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs*. Springer Science & Business Media.

Delignette-Muller, M. L. and Dutang, C. (2015) fitdistrplus: An r package for fitting distributions. *Journal of Statistical Software*, **64**, 1–34.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L. and Lin, S. M. (2010) Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K. and Beck, S. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, **38**, 1378–1385.

Eden, A., Gaudet, F., Waghmare, A. and Jaenisch, R. (2003) Chromosomal instability and tumors promoted by DNA hypomethylation. *Science*, **300**, 455–455.

Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. (2009) Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.

Elliott, G., Hong, C., Xing, X., Zhou, X., Li, D., Coarfa, C., Bell, R. J., Maire, C. L., Ligon, K. L., Sigaroudinia, M., Gascard, P., Tlsty, T. D., Harris, R. A., Schalkwyk, L. C., Bilenky, M., Mill, J., Farnham, P. J., Kellis, M., Marra, M. A., Milosavljevic, A., Hirst, M., Stormo, G. D., Wang, T. and Costello, J. F. (2015) Intermediate dna methylation is a conserved signature of genome regulation. *Nature Communications*, **6**, 6363.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. and Bernstein, B. E. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.

Feinberg, A. P., Koldobskiy, M. A. and Göndör, A. (2016) Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature Reviews Genetics*, **17**, 284.

Fisher, R. A. (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, **10**, 507–521.

Fortin, J.-P. and Hansen, K. D. (2015) Reconstructing a/b compartments as revealed by hi-c using long-range correlations in epigenetic data. *Genome Biology*, **16**, 180.

Fortunato, S. (2010) Community detection in graphs. *Physics reports*, **486**, 75–174.

Fortunato, S. and Barthelemy, M. (2007) Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, **104**, 36–41.

Frigola, J., Song, J., Stirzaker, C., Hinshelwood, R. A., Peinado, M. A. and Clark, S. J. (2006) Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band. *Nature Genetics*, **38**, 540.

Gaidatzis, D., Burger, L., Murr, R., Lerch, A., Dessus-Babus, S., Schübeler, D. and Stadler, M. B. (2014) Dna sequence explains seemingly disordered methylation levels in partially methylated domains of mammalian genomes. *PLoS Genetics*, **10**, e1004143.

Gao, Y. and Teschendorff, A. E. (2016) Epigenetic and genetic deregulation in cancer target distinct signaling pathway domains. *Nucleic Acids Research*, **45**, 583–596.

Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M. A. and Malinverni, R. (2015) regioner: an r/bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*, **32**, 289–291.

Gesmann, M. and de Castillo, D. (2011) Using the google visualisation api with r. *The R Journal*, **3**, 40–44.

Gillespie, C. S. (2014) Fitting heavy tailed distributions: the powerlaw package. *arXiv preprint arXiv:1407.3492*.

Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry, R. A. and Feinberg, A. P. (2011) Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, **43**, 768–775.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. and Glass, C. K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular Cell*, **38**, 576–589.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *Journal of Classification*, **2**, 193–218.

Jaffe, A. E., Feinberg, A. P., Irizarry, R. A. and Leek, J. T. (2011) Significance analysis and statistical dissection of variably methylated regions. *Biostatistics*, **13**, 166–178.

Jenkinson, G., Abante, J., Feinberg, A. P. and Goutsias, J. (2018) An information-theoretic approach to the modeling and analysis of whole-genome bisulfite sequencing data. *BMC Bioinformatics*, **19**, 87.

Jenkinson, G., Abante, J., Koldobskiy, M. A., Feinberg, A. P. and Goutsias, J. (2019) Ranking genomic features using an information-theoretic measure of epigenetic discordance. *BMC Bioinformatics*, **20**, 175.

Jenkinson, G., Pujadas, E., Goutsias, J. and Feinberg, A. P. (2017) Potential energy landscapes identify the information-theoretic nature of the epigenome. *Nature Genetics*, **49**, 719.

Jones, P. A. (2012) Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics*, **13**, 484.

Keshet, I., Schlesinger, Y., Farkash, S., Rand, E., Hecht, M., Segal, E., Pikarski, E., Young, R. A., Niveleau, A., Cedar, H. and Simon, I. (2006) Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nature Genetics*, **38**, 149.

Kim, S., Li, M., Paik, H.-i. H., Nephew, K. P., Shi, H., Kramer, R., Xu, D. and Huang, T. H.-M. (2008) Predicting dna methylation susceptibility using cpg flanking sequences. In *Pacific Symposium on Biocomputing*, vol. 13, 315–326.

Kribelbauer, J. F., Laptenko, O., Chen, S., Martini, G. D., Freed-Pastor, W. A., Prives, C., Mann, R. S. and Bussemaker, H. J. (2017) Quantitative analysis of the dna methylation sensitivity of transcription factor complexes. *Cell Reports*, **19**, 2383–2395.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. and Marra, M. A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Research*, **19**, 1639–1645.

Langfelder, P. and Horvath, S. (2008) Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Langfelder, P., Luo, R., Oldham, M. C. and Horvath, S. (2011) Is my network module preserved and reproducible? *PLoS Computational Biology*, **7**, e1001057.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K. and Irizarry, R. A. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, **11**, 733–739.

Li, Y., Xu, J., Ju, H., Xiao, Y., Chen, H., Lv, J., Shao, T., Bai, J., Zhang, Y., Wang, L., Wang, X., Ren, H. and Li, X. (2014) A network-based, integrative approach to identify genes with aberrant co-methylation in colorectal cancer. *Molecular BioSystems*, **10**, 180–190.

Li, Y., Zhu, J., Tian, G., Li, N., Li, Q., Ye, M., Zheng, H., Yu, J., Wu, H., Sun, J., Zhang, H., Chen, Q., Luo, R., Chen, M., He, Y., Jin, X., Zhang, Q., Yu, C., Zhou, G., Sun, J., Huang, Y., Zheng, H., Cao, H., Zhou, X., Guo, S., Hu, X., Li, X., Kristiansen, K., Bolund, L., Xu, J., Wang, W., Yang, H., Wang, J., Li, R., Beck, S., Wang, J. and Zhang, X. (2010) The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biology*, **8**, e1000533.

Libertini, E., Hamoudi, R. A., Heath, S., Lancashire, L., Garcia, A. R., Grassi, L., Downes, K., Ouwehand, W. H., Javierre, B.-M., Cairns, J., Wingett, S., Paul, D. S., Gut, M., Gut, I. G., Martens, J. H. A., Ivliev, A., Stunnenberg, H. G., Frontini, M., Spivakov, M., Fraser, P., Cutler, A., Wallace, C. and Beck, S. (2018) Dna methylation oscillation defines classes of enhancers. *bioRxiv*. URL: https://www.biorxiv.org/content/early/2018/10/10/262212.

Libertini, E., Heath, S. C., Hamoudi, R. A., Gut, M., Ziller, M. J., Czyz, A., Ruotti, V., Stunnenberg, H. G., Frontini, M., Ouwehand, W. H., Meissner, A., Gut, I. G. and Beck, S. (2016) Information recovery from low coverage whole-genome bisulfite sequencing. *Nature Communications*, **7**.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P. and Mesirov, J. P. (2011) Molecular signatures database (msigdb) 3.0. *Bioinformatics*, **27**, 1739–1740.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q. M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B. and Ecker, J. R. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

Love, M. I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, **15**, 550.

