UNIVERSITÀ DEGLI STUDI DI MILANO

PhD Course in Experimental Medicine and Medical Biotechnologies (XXXIII cycle)
Department of Medical Biotechnology and Translational Medicine (BIOMETRA)

Development of bioinformatic methods for the integration of transcriptomic and epigenomic analysis of colorectal cancer derived organoids.

PhD thesis of:
Federica Gervasoni
R12011

Tutor: Prof. Massimiliano Pagani
Supervisors: Raoul J.P. Bonnal and Michaela Fakiola, PhD
PhD Coordinator: Prof. Massimo Locati

A.A. 2019/2020

ABSTRACT

Cancer is characterized by pervasive epigenetic alterations with enhancer dysfunction as key driver of the tumor transcriptional deregulation and dependencies. In this work, we seek to unravel the chromatin landscape of human colorectal cancer (CRC) by exploiting the organoid model in order to identify a common epigenetic blueprint and investigate its relevance in other types of cancers.

To this extent, we generated a library of patient derived organoids (PDOs) from different subtypes of CRC representing the heterogeneity of this type of tumor. We reconstructed the epigenetic landscape of CRC and retrieved a catalogue of regulatory elements using a complete panel of the most common histone modifications. Next, we identified a conserved and tumor-specific enhancerome that is cancer cell intrinsic and independent of interpatient tumor heterogeneity. Interestingly, we also identified the transcriptional co-activator YAP and TAZ as key regulators of the conserved CRC gained enhancerome. Reaching beyond CRC, we took advantage of ATAC-seq data of diverse tumor malignancies to demonstrate that our CRC-enhancer blueprint was a conserved feature of epigenetic deregulation in human cancer pathology. We next sought to depict the cancer epigenetic deregulation at single-cell resolution demonstrating the specificity of our cancer regulatory blueprint for malignant cells in different types of cancer, suggesting a key role of the cancer regulatory blueprint in tumorigenesis and maintenance of the cancer cell state.

Despite the considerable genetic and clinical heterogeneity of CRC, our work suggests a common layer of YAP/TAZ-mediated epigenetic deregulation in cancer and provides a detailed epigenetic resource of critical importance for identifying therapeutic targets with enhanced precision.

TABLE OF CONTENTS

1. Introduction

1.1 Cancer

Cancer represents the second leading cause of death worldwide[1]. It is the result of the accumulation of genetic and epigenetic alterations leading to changes in cell functions. Specific patterns of alterations are associated with environmental factors, such as tobacco smoke, mutagenic chemicals, ultraviolet light, viruses and bacteria[2]. These changes occur mainly in genes that positively (*proto-oncogenes*) or negatively (*tumor suppressor genes, TSG*) regulate cell division and are involved in DNA repair mechanisms (*DNA repair genes*). The dysfunction of these genes is responsible of alterations in cell cycle and apoptosis resulting in an uncontrolled cell proliferation[1]. The abnormal growth of cells leads to the formation of masses of tissue called tumors. One of the main characteristics of tumors is the phenotypic and functional heterogeneity which is caused by genetic alterations and environmental stimuli. Both intra-tumor (tumor by tumor) and inter-tumor (within a tumor) heterogeneity makes cancer difficult to diagnose and treat efficiently[3,4].

1.1.1 The hallmarks of cancer

In the last years, the puzzling complexity and heterogeneity underlying cancer have raised a number of questions pointing in the same direction. Are there specific features shared by all types of tumors? Which is the relationship between the surrounding microenvironment and the tumor? An initial answer to these questions was offered with the influential reviews by Hanahan and Weinberg in 2000 and 2011, in which they described the "Hallmarks of cancer"[5–8] (**Figure 1**). During the long process of tumor development and malignant progression, tumors acquire eight essential features that support uncontrolled cell proliferation, escape from cell death and invasiveness. So far, the known hallmarks of cancer include:

Self-sufficiency in growth signals: is the capability of tumor cells to generate growth signal factors, overexpress transmembrane receptors and change the type of extracellular matrix receptors, acquiring independency from normal tissue stimuli and disrupting homeostatic balance.

Insensitivity to anti-growth signals: is the capability to circumvent anti-growth signalling circuits.

Evading apoptosis: it is an acquired ability of cancer to escape regulatory and effector components of programmed cell death signalling circuits (apoptosis).

Limitless replicative potential: It is the ability to avoid cell-autonomous programs that restrain cell-multiplication in order for a clone to expand in macroscopic fashion. This is the result of the previous mentioned capability and it is related to telomere maintenance.

Sustained angiogenesis: The limitless ability of cancer cells to proliferate requires continue supply of fuel (oxygen and nutrients) that is guaranteed by the growth of new blood vessels.

Tissue invasion and metastasis: primary tumors produce pioneer cells that escape the primary sites invading adjacent tissues with the purpose of reaching distant nutrient-rich sites and create new colonies (metastasis). The accomplishment of this step depends upon all the previous acquired features.

Deregulating cellular energetics: It is the ability of a tumor to reprogram and increase the production of energy according to cell proliferation.

Avoiding immune destruction: The immune system has emerged as a powerful barrier to obstruct tumor progression. However, tumor cells acquire the ability to reprogram the immune cells in charge of destroying them.

The acquisition of all these functional capabilities is enabled by *genome instability and inflammation.* Genome instability results in mutation of hallmark- key genes while inflammation caused by immune cells induces wound healing. One of the possible benefits of the hallmark of cancer is to identify common functional layers in tumor malignancies that can lead to new potential therapeutic targets[5–8].

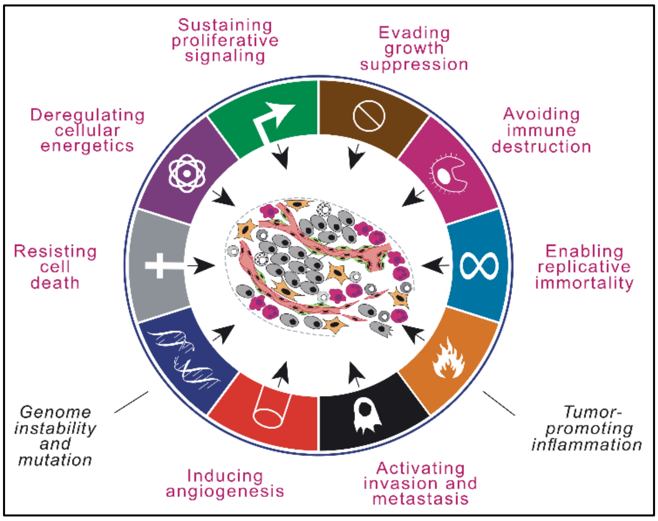

Sustaining proliferative signaling
Evading growth suppression
Deregulating cellular energetics
Avoiding immune destruction
Resisting cell death
Enabling replicative immortality
Genome instability and mutation
Tumor-promoting inflammation
Inducing angiogenesis
Activating invasion and metastasis

Figure 1 | Schematic representation of the eight specific features (hallmarks) necessary to manifest malignant disease (adapted from Hanahan & Weinberg, 2017[8]).

1.2 Colorectal cancer

1.2.1 Epidemiology and carcinogenesis

Colorectal cancer (CRC) is a heterogeneous class of malignant epithelial tumors originating in the colon or rectum. The term “malignant” is attributed when the mass penetrates through the muscular mucosa layer and reaches the submucosa. CRC is the third cause of death due to malignant neoplasia worldwide and its incidence is rising in developing countries, affecting approximately one million of people per year[9]. Despite the decreased mortality rate in the last decades, due to new therapeutic approaches, the CRC incidence and the age of onset are aggravating but the reasons are not completely clear. The incidence of CRC has a strong correlation with male sex and age. However, both environmental (*i.e.* obesity, sedentary lifestyle, tobacco smoke and alcohol) and hereditary risks (*i.e.* hereditary predisposition, bowel inflammatory disease and polyps) play a fundamental role in the development of CRC. The acquisition of genetic and epigenetic alterations in epithelial cells confer them a selective advantage resulting in tumor formation. These hyper-proliferative cells give rise to benign adenoma, which eventually can evolve in carcinoma and potentially also metastasis. CRC can be classified in three main groups according to its origin and expression:

Sporadic type accounts for 60-80% of CRC cases and is not associated to family risk. The onset of this CRC type is associated with environmental risk factors directly involved in tumorigenesis.

Familial type represents 20-40% of CRC cases, which have a family member of primary consanguinity affected by CRC[10].

Hereditary type concerns the remaining 5% of cases with inherited predisposition to CRC. It is characterized by the loss of function of key tumor suppressors genes and DNA repair genes. Usually, these genes are recessive and thus they need both alleles to be altered to give rise to the pathology; indeed, a mutated copy is inherited and, eventually, a somatic event induces the alteration also in the second allele, giving rise to the tumor (Loss of heterozygosity, LOH). This type has two tumor variants that can be distinguished by a person's predisposition to develop adenomatous polyps or not. In the FAP (familial adenomatous polyposis) and MAP (MUTYH (MYH)-associated polyposis)[11] category patients present multiple polyps which should be surgically removed to avoid cancer development. On the contrary, HNPCC (hereditary nonpolyposis CRC, or Lynch syndrome) is not associated with polyps. The HNPCC syndrome can be caused by mutations in DNA mismatch repair genes, including MLH1, MSH2, MSH6, PMS2, MLH3, PMS1, and TGFBR.

The progression of CRC from adenoma to carcinoma proceeds in a multi-step fashion, leading to the accumulation of genetic and epigenetic alterations in three fundamental gene categories orchestrating epithelial development and cellular differentiation: i) *tumor suppressor genes,* including adenomatous polyposis coli (APC), DCC, TP53, SMAD2, SMAD family member 4 and p16INK4. ii) *proto-oncogenes*, such as K-ras and N-ras and iii) *DNA repair genes*, such as MMR and MUTYH. As clinical and histopathological studies suggest, the vast majority of CRCs are generated from polyps of small dimension that gradually progress increasing the degree of dysplasia from early, middle and late adenomas to carcinoma (**Figure 2**). During the tumorigenic process, cells acquire proliferative and expansion capabilities compared to normal cells, generating a clone of altered epithelial cells. Inside each clonal population, there is a subsequent accumulation of alterations that confer further selective advantages. This results in a mosaic of genetically and phenotypically heterogeneous cells, each one characterised by specific differentiation degree, invasiveness, drug resistance and proliferation rate. The model of progressive transition from adenoma to carcinoma[12] has provided early support for the role of driver genetic aberrations in tumor suppressor genes and oncogenes spurring large-scale sequencing analyses that have enriched the list of recurrently mutated genes in CRC[13,14].

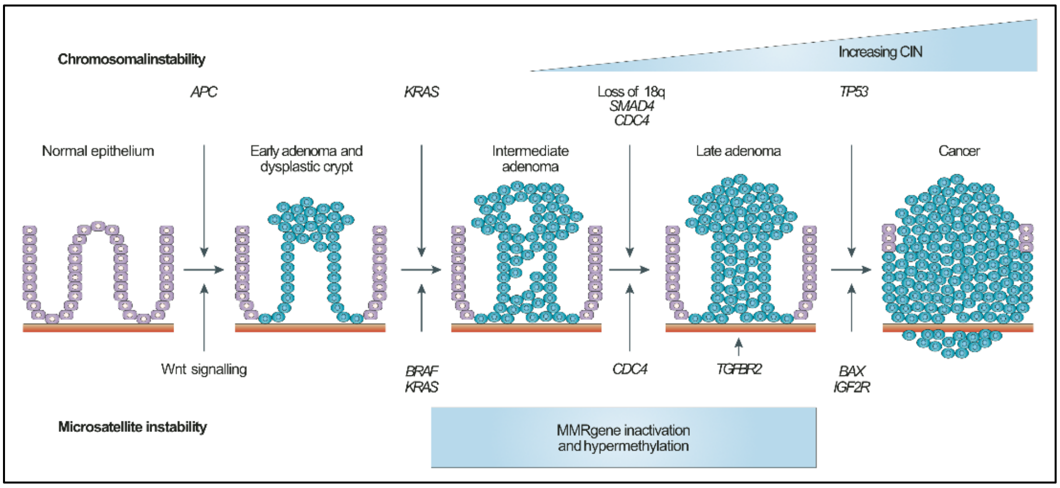


Figure 2 | Schematic representation of CRC progression. The figure reports the genes and growth factor pathways driving the progression. Adapted from Walther *et al.*, 2009[15].

1.2.2 CRC classification

"Colorectal cancer is not just colorectal cancer" [16], this sentence from Blank at al. in 2018 is of major importance to highlight one of the most important feature of CRC: tumor heterogeneity. CRC is characterized by diverse clinical and pathological features that have an impact on tumor progression, drug responses and resistance. For this reason, one of the main questions associated to the study of CRC is the identification of a class of tumors that share phenotypic and molecular features in order to design specific drug targets. Different types of classification have been developed in the past: the clinical classification (based on tumor location and histology), the Tumor Node Metastasis (TNM) classification (based on size and tumor microenvironment) and the canonical classification (based on the genetic background). However, with the advent of Next Generation Sequencing (NGS) technology the efforts are increasingly oriented toward the development of gene expression-based classifications, involving the analyses of transcriptomic data with clustering, neuronal network and other machine learning techniques. The interesting paradox is that even if there are a number of classification systems based on genetics and gene expression, the consistency between them is limited[17]. This incongruence highlights the urgent need to consider epigenetic events and mechanisms in the classification of cancers. Another important issue is that despite the quality and reliability of classification systems, which are improving together with analytical methods and technologies, it is becoming more and more clear that tumors do not fall within defined and compartmentalized subtypes, but they are instead characterized by a continuum of subtypes[18]. There is a plethora of different classification systems extensively discussed in a review published by Wang *et al.* 2019[19]. For the purpose of this thesis, the next paragraphs describe the CRC classification systems used in this work to stratify patients in specific groups.

1.2.2.1 Clinical classification

CRC can be divided based on the site of tumor origin into i) *right-sided* (originated from colon sections proximal to splenic flexure, *i.e.* caecum, ascending and transverse colon), ii) *left-sided* (originated distal from splenic flexure, *i.e.* descending colon and sigmoid colon), and iii) *rectum* (arise within 15 cm of the anal sphincter)[9]. The first two classes usually metastasize in liver and have better prognosis compared to the third class, which has higher rates of loco-regional relapse and lung metastases. The majority of CRC are classified as adenocarcinoma and subdivided in low-grade and high-grade. Some CRC can also be characterized by rare histological features, such as mucinous adenocarcinoma, adeno-squamous carcinoma, signet-cell carcinoma and medullary carcinoma[20].

1.2.2.2 Tumor, Node, Metastasis staging system

The first version of the Tumor, Node, Metastasis (TNM) classification was published in 1977 by the American Joint Committee on Cancer (AJCC) tumor-node-metastasis (TNM). The aim of this classification is to identify prognostic values fundamental for appropriate therapeutic decisions and clinical treatment. They evaluate three aspects using different parameters:
T parameter: describes the size of primary tumor (how deep the primary tumor has grown into the bowel lining). This parameter is described by five degrees of size, from T0 to T4.
N parameter: considers the possible involvement of lymph nodes. The starting value for this parameter is N0, a condition in which the regional lymph nodes are not affected. The parameter increases (N1, N2 ..Nn) with the degree of tumor spread to regional or more distant lymph nodes.
M parameter: refers to the presence of distal metastasis. If the parameter is M0 metastases are not present, otherwise the number increases.

Another reported parameter is the *grade (G)*, which describes how closely cancer cells resemble healthy cells when viewed under a microscope. For this parameter there are 4 degrees of severity: from G1, where cells are differentiated, to G4, where cells are very undifferentiated[21,22].

1.2.2.3 Canonical classification

One of the earliest classification systems, also called "canonical classification", describes CRC according to its genetic instability. The three features used to explain CRC development are: *chromosomal instability* (CIN), *microsatellite instability* (MSI) and *CpG island methylator phenotype* (CIMP).

Chromosomal instability (CIN): described by Fearon and Vogelstein[12], is characterized by cytogenetic changes, loss of allelic heterozygosity and mutation of tumor suppressor genes. This category includes the FAP tumours and 80% of sporadic tumours. The series of the events are initiated by the mutation or loss of APC (TSG), followed by mutations in the proto-oncogene KRAS (transition to adenoma), TP53 (transition from adenoma to carcinoma) and DCC (epithelial/mesenchymal interactions) (**Figure 2**).

Microsatellite instability (MSI): this category includes *Lynch syndrome* tumours and approximately 15% of the sporadic tumours. It is characterized by mutations in genes involved in mismatch error repairs (MMR) which lead to a state of genomic instability resulting in a *hypermutator* phenotype or *microsatellite instability.* The alteration of MMR genes causes the onset of new alterations predominantly located in repetitive DNA sequences generating the instability of microsatellite markers and promoting the contraction, insertion or deletion of repetitive elements. These events cause the activation of genes involved in the regulation of apoptosis (*e.g.* BAX and *Caspasa-5*), cellular growth (*e.g.* TGFBRII, IGFIIR) or MMR itself (*e.g.* MSH3, MSH6)[23].

Epigenetic alterations: More recently, epigenetic alterations have also been described as a key element in CRC tumorigenesis[24]. The inactivation of TGS, MSI or CIN repair genes generate epigenetic changes that create instability[25]. There are two types of epigenetic alterations: in the first type, *CpG island methylator phenotype* (CIMP), the widespread DNA hypermethylation of promoter-associated CpG islands is achieved though the addition of a methyl group (-CH3) to DNA nucleotides leading to the silencing of specific DNA regions and genes, including a range of tumor suppressor and DNA repair genes (*e.g.* MLH1, one of the MMR genes). The second type of changes can modify the acetylation of histone proteins with the addition of acetyl group (-CH3CHO) to the histone core promoting chromatin accessibility and thus increasing gene expression[24,25].

1.2.2.4 CRCassigner classification

The Sadanandam *et al.*[26] classification was one of the first classifications based on gene expression profiles. In particular, the authors considered the gene expression of 445 CRC patients from two public datasets. They identified six CRC subtypes using consensus non negative matrix factorization (NMF) clustering technique and created subtypes based on predominant cellular types that compose the crypt-like structure of CRC. The validation of their findings was done on seven independent studies and 51 cell lines. The six subgroups were defined as:

- *Goblet-like*: Goblet-marker genes (MUC2 and TFF3) are highly expressed. They are related to a crypt top signature and associated with a good prognosis.
- *Enterocyte*: Enterocyte genes are highly expressed. They have an intermediate disease-free survival (DFS).
- *Stem-like*: stem cell and WNT-signalling marker genes are highly expressed. They are related to a bottom crypt signature and associated to the shortest DFS.
- *Inflammatory*: High expression of chemokines and interferon (INF)-related genes. Usually enriched in MSI type and with intermediate DFS.
- *Transient amplifying*: They are characterized by patients with heterogeneous molecular features and are enriched for the MSS type.

The authors also reported subtype-specific sensitivity to drug therapies for CRC. Beneficial responses to FOLFIRI treatment were associated with patients classified to the stem-like and inflammatory subtypes, whereas cetuximab and cMET inhibitors could be effective for the transient-amplifying subtypes in metastatic settings.

