Modular dynamics of DNA co-methylation networks exposes the functional organization of colon cancer cells' genome

Authors:

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

Affiliations:

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.

* Correspondence:

Miguel A. Peinado, email: mpeinado@igtp.cat (Lead contact)

Izaskun Mallona, email: izaskun.mallona@gmail.com

$ Current affiliation:

SIB Swiss Institute of Bioinformatics, Zurich, Switzerland

Summary

DNA methylation dynamics is intrinsically interconnected with processes underlying the malignant properties of cancer cells. By applying network-based approaches in two series of colorectal cancers we dissected the long-range co-methylation structure finding consistent patterns of compartmentalization in both normal and tumor tissues. Large transchromosomal modules showed unique regulatory signatures and coalesced into a structured network and allowing simple patient stratification. Normal-tumor comparison revealed substantial remodeling of specific modules and migration of subsets of co-methylating sites denoted by functional aggregates, pointing out potential sources of epigenetic and phenotypic variability. We conclude that DNA methylation dynamics architecture embodies interpretable information that can be used as a proxy of the drivers and the phenotypes of malignant transformation.

Significance

DNA methylation is a key epigenetic mark directly involved in genome organization and regulation. DNA methylation profiles are variable and are extensively altered in most cancers. We show that DNA methylation variability follows a transchromosomal modular dynamics in both normal and colon cancer cells. The reshaping of the DNA methylation variability network in tumorigenesis exposes genomic and functional associations and points out both the mechanisms and the phenotypes of individual tumors. This information may be used for patient stratification and identification of disrupted pathways and therapeutic targets.

Highlights

- DNA methylation variability displays a modular architecture in normal and cancer.
- Coordinated transchromosomal variations supersede regional DNA methylation dynamics.
- Co-methylation network modularity evinces functional and structural features.
- Epigenetic rewiring can be used as patient stratifier.

Keywords

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

48 ## Running title

49 DNA co-methylation architecture in colon cancer

50

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 (Feinberg et al., 2016; Portela and Esteller, 2010). 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; Feinberg et al., 2016; Jones, 2012; Rodriguez et al., 2006; Schubeler, 2015). Direct comparison of the DNA methylation profiles in the tumor versus the paired normal tissue reveals both losses (hypomethylation) and gains (hypermethylation) of the epigenetic mark. The extent of the change may range from discrete sites and promoters to large regions (Feinberg et al., 2016; Frigola et al., 2006; Hansen et al., 2011; Jones, 2012; Portela and Esteller, 2010).

Neighboring CpGs have a higher chance of being similarly methylated (Barrera and Peinado, 2012; Eckhardt et al., 2006; Libertini et al., 2016; Shoemaker et al., 2010); 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 (Akulenko and Helms, 2013; Fortin and Hansen, 2015; Li et al., 2010; Salhab et al., 2018). Most studies about the functional impact of DNA methylation changes have focused the analysis on local effects on neighboring genes (Jones, 2012; Schubeler, 2015). More recently, taking advantage of the availability of genome-scale DNA methylation data from large 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), gene centered analyses (Gao and Teschendorff, 2017; Li et al., 2014) and modeling of DNA methylation variation (Jenkinson et al., 2017; Libertini et al., 2018; Rulands et al., 2018; Teschendorff and Relton, 2018).

We hypothesize that epigenetic phenotypes exposed by DNA methylation co-variation 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 a novel network-based synthetic analysis. Recent leading studies have proposed network based elucidation of molecular determinants of disease (Creixell et al., 2015; Chen et al., 2014; Liu et al., 2016). Our

rationale is that cells 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. Concomitantly, the linkage between the events and the emergence of relevant traits alters the epigenome. The heterogeneity of the samples, which might be classified in a wide assortment of states and transitions, challenges the final state-focused differential methylation analysis (the one resulting in regional hyper-or hypomethylations), whilst favoring more process-oriented, flexible co-variation analysis, which also unfolds variability. Here, we have scrutinized over 300,000 individual CpGs in two colon cancer datasets to extract and characterize the highly connected co-methylation modules. The structural and functional insights of epigenomic modules are dissected providing a framework to disentangle cancer cell's genome functional reorganization.

Results and Discussion

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 from the Colonomics cohort (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 (Table S1). 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 or presumably non significant. Finally, probes with missing data in any sample were eliminated (Table S1). No detectable batch effects were found (Figure S1). Next we sequentially calculated the bulk pairwise Spearman's correlations between any possible pair of probes adjusting for multiple testing (Table S2).

Co-methylations were detectable even at long distances (Figure 2A) and did not depend on local probe density (Figure S2A). Correlation coefficients ρ (a measure of association 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 (Figure 2), independently of the CpG location in open or closed chromatin compartments (Figure S3). The ρ distribution was not centered to 0 but shifted towards positive values, thus indicating a

trend towards co-methylation changes. That is, the detected changes in DNA methylation correspond to either the increase or decrease of the scrutinized CpGs altogether, and not in opposite directions (e.g. inverse associations). The trend to co-methylate was noticeably increased in cis, being close CpGs' distribution negatively skewed: co-methylations were enriched at short distances, whereas anti-methylations (negative correlations) were not (Figure 2A).

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 (see Supplemental Methods) 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 coefficient $\rho \geq 0.8$, 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 S2) (Fisher, 1915; Shakhbazov et al., 2016). The absence of notorious clustering of DNA methylation values (Figure S1) 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). To reinforce this, nonparametric Spearman correlations were run, which rely on DNA methylation ranks rather than values and are therefore more resistant to outliers (Croux and Dehon, 2010). As for the Infinium array design, neither the probes GC content (Figure S4) nor the dye channel (Figure S5) drove the correlations structure; probes mapping to multiple locations or overlapping to SNPs were filtered out (Price et al., 2013).

Anatomy of the co-methylating network in colorectal cancers

We selected the top-scoring correlations ($\rho \geq 0.8$) and assembled a network in which loci and correlations are represented by nodes (vertex) and links (edges) (Figure 1B). For the sake of simplicity, we only considered edges with positive correlations. The Colonomics tumor series network resulted in 63,130 nodes and 26 million connections (Table S3). 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 (Figure S6). The degree distribution did

not resemble power-law, lognormal nor exponential (goodness of fit, Kolmogorov Smirnov tests, $p \geq 0.05$; Figure S7), as there was not linear dependency between the cumulative frequencies and the connectivities, as in other biological quantities spanning several orders of magnitude and heavily skewed to the right (Newman, 2005). This structure was unaffected by loci features such as chromatin state (Figure S8) and genomic category (Figure S9); but lost when filtering out trans interactions, as probes placed at any distance in cis showed power-law compatible distributions (Figure S10). Interestingly, the 99th percentile of the most connected trans-comethylators presented homogeneous intermediate DNA methylation levels in normal samples (Figures S11 and S12), with an important enrichment of imprinted loci (n=53, 11%). Interestingly, partially methylated domains (PMDs) have been reported as loci with intermediate DNA methylation values and high variability (Lister et al., 2009), which is consistent to the top connected co-methylated probes; however, we did not find an enrichment in them: 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).

To test the reproducibility of the network, the analysis was repeated using an independent dataset, the COAD cohort from TCGA, consisting of 256 primary colon adenocarcinomas. In TCGA dataset, the DNA methylation β value calling procedure differs from Colonomics', and therefore reduces the chance of covariation artifacts arising due to the data processing bias. No batch effects were detected (Figure S1).

Near a quarter million probes fulfilled the variability criteria in TCGA colon tumors (Tables S6 and S7). The overall correlations distribution and the co-methylation decay with distance matched that of Colonomics' (Figure 2A and Figure S2B). 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 Colonomics' ρ values of each CpG pair against TCGA's. The linearity of both landscapes (Figure 2B) indicated a high concordance of the overall co-methylation levels.

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 Colonomics and TCGA

tumor co-methylation datasets showed a reproducible distribution of nodes' PageRanks mostly composed by lowly influential CpGs (Figure S13).

Next, we modeled the results as a network keeping the ρ cut-off at 0.8 even though the multiple testing adjusted significance cut-off at TCGA cohort admitted lowering it due to the larger sample size (Table S2). The network comprised 37 thousand nodes and around eight million edges (Table S3). As in the Colonomics cohort, the TCGA network degree distribution showed a long tail, indicating vast differences in connectedness (Figures S6 and S7).

Coalescent embedding of normal co-methylation networks in the tumor networks

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's one. The normal colon co-methylome network was built using 90 non-tumor tissues from the Colonomics cohort. Given the equivalent sample sizes, we maintained the ρ cutoff unaltered (Table S2). 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). Strikingly, negative correlations (cut-off $\rho < -0.8$) showed a >1,000-fold reduction and dropped from 7.6 million in tumor to less than 7 thousands in normal tissue (Table S7). In agreement with the associations found in tumors, co-methylations were underrepresented in active promoters (Table S8, Figure S14) and the co-methylation network' connectivities were not power-law distributed (Figures S6 and S7). Probes pairs correlation values showed partial agreement between the normal and tumor datasets (Figure 2B), with differences being more conspicuous at the node influence level, pointing to changes in network connectivity (Figure S13).