MacDonald, W. A., Sachani, S. S., White, C. R. and Mann, M. R. (2015) A role for chromatin topology in imprinted domain regulation. *Biochemistry and Cell Biology*, **94**, 43–55.

Mayor, R., Casadome, L., Azuara, D., Moreno, V., Clark, S., Capella, G. and Peinado, M. (2009) Long-range epigenetic silencing at 2q14. 2 affects most human colorectal cancers and may have application as a non-invasive biomarker of disease. *British Journal of Cancer*, **100**, 1534.

McGarvey, K. M., Van Neste, L., Cope, L., Ohm, J. E., Herman, J. G., Van Criekinge, W., Schuebel, K. E. and Baylin, S. B. (2008) Defining a chromatin pattern that characterizes dna-hypermethylated genes in colon cancer cells. *Cancer Research*, **68**, 5753–5759.

Michieletto, D., Chiang, M., Coli, D., Papantonis, A., Orlandini, E., Cook, P. R. and Marenduzzo, D. (2017) Shaping epigenetic memory via genomic bookmarking. *Nucleic Acids Research*, **46**, 83–93.

Newman, M. E. and Girvan, M. (2004) Finding and evaluating community structure in networks. *Physical review E*, **69**, 026113.

Ohm, J. E., McGarvey, K. M., Yu, X., Cheng, L., Schuebel, K. E., Cope, L., Mohammad, H. P., Chen, W., Daniel, V. C., Yu, W., M, B. D., Thomas, J., Kevin, P., J, S. S., Neil, W. D., G, H. J. and B, B. S. (2007) A stem cell–like chromatin pattern may predispose tumor suppressor genes to dna hypermethylation and heritable silencing. *Nature Genetics*, **39**, 237.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1999) The pagerank citation ranking: bringing order to the web.

Portela, A. and Esteller, M. (2010) Epigenetic modifications and human disease. *Nature Biotechnology*, **28**, 1057–1068.

Price, E. M., Cotton, A. M., Lam, L. L., Farré, P., Emberly, E., Brown, C. J., Robinson, W. P. and Kobor, M. S. (2013) Additional annotation enhances potential for biologically-relevant analysis of the illumina infinium humanmethylation450 beadchip array. *Epigenetics & Chromatin*, **6**, 4.

Quinlan, A. R. and Hall, I. M. (2010) Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Rodriguez, J., Frigola, J., Vendrell, E., Risques, R.-A., Fraga, M. F., Morales, C., Moreno, V., Esteller, M., Capellà, G., Ribas, M. and Peinado, M. A. (2006) Chromosomal instability correlates with genome-wide dna demethylation in human primary colorectal cancers. *Cancer Research*, **66**, 8462–9468.

Rodriguez, J., Muñoz, M., Vives, L., Frangou, C. G., Groudine, M. and Peinado, M. A. (2008) Bivalent domains enforce transcriptional memory of dna methylated genes in cancer cells. *Proceedings of the National Academy of Sciences*, **105**, 19809–19814.

Rulands, S., Lee, H. J., Clark, S. J., Angermueller, C., Smallwood, S. A., Krueger, F., Mohammed, H., Dean, W., Nichols, J., Rugg-Gunn, P., Kelsey, G., Stegle, O., Simons, B. D. and Reik, W. (2018) Genome-scale oscillations in dna methylation during exit from pluripotency. *Cell systems*, **7**, 63–76.

Saghafinia, S., Mina, M., Riggi, N., Hanahan, D. and Ciriello, G. (2018) Pan-cancer landscape of aberrant dna methylation across human tumors. *Cell reports*, **25**, 1066–1080.

Salhab, A., Nordstrom, K., Gasparoni, G., Kattler, K., Ebert, P., Ramirez, F., Arrigoni, L., Muller, F., Polansky, J. K., Cadenas, C., G Hengstler, J., Lengauer, T., Manke, T. and Walter, J. (2018) A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains. *Genome Biol.*, **19**, 150.

Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M. and Esteller, M. (2011) Validation of a dna methylation microarray for 450,000 cpg sites in the human genome. *Epigenetics*, **6**, 692–702.

Sanz-Pamplona, R., Lopez-Doriga, A., Pare-Brunet, L., Lazaro, K., Bellido, F., Alonso, M. H., Ausso, S., Guino, E., Beltran, S., Castro-Giner, F., Gut, M., Sanjuan, X., Closa, A., Cordero, D., Moron-Duran, F. D., Soriano, A., Salazar, R., Valle, L. and Moreno, V. (2015) Exome Sequencing Reveals AMER1 as a Frequently Mutated Gene in Colorectal Cancer. *Clinical Cancer Research*, **21**, 4709–4718.

Schübeler, D. (2015) Function and information content of dna methylation. *Nature*, **517**, 321.

Shakhbazov, K., Powell, J. E., Hemani, G., Henders, A. K., Martin, N. G., Visscher, P. M., Montgomery, G. W. and McRae, A. F. (2016) Shared genetic control of expression and methylation in peripheral blood. *BMC Genomics*, **17**, 1.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**, 2498–2504.

Shoemaker, R., Deng, J., Wang, W. and Zhang, K. (2010) Allele-specific methylation is prevalent and is contributed by cpg-snps in the human genome. *Genome Research*, **20**, 883–889.

Silva-Martínez, G. A., Zaina, S. and Lund, G. (2017) Array probe density and pathobiological relevant cpg calling bias in human disease and physiological dna methylation profiling. *Briefings in Functional Genomics*, **17**, 42–48.

Smyth, G. K. (2005) Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, 397–420. Springer.

Sole, X., Crous-Bou, M., Cordero, D., Olivares, D., Guino, E., Sanz-Pamplona, R., Rodriguez-Moranta, F., Sanjuan, X., de Oca, J., Salazar, R. and Moreno, V. (2014) Discovery and validation of new potential biomarkers for early detection of colon cancer. *PloS One*, **9**, e106748.

Teschendorff, A. E. and Relton, C. L. (2018) Statistical and integrative system-level analysis of dna methylation data. *Nature Reviews Genetics*, **19**, 129.

Visakh, R. and Nazeer, K. A. (2019) Multi-network approach to identify differentially methylated gene communities in cancer. *Gene*, **697**, 227–237.

Wang, J., Duncan, D., Shi, Z. and Zhang, B. (2013) Web-based gene set analysis toolkit (webgestalt): update 2013. *Nucleic Acids Research*, **41**, W77–W83.

Wang, Y., Liu, T., Xu, D., Shi, H., Zhang, C., Mo, Y.-Y. and Wang, Z. (2016) Predicting dna methylation state of cpg dinucleotide using genome topological features and deep networks. *Scientific Reports*, **6**.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., Nitta, K. R., Taipale, M., Popov, A., Ginno, P. A., Domcke, S., Yan, J., Schübeler, D., Vinson, C. and Taipale, J. (2017) Impact of cytosine methylation on dna binding specificities of human transcription factors. *Science*, **356**.