1.2.2.5 Consensus molecular subtypes classification

The consensus molecular subtype (CMS) classification system was published in 2015 by Guinney *et al.*[27] with the support of the CRC Subtyping Consortium (CRCSC) and, so far, is the most widely-used by the scientific community. The CMS classification used a network based approach, based on six previous classification systems, to classify 18 datasets including 4151 primary tumors[19] (**Figure 3**). The result of this approach identified four distinct sub-groups which share molecular features and clinical characteristics:

CMS1– MSI Immune subtype (~14%): It is characterized by samples associated with MSI, high expression of DNA repair proteins, low prevalence of somatic copy number alterations (SCNAs), and mutations in the proto-oncogene BRAF. It is defined as the “immune” subtype since it shows an increase expression of immune-related genes, mainly associated with Th1 and cytotoxic T lymphocytes. The tumors usually derive from serrated lesions, located in the proximal regions of the colon. The patients in this class are associated with worse survival after relapse.

CMS2 – Canonical subtype (~37%): It is defined as canonical because it shows epithelial differentiation and up-regulation of WNT and MYC signalling pathways, classically involved in CRC. It is characterized by elevated CIN and tumors originate predominantly from tubular lesions located in the distal region of colon.

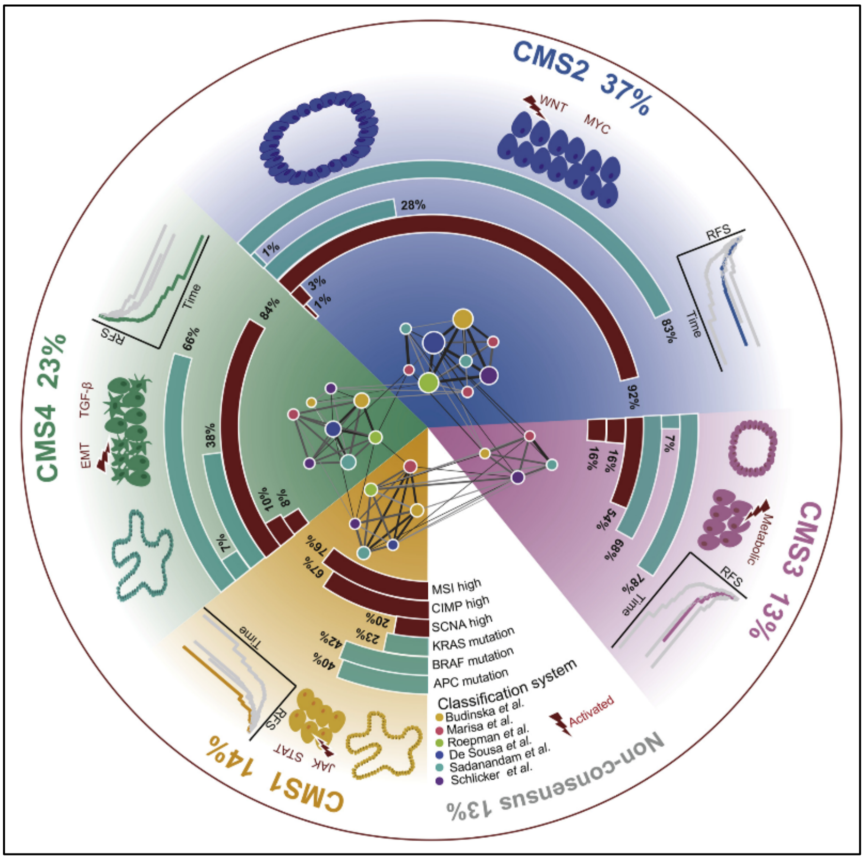


Figure 3 | Schematic overview of CMS system. The figure reports the six studies used to perform the network-based meta-analysis and identify the four CMS classes. Each slice of the circle plot also reports the molecular and phenotypical characteristics of the corresponding CMS, together with frequencies and DFS. From Wang *et al.* 2019[19].

CMS3 – metabolic subtype (~13%): It is characterized by enrichment in pathways involved in cellular metabolism and mutations in proto-oncogene KRAS that induce metabolic adaptation. It is represented by CIN tumors that show a distinctive genomic and epigenomic profile, such as lower levels of SCNAs and high prevalence of the TCGA-defined CIMP-low (CIMP-L) cluster. These are localized in the proximal and distal segments of the colon. They are associated with intermediate prognosis.

CMS4 –mesenchymal subtype (~23%): It is characterized by up-regulation of genes involved in epithelial to mesenchymal transition (EMT) and of multiple pathways involved in TGF-β signalling, matrix remodelling, angiogenesis and complement system. These tumours derived from serrated precursor lesions and are localized in the distal segments of colon. They tend to be diagnosed at later stages (III/IV) and are associated with worse relapse free and overall survival.

The CMS classification has been used in multiple studies to classify different pre-clinical models, including cell lines[28,29], patient-derived xenografts (PDX) and patient-derived organoids (PDOs). However, its reliability and sensitivity are slightly affected when assessing patient-derived models compared to primary tissues[30].

1.2.2.6 CRC intrinsic subtypes classification

The vast majority of gene expression classification systems are based on bulk RNA-seq data of primary tumor tissues, which represent a mixture of cell types, including stromal cells. The largest part of the stromal cells is composed by cancer-associated fibroblasts (CAFs) and their presence is directly correlated with tumor aggressiveness, with a dramatic effect on patient prognosis. Another important aspect is that stromal contamination can be a major source of gene expression variability, affecting also the CRC classification. For example, the gene signature of CMS4 could be severely affected by the expression of stromal- rather than cancer-related genes[31–33]. Furthermore, several models used to study cancer, such as cell lines, PDX and PDOs, lack the stromal component and thus their classification can be dramatically affected. In 2017, Isella *et al.*[31] developed a new classification system, using gene expression profiles of PDX, known as CRIS (ColoRectal cancer Intrinsic Signature). The use of PDX enabled a classification that was not influenced by stroma-derived transcripts but relied instead on cancer-cell intrinsic signatures. The authors defined five subtypes, characterized by different functional and phenotypic features:

CRISA: The majority of the tumors in this subtype are MSI and are characterized by mutations in BRAF or KRAS (proto-oncogenes). They show a secretory/mucinous histology with sustained inflammatory attributes and glycolytic metabolism. CRISA tumors show sensitivity to anti-metabolic therapies.

CRISB: These tumors are generally poorly differentiated and characterized by TGF-β signalling and EMT features.

CRISC: The tumors grouped in this subtype are distinguished by chromosomal instability and are KRAS wild-type. They show high activity of ERBB/EGFR pathway and have copy number gains of the MYC proto-oncogene (focal amplification of chromosome 8q.24.21). They have sensitivity to anti-EGFR antibodies.

CRISD: This group of tumors show a stem-like phenotype and high activity of WNT signalling pathway, coupled with strong enrichment of IGF2 amplification/overexpression and autocrine stimulation. These tumors do not have sensitivity to anti-EGFR therapies.

CRISE display WNT-related features, however in this case they are associated with TP53-mutated genotype and Paneth-like phenotypes.

The authors suggested that the CRIS system outperforms the CMS system in the assignment of samples to specific subtypes which due to the removal of stromal contamination are characterized by newly described cancer-intrinsic features. Overall, there is limited agreement between the CRIS and CMS classification systems. Most CMS1 samples were assigned to CRISA and B; CMS2 tumors were separated across CRISC, D and E; CMS3 contributed mostly to CRISA; and, finally, CMS4 samples were equally distributed across all five CRIS subgroups[31] (**Figure 4**).

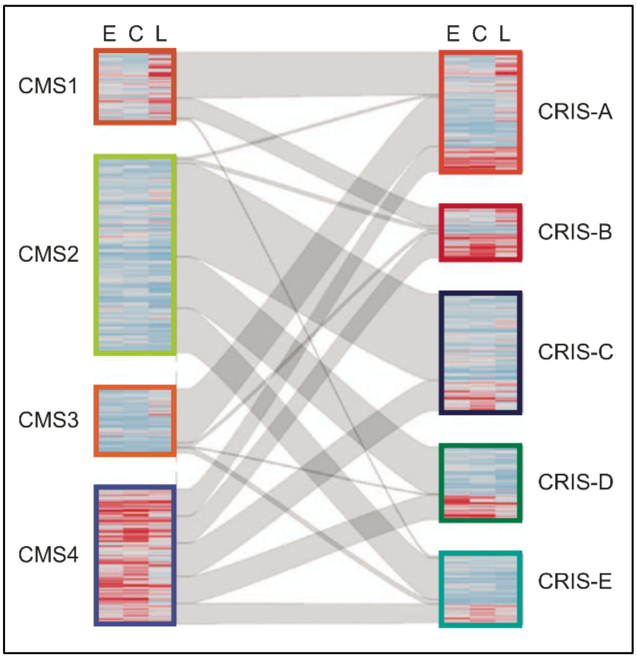


Figure 4 | Correspondence between samples classified with CMS and CRIS classification. Heatmaps report the expression levels of genes associated with endothelial cells (E), cancer-associated fibroblasts (C) and leucocytes (L). Adapted from Isella *et al.* 2017[31].

1.3 Organoids model

In the last decades, advancements in the anti-cancer field have resulted in the development of new therapeutic agents and improved patient survival. However, many of the developed treatments include conventional drugs aimed to treat the majority of patients. These conventional therapies have limitations due to the intrinsic heterogeneity of tumors which affects tumor growth rate, invasion, drug sensitivity and prognosis. To meet the urgent need for personalized anti-cancer treatments, a large number of pre-clinical models have been proposed. In the past years, most biomedical and cancer-related research involving human specimens has focused on *two-dimensional (2D) cultured cell lines*. Classical cell lines have low costs and have been extensively used for high-throughput screening of drugs[34,35] and cancer biomarkers[36]. However, cell lines can be created only from a reduced number of cancer subtypes, their initial establishment is difficult and leads to a dramatic genetic and phenotypic adaptation to the culture conditions[37]. In the 2D model, the heterogeneity of the tumor of origin is gradually lost because of genetic and epigenetic drifts during long-term culture. Furthermore, cell lines lack the normal counterpart that should be used as control. An alternative model system in cancer research is the PDX in which the tumor tissue is transplanted in an immunodeficient mouse. The PDX model preserves the tumor heterogeneity and the genomic stability across different passages. This model can also reproduce the cancer-stromal and cancer-matrix interactions. However, PDX are more efficiently generated for aggressive tumors[38], they are expensive and labour intensive. In addition, the creation of PDX can last four months making them difficult to use for personalized medicine and high-throughput screening[39,40].

More recently, *three dimensional (3D) organoid models* have been generated to overcome the series of disadvantages related to cell lines and PDX. The term "organoid" was first introduced in 1946 to describe a cystic teratoma[41]. Organoids are three-dimensional structures, derived from mammalian stem cells and grown *in vitro,* which are able to self-organize and recapitulate the key features of the *in vivo* original tissue[42]. The organoid model can represent tissue heterogeneity since it is able to maintain the complex spatial organization of the different cell types present in each tissue and to mimic the specific function of the derived organ. Organoids can be generated at relatively low cost, in as early as four weeks and they grow on micro-plates making them a perfect tool for high-throughput screening and personalized medicine. Organoids can be derived from embryonic stem cells (ESC), induced pluripotent stem cells (iPSC) and adult stem cells (ASC)[43]. Both PSC- and ASC- derived organoids need a source of extracellular matrix that acts as basal lamina for cultured cells. One of the most used mediums is the laminin- and collagen-rich Matrigel. Yet major differences exist between PSC- and ASC-derived organoids:

PSC-derived organoids: They can re-create tissue structures through processes specific to the embryonic development, making them a feasible tool for the study of *in vivo* development, embryo growth and cellular differentiation. Generally, PSC are expanded and undergo multi-step differentiation using step-specific growth factor cocktails, until they are totally differentiated. PSC-derived organoids can be generated in two or three months. They can be composed by mesenchymal, epithelial and even endothelial cells. This model is usually used to study development, genetic diseases, organs with partially regenerator capacity (*i.e.* brain and renal glomerulus) and infectious diseases. PSC-organoids can also provide a mechanistic vision of stem cell development and their related niche and, at the same time, their lineage commitment into differentiated states. In past, this model was used to elucidate the development of stomach, brain and pancreas[44,45] (**Figure 5**).

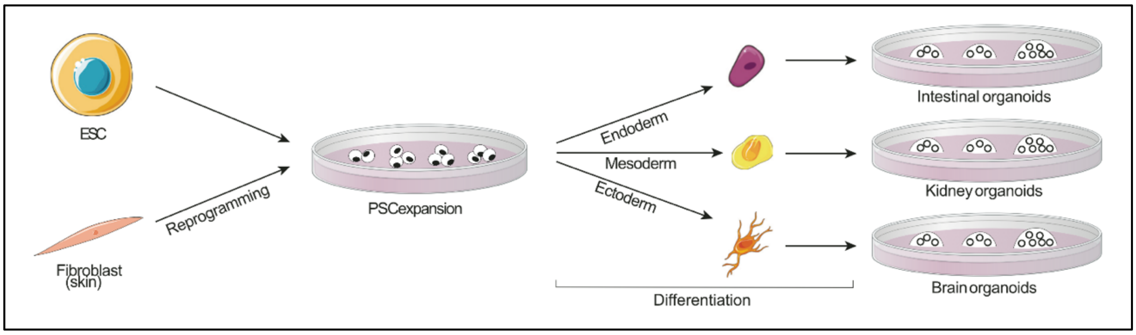


Figure 5 | Schematic representation of differentiation from pluripotent stem cell (PSC) toward each of the three germ layers (endoderm, mesoderm, and ectoderm). Specific growth factor cocktails are used to obtain PSC-derived organoids from the tissue of interest. Adapted from Schutgens & Clevers, 2019[40].

ASC-derived organoids: They recapitulate the adult tissue repair and can be generated from any tissue with regenerative capacity. They only represent epithelial cells, thus lack the stromal, nerves and vascular compartments. ASC-derived organoids mimic the structure and functions of the tissue of origin and they have a less complex organization compared to PSC-derived organoids. This model can be generated from both normal and tumor tissues in approximately seven days. This is of major importance because it is possible to have matched tumor and normal counterparts from the same patient in a short period of time, making them a powerful technique for personalized medicine (**Figure 6**).

Nowadays, organoid models have been generated for a number of different healthy tissues and tissue conditions. For example, various PDOs have been generated from normal and malignant epithelial tissues, such as colon[46,47], liver[48,49], pancreas[50] and prostate[51].

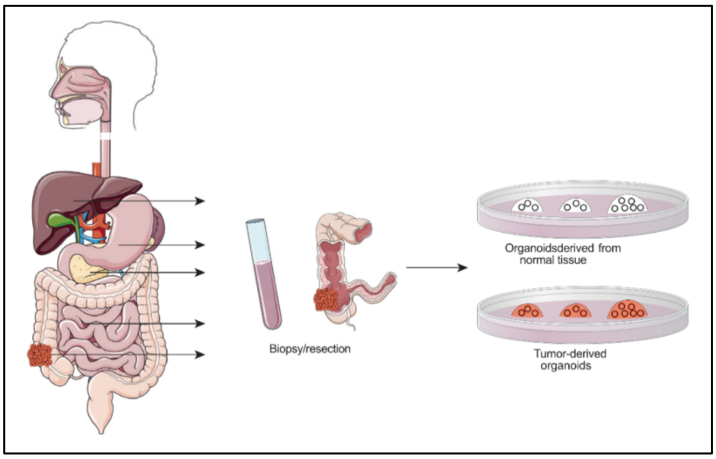


Figure 6 | ASC are obtained from biopsies or tissue resection during surgical operation on healthy tissue or tumours from different organs, such as liver, colon, pancreas. Obtained ASCs proliferate in semisolid matrices and can generate both normal and tumor derived organoid cultures. Adapted from Schutgens & Clevers, 2019[40].

1.3.1 Organoids as experimental tool

Although animal models are essential in biomedical research and cell lines are still extensively used for high-throughput screening, organoid models are emerging as a new tool acting as a bridge between cell lines and *in vivo* approaches in the study of cellular processes and gene functions in tumor and developmental biology. Both PSC- and ASC-derived organoids are promising technologies in translational research. Their applications in the general field of biology and developmental studies have been extensively reviewed elsewhere[42,52]. Here, I will focus on their application in cancer-related studies[39] and specifically in CRC. Cancer is a multi-step disease that occurs through temporal accumulation of cancer-specific alterations in normal cells. The organoid model coupled with gene-editing strategies, such as CRISPR-Cas9 gene-editing technologies or lentiviral and retroviral infections, has offered detailed insights into the genetic mutations that favour the tumor initiation and progression[53–55]. This 3D model has also been used to investigate the complex relationship between genetic changes and niche factors during cancer development[33]. Another fundamental process to understand in the cancer field is the invasion of cells from the primary tumor site to distal sites, known as metastasis. Using the organoids model it is possible to investigate the mechanisms involved in the initiation and inhibition of cancer invasion[56–60].

The research fields of major relevance for the study of tumors, where organoid models have been applied are: anoikis, the apoptosis state induced by the lack of cancer-matrix interactions[47], and tumor dormancy, the state in which cells stop to divide but survive in a quiescent state waiting for favourable conditions to start again their proliferation[61]. As mentioned above, the cancer field requires an appropriate/physiological tool to study tumor biology and develop new approaches taking into consideration the patient-to-patient heterogeneity, at both genetic and epigenetic level. All together these efforts have the potential to lead to the development of personalized anti-cancer therapies exploiting the organoids model to perform drug screening of different types of cancer[62–65]. For instance, hepatocellular carcinoma (HCC) derived organoids have been used to test patient-specific sensitivity to Sorafenib which is the only treatment option for advanced HCC[63]. In 2018, Sachs *et al.*[64] generated a living biobank of breast cancer organoids (>100 primary and metastatic breast cancers) representing the tumor heterogeneity and they used it to study the dose-response of a set of drugs targeting the HER signalling pathway. Yet another work, published by Yao *et al.*, 2020[66], where they generated a biobank of 80 tumor organoids from treatment-naive CRC patients and demonstrated that the usage of this model could improve patient-specific treatments and determine which patient is sensitive to irradiation, 5-Fu, or CPT-11 treatment. Moreover, the possibility to co-culture organoids with other cell populations constituting the tumor microenvironment, such as fibroblasts[67,68] and in particular immune system cells[69–71], could provide new insight into immunotherapeutic studies. Finally, the organoids model has also been used for clinical trials (https://clinicaltrials.gov)[72].