We repeated the analysis with TCGA normal colon samples. It should be noted that this dataset only includes 38 normal samples (Table S2), 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 (ρ = 0.96, Table S2) 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 S4 and Table S9), whose differences are especially conspicuous at short ranges (Figure 2A) and, importantly, in nodes connectivity.

Co-methylating networks display a modular structure in normal and tumor tissue

Colonomics tumor network had two noticeable giant components (Figure 3), that were not present in the normal tissue, indicating a major restructuration of co-methylation architecture associated with malignant transformation, as we will discuss below. TCGA tumor and normal co-methylation networks replicated Colonomics overall networks structure (Figure 3).

In order to dig into the network preferential attachment, 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 (Fortunato, 2010; Newman and Girvan, 2004). 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 S3), and using the Clauset's fast greedy method (Clauset et al., 2004) we partitioned it into 3,270 modules ranging from two to 18,727 nodes. Interestingly, the normal tissue network exhibited a higher co-methylome modularity (0.62) (Table S3) and network segmentation resulted in 1,265 modules ranging from two to 17,758 nodes. The co-methylation modules retained tight correlation structure after subtracting purity effects (Zheng et al., 2017) (Figure S16).

The vast majority of the small modules were, in fact, composed by sets of probes located at close distance from each other (e.g. at CpG islands), so we discarded them and focused in transchromosomal modules, with at least 10 members and located at least 1 Mbp apart or placed in different chromosomes. The number of transchromosomal modules was 32 (1%) in the tumor cohort and 18 (1.4%) in the normal cohort (Tables S4 and S5).

In agreement with Colonomics's results, TCGA tumor co-methylation network was also modular (modularity score 0.41) (Table S3) and segmentation produced 3,421 modules ranging from two to 8,981 nodes. The application of size and co-location filters reduced the number of transchromosomal modules to 35.

Next, we evaluated the degree of conservation of the whole network partitioning into modules across the four datasets using the adjusted Rand statistic. In this test, the distance measure can be interpreted as a probability, being zero when the congruence is expected by chance and one when the matching is perfect. It should be noted that the Rand statistic renders negative values when finding anti-associations (Hubert and Arabie, 1985). Networks clustering on adjusted Rand's distance indicates that modules memberships separate tumor's from normal's networks in both datasets (Figure S15A), in line with the similarities in nodes population (Figure S15B), and their spatial co-methylation patterns (Figure S2B). Overall conservation of co-methylomes structure and connectivity as well as the differences between normal and tumor samples was noticeable by visual inspection (Figures 3, S17 and S18). 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 (Appendix 1, Figure S18).

Module preservation across tissue types and cohorts was also evaluated by cross-tabulation of the number of shared CpGs (Table S10). Twelve Colonomics tumor modules had one or more counterparts in the normal tissue network, and a similar number in TCGA tumor cohort (Fisher's exact test, $p<0.0001$) (Tables S10, S11). Strikingly, the five-top sized Colonomics tumor modules partially matched to multiple TCGA's modules (Table S11). This result is in concordance with the resolution limit of modularity-optimizing module detecting algorithms, which tend to aggregate modules into few giant components, disregarding their inner complexity (Fortunato and Barthelemy, 2007).

Co-methylating module membership evinces 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. It should be noted that a large subset of HumanMethylation450k probes are located in promoters and promoter-related features. Thus, the specific design of the HumanMethylation450k array may introduce biases as it oversamples TSS-related features, in which clusters of probes are located, and disregards other genomic compartments (Bibikova et al., 2011; Sandoval et al., 2011; Silva-Martinez et al., 2017). Moreover, coordinated co-methylation is expected among neighboring CpGs within each one of these genomic elements (Barrera and Peinado, 2012; Gaidatzis et al., 2014; Kim et al., 2008; Libertini et al., 2016; MacDonald et al., 2015; Wang

et al., 2016). Therefore, enrichment analysis for genomic features and compartments were corrected according to the HumanMethylation450k array background.

The tumor modules displayed important differences in feature enrichment, including chromatin states, genomic categories, CpG islands and association with known motifs (Figure S19 and Appendix 1). In line with the weighted gene co-expression network analysis, in which functional signatures can be told apart by mining gene co-expression (Horvath et al., 2012), the module-specific co-regulation patterns denoted by distinctive abundance of genomic and functional states (Figure S19; Appendix 1) of co-methylation modules pointed out the existence of a latent structure. Among the multiple features analyzed, a striking global enrichment of inactive promoters was observed in a large number of modules (Figure S20), pointing out potential clusters of co-regulated genes.

Next, we explored the overlapping of the co-methylation modules with regions of DNA methylation variability previously reported in colon cancer (Hansen et al., 2011). Interestingly, seven out of 32 Colonomics tumor modules significantly overlie tumor hypermethylated blocks (Figures S21 and S22 and Table S12). Regarding other types of DNA methylation variability reported by Hansen, 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, suggesting that they might reflect different mechanisms.

A comprehensive summary of structural and functional feature enrichment for each co-methylation module is shown in Appendix 1 with the top associations listed in Table S13). 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 underepresentation of co-methylations within active promoters (Figure S14).

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 S14). Enriched motifs included

ETS and RUNX families (modules 1, 3, 327), FOS family members (Fra1, Atf3, BATF, Fosl2, AP-1, Jun-AP1; modules 2 and 4), FOXA1-related (FOXA1, HNF4a, FOXMA; module 2), GC box (KLF5 and KLF4; module 2), C/EBP (module 3), PAX7 and MYF5 (module 5), homeobox (modules 5, 83), MADS (module 152), ASCL1 (module 598) (Appendix 1).

In summary, the functional signatures of DNA co-methylation modular architecture evince the putative mediators of epigenetic remodeling and signaling reprogramming in colorectal cancer. We postulate that coordinated DNA methylation changes at interspersed sites (here identified as belonging to the same module) regulate signaling pathways and biological functions. This hypothesis is supported by recent studies demonstrating the DNA methylation mediated binding of transcription factors to specific sites with a direct impact in gene regulation (Kribelbauer et al., 2017; Yin et al., 2017).

Modeling of module's DNA methylation variation allows categorization and study of feature associations in new samples

As shown above, modules depict shared patterns of co-methylations (network edges) which emerge from structured DNA methylation levels among loci (network nodes). To dissect the latter, we applied a samples stratification procedure based on the DNA methylation status of their CpGs (Supplemental methods) that results in the partition of each module into two to three DNA methylation profiles (Figure S23). The putative effects of tumor purity to DNA methylation levels were identified and subtracted (Figure S24).

This methodology provides with a powerful tool to explore potential correlates of DNA methylation profiles with molecular and biological features, including clinical data, and importantly, enabling the model to classify new samples and to make predictions without computing new correlations. An in-depth exploitation of this approach is beyond the scope of this paper, but as a proof of concept, we evaluated whether module cluster membership conveyed gene expression signatures to tumor samples. The pairwise differential expression between samples belonging to different module clusters was computed (adjusted $p < 0.1$ cut-off, Figure S25) in both cohorts of colon tumors. TCGA cohort consistently exhibited a higher number of differentially expressed genes. This result may be explained by the larger size of this series and the use of RNA-Seq, that has more sensitivity than the microarrays (Zhao et al., 2014) applied in the Colonomics. Nevertheless, the overall gene over- and down-expression trends were maintained across cohorts (Figure S26). The top 50 significant

differentially expressed genes in both cohorts are listed in Appendix 2 for each one of the 32 Colonomics tumor modules.

Dynamics of co-methylation modules reveals epigenetic rewiring of defined genomic compartments in cancer

As noted above, the tumor co-methylation network displayed a striking disjoint structure visualized as two giant compartments (Figure 3). The emerging large compartment, not present in normal tissue, spanned multiple modules (Figure 3 and Figure S27) and was funneled by DNA methylation negative correlations between modules (Figure 4A). The coordinated inversion of DNA methylation variation affected hundreds or even thousands of sites throughout the whole genome (Figure 4B). The pervasive nature of anticorrelations overcame age, gender, tumor stage and anatomical site potential effects on modules' DNA methylation variation (Figure S28A). Loci with copy number alterations also conveyed the module-specific DNA methylation ranks mimicking the profiles along balanced regions (Figure S28B).

To further dissect the co-methylation dynamics we analyzed the modules preservation between normal and tumor. Significant equivalences were found for most modules (Figure S29): normal modules N2 and N3 largely overlapped with tumor modules T1 and T2 respectively (Figure S29), which suggests the preservation of module's structure and co-methylation links. A large overlap of associated genes among modules was also observed (Figure S29C). At the functional level, the preserved modules showed specific enrichments. For instance, normal module N2 intersection with tumor module T1 (N2∩T1) showed enrichment for RNA transcription and metabolic processes and DNA binding functions (Figure S30) and a high proportion of probes were located in CpG islands (77%) with a clear trend towards tumor hypermethylation (Figure 5B, Figure S31).