Zhang, H., Meltzer, P. and Davis, S. (2013) Rcircos: an r package for circos 2d track plots. *BMC Bioinformatics*, **14**, 244.

Zhang, J. and Huang, K. (2017) Pan-cancer analysis of frequent dna co-methylation patterns reveals consistent epigenetic landscape changes in multiple cancers. *BMC Genomics*, **18**, 1045.

Zhang, L., Xie, W. J., Liu, S., Meng, L., Gu, C. and Gao, Y. Q. (2017) Dna methylation landscape reflects the spatial organization of chromatin in different cells. *Biophysical Journal*, **113**, 1395–1404.

Zheng, X., Zhang, N., Wu, H.-J. and Wu, H. (2017) Estimating and accounting for tumor purity in the analysis of dna methylation data from cancer studies. *Genome Biology*, **18**, 17.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., Benner, C. and Chanda, S. K. (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*, **10**, 1523.

Zhu, Y., Qiu, P. and Ji, Y. (2014) Tcga-assembler: open-source software for retrieving and processing tcga data. *Nature Methods*, **11**, 599–600.

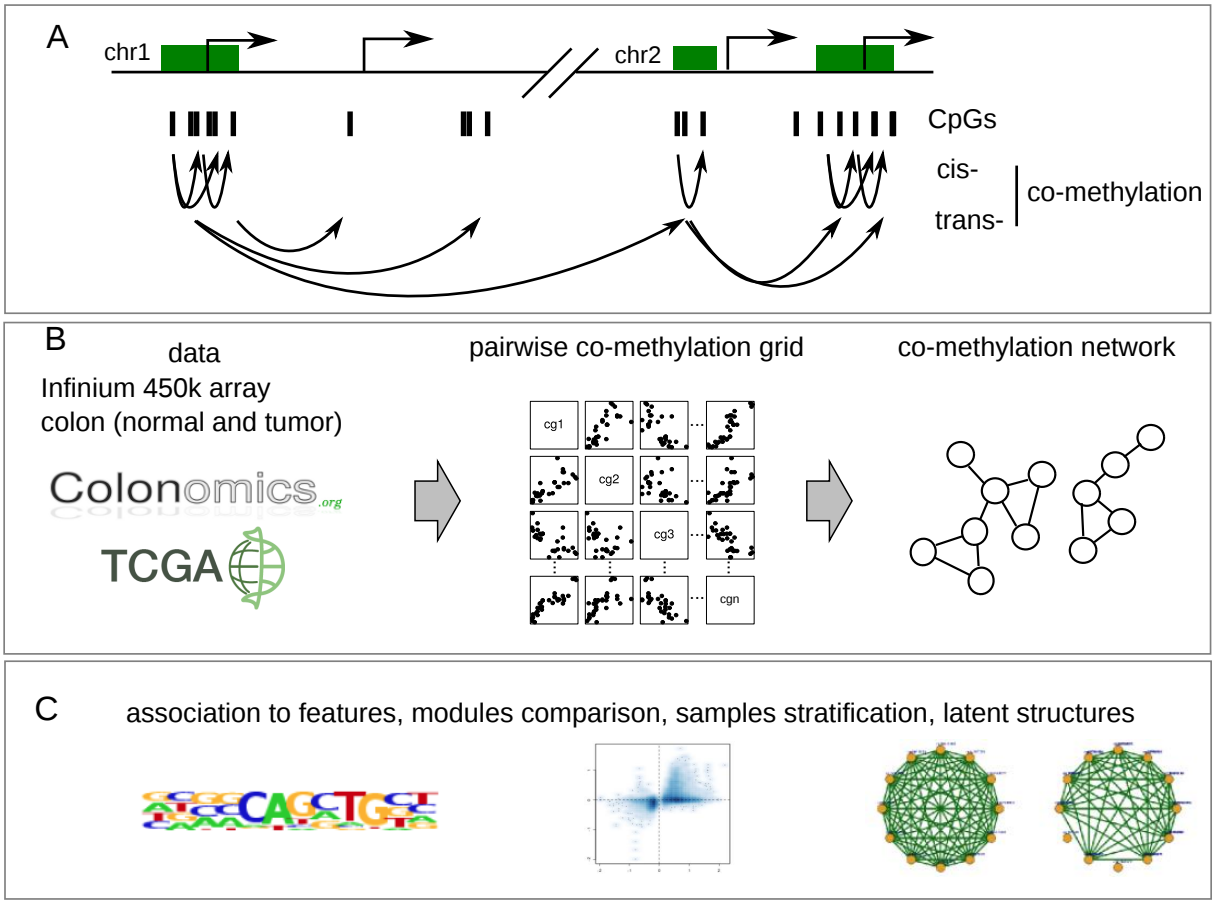


Figure 1: Co-methylation analysis framework. A, CpGs co-methylation occurs at close (cis) and long distances (trans). B, DNA co-methylation networks describe the distance-agnostic co-methylation structure in colon tumors and normals. C, co-methylation networks associate to molecular signatures.