1.3.2 Intestinal organoids

1.3.2.1 Intestine cells functions and composition

The intestine has a high turnover rate, making it perfect to study mechanisms associated to proliferation and cellular differentiation. It is organized in different parts that guarantee the efficiency of the intestinal function which is to digest and absorb nutrients. This organ contains also a barrier against genotoxic substances, bacterial flora and related metabolites. The *small intestine (SI)* is the first part of the intestinal track and it is characterized by finger-like protrusion, which are called *villi*, surrounded by depressions, which are called *crypts.* Since the SI tract is responsible for the absorption of nutrients, below the epithelium there is a complex network of capillaries and lymphatic vessels which deliver the nutrients first to the liver and then to the whole body. The following track, the *large intestine (LI)* is composed by the cecum, different tracts of the colon, sigma, rectum and anus. Since the function of the LI is to compact and secrete stools, it does not present villi structure, but the internal epithelium is disposed in order to form multiple crypts associated to a flat surface. Both tracts of the intestine have shared functions and thus cellular organization of the crypts, which show three distinct zones comprising different epithelial cell types: multipotent intestinal stem cell zone (ISC, historically known as crypt base columnar cells - CBCs), proliferative zone and differentiated compartment. The intestinal epithelium is one of the most actively cycling; it is estimated that intestinal epithelial cells (IEC) are completely replaced every 4/5 days, through a process of renewal and differentiation. The high turnover of the crypts is driven by gene Leucine-rich repeat containing G protein-coupled receptor 5 positive cells (LGR5+) which are localized in the ISC zone. LGR5+ are located at the bottom of the crypt and have a key role in the maintenance and regeneration of the intestinal epithelium. ISC are the precursors of cycling progenitor cells, also known as *transiently-amplifying* (TA). TA cells have the highest proliferative rate and they go through 4-5 cell divisions before differentiating into one of the epithelium-specific lineages. The differentiation process is orchestrated upon the activation of different signalling pathways and moves from the bottom to the apex of the crypts generating IEC lineages which are characterized by specialized functions. The maintenance of the stem cell compartment and the regulation of differentiation are controlled by the WNT, Notch, bone morphogenetic protein (BMP) and epithelial growth factor (EGF) pathways. Two major groups of differentiated lineages are generated: absorptive type (*i.e.* Enterocytes and M cells) and secretory type (*i.e.* Goblet, Paneth, enteroendocrines and tuft cells) (**Figure 7**).

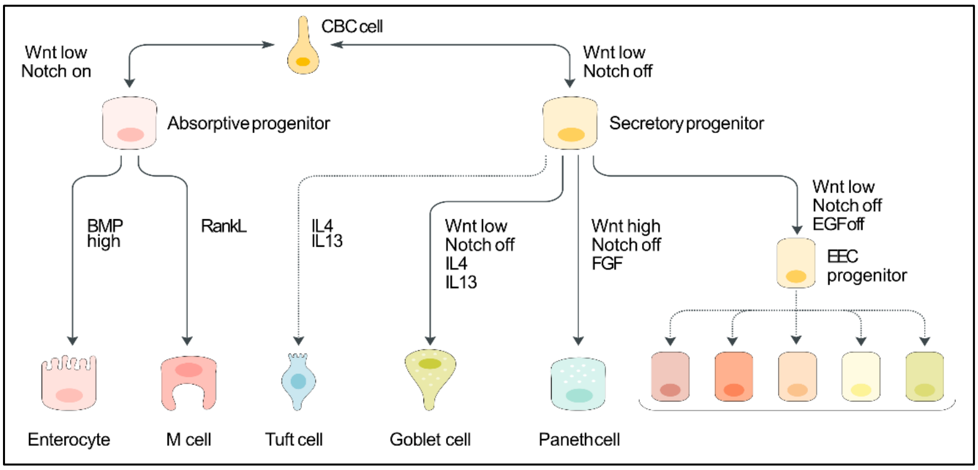


Figure 7 | Schematic representation of the intestinal cell lineage development. Adapted from Gehart & Clevers, 2019[73]. CBC: crypt base columnar, BMP: bone morphogenetic protein, EEC: enteroendocrine cell, RANKL: receptor activator of nuclear factor κ-B (RANK) ligand, FGF: fibroblast growth factor.

Enterocytes are columnar basal cells involved in the absorption of nutrients and water. They are the most prominent cell type in the crypt-villus axis in the small intestine and also in the colon, where they are called colonocytes. They can secrete antimicrobial peptides. IECs, and in particular enterocytes, express a variety of innate immune system receptors, including Toll-like receptors (TLRs). These cells are polarized and the TLRs are disposed on the basolateral and apical surface. The TLRs expressed on the surface allow the enterocytes to recognize microbes and their components and to trigger immune responses to the site of infection. Furthermore, enterocytes can act themselves as physical barrier and they can also be expelled to prevent pathogens breaching the epithelial barrier[74,75].

M cells are involved in antigen uptake and delivery to the immune cells. They are exclusively present in the small intestine and are localized to the follicle-associated epithelium (FAE) overlying Payer's paches[74,75]. Payer's patches are lymphoid follicles enriched is B cells, T cells and mononuclear cells distributed along the intestine.

Goblet cells: play a fundamental role in the establishment of intestinal tolerance and in prompting the mucosal immune response. Goblet cells facilitate luminal antigen delivery to dendritic cells (DC) by uptaking small molecular weight antigens and delivering them directly to DC via the goblet associated passage (GAP)[76]. The GAP system allows goblet cells to maintain intestinal tolerance to the microbiota. Another function of goblet cells is to secrete mucins (such as MUC2) and other antimicrobial proteins (AMP, such as Agr2, ZG16, CCLA1) in order to build the outer and inner mucous layers which constitute a protective barrier. This type of cells is present both in small intestine and colon[74,75].

Paneth cells: they are localized at the bottom of the crypt and do not migrate to the top. Paneth cells are involved in two main functions, intestinal defence and maintenance of the stem cell niche. In fact, Paneth cells secrete abundant antimicrobial proteins (*i.e.* defensins, cryptidins, lysozyme) that mix with Goblet- derived mucus to actively counteract bacterial infections. At the bottom of the crypts, Paneth cells surround stem cells providing them WNT ligands, EGF and Notch stimuli for stemness maintenance. Paneth cells are present in the small intestine but are instead missing in the colon track[74,75]. However, cells with analogous function and presenting an intermediate genetic signature between Paneth and Goblet cells have been identified also in the large intestine[77] and are known as deep crypt secretory cells marked by the regenerating family member 4 (REG4)[78].

Enteroendocrine cells: are cells involved in hormone secretion upon stimulation. This population is subdivided in multiple types according to the secreted hormones, including enterochromaffin cells (5-HT/serotonin), D cells (somatostatin) and G cells (gastrin)[79]. Enteroendocrine cells are present along the entire intestinal tract.

Tuft cells are very rare cells with a fundamental role in the elimination of helminths and the expansion of type 2 innate cells. They can be subdivided in two classes according to the expression of different genes (one type expresses the epithelial cytokine TSLP and the other the immune marker CD45)[80].

Intestinal crypts have a hierarchical organization that renders them a perfect stem cell model to study the signalling pathways orchestrating stem cell niche maintenance. The comprehension of this complex network of pathways promoted the efficient generation of 3D organoid models from this tissue. All together these pathways are tightly regulated by microenvironmental stimuli, deriving from the epithelium itself but also from external mesenchymal cells (fibroblasts, immune cells, enteric neurons, etc.) (**Figure 8**). A plethora of epithelial secreted molecules, growth factors and cytokines control ISCs proliferation and differentiation. The activity of ISCs can be modulated by external perturbation, including inflammation, toxins, radiations and chemotherapy, to compensate cellular loss and enable epithelial regeneration.

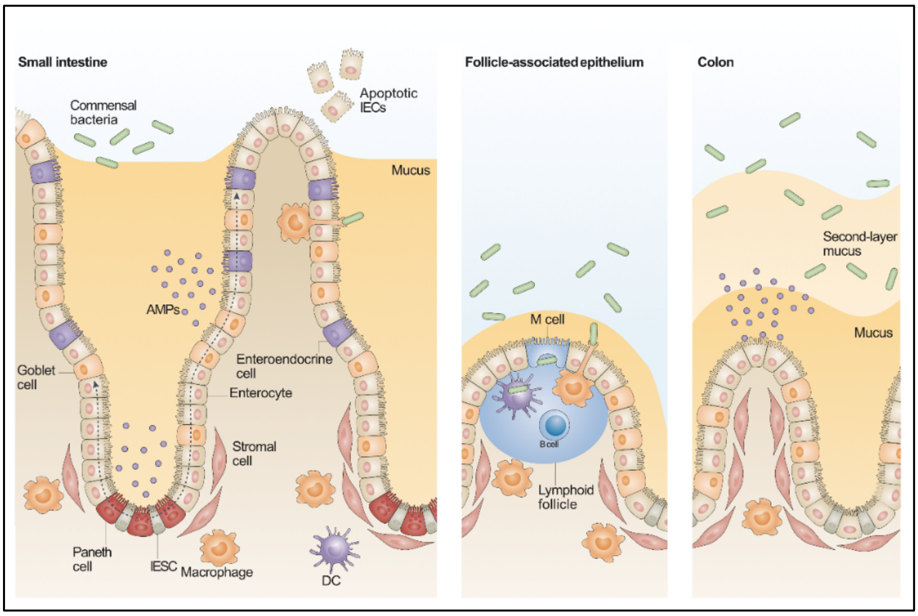


Figure 8 | Schematic representation of the crypt-villi axis in the small intestine and colon. Stem cells and the six different cell lineages are reported together with surrounding microbiota, stromal and immune cells. Adapted from Peterson & Artis, 2014[74].

1.3.2.2 Development of intestinal organoids

The comprehensive study of the mechanisms and composition of the intestinal crypts were fundamental for the creation of *in vitro* mini gut. In fact, Sato *et al.*[81] were the first to develop ASC-derived organoid cultures accurately reproduce *in vitro* intestinal epithelium able to self-renew for more than one year. The development of this system was the result of three key discoveries: i) In 1998, Korinek *et al.*[82] demonstrated that the WNT pathway is essential for the maintenance of the stem niche in the small intestine. ii) In 2007, Barker *et al.*[83] identified the LGR5 as the marker for intestinal and colonic stem cells and a target of WNT. Their results demonstrated that *in vivo* stem cells are able to proliferate and have an unlimited self-renewal capability. iii) The ectopic expression of R-spondin, which is a unique agonist of WNT and a ligand of LGR5, was shown to induce hyperplasia on mouse crypts[84]. Together these findings have led to the first serum-free 3D organoid culture derived from Lgr5$^+$ stem cells from mouse small intestine. Interestingly, this system needs few essential factors: an extracellular laminin-rich matrix (Matrigel, which mimics extracellular matrix functions), R-spondin, EGF and Noggin, an inhibitor of BMP signalling pathway. These organoids were able to recapitulate the architecture of the intestinal epithelium: first, single stem cells create "villus-like" cystic structures with a single central lumen; then, this cyst projects some "crypt-like" structure toward the outside[85]. Importantly, the localization of each cell type resembles their respective localization *in vivo.* Therefore, ISC and Paneth cells are located in the lower part of the buds (crypt-like), whilst mature enterocytes migrate to the central area of the cyst. Paneth and goblet cells secrete their products in the internal lumen. In 2011, Sato *et al.* published also the protocol to establish long-term organoids derived from human intestinal, adenoma and adenocarcinoma[46] (**Figure 9**).

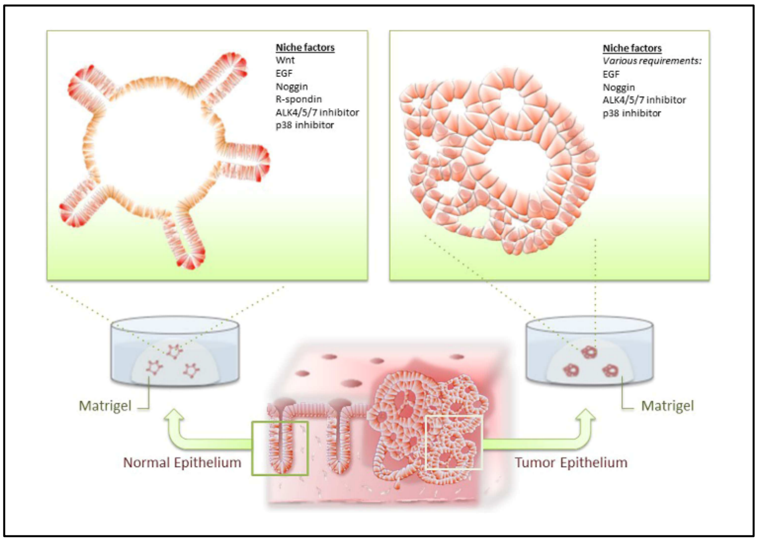


Figure 9 | Organoid culture from normal or tumor colon epithelium. The different structures and the niche factors required for their growth are reported. Adapted from Otha and Sato, 2014[85].

These organoids need other factors in addition to the previously mentioned: p38 and TGF-β inhibitors and the addition of Wnt3a, a ligand relevant for Wnt pathway activation. Below there is a brief report of the components for the culture of colon organoids[86] (**Figure 10**):

Wnt3a/R-spondin: Wnt pathway has a key role in the stemness maintenance, proliferation rate and Paneth cell differentiation. Wnt ligands, which are secreted by Paneth cells and mesenchymal cells *in vivo*, trigger an intracellular signalling cascade upon the binding between Frizzled and LRP5/6 receptors. This results in an accumulation of β-catenin in the cytoplasm followed by the translocation inside the nucleus where it binds to TCF transcription factors activating a cellular program that maintains ISC in an undifferentiated state. R-spondin is a secreted protein that binds to LGR5 receptor and inhibits the degradation of Wnt receptors, thus enhancing Wnt signalling. Colon organoids require the addiction of Wnt3a and R-spondin because of the lack of Paneth cells that produce them physiologically.

Noggin: Noggin is a secreted glycoprotein which is an antagonist of the BMP family members. The inhibition of BMP proteins, using Noggin or Gremlin, increments the number of ISCs present *in vivo* and it is essential for long-term culture organoids[85].

EGF: Intestinal organoids require the activation of KRAS and PI3K/Akt pathway mediated by EGF to sustain their formation and growth. EGF can also be substituted by HB-EGF or IGI1 since they activate the same pathways[87].

ALK4/5/7. TGF-β hampers organoid proliferation. For this reason, the production of TGF-β must be inhibited using a family of type I receptors of TGF-β that inhibit the function of the ligand.

Notch signalling: ISCs require a constant engagement of Notch ligands (DII1 and DII4) expressed on nearby cells. The inhibition of Notch results in the up-regulation of these ligands. To generate intestinal organoids, ISCs need a constant connection with cells expressing ligands of Notch[88].

Normal and malignant intestinal organoids require the same protocol to grow, except for Wnt3 and R-spondin which is not needed for tumor organoids[46]. In fact, most of CRCs are characterized by the presence of mutations in the APC gene that constitutively activate the Wnt signalling pathway. The use of this 3D model allowed to recreate *in vitro* tumors that recapitulate their original morphological, structural and functional composition.

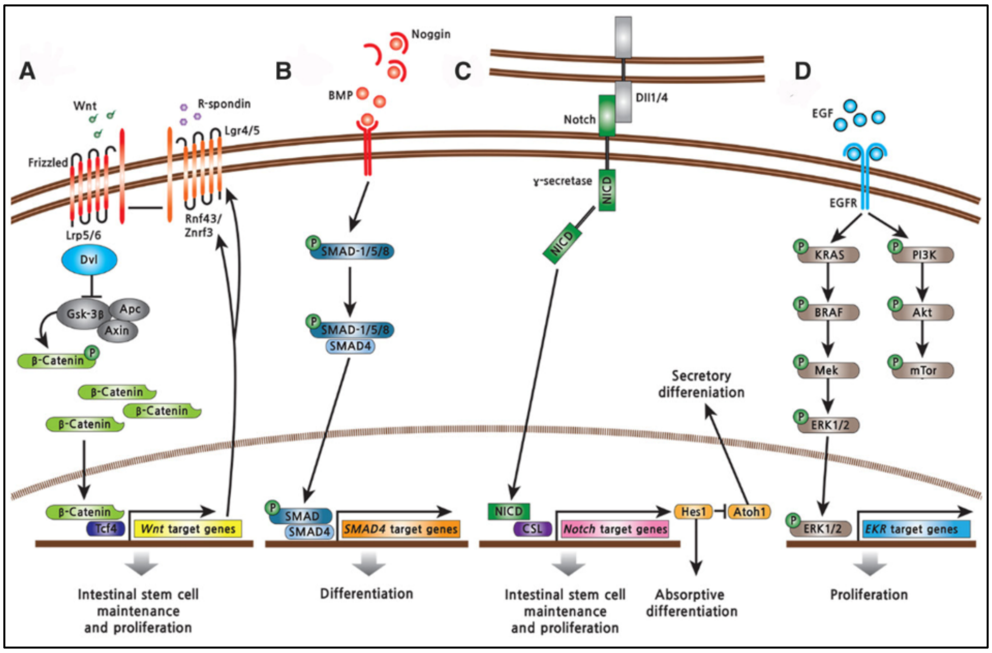


Figure 10 | Signalling pathways regulating intestinal stem cells (ISCs). Adapted from Hong, Dunn, Stelzner, & Martín, 2017[89]

1.4 The Epigenetics

Epigenetics is a branch of biology which studies potentially stable and heritable changes that alter gene activities without changing the underlying DNA sequence[90,91]. Many types of epigenetic processes have been identified including DNA-methylation, histone modifications, *cis*-regulatory elements, nucleosomes remodelling and non-coding RNAs (miRNA, siRNA, piRNA, lncRNA). All these mechanisms are essential for physiological cell fate transition and maintenance of tissue specific gene expression. The next sessions will focus on the specific epigenetic processes studied in this project (*i.e.* histone modifications and *cis*-regulatory elements).

1.4.1 Chromatin structure and organization in normal cells

Chromatin is a regulated macromolecular complex composed by repetition of nucleosome units (**Figure 11**) made by a DNA segment of ~146 bp wrapped around an octamer core of four histone proteins (H2A, H2B, H3, H4). Gene activity can be modified through the interaction between chromatin and regulatory factors[92]. The interaction between the DNA and the core histone modifications is guaranteed by non-covalent bonds between the phosphate residues located in the minor groove and lysine, arginine side chain and main chain amide nitrogen of the proteins. Core histones display structurally undefined and evolutionary conserved "tail" domains[93]. The amino terminal portion of the histones contains specific residues subjected to post-translational modifications, such as methylation, acetylation, ubiquitylation, sumoylation and phosphorylation that correlate with changes in transcriptional regulation[94]. The interplay of these modifications creates an "epigenetic landscape" that can define cellular identity, different developmental stages and disease conditions, including cancer.

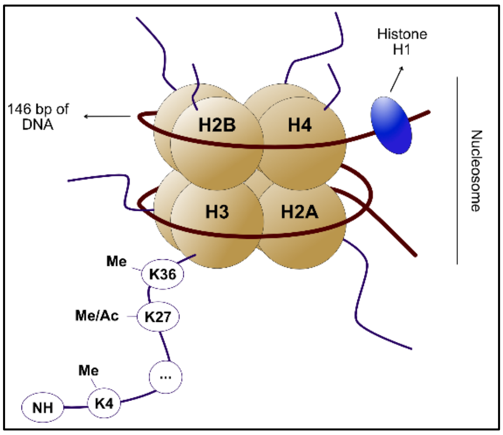


Figure 11 | Schematic representation of nucleosome structure.

1.4.2 Histone modifications

The combinatorial pattern of histone modifications, also known as "histone code", affects chromatin organization and activity by determining its accessibility and the recruitment of epigenetic modifiers.

The plethora of the epigenetic players are classified in *writers,* which are enzymes involved in the addition of covalent modifications; *readers,* which are able to interpret the histone code and *erasers,* which can remove the chemical modifications[95] (**Figure 12**). These enzymes can also be classified in two main categories; one involved in epigenetic activation and the other in repression, depending on the type and the position of the modifications.