Next we analyzed the dynamics of CpG sites between normal and tumor modules. For the sake of simplicity only probes in the four largest modules in the Colonomics normal and tumor series were considered for differential module membership. A particular case was the scattering of normal module N1 probes into different tumor modules (Figure 5A), including the hijacking of several hundreds of sites by modules with inverse correlations, e.g.: tumor modules T1 and T2 (Figure S27). The subsets of probes flowing from the normal module N1 to each one of the tumor modules (intersections between normal module N1 and the tumor modules T1 to T4) were associated to subsets of genes with limited overlap (Figure 5C) and

displayed distinctive genomic features in regard to gene regulation: N1∩T1 members were enriched in CpG islands, while N1∩T2 were depleted and N1∩T3 were frequently near the TSS (Figure 5D). The tumor-normal DNA methylation signatures were in concordance with the preferential genomic location of probes: members of the intersection N1∩T1 were characterized by the prevalence of hypermethylations in the tumor, while the rest showed a clear trend towards hypomethylation (Figure 5E and Figure S31).

As a whole, our analysis points out an overall preservation of co-methylation modules in the normal-tumor transformation concomitantly with an important dispersal of subsets of sites with distinctive features into tumor modules. The tumor redefined modular landscape appears to have biological insights: the subsets of sites flowing from one module to another (denoted here as normal-tumor module intersections) display differential enrichments in functional and biological processes involved in cancer transformation (Figure 5E and Appendix 2). Some of the affected signaling pathways, including polycomb regulation, chromatin binding and genes defining epithelial-mesenchymal transition appear as the usual suspects contributing to the epigenetic reshaping of genomic compartments and the functional reprogramming of cancer cells (Appendix 2).

Surfing the co-methylating networks pinpoints functional sites

Beyond the remarkable functional and structural features of co-methylation modules revealed by this analysis, it has not escaped our attention that the stored data provide an excellent resource to carry out an insightful tracing of individual correlations. A detailed analysis of the data at this level is beyond the scope of this paper, but as a simple shortcut to navigate the co-methylating network and their associated functional features we developed a web tool "*corre*" (available at http://maplab.cat/corre). To illustrate the discovery potential 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 20kb upstream and downstream of the gene in both normal and tumor samples (Figure 6). 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 6E), 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 S32). Another interesting result was in regard to cg03699182 probe (Figure 6D, arrowhead) located in the CpG island of the INHBB promoter presented 42 co-methylations (ρ >0.8) 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 S32). 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., 2018).

Final considerations

Dissection of DNA methylation encoded information offers a far-reaching gamut of insights into genome biology (Jones, 2012; Schubeler, 2015), including the inference of genome architecture as demonstrated by recent studies (Fortin and Hansen, 2015; Jenkinson et al., 2017; Jorda et al., 2017; Raineri et al., 2018; Zhang et al., 2017). As a new inquiry, here we report a novel and robust analysis of coordinated DNA methylation dynamics in non-contiguous CpGs in two cohorts of colon normal and cancer tissues. This application provides a reproducible and synthetic network representation of cell's epigenome meta-structure and unveils modules or genomic territories of highly connected loci. The co-methylation modules comprise regions displaying common structural and functional features pointing out putative drivers of variability. Despite wide overlapping between normal and tumor tissue networks, striking differences in connectivity reveal specific patterns of functional rewiring and convey gene expression signatures with a potential impact on cancer cell biology.

Our data present a remodeled epigenetic landscape of colon cancer cells outlined by coordinated DNA methylation variations superseding the stochastic nature of DNA methylation dynamics (Jenkinson et al., 2017; Landan et al., 2012; Pujadas and Feinberg, 2012). The model can be visualized by a scrambled Rubik's cube resulting from just of a few flips (Figure S33A). In our case, cube's pieces correspond to the set of loci with coordinated methylation, and the axes would be the mechanisms flipping one or more modules (Figure S33B). This metaphor has two important corollaries with the corresponding epigenetic representations:

- Pieces linked by connectors move together, which implies that any specific scrambled conformation is the result of specific flips. Moreover, not all the arrangements are possible unless the cube is disassembled and reassembled. Similarly, cancer cell methylome dynamics is determined by sequential activation/inactivation of a limited number of mechanisms affecting genomic regulation. Interestingly, chromosomal rearrangements would provide an additional level of reshuffling that would be equivalent to reassembling Rubik's cube.
- The scrambled cube may be solved by predictable flips that do not imply the reversal of the flips that generated it. Currently we can only speculate, but this means that knowing the mechanisms governing epigenetic programs, it would be possible to design a strategy to reconstruct a "normal" epigenome by just turning on/off the appropriate switches and in the right sequence.

Summing up, our approach aims to offer a contextual view of the cancer epigenetic landscape to better define their nature and their eventual impact on the disease. The use of DNA co-methylation architecture to portrait the complex genome regulation scenario aims to provide a feasible surrogate marker that can be easily assessed in prospective clinical settings (e.g. response to treatment).

Materials and methods

Two colon cancer datasets were used. Colonomics (http://www.colonomics.org) series 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), 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 Array, gene expression by RNA-Seq counts and somatic mutations by exome sequencing.

A scheme summarizing 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 (primary dataset) and normal adjacent tissue, as well and both the 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 using previously described approaches (Clauset et al., 2009; Csardi and Nepusz, 2006; Cullen and Frey, 1999; Delignette-Muller and Dutang, 2015; Gillespie, 2014; Saha et al., 2017; Zhang and Horvath, 2005), from which highly connected modules according to the fast greedy community detection algorithm were isolated (Clauset et al., 2004; Csardi and Nepusz, 2006). Module members were further classified into major DNA methylation clusters using kNN (Venables and Ripley, 2002) taking into account not only the co-methylation but the purity-corrected DNA methylation status (Aran et al., 2015; Zheng et al., 2017) of their members (i.e. consistently lowly or highly methylated) (Chang et al., 2010; Wang et al., 2007). Next, modules and/or profiles were functionally annotated according to public datasets (Aryee et al., 2014; Fortin and Hansen, 2015; Hansen et al., 2011; Lister et al., 2009), molecular features databases (Heinz et al., 2010; Liberzon et al., 2011) and expression signatures (Gel et al., 2016; Love et al., 2014; Quinlan and Hall, 2010; Smyth, 2005). Modules characterization, including reproducibility assessment, consisted in mutual profiles comparison and differential expression analysis among different cohorts (Akdemir and Chin, 2015; Hubert and Arabie, 1985; Krzywinski et al., 2009; Langfelder et al., 2011; Shannon et al., 2003).

A Web application "Corre" has been implemented to facilitate browsing the DNA co-methylation events of investigator's favorite locus or gene in the both the Colonomics and TCGA COAD datasets. 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 (Conway et al., 2016; Ernst et al., 2011; Gesmann and de Castillo, 2011; Zhang et al., 2013). Source code is available at https://bitbucket.org/imallona/corre under the GPL terms. Corre can be accessed freely and without registration at http://www.maplab.cat/corre.

Extended methods are available in supplementary material.

Acknowledgements

We thank Iñaki Martínez de Ilarduya for his excellent technical support and Francisco Chen for help parallelizing the sequential correlations. The data presented here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Declarations

Authors' contributions

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

Samples, Ethics approval and consent to participate

Colonomics samples were collected at the Bellvitge Hospital. The Clinical Research Ethics Committee (CEIC) of the Bellvitge Hospital approved the study protocol, and all individuals provided written informed consent to participate and for genetic analyses to be done on their samples. The approval number is PR178/11. Additional information about the study and patient samples can be found at http://www.colonomics.org. TCGA data were obtained at http://cancergenome.nih.gov/.

Availability of data and material

The code is available at: https://bitbucket.org/imallona/correlations under the GPL v3 terms. The *Corre* Web tool can be freely accessed at http://maplab.cat/corre; its source code is available at https://bitbucket.org/imallona/corre under the GPL v3 terms. Colomics data may be accessed at http://colonomics.org. TCGA data were obtained at http://cancergenome.nih.gov/.

Competing interests

MAP is cofounder and equity holder of Aniling, a biotech company with no interests in this paper. 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 to MAP), the Agency for Management of University and Research Grants (AGAUR) of the Catalan Government grants 2017SGR723 and 2017SGR529. The funding agency had no role in the design of the study, collection, analysis, interpretation of data nor manuscript writing.

Figure Legends

Figure 1. Co-methylation analysis framework. **A**, CpGs co-methylation occurs at close (cis) and long distances (trans). **B**, DNA co-methylation networks display a modular structure in normal and tumor samples. **C**, Co-methylation modules display differential genomic and functional signatures. **D**, Analysis of normal-tumor co-methylation dynamics points out cancer pathways and mechanisms.

Figure 2. A, Correlation distribution in normal and tumor samples among probes located in chromosome 10. The bell-shaped distribution and shifted towards positive values at the tumor cohort (whole chromosome 10); 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 ($\rho \geq 0.5$). **B**, Correlations replication across cohorts. For each cohort, a pairwise correlation analysis was conducted for any probe with $sd \geq 0.05$. The correlation coefficient rho (X and Y axes) for each probe pair was plotted to check whether the co-methylation landscape was reproduced. Analysis was restricted to chromosome 10 Infinium450K probes.