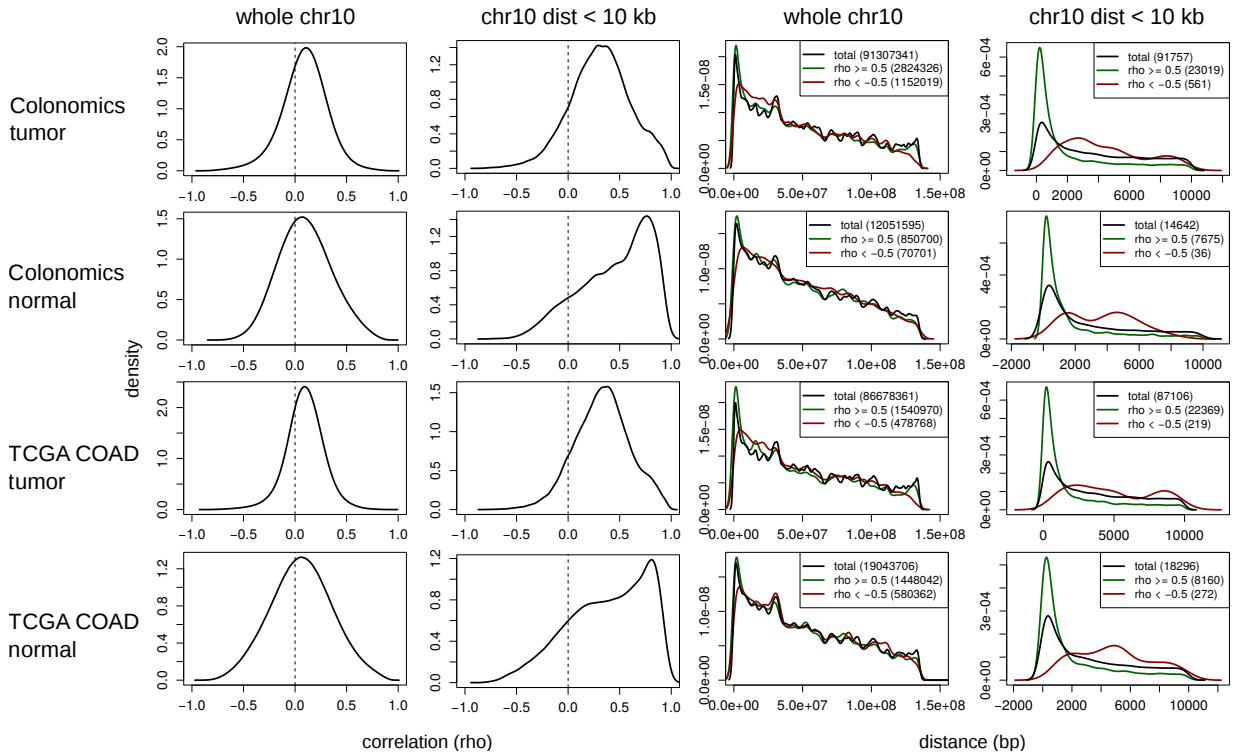


Figure 2: Co-methylation and distance effects. Correlations distribution is bell-shaped and shifted towards positive values at the tumor cohort (whole chr10); the trend to co-methylate is noticeably increased in cis (chr10 probes located at less than 10 kbp). Red: negative correlations ($\rho < -0.5$); green: positive correlations ($\rho \geq 0.5$).

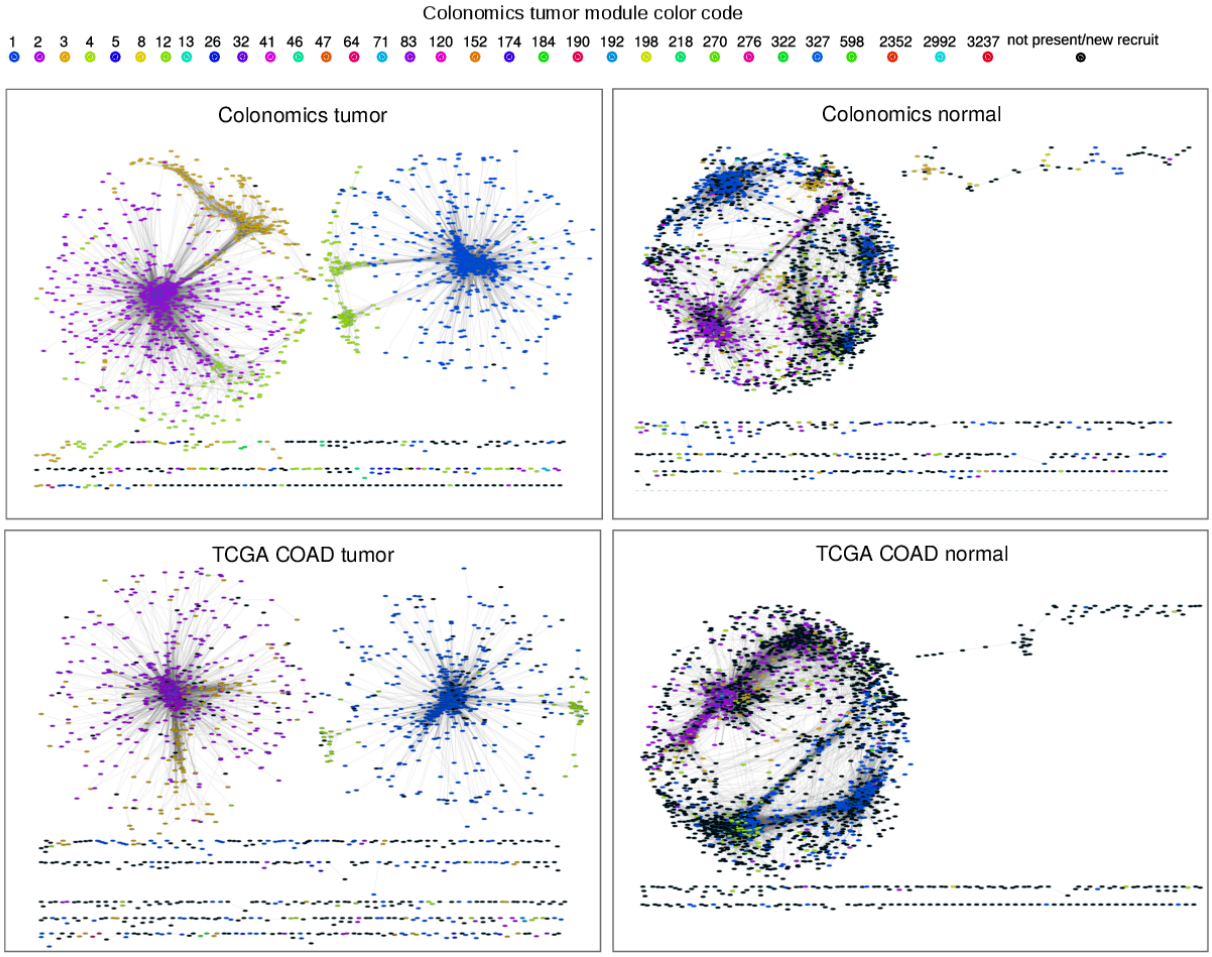


Figure 3: Co-methylation network module memberships comparison. Networks, built independently for each dataset, highlight the nodes (CpGs) by the trans module they belong to at Colonomics tumor (black indicates new recruits or, for Colonomics tumor, non-trans module membership). Graphs are limited to a random sample of 5000 nodes and solitary nodes are not plotted; network layout was calculated by $1 - \rho$ (edges) weighted springs.