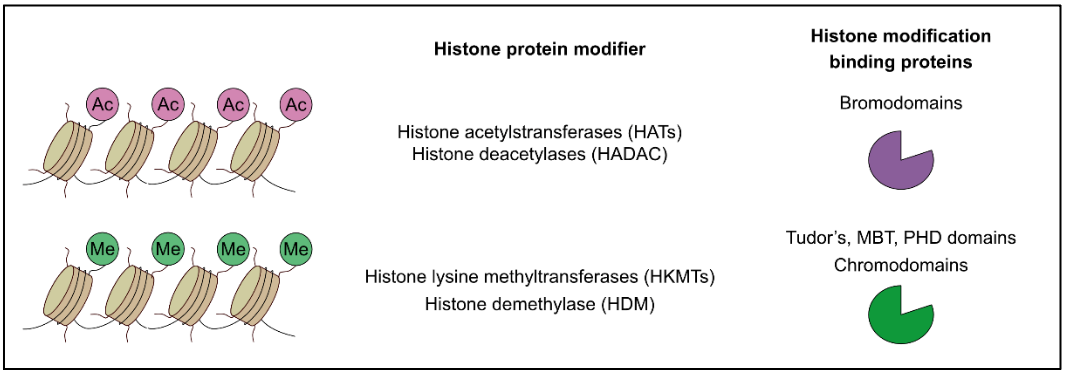

Histone protein modifier
Histone modification binding proteins
Ac
Ac
Ac
Ac
Histone acetylstransferases (HATs)
Histone deacetylases (HADAC)
Bromodomains
Me
Me
Me
Me
Histone lysine methyltransferases (HKMTs)
Histone demethylase (HDM)
Tudor's, MBT, PHD domains
Chromodomains

Figure 12 | Overview of proteins actively involved in the addition, interpretation and removal of post-translational modifications. In particular, the upper panel reports proteins involved in acetylation and the bottom panel the proteins involved in methylation.

1.4.2.1 Histone acetylation

The balance of histone acetylation and deacetylation is a key process associated with several regulatory mechanisms, such as transcription, chromatin dynamics, DNA repair and differentiation. Histone acetyltransferases (HATs, *writers*) catalyse histone acetylation whereas histone deacetylases (HDACs, *eraser*) perform the reverse reactions (**Figure 12**). Histone acetylation occurs by the enzymatic addition of an acetyl group (-COCH3) from an Acetyl-CoA, mediated by HATs[96]. The acetylation of histones mediated by HATs leads to a reorganization of the chromatin which becomes more accessible and available for transcription factor binding. HATs are classified into type A and type B according to cellular localization. Type A are found predominantly in the nucleus and they catalyse gene expression processes. Type A can be further subdivided into five families according to conformational structures and amino-acid sequence homology, such as GNAT, CBP/p300, MYST, TAF1, p600[97]. Type B are found in the cytoplasm and acetylate newly translated histones but not those already deposited onto chromatin. Acetylated lysine residues can be specifically bound by the reader of acetyl-lysine binding domain: Bromodomain (BRD). BRD4 is recruited by other transcriptional regulators at the promoters and enhancers of many genes boosting their expression[98]. BET inhibitors, such as JQ1, are able to dissociate BRD4 from the acetylated histone of promoters and enhancers leading to downregulation of gene transcription[98,99]. Dysfunction of BRD proteins has been associated with several diseases and also with cancer onset[100]. A plethora of specific acetylated sites have been identified in each of the core histones, mainly located toward the N-terminal. This work is focused in particular on the most known acetylated histone, the acetylation of the 27th lysine residue of H3. H3K27ac is localized in the proximity of the transcription start site (TSS) of actively transcribed genes in the co-presence of H3K4me3. Distal regulatory elements (*i.e.* enhancers) also show increased levels of H3K27ac in combination with mono-methylation of H3K4[96] (**Figure 13**).

1.4.2.2 Histone methylation

During methylation events an alkylation reaction occurs, replacing a methyl group to a hydrogen atom. This reaction is catalysed by methyltransferases (MT, *writers*) which use a high energy methyl donor, the S-adenosylmethionine (SAM). Writers cooperate with histone demethylases (HDM, *eraser*) to remove methyl group, and methyl *readers* to recognise them[101]. Histone lysine methyltransferases (HKMTs) are classified in two main groups based on the presence or absence of the SET (Su(var)3–9, Enhancer of Zeste, and Trithorax) catalytic domain[102], which harbours the enzymatic activity. HDM is composed by two families, the LSD1 family and JmJC domain-containing family. Histone methylation adds a level of complexity inducing structural changes and influencing the chromatin folding via electrostatic mechanism. Usually, methylation occurs mainly on the side chains of Lysine and Arginine residues. Lysine can be mono (me1), di (me2), tri-methylated (me3) and Arginine can be mono– and di–methylated[101]. Unlike acetylation, none of these modifications change the electric charge of the histone proteins. Therefore, the main methylation functions are accomplished by effector molecules that specifically recognise the methylated sites. The methylation *readers* are classified in different classes, including PHD, chromo, Tudor, PWWP, WD40, BAH, ADD, ankyrin repeat, MBT and zn-CW domains[102] (**Figure 12**). Histone methylation can be associated with activation or repression according to the position of the modified residues (**Figure 13**).

Tri-methylation of H3K4 is predominantly spread around the TSS of active genes and promotes transcription through interaction with RNA polymerase II (RNA pol II) (**Figure 13**). The distribution of H3K4me3 is highly correlated with unmethylated CpG islands. SET1 family and MLL proteins are involved in the regulation of H3K4me3[103].

Mono-methylation of H3K4 is spread inside genic regions or in intergenic regions (**Figure 13**) and is recognised by MLL3/MLL4. The occurrence of H4K4me1 determines the presence of putative enhancer regions (*primed* enhancers). The co-localization of H3K4me1 and H3K27ac or H3K27me3 indicates the presence of *active* or *poised* enhancers, respectively[104].

Tri-methylation of H3K36 tends to spread along the gene body and to increase toward the 3’ of the gene (**Figure 13**). H3K36me3 is regulated by the SETD2 methyltransferase which is recruited by RNA Pol II during transcriptional elongation. In fact, the H3K36me3 level is higher on exons of actively transcribed genes supporting the transcriptional machinery and preventing spurious transcription[105].

Tri-methylation of H3K27 is spread along the gene body of repressed genes (**Figure 13**) and it is regulated by the Polycomb group (PcG) of proteins. PcG antagonizes transcriptional activation, therefore H3K27ac and H3K27me3 are mutually exclusive at the promoter level[106].

1.4.3 Nucleosome free regions

The regulation of gene activity is further influenced by chromatin organization and nucleosome remodelling. Transcriptional regulatory elements, such as promoters and enhancers, show low nucleosome occupancy resulting in nucleosome free regions (NFR)[107] (**Figure 13**). The modulation of accessible regulatory sites is orchestrated by ATP-dependent chromatin remodelling complexes which can shift and remove nucleosomes. The increased accessibility of regulatory regions is tightly correlated with gene activation since it facilitates transcriptional machinery assembly and the binding of transcription factors[108].

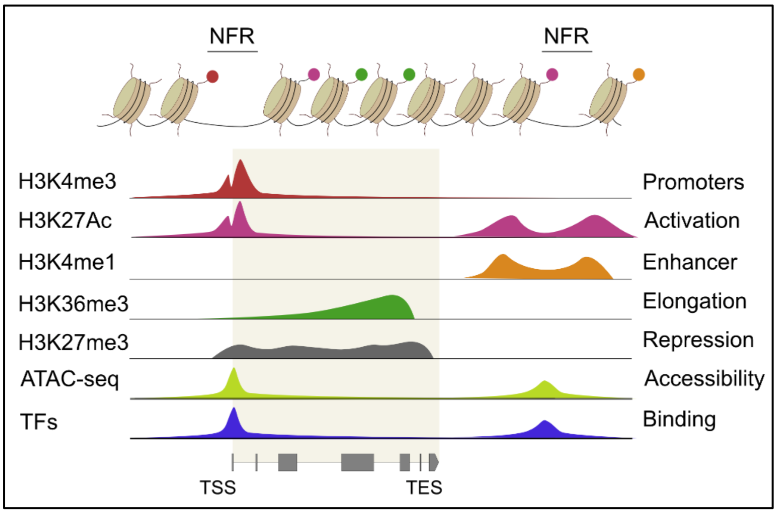


Figure 13 | Schematic overview of a genic region (grey box). Each line reports the name and the role of a histone modification, followed by the accessible regions detected by ATAC-seq, and finally the profile of a representative transcription factor (TF).

1.4.4 Enhancer regulatory elements

Enhancers are key non-coding regulatory elements orchestrating gene regulation in human development, homeostasis and disease. They can be located in genic or intergenic regions and their role is to enhance the transcription of *cis*-located targets (**Figure 14**). Genes can have multiple enhancers and each enhancer can act in different ways to modulate the transcriptional rates according to environmental stimuli. Their action is independent of the orientation of the enhancing region. This *cis*-acting regulatory elements are nucleosome free regions flanked by histones carrying H3K27ac and H3K4me1 modifications. Inside each enhancer there is a cluster of transcription factor binding motifs on which regulatory factors specifically bind to accomplish their activity[109]. The distance between an enhancer and its cognate promoter can be from a few to ten Kb and their interaction is achieved by a loop in the 3D space (**Figure 14**), making them difficult to map. For this reason, chromosome conformation capture techniques (*i.e.* 3C, 4C, Hi-C) are used to systematically annotate the interaction between regulatory sequences and their target promoters[110,111].

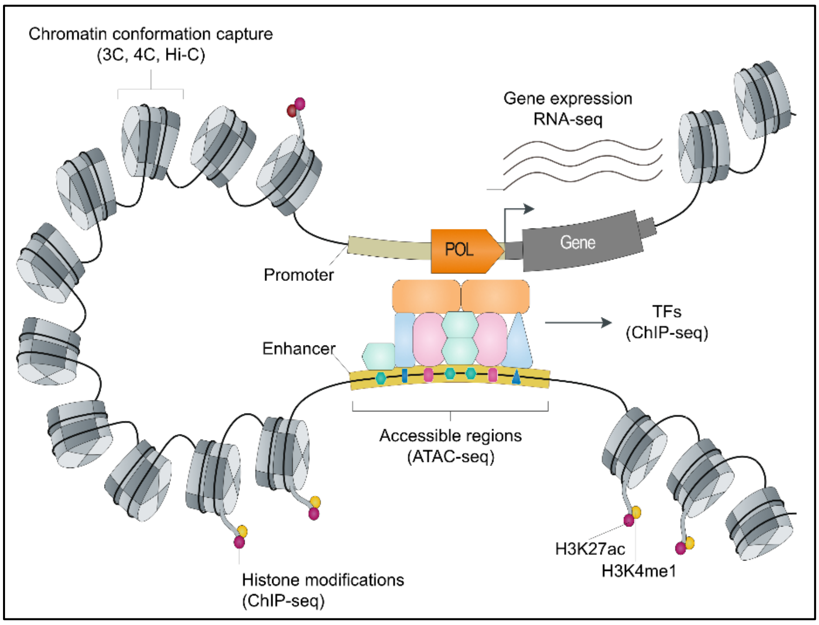


Figure 14 | Schematic representation of enhancer-promoter regulation. The enhancer element is bound by a cluster of transcription factors and it loops in the 3D space to interact with its target gene.

Enhancers are characterized by high heterogeneity between tissues and species, highlighting their importance in gene regulation. To delineate their role, several consortiums have been created in order to collect information about regulatory elements across a wide range of tissues and species, such as the Encyclopaedia of DNA elements (ENCODE)[112], the Roadmap epigenomic project[113,114], the Blueprint epigenome[115] and the TCGA consortiums[116].

1.4.5 Epigenetics in cancer

Cancer has been typically considered as a genetic disease characterized by mutations in genes that control cell proliferation and apoptosis. In “The hallmarks of cancer”[5], it was assumed that malignant mechanisms are fundamentally rooted in genetic alteration[5,6]. However, epigenetic deregulation has recently emerged as a new paradigm of cancer influencing cancer initiation and progression (**Figure 15**). The epigenetic mechanisms can contribute in different ways to each hallmark: global changes in DNA methylation, chromatin states and cis-regulatory elements, as well as genetic aberrations in chromatin proteins characterize more than 50% of human cancers[117,118]. Enhancers, identified primarily via global mapping of histone modifications, are critical cis-regulatory elements for cell fate determination and tissue-specific gene regulation[119]. Gains and losses of cell-type specific enhancer activity contribute to cell reprogramming towards tumor growth and metastasis[120,121], rendering enhancer dysfunction a promising biomarker of diagnosis and a critical target for therapeutic intervention[122]. For instance, Roe *et al.* in 2017, exploiting the organoids model, illustrated how the metastasis transition in pancreatic ductal adenocarcinoma (PDA) was mediated by massive alteration of the enhancer activity that was in turn driven by the pioneer factor FOXA1[123]. A plethora of different studies were published in order to identify both "private" enhancer regions specific for different tumor subtypes[124] as well as conserved enhancers across diverse tumor subtypes[125,126] and the putative transcription factors orchestrating their aberrant activities.

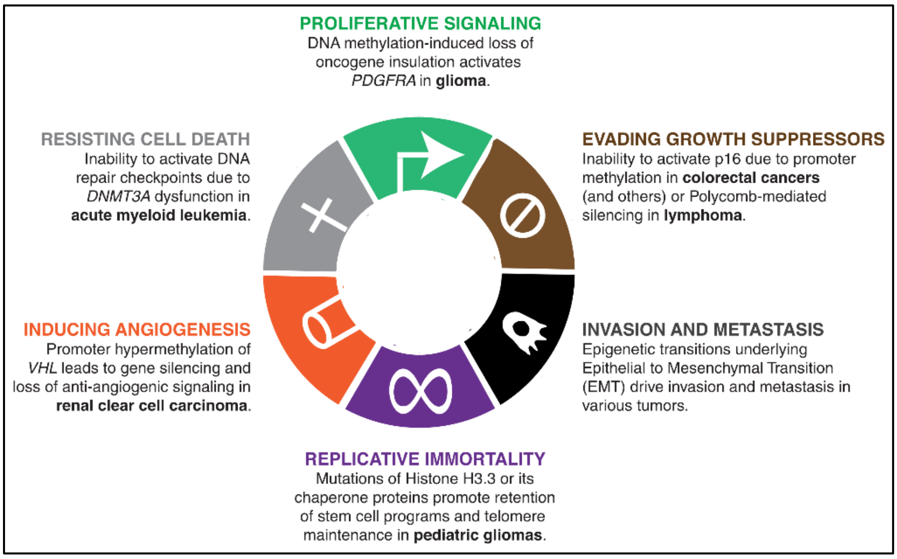

PROLIFERATIVE SIGNALING
DNA methylation-induced loss of oncogene insulation activates *PDGFRA* in **glioma**.
RESISTING CELL DEATH
Inability to activate DNA repair checkpoints due to *DNMT3A* dysfunction in **acute myeloid leukemia**.
EVADING GROWTH SUPPRESSORS
Inability to activate p16 due to promoter methylation in **colorectal cancers** (and others) or Polycomb-mediated silencing in **lymphoma**.
INDUCING ANGIOGENESIS
Promoter hypermethylation of *VHL* leads to gene silencing and loss of anti-angiogenic signaling in **renal clear cell carcinoma**.
INVASION AND METASTASIS
Epigenetic transitions underlying Epithelial to Mesenchymal Transition (EMT) drive invasion and metastasis in various tumors.
REPLICATIVE IMMORTALITY
Mutations of Histone H3.3 or its chaperone proteins promote retention of stem cell programs and telomere maintenance in **pediatric gliomas**.

Figure 15 | Each classic hallmark of cancer can be caused by chromatin aberration. Taken from (Flavahan, Gaskell, & Bernstein, 2017[92]).

1.4.6 Transcriptional addiction

Transcription factors (TFs) are families of proteins that recognise and bind specific DNA sequences (motifs) creating a cooperative network to synergistically regulate chromatin organization and gene transcription. This complex regulatory network is usually orchestrated by a "master" regulator which is located at the top of the regulatory hierarchy and regulates multiple downstream genes either directly or through a cascade of gene expression changes[127]. The correct regulation of master transcription factors themselves is fundamental for the maintenance of normal development and homeostatic conditions[128]. Changes in TF expression can lead to the onset of several pathologies, including cancer[129]. The role of TFs, and of master regulators in particular, is also of interest in light of the emerging paradigm of transcriptional addiction in cancer[130] (**Figure**

16). This concept describes the dependence of cancer cells on transcriptional regulators, including chromatin regulators and TFs, as a result of their uncontrolled proliferation and growing need for the basal transcriptional machinery[98,130,131].

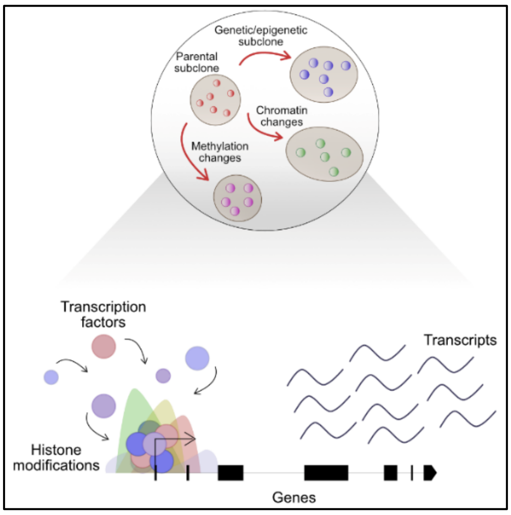


Figure 16 | Schematic representation of cancer transcriptional addiction. The high demand of chromatin regulators boosts the transcriptional rate of target genes.

1.5 Multi-omics approach to integrate transcriptomic and epigenomic data

Understanding how epigenomics and which specific epigenetic alterations can mediate the aberrant transcriptional programs of cancer cells can be greatly facilitated by the investigation of omics data and the development of integrated multi-omics approaches. Many of the available omics methods are based on the advent of next generation sequencing (NGS), one of the most important technological revolutions that kicked off modern biology. NGS, also known as high-throughput sequencing, is a term that comprehends a number of different modern sequencing technologies that have revolutionized the study of genomics and molecular biology over the last two decades. The low cost, high speed and rapid improvement of NGS technologies make them the perfect tool to study the functional annotation of the human genome, boosting the genetic, transcriptomic and epigenomic fields of research. The pioneers of sequencing technologies were Maxam and Gilbert[132] in 1977 with the creation of the chemical chain termination method for DNA sequencing, followed by Sanger' dideoxy method[133]. Sanger sequencing is considered as the first generation of sequencing and it has been used for more than thirty years. In 1990, the creation of these new methods led to the initiation of the largest international scientific research project known as the Human Genome Project promoted by the National Institutes of Health (NIH) in USA. This project lasted for thirteen years with a cost of three billion dollars and its aim was to reconstruct the entire human genome sequence and catalogue the position and functions of all the genes in the genome.

In 2004, the National Human Genome Research Institute (NHGRI) promoted a program to reduce the cost of the human genome sequencing to one thousand dollars over ten years. In the next few years, a series of second-generation DNA sequencing technologies have arisen with the capability to perform a large number of parallel sequencing reactions on a micrometer scale. They included 454 Life Science (later purchase by Roche), Solexa/Illumina sequencing platform by Illumina and SOLiD by Applied Biosystem (now Life Technologies). The advantages of these new systems were that they didn't require bacterial cloning of the fragments and electrophoretic technology platforms. Nowadays, the most widely-used technologies are Illumina platforms (also used in this study) and the costs have been dramatically reduced.