Figure 3. Modular structure of colorectal tumor and normal tissue co-methylation networks. The networks were built independently for each dataset, but the nodes (CpGs) are colored using the Colonomics Tumor module membership. Nodes with no cross-representation are shown in black. Graphs are limited to a random sample of 5,000 nodes and solitary nodes are not plotted; network layout was calculated by $1 - \rho$ (edges) weighted springs.

Figure 4. The two largest tumor modules show opposed DNA methylation dynamics. A, Correlation of three randomly picked CpGs from modules 1 and 2 in 92 Colonomics tumor samples. B, DNA methylation ranks of modules 1 and 2 in four Colonomics tumor samples. Patients labeled with color codes as depicted in panel A.

Figure 5. **A**, Sankey diagram depicting balanced probes overlap between normal and tumor modules in the Colonomics datasets. Only intersections with >300 probes are annotated, and the number of associated genes is indicated. **B**, Distribution of probes according to the mean DNA methylation values in normal tissue (Y-axis) against the tumor-normal delta value (X-axis). Only probes overlapping in normal and tumor modules (intersections) are represented. A decomposed version of this figure is shown in Figure S31. **C**, Circos representation of modules associated genes overlap (purple connectors) and enriched terms sharing (blue connectors). **D,** Genomic context enrichment of normal module 1 intersections with tumor

559 modules. **E**, Gene set functional enrichment of normal module 1 intersections with the four
560 largest tumor modules.

561 **Figure 6**: Illustrative example of the *Corre* web tool usage. This application queries a user
562 selected Infinium array probe (i.e.: cg25924274) or a set of gene associated probes (i.e.:
563 *INHBB*) and renders graphs displaying the feature distribution of the anchor (preselected site)
564 and correlating CpGs, including DNA methylation levels, genomic element category, HMM
565 chromatin states, etc. In addition, tables containing genetic and functional information on the
566 correlating sites may be downloaded for further analysis. **A**, UCSC genome browser
567 representation of the region encompassing the preselected *INHBB* gene. **B**, The tool renders
568 graphs showing relevant features (see legends) for each one of the gene associated probes
569 (anchor CpGs). **C**, Distribution of DNA methylation levels in anchor CpGs and the
570 correlating sites. **D**, Sum of correlating sites in normal and tumor tissues for each anchor CpG
571 represented by chromatin state frequency. Positive (+) and negative (-) correlating sites show
572 distinct chromatin state profiles in the tumors. **E**, Genomic distribution of cg11513884 co-
573 methylating CpGs located in chromosomes 2 (red), 7 and 8 (blue).

574

575

References

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

Akulenko, R., and Helms, V. (2013). DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. Hum Mol Genet *22*, 3016-3022.

Aran, D., Sirota, M., and Butte, A. J. (2015). Systematic pan-cancer analysis of tumour purity. Nat Commun *6*, 8971.

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 Res *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., *et al.* (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., *et al.* (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). 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., *et al.* (2014). Large differences in global transcriptional regulatory programs of normal and tumor colon cells. BMC Cancer *14*, 708.

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C., *et al.* (2015). Pathway and network analysis of cancer genomes. Nat Methods *12*, 615-621.

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.

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).

Chang, F., Qiu, W., Zamar, R. H., Lazarus, R., and Wang, X. (2010). Clues: an R package for nonparametric clustering based on local shrinking. Journal of Statistical Software *33*, 1-16.

Chen, J. C., Alvarez, M. J., Talos, F., Dhruv, H., Rieckhof, G. E., Iyer, A., Diefes, K. L., Aldape, K., Berens, M., Shen, M. M., and Califano, A. (2014). Identification of causal genetic

drivers of human disease through systems-level analysis of regulatory networks. Cell *159*, 402-414.

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

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., *et al.* (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet *38*, 1378-1385.

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

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

Feinberg, A. P., Koldobskiy, M. A., and Gondor, A. (2016). Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. Nat Rev Genet *17*, 284-299.

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 Biol *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. Proc Natl Acad Sci USA *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. Nat Genet *38*, 540-549.

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

Gao, Y., and Teschendorff, A. E. (2017). Epigenetic and genetic deregulation in cancer target distinct signaling pathway domains. Nucleic Acids Res *45*, 583-596.

Gel, B., Diez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M. A., and Malinverni, R. (2016). 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. R Journal *30*, 40-44.

Gillespie, C. S. (2014). Fitting heavy tailed distributions: the poweRlaw package. arXiv preprint arXiv:14073492.

Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., *et al.* (2011). Increased methylation variation in epigenetic domains across cancer types. Nat Genet *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. Mol Cell *38*, 576-589.

Horvath, S., Zhang, Y., Langfelder, P., Kahn, R. S., Boks, M. P., van Eijk, K., van den Berg, L. H., and Ophoff, R. A. (2012). Aging effects on DNA methylation modules in human brain and blood tissue. Genome Biol *13*, R97.

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

Jenkinson, G., Pujadas, E., Goutsias, J., and Feinberg, A. P. (2017). Potential energy landscapes identify the information-theoretic nature of the epigenome. Nat Genet *49*, 719-729.

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet *13*, 484-492.

Jorda, M., Diez-Villanueva, A., Mallona, I., Martin, B., Lois, S., Barrera, V., Esteller, M., Vavouri, T., and Peinado, M. A. (2017). The epigenetic landscape of Alu repeats delineates the structural and functional genomic architecture of colon cancer cells. Genome Res *27*, 118-132.

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. Nat Genet *38*, 149-153.

Kim, S., Li, M., Paik, H., Nephew, K., Shi, H., Kramer, R., Xu, D., and Huang, T. H. (2008). Predicting DNA methylation susceptibility using CpG flanking sequences. Pac Symp Biocomput, 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 Rep *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 Res *19*, 1639-1645.

Landan, G., Cohen, N. M., Mukamel, Z., Bar, A., Molchadsky, A., Brosh, R., Horn-Saban, S., Zalcenstein, D. A., Goldfinger, N., Zundelevich, A., *et al.* (2012). Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nat Genet *44*, 1207-1214.

Langfelder, P., Luo, R., Oldham, M. C., and Horvath, S. (2011). Is my network module preserved and reproducible? PLoS Comput Biol *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. Nat Rev Genet *11*, 733-739.

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

Li, Y., Zhu, J., Tian, G., Li, N., Li, Q., Ye, M., Zheng, H., Yu, J., Wu, H., Sun, J., *et al.* (2010). The DNA methylome of human peripheral blood mononuclear cells. PLoS Biol *8*, e1000533.

Libertini, E., Hamoudi, R. A., Heath, S., Lancashire, L., Rubio Garcia, A., Grassi, L., Downes, K., Ouwehand, W. H., Javierre, B. M., Cairns, J., *et al.* (2018). DNA methylation oscillation defines classes of enhancers. bioRxiv.

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., *et al.* (2016). Information recovery from low coverage whole-genome bisulfite sequencing. Nat Commun *7*, 11306.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, 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., *et al.* (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315-322.

Liu, X., Wang, Y., Ji, H., Aihara, K., and Chen, L. (2016). Personalized characterization of diseases using sample-specific networks. Nucleic Acids Res *44*, e164.

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. Biochem Cell Biol *94*, 43-55.

Mayor, R., Casadome, L., Azuara, D., Moreno, V., Clark, S. J., Capella, G., and Peinado, M. A. (2009). Long-range epigenetic silencing at 2q14.2 affects most human colorectal cancers and may have application as a non-invasive biomarker of disease. Br J Cancer *100*, 1534-1539.

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 Res *68*, 5753-5759.

Michieletto, D., Chiang, M., Coli, D., Papantonis, A., Orlandini, E., Cook, P. R., and Marenduzzo, D. (2018). Shaping epigenetic memory via genomic bookmarking. Nucleic Acids Res *46*, 83-93.

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary physics *46*, 323-351.

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., *et al.* (2007). A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat Genet *39*, 237-242.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: bringing order to the web. Technical Report Stanford InfoLab.

Portela, A., and Esteller, M. (2010). Epigenetic modifications and human disease. Nat Biotechnol *28*, 1057-1068.

Price, M. E., Cotton, A. M., Lam, L. L., Farre, 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.

Pujadas, E., and Feinberg, A. P. (2012). Regulated noise in the epigenetic landscape of development and disease. Cell *148*, 1123-1131.

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

Raineri, E., Serra, F., Beekman, R., Garcìa Torre, B., Vilarrasa-Blasi, R., Martin-Subero, I., MartÍ-Renom, M., Gut, I., and Heath, S. (2018). Inference of genomic spatial organization from a whole genome bisulfite sequencing sample. bioRxiv.