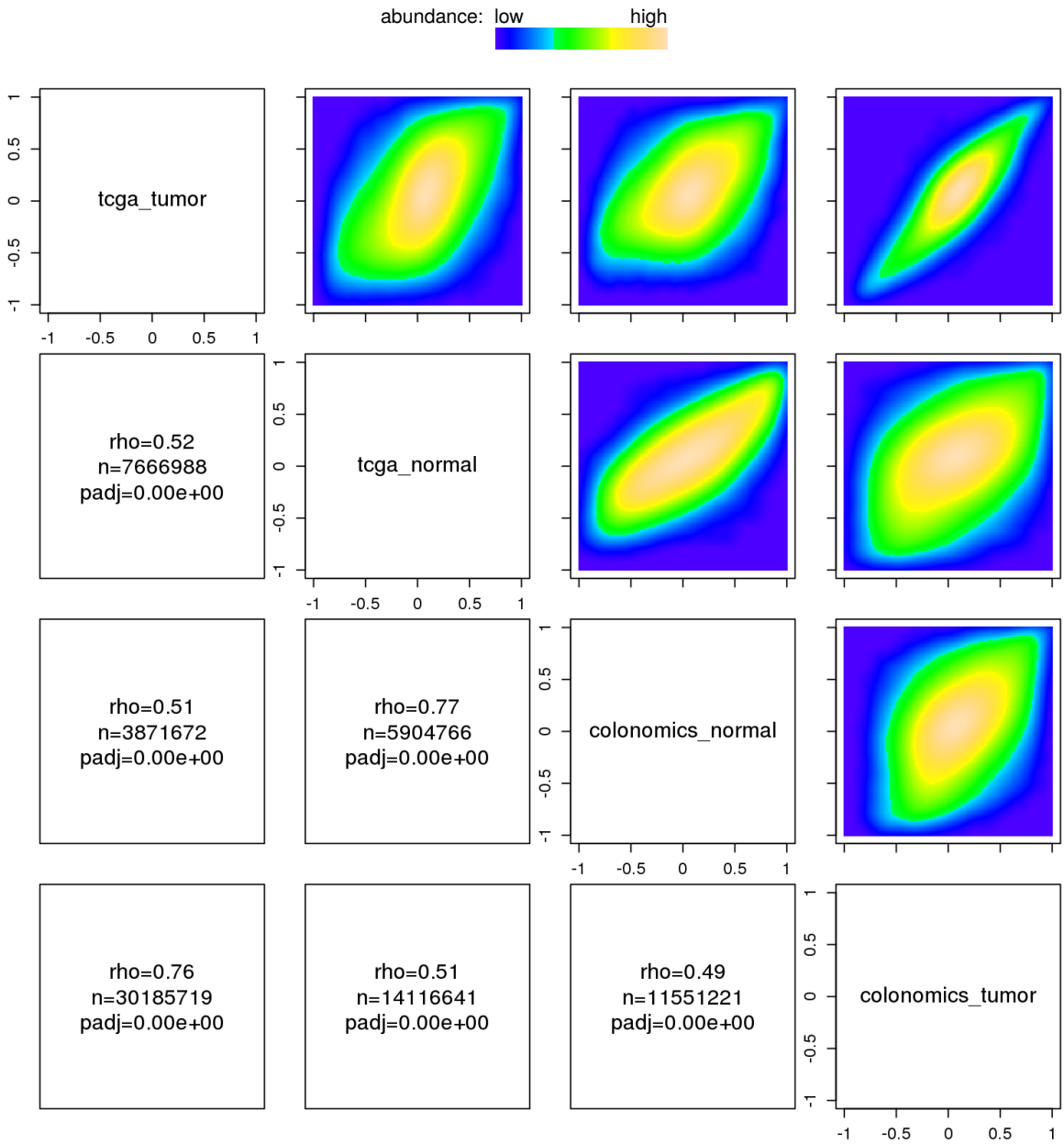


Figure 4: Correlations replication across cohorts. For each cohort, we conducted an exhaustive pairwise correlation analysis for any probe with at least $sd \geq 0.05$. Then we pairwise plotted the correlation coefficient rho (X and Y axis) for each probe pair to check whether the co-methylation landscape was reproduced. Analysis was restricted to chr10.