However, despite the major impact in biology, NGS technologies have also some limitations:

- Relatively short reads. NGS read refers to the number of base pairs of DNA fragment which are sequenced and converted into a string of letters. Along the human genome there are repeated sequences which are longer than the length of NGS reads causing misassembling and gaps.
- Larger structural variations (SV) are difficult to characterize. This is a major limitation since SV are implicated in a large number of genetic diseases.
- NGS methods rely on PCR amplifications. This can be a problem with GC rich regions since they are inefficiently amplified.

To fill these gaps, a third-generation sequencing (TGS) or long read sequencing has been developed. The characteristic of TGS is to sequence single molecules without DNA amplification and to obtain longer reads (up to 10 Kb long reads). One of the first TGS technologies developed was the single molecule real time (SMRT) platform from by Pacific Biosciences in 2011 and more recently (2014) the nanopore sequencing from by Oxford Nanopore Technologies (ONT).

1.5.1 Transcriptomic profile

1.5.1.1 Bulk RNA-seq

Transcriptomics refers to the study of the abundance and composition of the cells' transcriptome. The mRNA transcribed is highly dynamic and varies across different cell types, cell states and regulatory mechanisms. The advent of whole-transcriptome sequencing has provided many advantages in transcriptome quantification over existing approaches, such as hybridization-based techniques also known as microarrays[134]. One of the advantages of RNA-seq is that is has very low background noise and no upper limit in quantification, unlike microarrays that lack sensitivity for very lowly expressed genes (due to low fluorescence signal detection) or highly expressed ones (due to probe saturation). In fact, with RNA-seq it is possible to detect a large dynamic range of expression levels, up to 8000-fold dynamic range[135]. The rapid development of new approaches in transcriptomic studies (*i.e.* unique molecular identifier[136], PCR-free techniques[137]) further expands the potential applications of this technology. Moreover, in the last few years, the cost for the preparation and sequencing of RNA-seq samples has been dramatically reduced. This provides the possibility to sequence a larger number of samples at higher depth with a reduced cost. Depending on the sequencing technology and bioinformatic methods being employed, RNA-seq has applications that reach beyond the mere quantification of gene expression levels:

Splicing alternative: pair-ended RNA sequencing can reveal connectivity among multiple exons, providing information on splicing and alternative splicing events, which characterize complex transcriptomes.

De novo transcript assembly: RNA-seq can also be used to discover new transcripts that have not been identified yet, including small RNAs, lincRNAs, circular RNA etc[138].

Epitranscriptomic: RNA modifications could control the function of different class of RNAs, such as mRNA, small RNAs, linc-RNAs, and such modifications can also affect gene expression[139,140].

Disease classification: With the advent of high throughput techniques, it has become possible to stratify a disease condition according to gene expression subtypes[27,141], as described in section 1.2.2 for CRC.

Integration: Cancer cells are characterized by aberrant transcriptional profiles which can be the result of genetic and epigenetic changes. Understanding how and which epigenetic alterations can affect the gene expression programs of cancer cells can be greatly facilitated by the investigation of omics data and the development of integrated multi-omics approaches. Transcriptomic studies have facilitated the investigation of different biological mechanisms and the study of a plethora of diseases, including cancer. In recent years, the field of cancer NGS transcriptomics has rapidly evolved leading to scientific collaborations, or consortia, with the aim to deeply characterize the transcriptional profiles across different types of cancer. Among the projects that mainly contributed to the profiling of cancer transcriptome, the most renowned are the Genotype–Tissue Expression (GTEx) project and the Human Protein Atlas (HPA) for the sequencing of normal tissues; The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) for the sequencing of different types of tumor tissues and finally the Encyclopedia of DNA Elements (ENCODE) and Genentech for the sequencing of cell lines.

1.5.1.2 Single cell RNA-seq

Recently, a new technology able to detect the transcriptome of single cells in a tissue has been developed. Previously, RNA-seq was typically performed in "bulk"; each library was capturing different populations of cells and thus the data represented the average gene expression profile of thousands of diverse cell types belonging to the sampled population. To perform a targeted study there was the need to select a specific population, *e.g.* by cell sorting, leading to several challenges due to the limited amount of starting material. To address these limitations, a technology, termed single cell RNA-seq (scRNA-seq) was developed that is able to capture the variability of gene expression inside heterogeneous populations. With scRNA-seq we are now able to detect and analyse the transcriptome of each single cell present in a heterogeneous tissue. Currently there are several protocols which vary according to the quantification method and cell capture. The scRNA-seq platforms can be split in two main groups: i) *full-length transcript sequencing approaches, i.e.* SMART-seq2[142], and ii) *3'end sequencing approaches*, *i.e.* CEL-seq[137] and Drop-seq[143], Chromium[144]. Commercial kits are also available, such as the 10X Genomic Chromium platform which has been used in this study. The higher noise of scRNA-seq data compared to bulk-RNA-seq data has raised new computational challenges in bioinformatics analysis. For instance, although the primary analysis of scRNA-seq[145] includes the basic steps and tools used for bulk RNA-seq (Quality control, alignment and quantification), the downstream analyses require the adaptation of existing methods or the development of new ones[146].

1.5.2 Epigenomic profile

1.5.2.1 ChIP-seq, ATAC-seq and ChIPmentation

The increasing power of NGS technologies has encouraged the development of several approaches for the study of key regulatory elements, including the post-translational modification of histone proteins (indicative of the presence of promoter, enhancer, or repressed regions) and transcription factor binding. One of the most commonly-used techniques to untangle the epigenetic landscape of different cell populations, such as cancer, is chromatin immunoprecipitation followed by sequencing (ChIP-seq). Briefly, chromatin is immuno-precipitated subsequently to the binding of an antibody targeting a specific post-translational modification of histones or a DNA-bound transcription factor, as extensively described in this review[147]. A major limitation of this technique is the need of a large amount of starting cells to perform the protocols. The limited number of cells for some populations (*i.e.* human primary cells) raised the need to create new methods that require low input material (500-50,000 cells). Among these, one of the first applications to be designed in the epigenetic field was ATAC-seq (Assay for Transposase-Accessible Chromatin)[148]. This technique allows to map the chromatin accessibility patterns similar to DNAse-seq and MNase-seq but using less cells. ATAC-seq takes advantage of a hyperactive enzyme, Transposase 5 (Tn5), which is able to directly insert sequencing adapters in open chromatin regions and nucleosome free regions proving information on the accessibility of the chromatin. This kicks off a plethora of new epigenetic techniques, such as ChIPmentation[149], which exploited the hyperactive Tn5 solution to get post-translational modification of histones or transcription factor binding information from a limited amount of starting material. Overall, these epigenomic techniques are qualitative approaches and not quantitative. They require a very careful experimental design, and the selection of the correct concentrations, material, number of replicates and sequencing depth. Since these technologies are extremely challenging the field of single cell epigenomics is starting to grow albeit at a slow pace[150].

2. Aims of the project

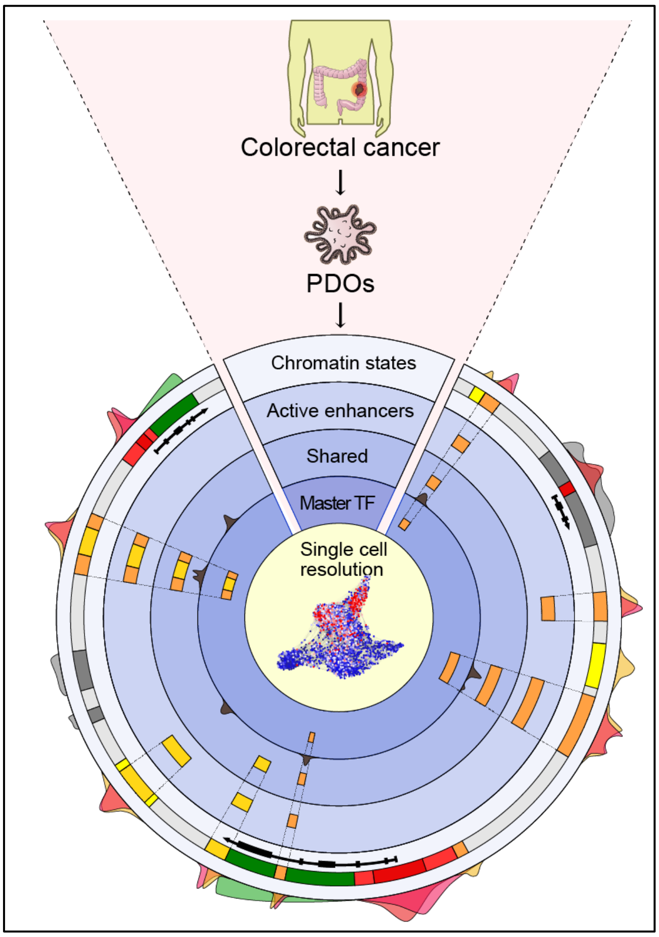


Figure 17 | Graphical representation of the work.

CRC is, historically, the leading paradigm of the concept of "cancer as genetic disease". And yet what remains largely unclear in this "mutation centered" view are the downstream transcriptional effects of the CRC mutations. Also unclear is how the CRC genetic complexity converges to just few transcriptional subtypes that exist despite the observed heterogeneity in tumor phenotypic states. These outstanding issues highlight the need of a paradigm shift: re-assessing inter-patient variability through the prism of a shared regulatory architecture and associated underlying mechanisms.

With this background in mind, in this work, we seek to unravel the chromatin landscape of human CRC by exploiting the organoid model in order to identify a common epigenetic blueprint and investigate its relevance in other types of cancers (**Figure 17**).

The project aims to develop an integrated multi-omics approach in order to pursue the following specific goals:

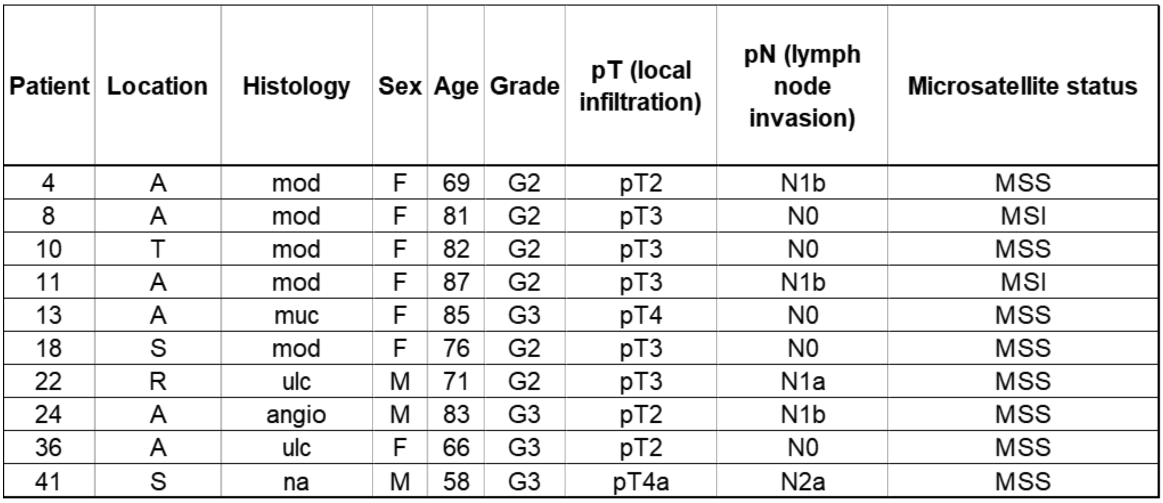
- Generate a balanced *ex vivo* library of PDOs that reflects the different clinical and molecular subtypes of CRCs.

- Generate and combine several chromatin maps for multiple histone modifications to unravel, for the first time, the epigenetic mosaic of CRC, accurately defining both active and repressed regulatory elements through *de novo* chromatin state discovery
- Exploit the interpatient variability inherent in our balanced PDOs library, to identify an aberrant CRC enhancerome largely shared by different CRCs, irrespectively of tumor subtype.
- Identify the master transcription factor orchestrating this shared regulatory architecture.
- Verify if a fraction of the CRC aberrant enhancers could be shared also by other types of cancers resulting in a targetable pan-cancer epigenetic fil rouge at the roots of cancer cell transcriptional addiction and tumor maintenance.
- Investigate at single cell resolution the specificity in malignant cells of the genes regulated by the core regulatory blueprint.

3. Methods

3.1 Human specimen

Primary CRC tissues were obtained from San Gerardo Hospital (Department of Surgery), Monza and UO Chirurgia Epatobiliopancreatica e Digestiva Ospedale San Paolo, Milan following ethical approval from their Institutional Review Boards. Informed consent was obtained from all patients prior to acquisition of the samples. Clinical details on patients are reported in **Table 1**. Samples were confirmed to be tumor or normal based on pathologist assessment and were obtained prior to treatment. MSI-MSS status was determined according to standard experimental procedures[151].



Patient	Location	Histology	Sex	Age	Grade	pT (local infiltration)	pN (lymph node invasion)	Microsatellite status
4	A	mod	F	69	G2	pT2	N1b	MSS
8	A	mod	F	81	G2	pT3	N0	MSI
10	T	mod	F	82	G2	pT3	N0	MSS
11	A	mod	F	87	G2	pT3	N1b	MSI
13	A	muc	F	85	G3	pT4	N0	MSS
18	S	mod	F	76	G2	pT3	N0	MSS
22	R	ulc	M	71	G2	pT3	N1a	MSS
24	A	angio	M	83	G3	pT2	N1b	MSS
36	A	ulc	F	66	G3	pT2	N0	MSS
41	S	na	M	58	G3	pT4a	N2a	MSS

Table 1 | Clinical information of patients used in the study. Location abbreviation. A: ascending colon; T: transversal colon; S: sigmoid colon; R: rectum. Histology abbrev. mod: moderately differentiated adenocarcinoma; muc: mucinous adenocarcinoma; ulc: ulcerated; angio: angioinvasive; na: not available. Sex abbrev. M: male; F: female. Microsatellite status abbrev. MSI: Microsatellite instable; MSS: Microsatellite stable.

3.1.1 Isolation of human primary tissues

Primary colonic normal and tumoral tissues were processed according to a previously published protocol[152]. The detailed experimental protocol and materials used for this procedure are reported in the **Appendix**.

3.2 Patient derived colorectal cancer organoids culture

PDOs were established and maintained as previously described[152]. The detailed experimental protocol and materials used for this procedure are reported in the **Appendix**.

3.2.1 Whole mount staining of PDOs

Isolated organoids embedded in Matrigel in μ-Plate Angiogenesis 96 Well (Ibidi) were fixed in 4% paraforlmaldehyde in PBS for 1 hour, at 4°C. The whole mount staining protocol was performed as previously described[153], with some modifications. The detailed experimental protocol and materials used for this procedure are reported in the **Appendix**.

3.2.2 Immunohistochemistry and in situ hybridization

CRC primary tissues and PDOs were immunohistochemically stained on formalin-fixed, paraffin-embedded or on fresh OCT-embedded tissue and PDOs sections as previously described[155]. anti-YAP (13584-I-AP; Spring Bioscience) and anti-WWTR1/TAZ (HPA007415; Sigma) were used as primary polyclonal antibodies. RNAscope Duplex Detection Kit (Chromogenic) was used to process RNA in situ detection from tissue sections (formalin fixed, paraffin embedded) according to the manufacturer's instructions (Advanced Cell Diagnostics). RNAscope probe was FOXQ1 (NM_033260.3, region 694 - 2197), which was detected using the HRP-based Green detection reagent.

3.3 RNA-seq and ChIP-seq preparation

3.3.1 RNA isolation and bulk RNA-seq library construction

To perform RNA-seq analysis, CRC PDOs, primary normal and tumor tissues were lysed in TRIzol reagent (Thermo Fisher) and processed for total RNA extraction with PureLink™ RNA Mini Kit (Thermo Fisher), according to manufacturer's instructions. The PDOs samples were collected at different time points, from early (<5 splits) to late passages (>5 splits). The quality of RNA was checked using RNA Integrity Number (RIN) value with RNA6000 assay (Agilent). In this study, only samples with RIN > 7.0 were used. RNA-seq libraries were constructed according to the TruSeq mRNA Stranded preparation kit (Illumina, San Diego, USA) and sequenced at HiSeq2500.

3.3.2 Chromatin Immunoprecipitation (ChIP) assay and library construction

For ChIP experiments, matrigel droplet containing ~0.3 x 10^6 organoid cells/well was dissolved using Cell Recovery Solution (Matrisperse Cell Recovery Solution - Sacco-L004419 CPB40253), following the indicated procedure. Organoids pellet was fixed rocking at room temperature and quenched. PBS-washed organoid pellets were lysed and incubated for at 4°C. Lysed chromatin was sheared at 200–500 bp fragments using Covaris® M220 focused-ultrasonicator. For organoids and crypts chromatin was incubated with antibody overnight at 4 °C on wheel. Antibody/antigen complexes were recovered and washed, followed by reverse crosslinking overnight. The washed immunocomplexes were incubated with ChIP elution buffer for reverse crosslinking. The immunoprecipitated DNA was then purified and eluted. The detailed experimental protocol and materials used for this procedure are reported in the **Appendix**. ChIP-seq libraries were constructed with TruSeq ChIP Library Preparation Kit (Illumina), according to the manufacturer's instructions and sequenced on the Illumina HiSeq2500 platform.

3.3.3 ChIPmentation assay and library preparation

ChIPmentation was carried out as previously described[149] with small modifications in cell lysis and washes after recovering. Briefly, the crosslinked pellet was lysed in buffer I in ice. The pellet was recovered and lysed with buffer II at room temperature and sonicated in lyses buffer III using Covaris® M220 focused-ultrasonicator. Sonicated chromatin was incubated with anti-WWTR1 (Sigma Aldrich, HPA007415) or anti-YAP1 (abcam 52771) overnight at 4 °C on the wheel. For control libraries, an immunoprecipitation with nonspecific IgG rabbit antibody was used. Antibody/antigen complexes were recovered with blocked ProteinG-Dynabeads (Invitrogen) and washed with low salt wash buffer (twice), high salt buffer (twice) and once with Tris pH8. Beads were then resuspended and incubated in tagmentation reaction containing Tagment DNA Enzyme from the Nextera DNA Sample Prep Kit (Illumina). Beads were then washed and incubated with elution buffer plus Proteinase K (NEB) to revert formaldehyde cross-linking. The detailed experimental protocol and materials used for this procedure are reported in the **Appendix**. Library preparation for ChIPmentation was performed using custom Nextera primers as described for ATAC-seq and enriched libraries were purified using 1.8V of SPRI AMPure XP beads and sequenced with Illumina HiSeq2500.

3.3.4 10X based single cell library preparation

Primary colonic tumoral tissues were processed as described above (**Isolation of human primary tissues**) and previously[152]. The tumoral tissue cell suspension was further reduced to single-cell level through incubation with TrypLE express (Thermo scientific) at 37°C, pipetting every 2-3 min, up to 20 min. Single-cell suspension was washed, filtered with a 40 µm cell strainer and then loaded into the Chromium System (10x Genomics), targeting 10,000 cells. Following the manufacturer instructions, barcoded sequencing libraries were generated using Chromium Single Cell 3' v2 Reagent Kit and sequenced on an Illumina HiSeq 4000 platform.