Rodriguez, J., Frigola, J., Vendrell, E., Risques, R. A., Fraga, M. F., Morales, C., Moreno, V., Esteller, M., Capella, G., Ribas, M., and Peinado, M. A. (2006). Chromosomal instability correlates with genome-wide DNA demethylation in human primary colorectal cancers. Cancer Res *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. Proc Natl Acad Sci U S A *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., *et al.* (2018). Genome-Scale Oscillations in DNA Methylation during Exit from Pluripotency. Cell Syst *7*, 63-76 e12.

Saha, A., Kim, Y., Gewirtz, A. D. H., Jo, B., Gao, C., McDowell, I. C., Engelhardt, B. E., and Battle, A. (2017). Co-expression networks reveal the tissue-specific regulation of transcription and splicing. Genome Res *27*, 1843-1858.

Salhab, A., Nordström, K., Kattler, K., Ebert, P., Ramirez, F., Arrigoni, L., Müller, F., Cadenas, C., Hengstler, J., Lengauer, T., *et al.* (2018). Partially methylated domains are hallmarks of a cell specific epigenome topology. bioRxiv.

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., *et al.* (2015). Exome Sequencing Reveals AMER1 as a Frequently Mutated Gene in Colorectal Cancer. Clin Cancer Res *21*, 4709-4718.

Schubeler, D. (2015). Function and information content of DNA methylation. Nature *517*, 321-326.

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*, 278.

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 Res *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 Res *20*, 883-889.

Silva-Martinez, 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. Brief Funct Genomics.

Smyth, G. K. (2005). Limma: linear models for microarray data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, eds. (New York: Springer), pp. 397–420.

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

794 Teschendorff, A. E., and Relton, C. L. (2018). Statistical and integrative system-level analysis
795 of DNA methylation data. Nat Rev Genet *19*, 129-147.

796 Venables, W. N., and Ripley, B. D. (2002). Modern Applied Statistics with S, Fourth edn
797 (New York: Springer).

798 Wang, X., Qiu, W., and Zamar, R. H. (2007). CLUES: A non-parametric clustering method
799 based on local shrinking. Computational Statistics & Data Analysis *52*, 286-298.

800 Wang, Y., Liu, T., Xu, D., Shi, H., Zhang, C., Mo, Y. Y., and Wang, Z. (2016). Predicting
801 DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep
802 Networks. Sci Rep *6*, 19598.

803 Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K.,
804 Kivioja, T., Dave, K., Zhong, F., *et al.* (2017). Impact of cytosine methylation on DNA
805 binding specificities of human transcription factors. Science *356*.

806 Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression
807 network analysis. Statistical applications in genetics and molecular biology *4*, 1128.

808 Zhang, H., Meltzer, P., and Davis, S. (2013). RCircos: an R package for Circos 2D track
809 plots. BMC Bioinformatics *14*, 244.

810 Zhang, J., and Huang, K. (2017). Pan-cancer analysis of frequent DNA co-methylation
811 patterns reveals consistent epigenetic landscape changes in multiple cancers. BMC Genomics
812 *18*, 1045.

813 Zhang, L., Xie, W. J., Liu, S., Meng, L., Gu, C., and Gao, Y. Q. (2017). DNA Methylation
814 Landscape Reflects the Spatial Organization of Chromatin in Different Cells. Biophys J *113*,
815 1395-1404.

816 Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-
817 Seq and microarray in transcriptome profiling of activated T cells. PLoS ONE *9*, e78644.

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

820 Zhu, Y., Qiu, P., and Ji, Y. (2014). TCGA-assembler: open-source software for retrieving and
821 processing TCGA data. Nat Methods *11*, 599-600.

822

823

Figure 1

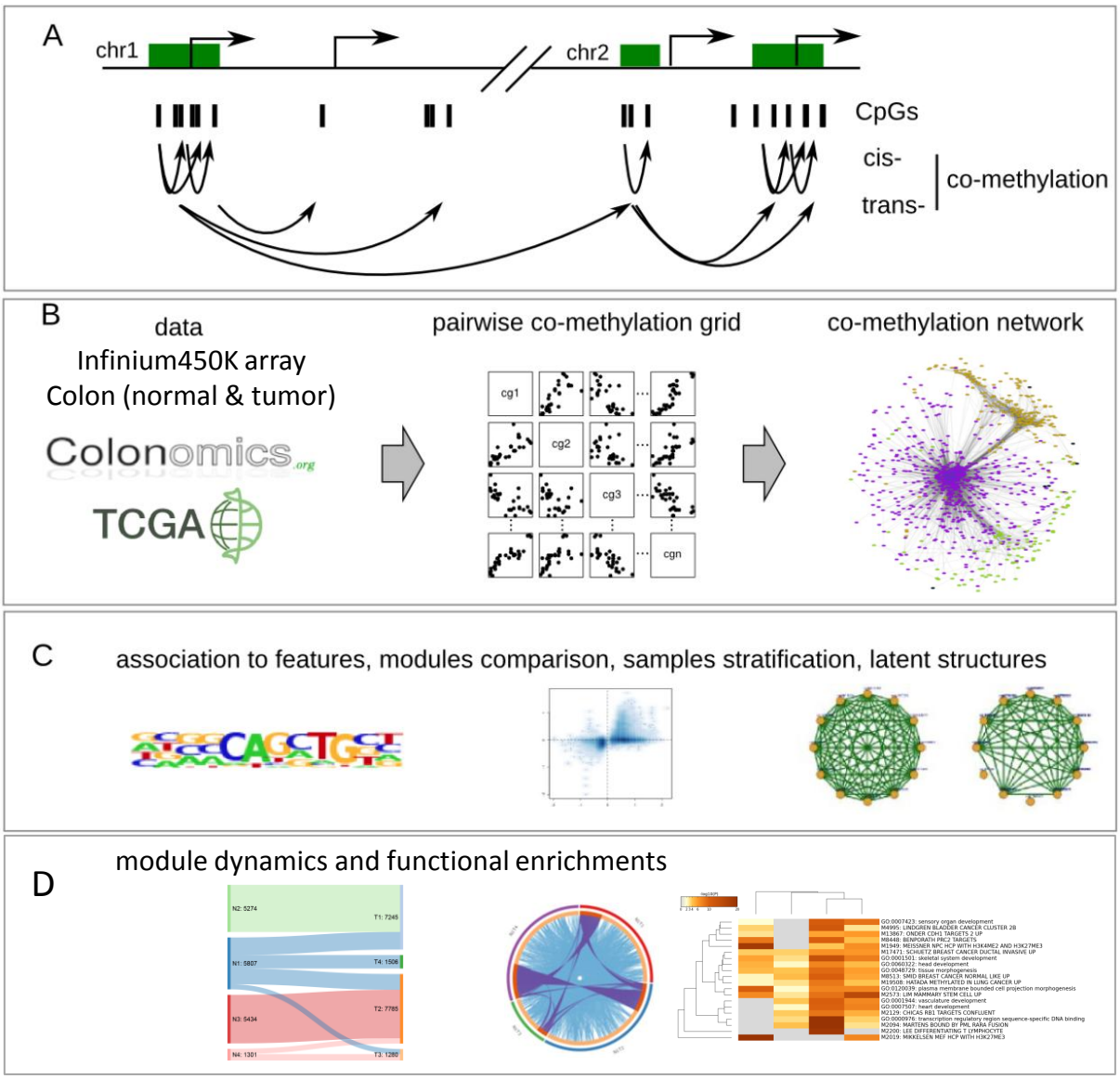
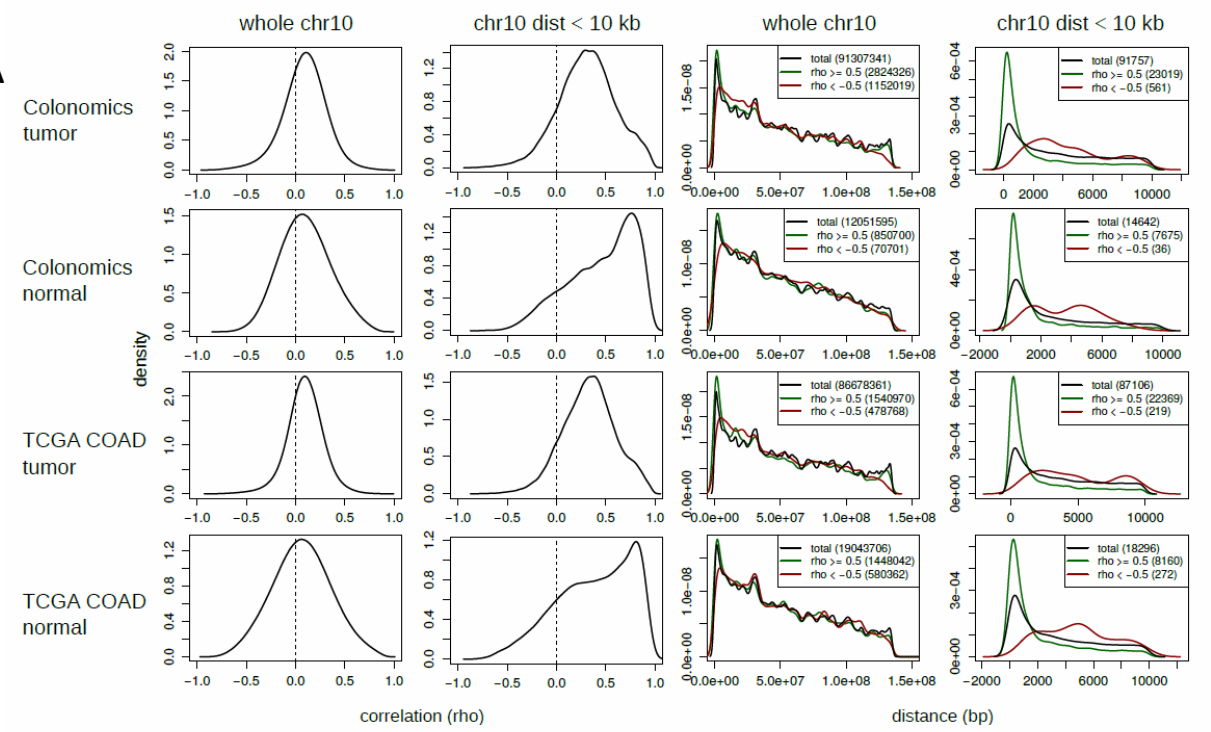