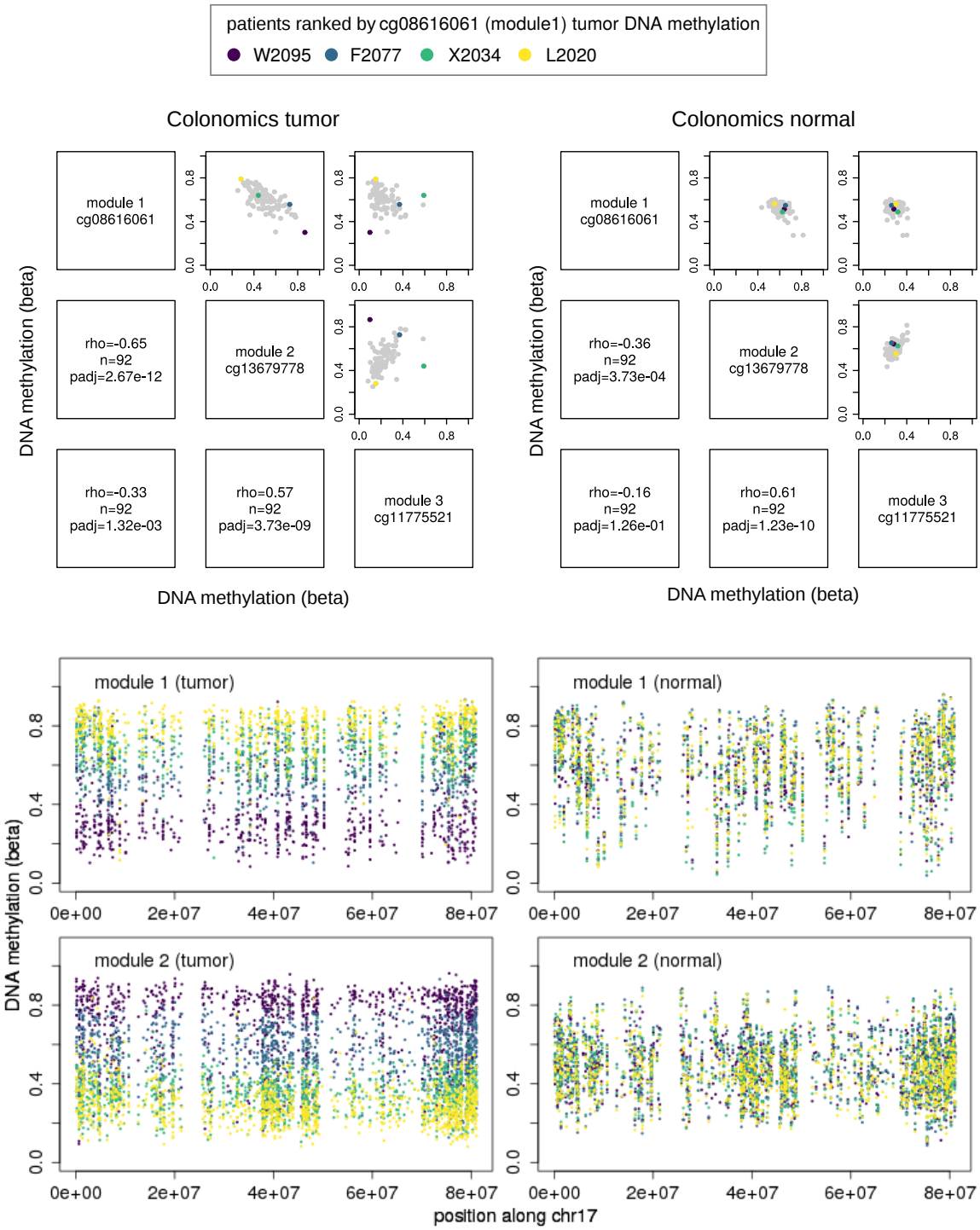
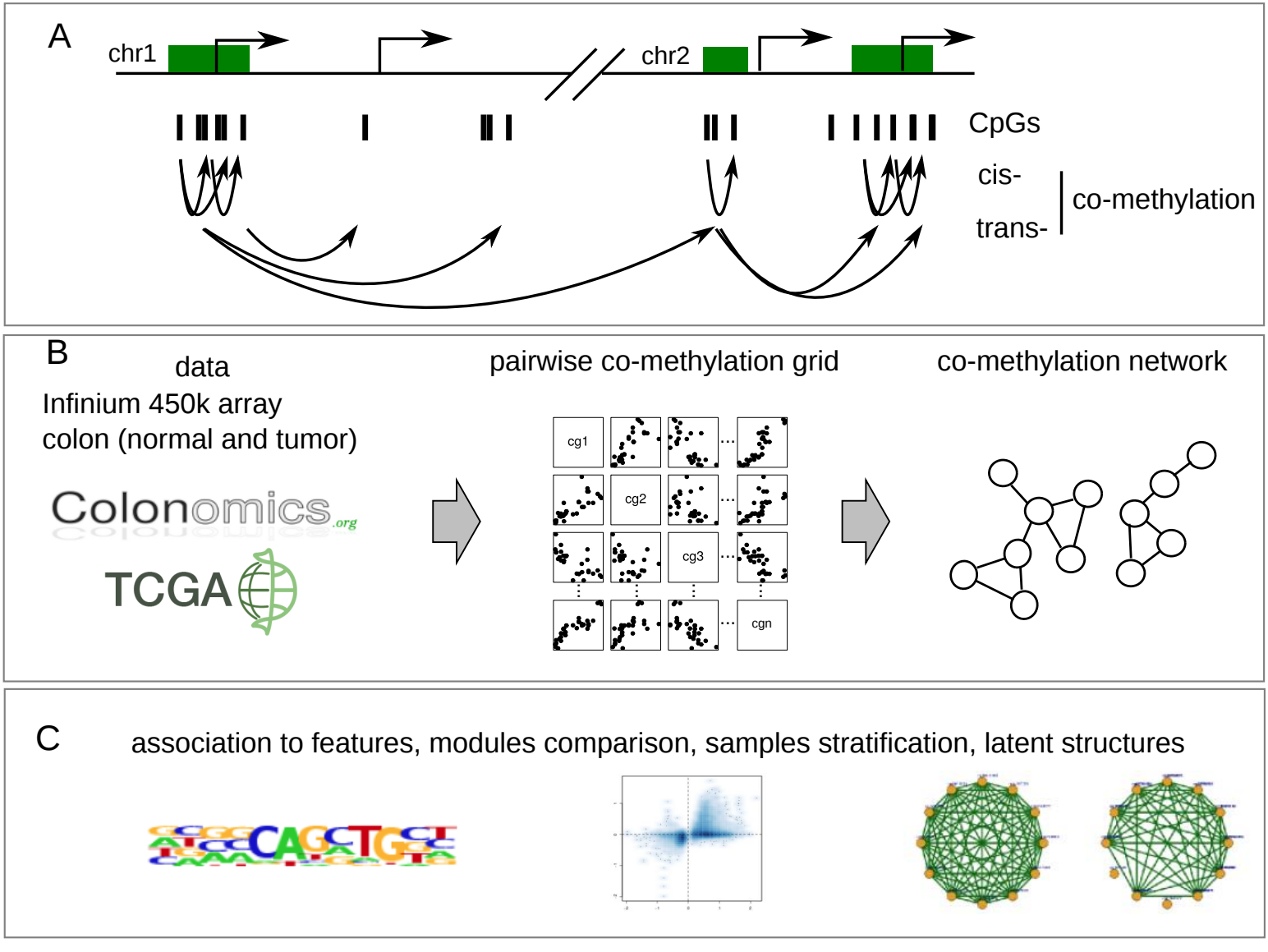
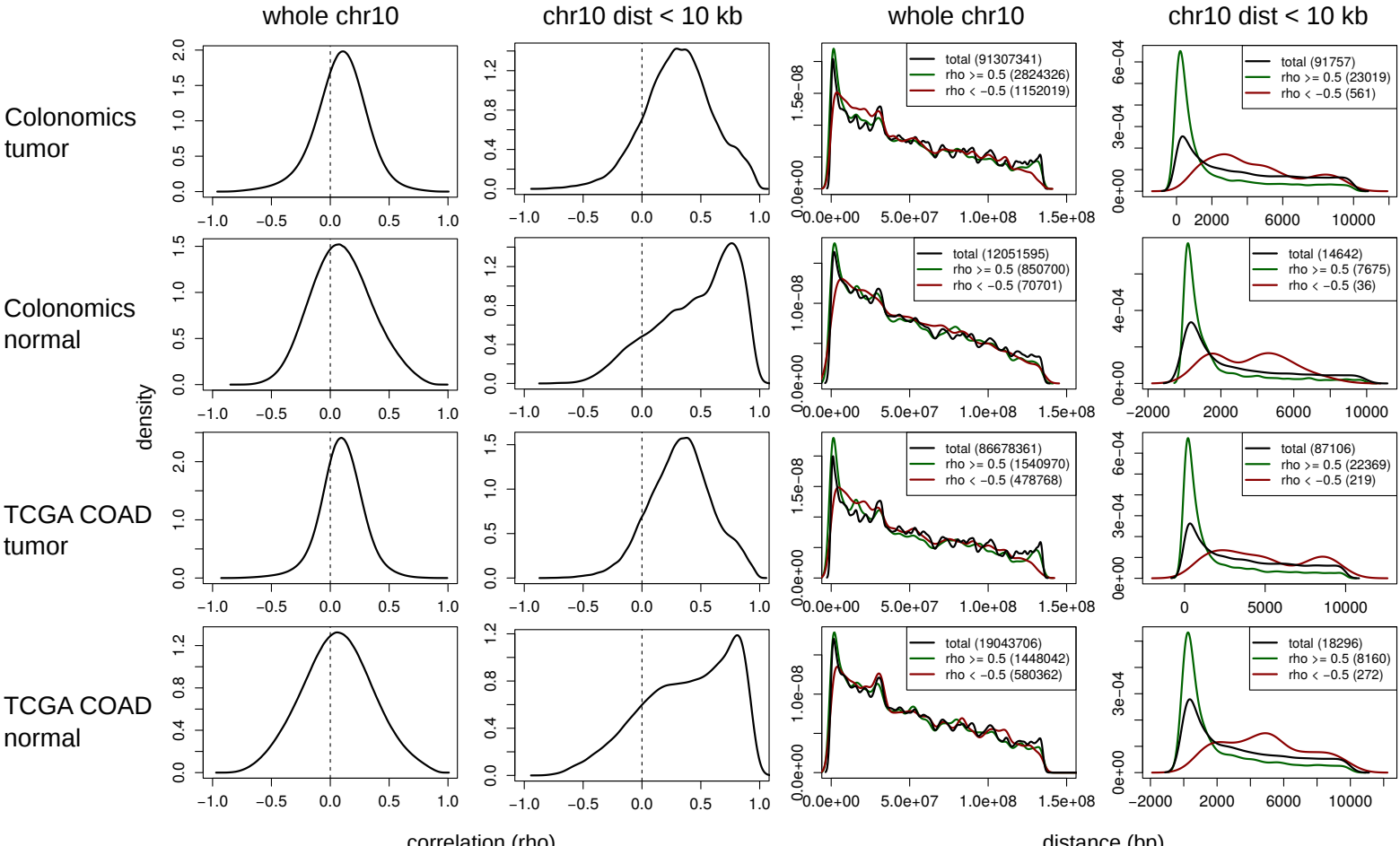
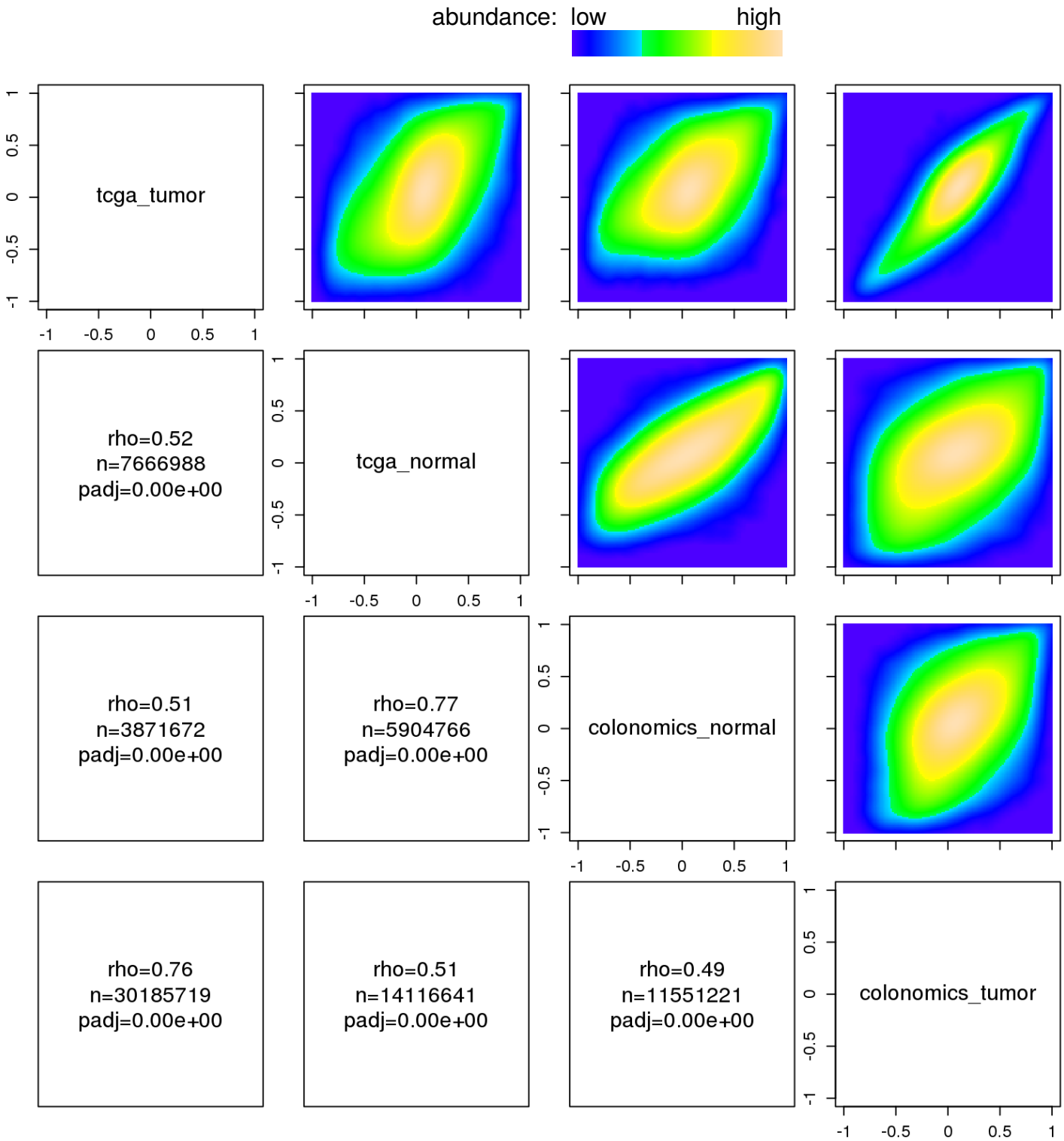
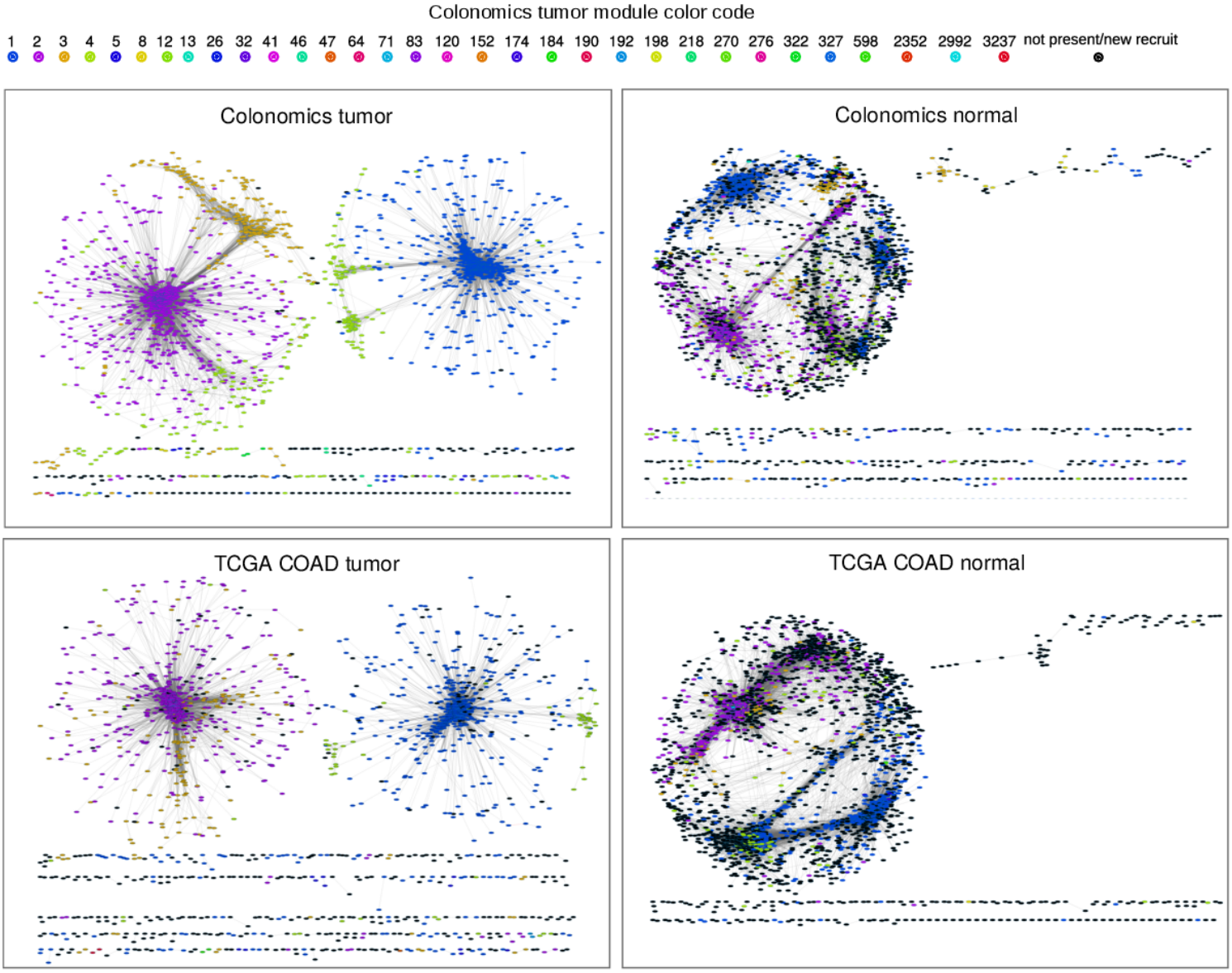