3.4 Scalability and reproducibility of bioinformatics pipelines

To ensure the scalability and reproducibility of the bioinformatics pipelines used for the primary analysis of both RNA-seq and ChIP-seq data (including quality control of sequencing data, alignment, quantification, peak calling and generation of coverage tracks), we developed our pipeline integrating Nextflow[156] and container technologies, including Docker (https://docs.docker.com/) and Singularity (https://singularity.lbl.gov/) (**Figure 18**). Nextflow is a pipeline manager which guarantees the execution and reproducibility of custom and publicly available pipelines. Nextflow is able to work with pipelines written in different programming languages making it a flexible and powerful tool to package an entire scientific workflow. The integration of Nextflow with container technology allows full control of the computing environment, including consistency of the packages' versions. In fact, Docker is able to produce an *Image* which includes all the libraries and data needed for the analysis. This *Image* can be shared with other users and can run on any major Linux operating system. Ultimately, the integration of this data-driven toolkit for computational analysis with the container solutions enable truly reproducible analyses.

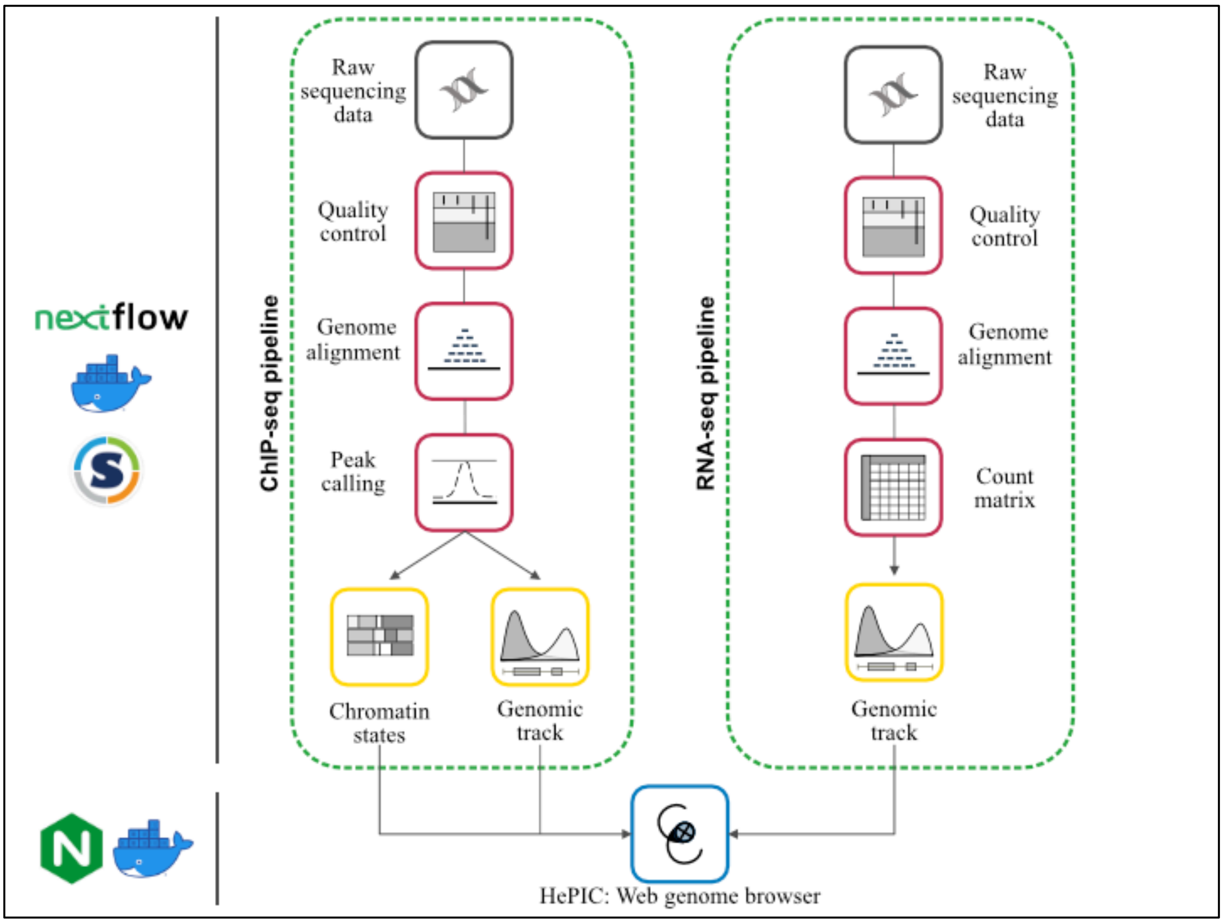


Figure 18 | Schematic representation of the bioinformatics workflow of this study for which Nextflow and container technologies were employed.

3.5 Transcriptomic data analysis

3.5.1 Processing and quality control

FastQC v0.11.769 and MultiQCv1.5 (http://www.bioinformatics.babraham.ac.uk/projects/) were used to perform the quality control of the sequenced reads. The reads were trimmed using BBDuk (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/). The trimming step is executed to remove adapters and low-quality reads that can affect the alignment on the genome. Then, the reads were aligned to the human reference genome hg38 (GENCODE Release 25 basic gene annotation) using STARv2.5.3a[157]. The alignment is a key step that assigns each read to its exact location along the genome (**Figure 19**). featureCounts4-Subreadv1.6.2[158] was used to perform the quantification of the reads. A raw count matrix that includes all the samples was created using a custom bash script (**Figure 18**).

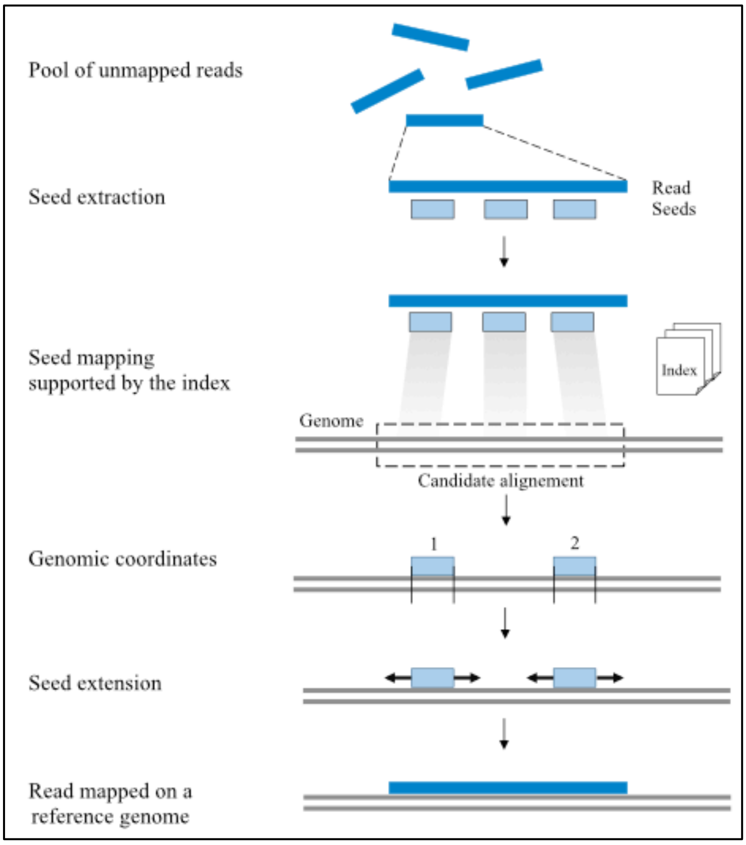


Figure 19 | STAR takes as reference the index of the genome build from a genome fasta file. The genome index is a compression of the reference genome obtained by applying Burrows–Wheeler transforms (BWT) which performs a reversible permutation of characters in a text. STAR algorithm is divided in two major steps: i) The "Seed Search", a sequential search for a Maximal Mappable Prefix (MMP), which is defined as the longest substring of a read that perfectly matches one or more substrings of a reference genome. This will become the first seed to be placed on the genome. Since, the algorithm will subsequently search only for the unmapped region of the read, this method increases the speed of the tool. ii) "Clustering, Stitching and Scoring"; the seeds are clustered together around a selected set of seeds, called "anchors". If these seeds fall into a defined genomic window, they are stitched together through a frugal dynamic programming algorithm**.**

The mitochondrial genes were removed from the downstream analyses. Normalization and differential analysis were performed using DESeq2 package[159] (version 1.22.2) and R version 3.5[160]. Principal component analysis was performed using the R function *prcomp* considering the 500 most variable transcripts with the parameters center=TRUE, scale=TRUE. The 500 highly variable genes were manually inspected and immune infiltrate genes present were removed. Genes were considered differentially expressed with a padj ≤ 0.01. A single organoid for each patient was chosen for downstream differential expression analyses in order to keep even sample sizes across the three tissue populations. Heatmap and hierarchical clustering were performed using the *pheatmap* function with default parameters (clustering_distance_cols="euclidean", clustering_method="complete"). Gene set enrichment analysis was performed using the GSEApy[161] package using the pre-ranked module with default parameters (permutation_num=1000). For the visualization of RNA-seq tracks, the normalized coverage tracks were generated using the *bamCoverage* function of deeptools[162]. The command lines used for this analysis are reported in the **Appendix**.

3.5.2 Tumor primary tissue classification

The classification of CRC primary tissues was performed following the Consensus Molecular Subtype classification[27], the CRC intrinsic subtypes classification[31] and the Sadanandam[26] classification. The classification of primary tumor samples was made using CMScaller[163] R package (https://github.com/peterawe/CMScaller) using default parameters (FDR=0.5, seed=1).

3.6 Epigenomic data analysis

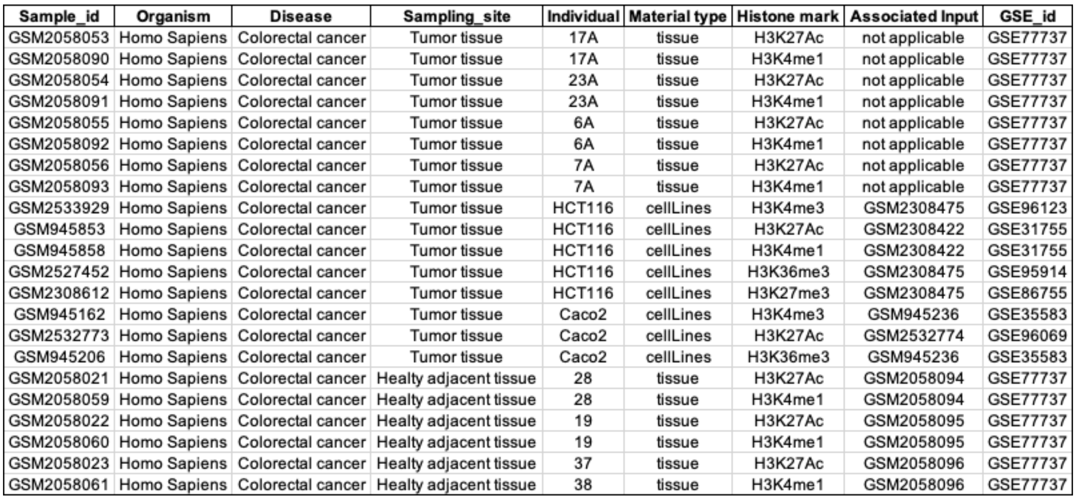
3.6.1 ChIP-seq processing and quality control

The quality control of the reads was performed using FastQC v0.11.7[164] and MultiQCv1.5 (http://www.bioinformatics.babraham.ac.uk/projects/). The reads were aligned to the human hg38 reference genome (GENCODE Release 25 basic gene annotation) using Bowtie v1.2.2[165], sorted using SAMtoolsv1.8[166] and directly converted into binary files (BAM). PCR duplicate reads were marked and removed using SAMtoolsv1.8. The peaks were called with MACS2 v2.1.0[167] using matched input DNA as a control. Peaks overlapping ENCODE blacklisted regions hg38 were removed. Peaks found in un-placed and un-localized scaffolds were removed. For the visualization of ChIP-seq tracks, Bedgraph tracks were generated using MACS2 *bdgcmp* function, converted into bigwig using UCSC bedClip and *bedGraphToBigWig* functions. The pyGenomicTrack[168] tool was used for the visualization of the tracks (**Figure 18**). The command lines used for this analysis are reported in the **Appendix**.

3.7 Analysis of publicly available data

3.7.1 Histone modifications for normal and tumor CRC tissues

ChIP-seq datasets (**Errore. L'origine riferimento non è stata trovata.**) of normal and tumor colon tissues (GSE77737), and CRC cell lines HCT116 and Caco2 (ENCODE) were reanalysed and processed using the same pipeline described above. These data were subsequently used for ChromHMM analysis.



Sample_id	Organism	Disease	Sampling_site	Individual	Material type	Histone mark	Associated Input	GSE_id
GSM2058053	Homo Sapiens	Colorectal cancer	Tumor tissue	17A	tissue	H3K27Ac	not applicable	GSE77737
GSM2058090	Homo Sapiens	Colorectal cancer	Tumor tissue	17A	tissue	H3K4me1	not applicable	GSE77737
GSM2058054	Homo Sapiens	Colorectal cancer	Tumor tissue	23A	tissue	H3K27Ac	not applicable	GSE77737
GSM2058091	Homo Sapiens	Colorectal cancer	Tumor tissue	23A	tissue	H3K4me1	not applicable	GSE77737
GSM2058055	Homo Sapiens	Colorectal cancer	Tumor tissue	6A	tissue	H3K27Ac	not applicable	GSE77737
GSM2058092	Homo Sapiens	Colorectal cancer	Tumor tissue	6A	tissue	H3K4me1	not applicable	GSE77737
GSM2058056	Homo Sapiens	Colorectal cancer	Tumor tissue	7A	tissue	H3K27Ac	not applicable	GSE77737
GSM2058093	Homo Sapiens	Colorectal cancer	Tumor tissue	7A	tissue	H3K4me1	not applicable	GSE77737
GSM2533929	Homo Sapiens	Colorectal cancer	Tumor tissue	HCT116	cellLines	H3K4me3	GSM2308475	GSE96123
GSM945853	Homo Sapiens	Colorectal cancer	Tumor tissue	HCT116	cellLines	H3K27Ac	GSM2308422	GSE31755
GSM945858	Homo Sapiens	Colorectal cancer	Tumor tissue	HCT116	cellLines	H3K4me1	GSM2308422	GSE31755
GSM2527452	Homo Sapiens	Colorectal cancer	Tumor tissue	HCT116	cellLines	H3K36me3	GSM2308475	GSE95914
GSM2308612	Homo Sapiens	Colorectal cancer	Tumor tissue	HCT116	cellLines	H3K27me3	GSM2308475	GSE86755
GSM945162	Homo Sapiens	Colorectal cancer	Tumor tissue	Caco2	cellLines	H3K4me3	GSM945236	GSE35583
GSM2532773	Homo Sapiens	Colorectal cancer	Tumor tissue	Caco2	cellLines	H3K27Ac	GSM2532774	GSE96069
GSM945206	Homo Sapiens	Colorectal cancer	Tumor tissue	Caco2	cellLines	H3K36me3	GSM945236	GSE35583
GSM2058021	Homo Sapiens	Colorectal cancer	Healty adjacent tissue	28	tissue	H3K27Ac	GSM2058094	GSE77737
GSM2058059	Homo Sapiens	Colorectal cancer	Healty adjacent tissue	28	tissue	H3K4me1	GSM2058094	GSE77737
GSM2058022	Homo Sapiens	Colorectal cancer	Healty adjacent tissue	19	tissue	H3K27Ac	GSM2058095	GSE77737
GSM2058060	Homo Sapiens	Colorectal cancer	Healty adjacent tissue	19	tissue	H3K4me1	GSM2058095	GSE77737
GSM2058023	Homo Sapiens	Colorectal cancer	Healty adjacent tissue	37	tissue	H3K27Ac	GSM2058096	GSE77737
GSM2058061	Homo Sapiens	Colorectal cancer	Healty adjacent tissue	38	tissue	H3K4me1	GSM2058096	GSE77737

Table 2 | Details of publicly available ChIP-seq samples used for ChromHMM analysis.

3.7.2 Capture-HiC of CRC

Capture Hi-C performed on the human colon cancer HT29 cell line and published by Orlando *et al.*[169] was used to annotate enhancer regions to their target genes; available at the European Genome-phenome Archive (EGA) under the accession code EGAS00001001946. In this work they used capture Hi-C (CHi-C) to catalogue the regulatory landscape of CRC through 19023 promoter fragments.

3.7.3 ATAC-seq data from TCGA consortium

To identify potential pan-cancer regulatory regions, pan-cancer ATAC-seq peaksets from the TCGA consortium[116] were used (https://gdc.cancer.gov/about-data/publications/ATACseq-AWG/). The data were not re-analysed; the consensus peakset, coverage tracks and count table were directly downloaded from the portal and used.

3.7.4 H3K27ac data from various types of cancer

ChIP-seq data of H3K27ac occupancy derived from different primary tumor types (**Table 3**) were re-analysed as described before and used to further validate the YAP/TAZ regulated pan-cancer core of enhancers. See method session “Analysis of publicly available H3K27ac ChIP-seq datasets” for further details.