A
chr1
chr2
CpGs
cis-
trans-
co-methylation
B
data
Infinium450K array
Colon (normal & tumor)
Colonomics.org
TCGA
pairwise co-methylation grid
cg1
cg2
cg3
cgn
co-methylation network
C
association to features, modules comparison, samples stratification, latent structures
D
module dynamics and functional enrichments

Figure 2

A


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)

B

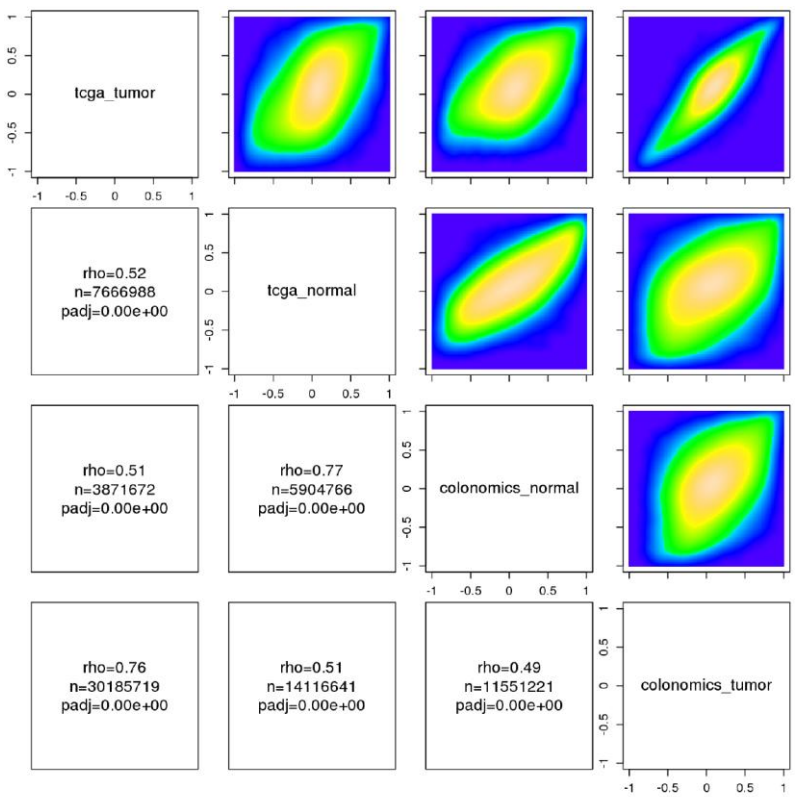

tcga_tumor
rho=0.52
n=7666988
padj=0.00e+00
tcga_normal
rho=0.51
n=3871672
padj=0.00e+00
rho=0.77
n=5904766
padj=0.00e+00
colonomics_normal
rho=0.76
n=30185719
padj=0.00e+00
rho=0.51
n=14116641
padj=0.00e+00
rho=0.49
n=11551221
padj=0.00e+00
colonomics_tumor

Figure 3

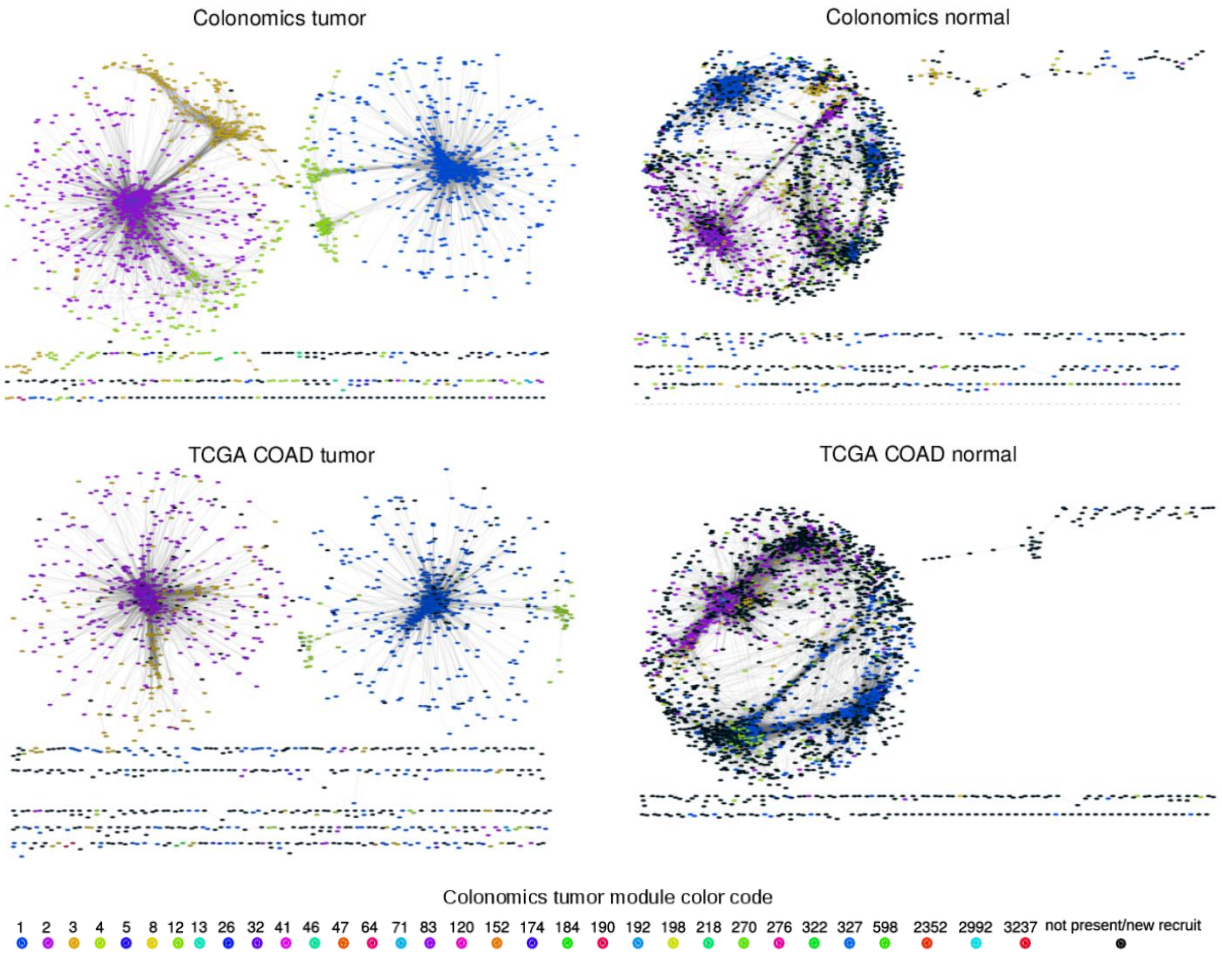

Colonomics tumor
Colonomics normal
TCGA COAD tumor
TCGA COAD normal
Colonomics tumor module color code
1 2 3 4 5 8 12 13 26 32 41 46 47 64 71 83 120 152 174 184 190 192 198 218 270 276 322 327 598 2352 2992 3237 not present/new recruit