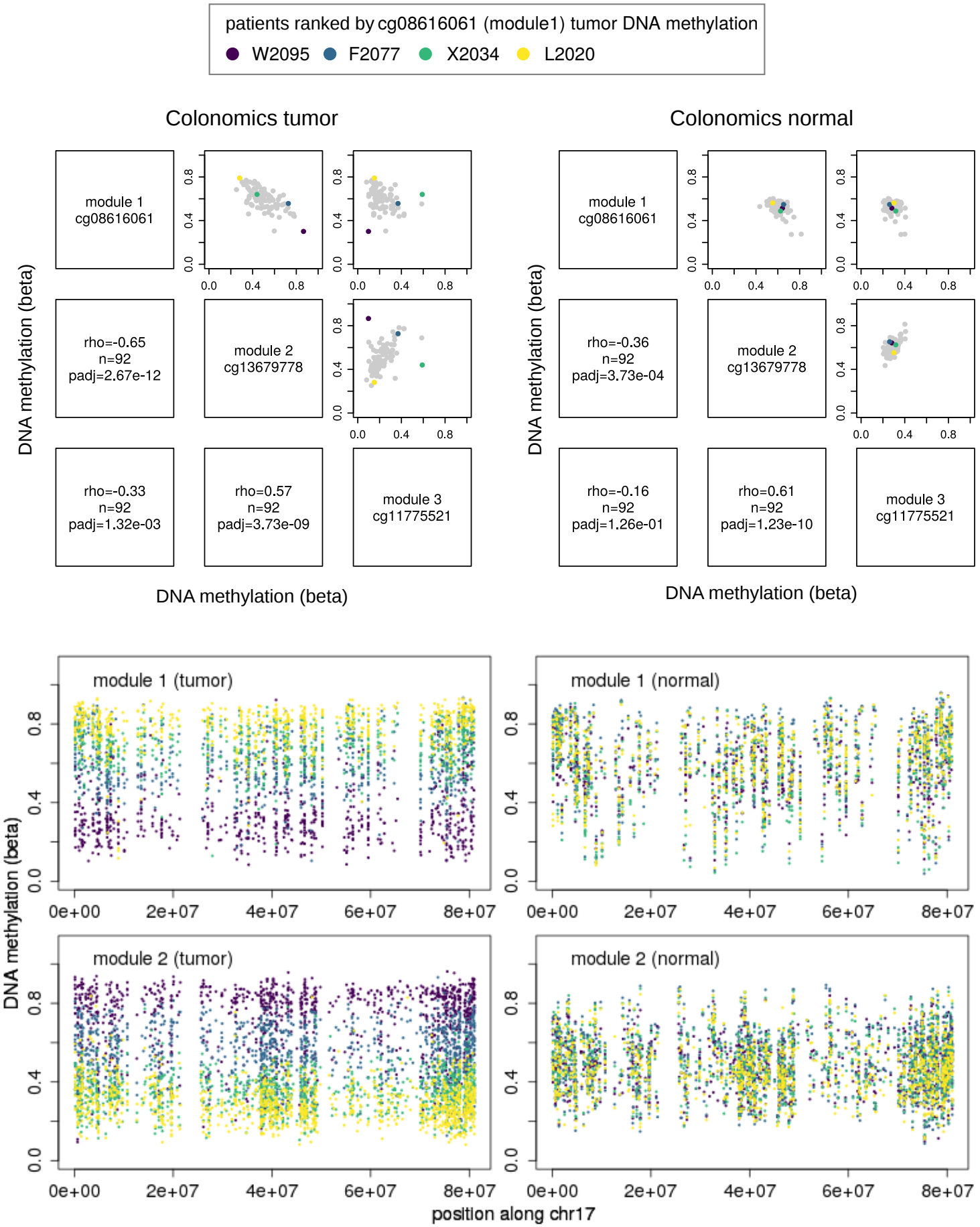
Figure 5: Opposed DNA methylation ranks. Tumor samples show opposed DNA methylation ranks at module 1 and module 2. Colored dots depict four patients sorted according to their tumor DNA methylation values at module 1.