Organism_part	Disease	Sampling_site	Individual	Material_type	Developmental_stage	Histone_mark	Associated_Input_id	Reference study
tumor_gastric	gastric tumor	Tumor tissue	CHG018	tissue	not applicable	H3K27Ac	GSM1252273	GSE51776
tumor_gastric	gastric tumor	Tumor tissue	CHG026	tissue	not applicable	H3K27Ac	GSM1252281	GSE51776
tumor_gastric	gastric tumor	Tumor tissue	CHG034	tissue	not applicable	H3K27Ac	GSM1252289	GSE51776
tumor_gastric	gastric tumor	Tumor tissue	CHG093	tissue	not applicable	H3K27Ac	GSM1252313	GSE51776
normal_gastric	gastric tumor	Healty adjacent tissue	CHG022	tissue	not applicable	H3K27Ac	GSM1252277	GSE51776
normal_gastric	gastric tumor	Healty adjacent tissue	CHG030	tissue	not applicable	H3K27Ac	GSM1252285	GSE51776
normal_gastric	gastric tumor	Healty adjacent tissue	CHG038	tissue	not applicable	H3K27Ac	GSM1252293	GSE51776
normal_gastric	gastric tumor	Healty adjacent tissue	CHG089	tissue	not applicable	H3K27Ac	GSM1252309	GSE51776
normal_gastric	gastric tumor	Healty adjacent tissue	CHG097	tissue	not applicable	H3K27Ac	GSM1252317	GSE51776
breast	breast tumor	Tumor tissue	B1_H3K27ac_FA	tissue	not applicable	H3K27Ac	GSM3149117	GSE114737
breast	breast tumor	Tumor tissue	B2_H3K27ac_FA	tissue	not applicable	H3K27Ac	GSM3149117	GSE114737
breast	breast tumor	Tumor tissue	B3_H3K27ac_FA	tissue	not applicable	H3K27Ac	GSM3149117	GSE114737
breast	breast tumor	Tumor tissue	B4_H3K27ac_FA	tissue	not applicable	H3K27Ac	GSM3149117	GSE114737
endometrium	endomitrium tumor	Tumor tissue	E1_H3K27ac_FA	tissue	not applicable	H3K27Ac	GSM3149119	GSE114737
endometrium	endomitrium tumor	Tumor tissue	E2_H3K27ac_FA	tissue	not applicable	H3K27Ac	GSM3149119	GSE114737
endometrium	endomitrium tumor	Tumor tissue	E3_H3K27ac_FA	tissue	not applicable	H3K27Ac	GSM3149119	GSE114737
bone	osteosarcoma	Tumor tissue	Patient_27252-1_H3K27ac	tissue	not applicable	H3K27Ac	GSM2870621	GSE74230
bone	osteosarcoma	Tumor tissue	Patient_33010-1_H3K27ac	tissue	not applicable	H3K27Ac	GSM2870627	GSE74230
bone	osteosarcoma	Tumor tissue	Patient_P10_H3K27ac	tissue	not applicable	H3K27Ac	GSM2870639	GSE74230
bone	osteosarcoma	Tumor tissue	Patient_P2_H3K27ac	tissue	not applicable	H3K27Ac	GSM2870645	GSE74230
uterus	not applicable	Healty tissue	uterus_female_adult_53yrs	tissue	not applicable	H3K27Ac	GSM2701786	Encode
liver	not applicable	Healty tissue	male_adult_32years_liver_tissue	tissue	not applicable	H3K27Ac	GSM1059458	Roadmap
adrenal gland	not applicable	Healty tissue	male_adult_34years_adrenal_gland_tissue	tissue	not applicable	H3K27Ac	GSM896167	Roadmap
adrenal gland	not applicable	Healty tissue	female_adult_30years_adrenal_gland_tissue	tissue	not applicable	H3K27Ac	GSM1013168	Roadmap
pancreas	not applicable	Healty tissue	female_adult_30years_pancreas_tissue	tissue	not applicable	H3K27Ac	GSM1013172	Roadmap
pancreas	not applicable	Healty tissue	male_adult_34years_pancreas_tissue	tissue	not applicable	H3K27Ac	GSM906419	Roadmap
pancreas	not applicable	Healty tissue	female_adult_53years_body_of_pancreas_tissue	tissue	not applicable	H3K27Ac	GSM2701039	Encode
thyroid gland	not applicable	Healty tissue	female_adult_53years_thyroid_gland_tissue	tissue	not applicable	H3K27Ac	GSM2700106	Encode
thyroid gland	not applicable	Healty tissue	male_adult_37years_thyroid_gland_tissue	tissue	not applicable	H3K27Ac	GSM2534428	Encode
thyroid gland	not applicable	Healty tissue	male_adult_54years_thyroid_gland_tissue	tissue	not applicable	H3K27Ac	GSM2527552	Encode
uterus	not applicable	Healty tissue	female_adult_51year_uterus_tissue	tissue	not applicable	H3K27Ac	GSM4051146	Encode
body of pancreas	not applicable	Healty tissue	male_adult_54years_body_of_pancreas_tissue	tissue	not applicable	H3K27Ac	GSM2527460	Encode
body of pancreas	not applicable	Healty tissue	female_adult_51year_body_of_pancreas_tissue	tissue	not applicable	H3K27Ac	GSM4250622	Encode
body of pancreas	not applicable	Healty tissue	male_adult_37years_body_of_pancreas_tissue	tissue	not applicable	H3K27Ac	GSM2534589	Encode
thyroid gland	not applicable	Healty tissue	female_adult_51year_thyroid_gland_tissue	tissue	not applicable	H3K27Ac	GSM4247353	Encode
adrenal gland	not applicable	Healty tissue	male_adult_37years_adrenal_gland_tissue	tissue	not applicable	H3K27Ac	GSM2534395	Encode
adrenal gland	not applicable	Healty tissue	male_adult_54years_adrenal_gland_tissue	tissue	not applicable	H3K27Ac	GSM2534495	Encode
liver	not applicable	Healty tissue	female_adult_53years_liver_tissue	tissue	not applicable	H3K27Ac	GSM2527674	Encode

Table 3 | H3K27ac ChIP-seq samples of primary tumors and normal tissues used in the analysis of pan-cancer regions.

3.7.5 scRNA-seq from primary CRC and LUAD tissues

Two single cell RNA sequencing studies from primary CRC170 and Lung Adenocarcinoma (LUAD)171 tissues were used in this work to confirm our finding on the epigenetic regulation at single cell resolution. Raw filtered count matrix and cell annotation were downloaded directly from Gene Expression Omnibus (GEO) using the GSE132465 and GSE131907 project, respectively. The raw data were imported in Scanpy172 and analysed as extensively described later on in the methods.

3.8 HePIC: web genome browser

Figure 20 | Hepic Logo

HePIC (Human EPigenetic CRC) is a web app developed for the interrogation of the omic data produced in this work (it will be available at http://hepic.homic.eu following publication of the current work) (**Figure 20**). This web app consists of a web genome browser (https://github.com/igvteam/igv.js/) that allows interactive visualization and integration of the epigenomic (ChIP-seq on histone marks and ChromHMM tracks) and transcriptomic (RNA-seq) data analysed in this study. This application was build integrating a web server (NGINX) with a container technology (Docker).

3.9 Downstream analyses

3.9.1 Density and heatmap plot for each histone modification

Filtered and sorted BAM files were used to generate normalized coverage tracks using the *bamCoverage* function from deepTools[162] suite. The average signal profile and the heatmap plot along the genebody were calculated using *computeMatrix* scale-regions with default parameters and GENCODE Release 25 basic gene annotation. The command lines used for this analysis are reported in the **Appendix**.

3.9.2 Correlation analysis of histone marks

To obtain the correlation heatmap of all the histone modifications among the ten PDOs, a consensus peakset was generated using DiffBind v2.6.6[173] and merging together only peaks detected in at least two tracks. Then, a count matrix of 180250 peaks x 48 samples was created by counting the number of reads per peak for each sample using the *dba.count* function with default parameters. The correlation heatmap and the PCA were produced using *dba.plotHeatmap* (distMethod="pearson") and *dba.plotPCA* respectively, with default parameters.

3.9.3 *De novo* chromatin state characterization

De novo chromatin stated characterization of all PDOs was performed using a multivariate Hidden Markov Model approach (ChromHMM v1.12[174]) (**Figure 18, Figure 21**) considering five histone modifications (H3K4me3, H3K27ac, H3K4me1, H3K36me3 and H3K27me3) across 10

PDOs and publicly available data (**Errore. L'origine riferimento non è stata trovata.**), using default parameters. The datasets were down-sampled to a maximum depth of 45 million reads (the median read depth over all samples considered in this analysis). The read counts for all the considered samples were computed in non-overlapping 200-bp bins across the entire genome. The binarization was performed comparing ChIP-seq read count to corresponding input DNA as control to reduce the technical noise. Several models were trained in parallel considering 8, 10 and 12 number of states. The 8-state model was chosen for downstream analysis since it captured the key interaction between histone marks and because it was the model with minimal redundancy. The command lines used for this analysis are reported in the **Appendix**.

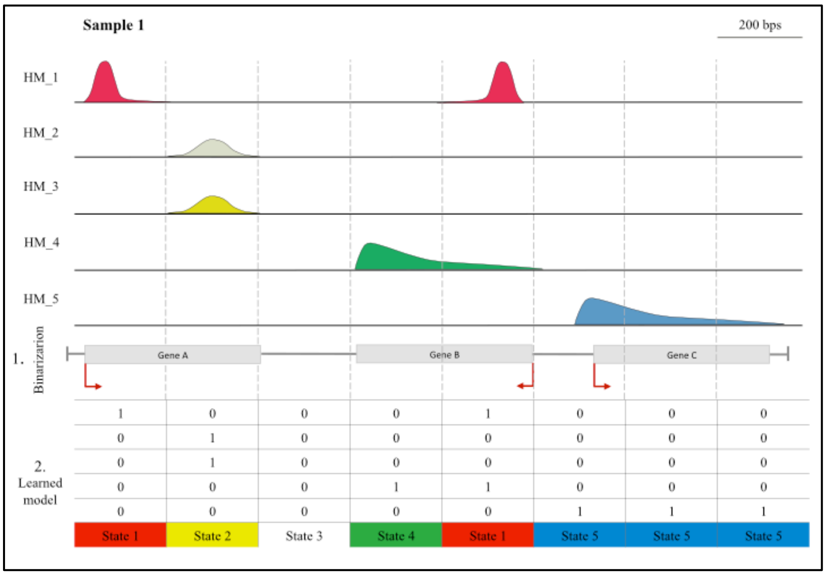


Figure 21 | ChromHMM pools together multiple epigenomes and it is able to summarize the combinatorial pattern of different HMs across different cell populations. ChromHMM uses Hidden Markov Model approach which is a machine learning technique able to reconstruct a model based on the probability distribution of a series of observations.

3.9.4 Overlap of ChromHMM states and COAD ATAC-seq from TCGA

To validate our model, the probability of detecting previously reported open chromatin regions for colon cancer within each chromatin state was estimated. To this end, ATAC-seq data for colon adenocarcinoma (COAD) was downloaded from the TCGA site (https://gdc.cancer.gov/about-data/publications/ATACseq-AWG/). The number of ATAC-seq peaks inside each ChromHMM state was defined by overlapping the regions of each ChromHMM state with the ATAC-seq peak summits. Since each ATAC-seq peak was reduced to the summit of the peak, the length of each ATAC-seq peak corresponded to 1 bp. Then, a conditional probability was calculated to estimate the probability of identifying open chromatin regions in each chromatin state across the ten PDOs. The probability p(A | B) is the probability that the event A will occur given the knowledge that an event B has already occurred. The conditional probability of A given B is defined as the quotient of the probability of the joint event A and B (both events A and B occur together) and the probability of B.

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

For *i*= 1..n where n is the numbers of chromatin states,
p(A), p(B) and p(A ∩B) were defined as follows:

p(A) = total length of ATAC-seq peaks / total length of the genome

p(B) = total length of ChromHMM$_i$ state / total length of the genome

p(A ∩ B) = total length of ATAC-seq peaks overlapping ChromHMM$_i$ state/total length of the genome

With p(A|B) defined as the probability of finding ATAC-seq peaks in each ChromHMM state:

p(A|B) = total length of ATAC-seq peaks overlapping ChromHMM$_i$ state/total length of ChromHMM*i* state.

3.9.5 Identification of highly active enhancers

To identify the highly active enhancer elements, all the "Active Enhancer" and "Flanking Active Enhancer" regions from the ten PDOs and the five normal colon tissue ChromHMM states were selected. These two states are defined by the co-presence of high levels of H3K27ac and H3K4me1 signal. The pool of active enhancer regions was filtered excluding all the regions with a length less than 200 bp and all the regions that fall within a window of 5 kb around (upstream and downstream) the TSS (considering known genes annotated in GENCODE Release 25 basic gene annotation). Then, a consensus peakset was built using DiffBind, as previously described (Correlation analysis of histone marks). The number of H3K27ac reads in the consensus peakset was counted generating a count matrix of 33131 regions X 15 samples. Differential analysis was performed using DESeq2 package[175] (version 1.22.2) considering as differentially activated all the regions with a padj ≤ 0.01 and a $|\log2FC| \geq 2$.

3.9.6 Enhancer conservation across patients

A master list of enhancer regions across all the ten PDOs samples was produced merging together the "Active Enhancer" and "Flanking Active Enhancer" (**Identification of highly active enhancer**s) states using BEDTools[176]. Enhancer regions found only in one patient were considered in this analysis. To assess if an enhancer was conserved among different patients the enhancerome of each patient was intersected with the master list of enhancers (n=33,131) as described above. A matrix of presence/absence for each region across PDOs was created by a custom script in Python. The matrix reported in one dimension the number of PDOs lines and in the other dimension the number of enhancers. The correspondence between an enhancer for a patient and the enhancer in the master list is reported as "1" and the absence as "0". To assess conservation, enhancers were stratified according to their frequency across PDOs and further filtered for enhancers differentially activated (gained) in PDOs compared to normal tissues.

3.9.7 Motif binding discovery

Motif binding discovery was performed within the accessible regions of the conserved gained enhancers. First, the ATAC-seq peakset for colon adenocarcinoma (COAD) was downloaded from the TCGA site (https://gdc.cancer.gov/about-data/publications/ATACseq-AWG/). To identify putative open chromatin regions inside the most conserved enhancers, gained enhancer regions, conserved in at least 80% of the patients (n=486), were overlapped with the COAD ATAC-seq peaks. The HOMER[177] *findMotifsGenome* function was used to evaluate the enrichment of known motifs in the exact size of the accessible regions (setting region size parameter to "given"). Transcription factor binding motifs encompassing the summit of TAZ peaks were identified with HOMER and MEME-chip[178] on 500 bp windows centred around TAZ peak summits.

3.9.8 Annotation of differentially activated enhancers

Differentially activated enhancers were annotated using chromosome conformation capture data on human CRC HT29 cell line from Orlando *et al.*[169]. To annotate the remaining differentially activated enhancers we extracted and merged all the "Active TSS" and "Flanking active TSS" regions from the ChromHMM states of the ten PDOs and the five normal tissues. Then, we created a txdb object using makeTxDbFromGFF (GenomicFeatures v1.30.321) including all the protein coding genes (GENCODE Release 25 basic gene annotation) with an active promoter (*n*=13802). Finally, we used the annotatePeakInBatch (output="both", PeakLocForDistance="middle") function of ChIPpeakAnno v3.12.7[179] to annotate the active enhancers to their nearest protein coding gene with an active promoter.

3.9.9 Functional enrichment analysis

We used over-representation analysis based on Fisher's exact test to assess the functional enrichment of biochemical and signalling pathways in the list of 495 tumor-upregulated genes annotated to gained enhancers. Functional enrichment analysis was conducted in R using the *fisher.test* function of the stats package on the 321 gene sets of the KEGG collection (downloaded from ConsensusPathDB; http://cpdb.molgen.mpg.de/) considering a genomic background of 21,528 unique gene symbols (given by the union of the 19,950 protein coding genes of the human hg38 reference GENCODE Release 25 and of the genes of all KEGG gene sets). P-values have been adjusted (*i.e.*, False Discovery Rate) using the *p.adjust* function of R stats package and the threshold for statistical significance set at FDR < 5%. The visualization of the functional enrichment analysis results was obtained in Cytoscape[180] using its EnrichmentMap and AutoAnnotate applications (with default parameters).

3.9.10 ChIPmentation data processing and quality control

The processing of ChIPmentation data was performed as previously described (**ChIP-seq processing and quality control**), with the difference that the adapters were removed before the alignment of the reads using BBDuk (command line parameters: ktrim=r k=23 mink=11 hdist=1). Peaks were called using MACS2 v2.1.0, with the associated ChIPmentation on the input as control (-p 0.001). Density plots and heatmaps were produced as described above (**Density and heatmap plot for each histone modification**), considering as regions all the promoter in GENCODE Release 25 basic gene annotation and the active enhancer ranges identified (n = 33,131). The command lines used for this analysis are reported in the **Appendix**.

3.9.11 Analysis of TAZ ChIPmentation data

To assess the preferential binding of TAZ along the genome, TAZ peaks were overlapped with the previously-defined ChromHMM states using BEDTools. Permutation analysis was performed to assess the enrichment of TAZ occupancy in gained CRC enhancers compared to a random distribution of enhancers. The BEDTools shuffle function was used to generate 1000 shuffle tracks of the gained enhancers, preserving the size of each gained enhancer in the input BED file. TAZ peaks were subsequently overlapped with i) all gained enhancers identified in PDOs, ii) the gained enhancers conserved in at least 5 patients, iii) those conserved in at least 8 patients, and finally iv) the shuffled tracks (control). In counting enhancer regions, a single count was considered for regions that overlapped multiple TAZ peaks. Finally, a Fisher exact test was performed considering as significantly enriched the comparisons with a P-value < 0.001.

3.9.12 Analysis of TCGA pan-cancer ATAC-seq data

To identify potential pan-cancer regulatory regions, pan-cancer ATAC-seq peaksets from the TCGA consortium[116] were used (https://gdc.cancer.gov/about-data/publications/ATACseq-AWG/). The pan-cancer peakset was overlapped with the YAP/TAZ-bound gained enhancers conserved in at least 8 patients (n=195). When multiple ATAC-seq peaks were assigned to a specific enhancer only the ATAC-seq peak with the highest normalized enrichment score was considered. Then, the normalized ATAC-seq insertion counts of the pan-cancer peaksets was downloaded from TCGA site (https://gdc.cancer.gov/about-data/publications/ATACseq-AWG/) and was used to produce a heatmap (*pheatmap*; clustering_distance_cols=euclidean, clustering_method=complete) of all the TCGA patients and the 195 enhancer regions of interest. To identify pan-cancer accessible regions, we performed hierarchical clustering with cluster_rows=TRUE directly using *pheatmap* package.

3.9.13 Analysis of H3K27ac ChIP-seq datasets from different cancer types

ChIP-seq data for H3K27ac were obtained from the Gene Expression Omnibus (**Table 3**). Raw sequencing reads were processed as described above (**ChIP-seq processing and quality control**). The detailed command lines used for this analysis are also reported in the **Appendix**. For each sample, the number of H3K27ac reads in the consensus peakset of ~33K active enhancers (**Identification of highly active enhancer**) was counted in DiffBind. Read counts across samples were normalized and corrected for potential batch effects using ComBat[181]. For each of the primary tumor and normal tissue samples, the mean H3K27ac normalized counts across all 46 pan-cancer enhancer regions was calculated and wilcoxon rank sum test was performed to determine the difference in H3K27ac intensities between primary tumors and normal tissues.

3.10 Single cell RNA-seq analysis

3.10.1 Single-cell RNA-seq data processing and quality control

The fastq files of primary CRC tissue were processed by the Cell Ranger software pipeline (version 3.0.1) provided by 10X Genomics. Alignment with STAR (human genome GRCh38 version 25), multiplexing, UMIs and cell filtering were performed using default parameters creating one count matrix of 19,702 genes across 4,299 cells. The matrix was then processed using the python package Scanpy (v1.4.2)[172]. First, genes detected in less than 0.1% of the total cells and cells with fewer than 200 expressed genes were removed. Low quality cells and outliers based on percentage of mitochondrial (MT) and ribosomal (RB) genes, total number of genes, and gene counts were detected according to the median absolute deviation (MAD)[182]. Cells were removed if the value for any of the above features was greater than the number of selected MAD above the median, with the number of MAD set to four for MT percentage, two for RB percentage, two for the gene number and three for count number. Furthermore, the immune cells were filtered out based on the detection of canonical marker genes. More in details, three groups of cells expressing high levels of natural killer (NK) cell marker genes (NKG7 and KLRB1), CD8+ T lymphocytes marker genes (CD8A and CD8B) and B lymphocytes marker genes (IGHG3 and IGLC3) were removed. The remaining 3,044 cells together with 19,460 genes constitute the matrix used for the downstream analysis. The matrix was normalised considering a scaling factor of 10^4 and log-transformed using *scanpy.pp.normalize_per_cell* (data, counts_per_cell_after=1e4) and *scanpy.pp.log1p* (data). Highly variable genes (HVG) were selected based on specific thresholds for mean expression and dispersion using *scanpy.pp.highly_variable_genes* (min_mean=0.08, max_mean=4, min_disp=0.7) and excluding mitochondrial and ribosomal genes. The cell cycle phase of each cell was evaluated by scoring individual cells for their expression of G1-, S-, and G2M-phase genes[183].