Figure 4

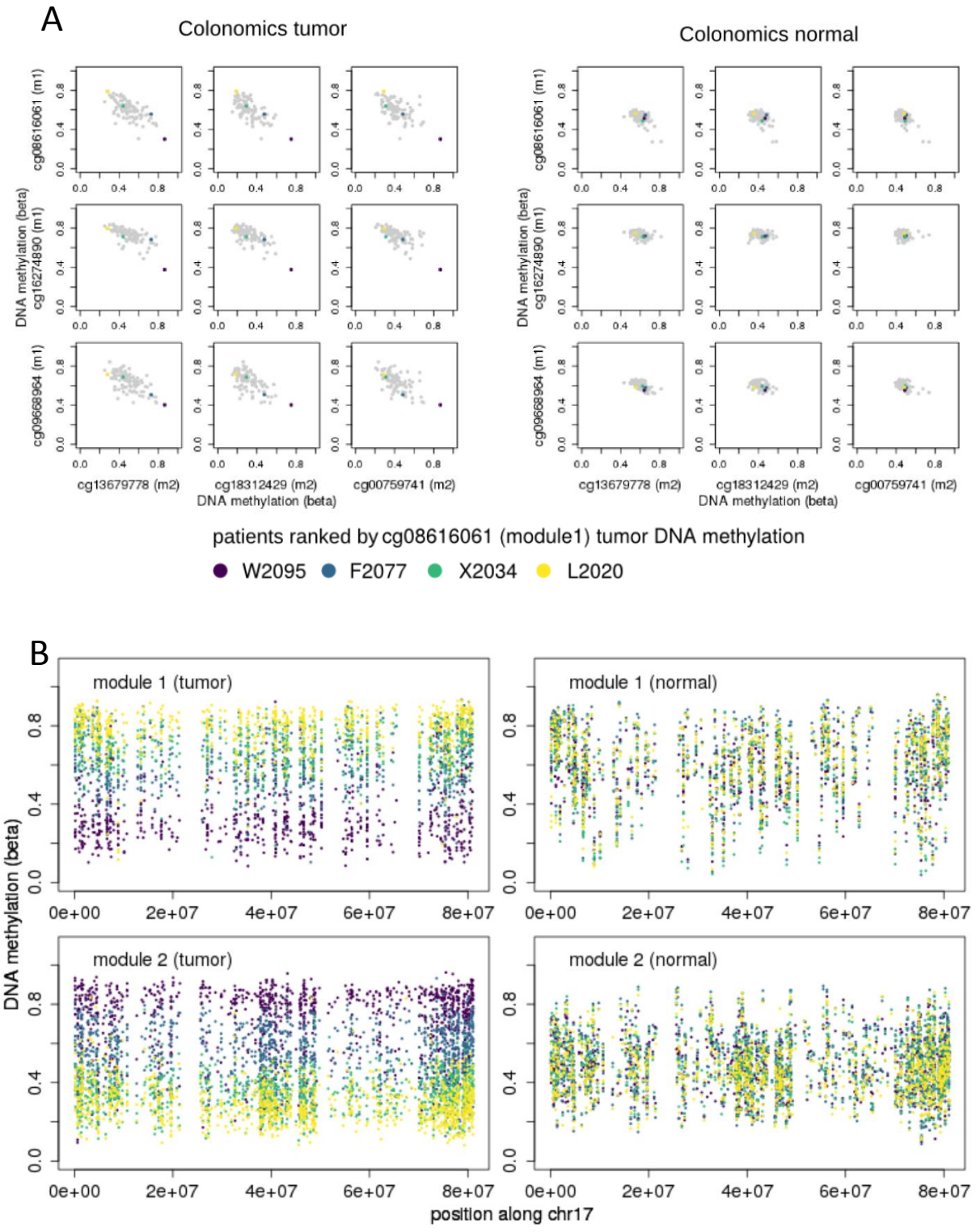
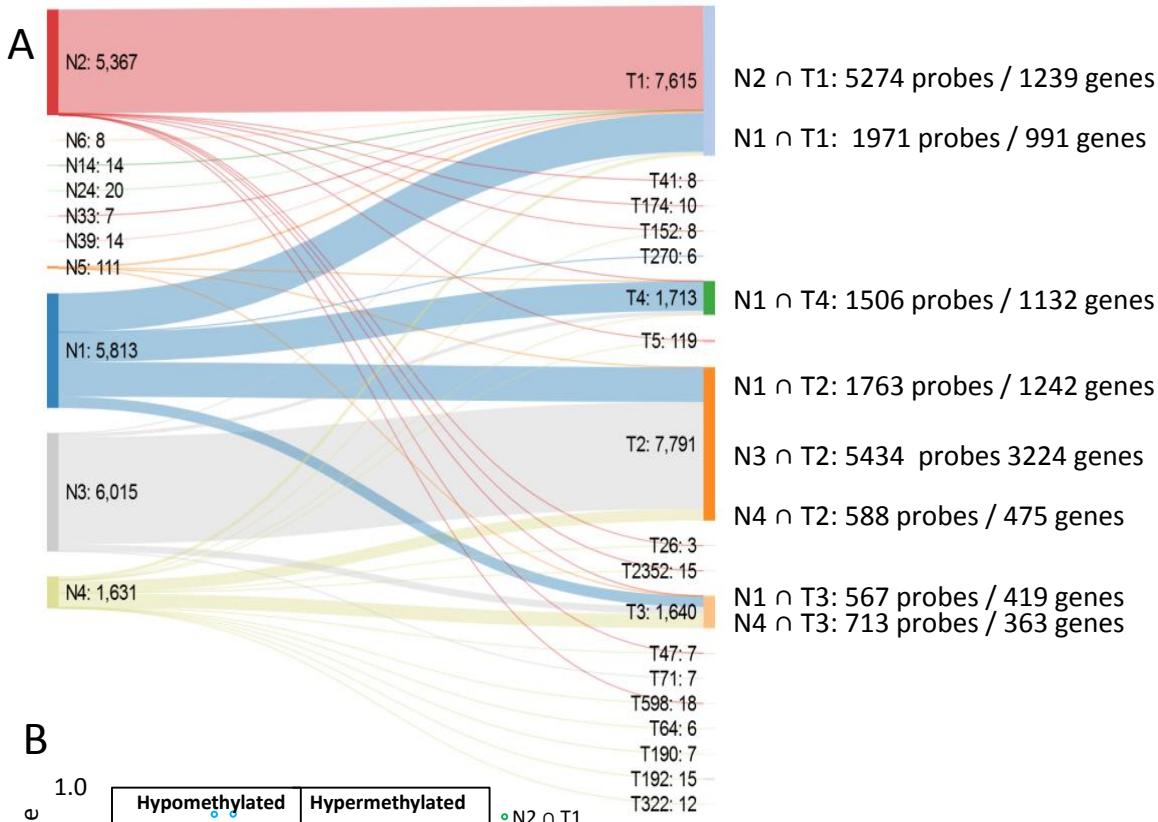
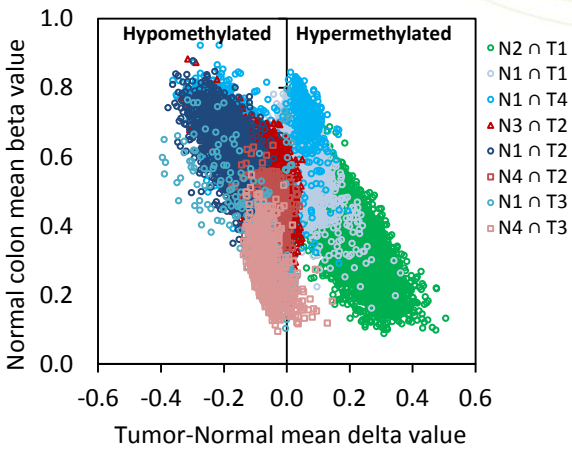
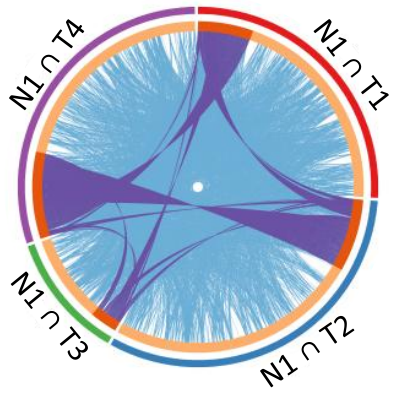
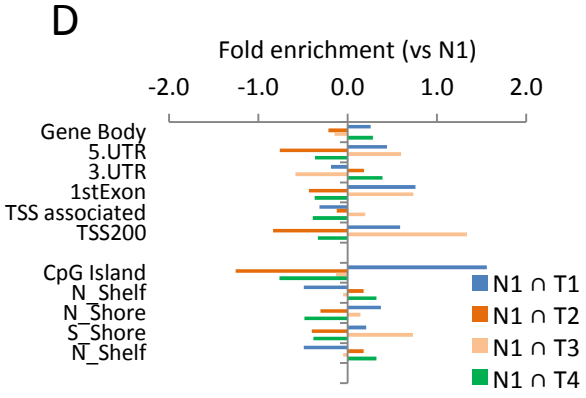

A
Colonomics tumor
Colonomics normal
cg08616061 (m1)
cg16274890 (m1)
cg09668964 (m1)
DNA methylation (beta)
cg13679778 (m2)
cg18312429 (m2)
cg00759741 (m2)
DNA methylation (beta)
patients ranked by cg08616061 (module1) tumor DNA methylation
W2095
F2077
X2034
L2020
B
module 1 (tumor)
module 1 (normal)
module 2 (tumor)
module 2 (normal)
DNA methylation (beta)
position along chr17