A
chr1
chr2
CpGs
cis-
trans-
co-methylation
B
data
Infinium 450k array
colon (normal and tumor)
Colonomics.org
TCGA
pairwise co-methylation grid
cg1
cg2
cg3
cgn
co-methylation network
C
association to features, modules comparison, samples stratification, latent structures


whole chr10
chr10 dist < 10 kb
whole chr10
chr10 dist < 10 kb
Colonomics tumor
Colonomics normal
TCGA COAD tumor
TCGA COAD normal
density
total (91307341)
rho >= 0.5 (2824326)
rho < −0.5 (1152019)
total (91757)
rho >= 0.5 (23019)
rho < −0.5 (561)
total (12051595)
rho >= 0.5 (850700)
rho < −0.5 (70701)
total (14642)
rho >= 0.5 (7675)
rho < −0.5 (36)
total (86678361)
rho >= 0.5 (1540970)
rho < −0.5 (478768)
total (87106)
rho >= 0.5 (22369)
rho < −0.5 (219)
total (19043706)
rho >= 0.5 (1448042)
rho < −0.5 (580362)
total (18296)
rho >= 0.5 (8160)
rho < −0.5 (272)
correlation (rho)
distance (bp)


abundance: low high
tcga_tumor
rho=0.52
n=7666988
padj=0.00e+00
tcga_normal
rho=0.51
n=3871672
padj=0.00e+00
rho=0.77
n=5904766
padj=0.00e+00
colonomics_normal
rho=0.76
n=30185719
padj=0.00e+00
rho=0.51
n=14116641
padj=0.00e+00
rho=0.49
n=11551221
padj=0.00e+00
colonomics_tumor
-1
-0.5
0
0.5
1


Colonomics tumor module color code
1 2 3 4 5 8 12 13 26 32 41 46 47 64 71 83 120 152 174 184 190 192 198 218 270 276 322 327 598 2352 2992 3237 not present/new recruit
Colonomics tumor
Colonomics normal
TCGA COAD tumor
TCGA COAD normal


patients ranked by cg08616061 (module1) tumor DNA methylation
W2095
F2077
X2034
L2020
Colonomics tumor
Colonomics normal
module 1
cg08616061
module 2
cg13679778
module 3
cg11775521
rho=-0.65
n=92
padj=2.67e-12
rho=-0.33
n=92
padj=1.32e-03
rho=0.57
n=92
padj=3.73e-09
rho=-0.36
n=92
padj=3.73e-04
rho=-0.16
n=92
padj=1.26e-01
rho=0.61
n=92
padj=1.23e-10
DNA methylation (beta)
module 1 (tumor)
module 1 (normal)
module 2 (tumor)
module 2 (normal)
position along chr17