Figure 5


A
N2: 5,367
N6: 8
N14: 14
N24: 20
N33: 7
N39: 14
N5: 111
N1: 5,813
N3: 6,015
N4: 1,631
T1: 7,615
T41: 8
T174: 10
T152: 8
T270: 6
T4: 1,713
T5: 119
T2: 7,791
T26: 3
T2352: 15
T3: 1,640
T47: 7
T71: 7
T598: 18
T64: 6
T190: 7
T192: 15
T322: 12
N2 ∩ T1: 5274 probes / 1239 genes
N1 ∩ T1: 1971 probes / 991 genes
N1 ∩ T4: 1506 probes / 1132 genes
N1 ∩ T2: 1763 probes / 1242 genes
N3 ∩ T2: 5434 probes 3224 genes
N4 ∩ T2: 588 probes / 475 genes
N1 ∩ T3: 567 probes / 419 genes
N4 ∩ T3: 713 probes / 363 genes


B
Hypomethylated
Hypermethylated
Normal colon mean beta value
Tumor-Normal mean delta value
N2 ∩ T1
N1 ∩ T1
N1 ∩ T4
N3 ∩ T2
N1 ∩ T2
N4 ∩ T2
N1 ∩ T3
N4 ∩ T3


C
N1 ∩ T4
N1 ∩ T1
N1 ∩ T3
N1 ∩ T2


D
Fold enrichment (vs N1)
Gene Body
5.UTR
3.UTR
1stExon
TSS associated
TSS200
CpG Island
N_Shelf
N_Shore
S_Shore
N_Shelf
N1 ∩ T1
N1 ∩ T2
N1 ∩ T3
N1 ∩ T4

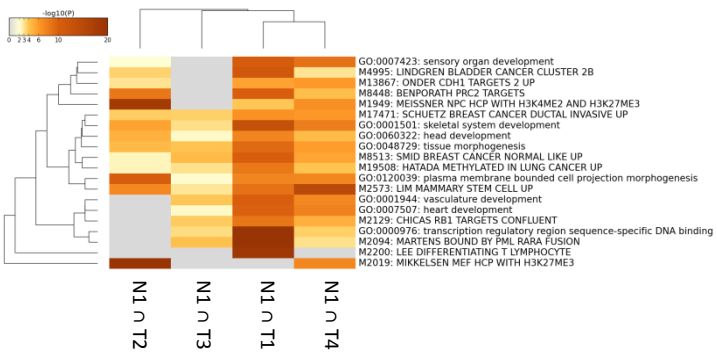
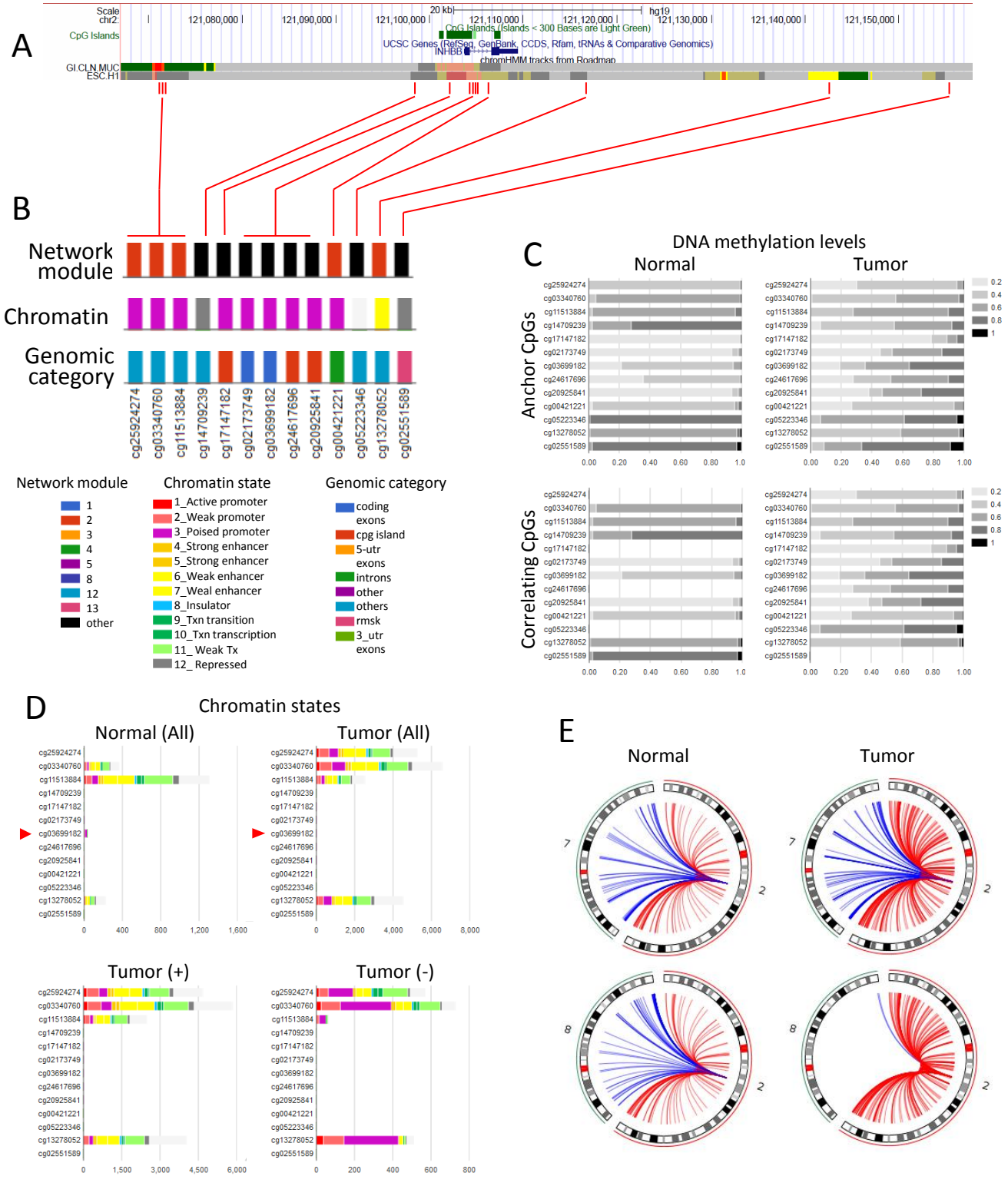

E
GO:0007423: sensory organ development
M4995: LINDGREN BLADDER CANCER CLUSTER 2B
M13867: ONDER CDH1 TARGETS 2 UP
M8448: BENPORATH PRC2 TARGETS
M1949: MEISSNER NPC HCP WITH H3K4ME2 AND H3K27ME3
M17471: SCHUETZ BREAST CANCER DUCTAL INVASIVE UP
GO:0001501: skeletal system development
GO:0060322: head development
GO:0048729: tissue morphogenesis
M8513: SMID BREAST CANCER NORMAL LIKE UP
M19508: HATADA METHYLATED IN LUNG CANCER UP
GO:0120039: plasma membrane bounded cell projection morphogenesis
M2573: LIM MAMMARY STEM CELL UP
GO:0001944: vasculature development
GO:0007507: heart development
M2129: CHICAS RB1 TARGETS CONFLUENT
GO:0000976: transcription regulatory region sequence-specific DNA binding
M2094: MARTENS BOUND BY PML RARA FUSION
M2200: LEE DIFFERENTIATING T LYMPHOCYTE
M2019: MIKKELSEN MEF HCP WITH H3K27ME3
N1 ∩ T2
N1 ∩ T3
N1 ∩ T1
N1 ∩ T4

Figure 6


A
Scale
chr2:
20 kb
hg19
121,080,000
121,090,000
121,100,000
121,110,000
121,120,000
121,130,000
121,140,000
121,150,000
CpG Islands
CpG Islands (Islands < 300 Bases are Light Green)
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
INHBB
chromHMM tracks from Roadmap
GI.CLN.MUC
ESC.H1
B
Network module
Chromatin
Genomic category
cg25924274
cg03340760
cg11513884
cg14709239
cg17147182
cg02173749
cg03699182
cg24617696
cg20925841
cg00421221
cg05223346
cg13278052
cg02551589
Network module
1
2
3
4
5
8
12
13
other
Chromatin state
1_Active promoter
2_Weak promoter
3_Poised promoter
4_Strong enhancer
5_Strong enhancer
6_Weak enhancer
7_Weal enhancer
8_Insulator
9_Txn transition
10_Txn transcription
11_ Weak Tx
12_ Repressed
Genomic category
coding exons
cpg island
5-utr exons
introns
other
others
rmsk
3_utr exons
C
DNA methylation levels
Normal
Tumor
Anchor CpGs
Correlating CpGs
0.00
0.20
0.40
0.60
0.80
1.0
0.2
0.4
0.6
0.8
1
D
Chromatin states
Normal (All)
Tumor (All)
Tumor (+)
Tumor (-)
0
400
800
1,200
1,600
2,000
4,000
6,000
8,000
1,500
3,000
4,500
200
600
E
Normal
Tumor
7
2
8