3.10.2 Dimensionality reduction and clustering

PCA was performed on scaled and centred values considering 1219 HVG. Unwanted sources of variation (*i.e.* number of detected counts and genes per cell, the percentages of mitochondrial and ribosomal counts and the cell cycle phase) were evaluated and regressed out using a linear regression as implemented in scanpy (*scanpy.pp.regress_out*). Initially, a K-Nearest Neighbour graph was constructed based on Euclidean distance in PCA space, thus refining the weight of the edges between two cells using Jaccard similarity (*scanpy.pp.neighbors* with n_neighbors=15, n_pcs=13). Finally, the Leiden algorithm was used to perform unsupervised clustering of the cells, with a resolution of 0.6 (*scanpy.tl.leiden*). Leiden-defined clusters were labelled based on previously reported marker genes for colonic epithelial cells[184] Progenitors cells were partitioned into two clusters (early and late) on the basis of decreasing expression levels of stem marker genes (OLFM4, SOX4) and increasing levels of marker genes for differentiated epithelial cells (FABP1, CA2).

3.10.3 Data visualization and trajectory analysis

Uniform Manifold Approximation and Projection (UMAP) and Force-directed graph were used for visualization of the data. The number of PCs used to calculate the embedding were the same as those used for the clustering. The force-directed graph was obtained using *scanpy.tl.draw_graph* with ForceAtlas2 as layout. Partition-based graph abstraction (PAGA) connectivity is based on the previously estimated clusters using Leiden algorithm and is calculated with default parameters using *scanpy.tl.paga* function.

3.10.4 Identification of differentially expressed genes

Differentially expressed genes (DEG) for each cluster against all other clusters were identified using *scanpy.tl.rank_genes_groups* implemented by Scanpy with default parameters. Upregulated DEG lists were used as ranked gene lists to perform GSEA as previously described

(RNA-seq QC and data analyses). Signature gene lists reported in Wang *et al.*[184] were used to verify the presence of normal-like epithelial cell populations.

3.10.5 Copy number variation analysis

To identify malignant cells in CRC primary tissue, large-scale CNVs were inferred from RNAseq data using the inferCNV package (https://github.com/broadinstitute/inferCNV) with default parameters (k_obs=2). CNVs are inferred for each cell based on a moving averaged pattern of expression profiles across large chromosomal intervals in comparison to a reference cell population as previously described[185].

3.10.6 Scoring cells using signature gene sets

Gene signature scores were calculated given a cell by gene expression matrix (M) and a geneset (g). For each cell in M, the fraction of genes from g that are expressed (expression levels >0) is computed. Similarly, an expression score for each cell in M is evaluated by summing up the expression levels of genes from g and dividing by the total sum of gene expression levels for all genes in the same cell. The two scores are then multiplied together to yield a combined score for each cell in M and the reciprocal of the negative logarithm of the combined score is computed. Following the mathematical equation:

Given a cell C as a vector of gene expression values [g_i, ..., g_c]

And a geneset G={ g_i, ..., g_G}

A co-expression score is computed as:

$$c\_score = \frac{\sum_{g \in G}[Cg > 0]}{|G|}$$

and an expression score is defined as

$$e\_score = \frac{\sum_{g \in G} Cg}{\sum C}$$

The two scores are then combined to yield a combined score

$$Combined\_score = c\_score * e\_score$$

Combined scores were created for a stemness gene set and the gene set related to the 46 pan-cancer enhancers.

3.10.7 Dendrogram

For each cell the average expression of the 46 genes associated to the pan-CRC enhancers was calculated. A correlation (method='kendall') on the average expression vectors of these genes and hierarchically clustering of the previously mentioned Leiden clusters was performed using python function *scipy.cluster.hierarchy.linkage* (method='complete', metric='euclidean').

3.10.8 scRNA-seq analysis of diverse primary tumor tissues

ScRNA-seq of primary CRC[170] and LUAD[171] samples were analysed as described above. For reproducibility, we used the clustering annotation reported by their reference papers[170,171].

3.11 Data availability

The RNA-seq and ChIP-seq generated during this study will be available, after the publication of the work, at the European Nucleotide Archive with accession numbers E-MTAB-8448 and E-MTAB-8416 respectively.

4. Results

4.1 Establishment of a balanced PDOs library representing colon tumor heterogeneity

In order to characterize the epigenetic landscape of human CRC we collected and generated several three-dimensional organoids from primary tumor surgical resection of patients with different clinical and molecular phenotypes. We then performed a histopathological and molecular characterization of PDOs, demonstrating the robustness of this model as surrogate of the primary tumor of origin, and finally we established their genome-wide epigenetic landscape (**Figure 22 A**). To obtain a pool of heterogeneous PDOs, we performed RNA-sequencing (RNA-seq) analysis of our primary tumors and classified them into distinct CRC molecular subtypes. To this end, we used markers for microsatellite instability and three recently published gene expression-based classification systems[26,27], including the classifier of the CRC intrinsic signatures (CRIS)[31] (**Figure 22 B**). Upon primary tumor analysis, we selected ten CRC organoid lines creating a diverse and balanced library that resembles the molecular diversity of the primary tumors. To determine the possible exploitation of this library as a model to study the CRC epigenomic landscape we characterized the PDOs at molecular level.

First, we evaluated whether our PDOs preserve the morphological characteristics and the deregulated architecture of crypt/villus-like structures typical of human colon cancer using 3D-immunofluorescence whole mount analysis. CRC PDOs showed disorganized epithelium polarity (Epcam and F-actin staining respectively, **Figure 22 C**, first lane), random distribution of cell proliferation (Ki67, **Figure 22 C**, first lane), displaced localization of enterocytes (FABP1, **Figure 22 C**, second lane) and presence of cytokeratin 20 positive cells (KRT20, **Figure 22 C**, third and fourth lane), faithfully recapitulating the common dysplastic features of human CRC. The presence of chromogranin A, a marker of enteroendocrine cells, specifically in PDO24 (**Figure 22 C**, third-fourth lane) but not in other PDOs (**Figure 22 C**, second lane) further indicated the heterogeneity of our library. Likewise, goblet-specific mucin 2 is absent in most organoids PDOs but massively produced in the organoid derived from the mucinous adenocarcinoma of patient 13 (**Figure 22 C** third and fourth lane), consistent with the histopathological features of the clinical specimen (**Table 1**).

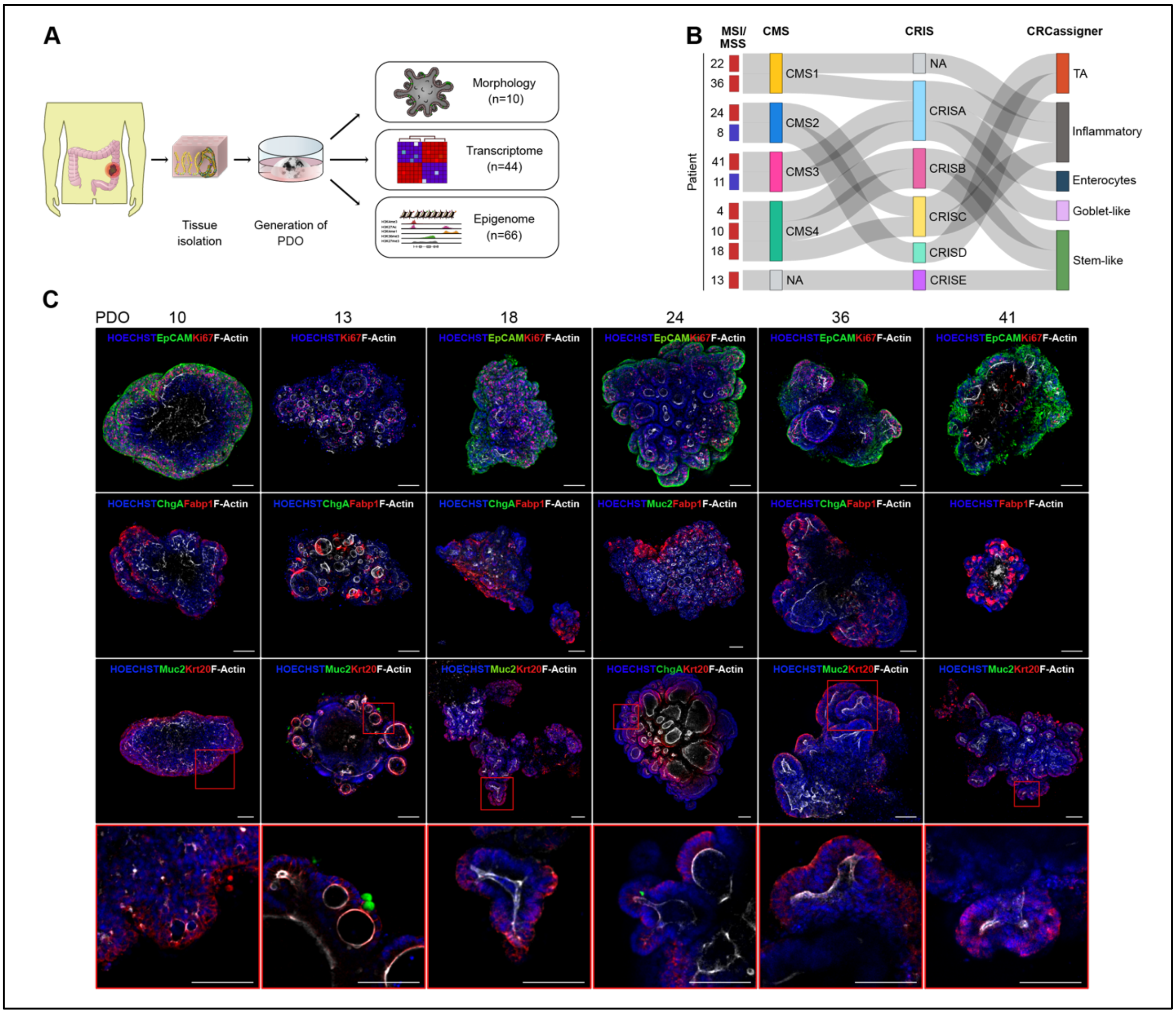


Figure 22 | Establishment of a heterogeneous PDOs library as a model of CRC that recapitulates the *in vivo* architecture of the primary tumors. (A) Graphical representation of the work. **(B)** Sankey plot showing the classification (MSI/MSS, CMS[27], CRIS[31] and CRCassigner of Sadanandam[26]) of the primary tumor tissues. **(C)** Representative confocal images of 3D-Immunofluorescence whole-mount analysis on CRC PDOs. Different markers of colon cell types are shown: polarity and structure (F-Actin, first lane), epithelium (EpCAM, first lane), proliferation (Ki67, first lane), absorptive cells (FABP1, second lane), enteroendocrine cells (ChgA, second and third lane), goblet cells (Muc2, third lane), and top epithelial crypt cells (KRT20, third lane). The fourth lane provides an enlargement of the boxed area in the third lane. Scale bars, 100 µm.

We then continued with the transcriptomic characterization of PDOs performing Principal Component Analysis (PCA) on all samples. We found that normal mucosa tissues derived from the same CRC patients are transcriptionally homogeneous and distinct from the nearby tumor tissues (**Figure 23 A**). On the contrary, the patient-driven heterogeneity of tumor tissues, evident on the second principal component (PC2), is preserved in PDOs which share the same PC2 spatial localizations as their parental tissues. Notably, multiple organoids from the same patient were grouped together indicating that PDOs remain transcriptionally stable during prolonged culture. Indeed, we showed the absence of significant gene expression alterations between early and late passage organoids (**Figure 23 B**), consistent with the lack of changes in culture morphology and proliferation rate. Focusing on the transcriptional comparison between PDOs and parental tumors, we found that 84% of expressed genes were concordant between PDOs and tumors (**Figure 23 C**, Venn diagram) with the expression levels across genes being well correlated (**Figure 23 C**, correlation plot). Using differential expression (DE) analysis between tumor and normal tissues, we identified the transcriptional changes associated to cancer development (tumor-related signature). To assess whether these are preserved in the organoids, we performed hierarchical clustering analysis revealing that the tumor-related signature groups PDOs together with primary tumors and separately from normal colon mucosa (**Figure 23 D**). To further validate whether the transcriptional profile of PDOs recapitulates that of tumor tissues, we performed gene set enrichment analysis (GSEA) using previously reported gene expression data sets from colon carcinoma patients[152], and showed that PDOs were significantly enriched for transcripts that were upregulated in colon carcinoma and downregulated in normal mucosa (**Figure 23 E**). We next evaluated whether the transcriptomic differences between primary tumors and PDOs are primarily due to the lack of a stromal component in PDOs. Focusing on the genes expressed in tumors but not in PDOs (**Figure 23 C**, Venn diagram, n=3,412), we found that they were enriched for gene signatures of stromal cells[31](**Figure 23 F**). We then investigated the expression distribution of the separate gene signatures for cancer-associated fibroblasts, endothelial cells and leucocytes[31] in PDOs and tumors. Consistent with previous reports[31,33], tumor tissues showed a prominent stromal component compared to PDOs. These findings indicate that PDOs retained the CRC gene signature but were deprived of stromal contamination, providing an advantage in deciphering the CRC molecular profiles inherent to cancer cells (**Figure 23 G**). Taken together, this data demonstrated that the histological subtypes and transcriptional signature of human CRC were conserved in our balanced library of PDOs, rendering it suitable for deciphering the common epigenetic blueprints of the colon cancer-cell intrinsic phenotype.

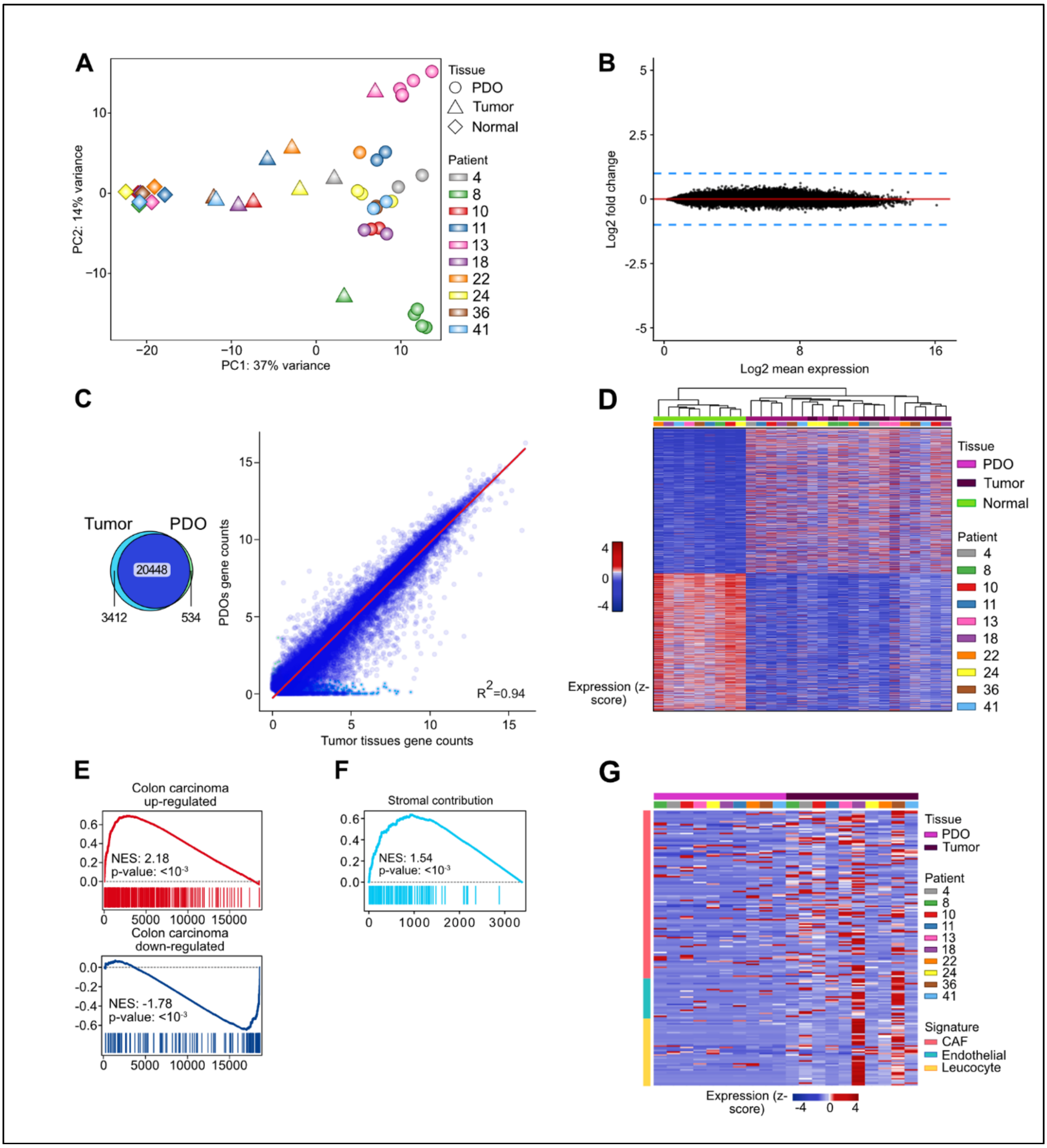


Figure 23 | PDOs library recapitulates the transcriptional profile of the primary tumors. (A) PCA on normalised gene counts from RNA-seq data distinguished normal colon mucosa, primary tumors and PDOs along the principal component 1. **(B)** MA plot of log2 mean gene expression over log2 fold-change showing the lack of differentially expressed genes between early and late passages of organoids. **(C)** Correlation between log2 mean normalized gene counts between primary tumors and PDOs. Venn diagram showing the concordance of genes expressed between tumors and PDOs.

(D) Hierarchical clustering analysis using differentially expressed genes (DEG) between tumor and normal colon tissues clustered PDOs together with parental tumors. Z-score normalized counts of DEG are represented as a heatmap. Tissue populations and patients are represented by color-coded bars above the heatmap. **(E)** GSEA showing enrichment of DEGs between PDOs and normal colon tissue in gene signatures that are up- (top) or down- (bottom) regulated in colon carcinoma clinical specimen compared to normal mucosa[152]. Normalized enrichment score (NES) and p-value are reported. **(F)** Genes expressed in primary tumors but not in PDOs (n=3,412), are enriched for gene signatures of stromal cells[31]. **(G)** Heatmap of Z-score normalized gene counts for stromal-related gene signatures[31] across primary tumors and PDOs. Tissue populations and patients are represented by color-coded bars above the heatmap. Stromal cell gene signatures are shown on the left side of the heatmap.

4.2 *De novo* chromatin state discovery reveals the epigenetic landscape of human CRC

Taking advantage of our molecularly diverse PDOs library, we sought to provide a systematic characterization of human CRC at the epigenomic level. The first step toward this aim was to perform a multi-factorial integrative analysis of genome-wide chromatin immunoprecipitation sequencing (ChIP-seq) for a core set of five histone modifications (H3K4me3, H3K27ac, H3K4me1, H3K36me3, and H3K27me3) on all PDOs. Consistent with a good enrichment signal, the genome-wide distribution of histone modifications (**Figure 24 A**) reflected their expected localization in relation to the gene body as well as TSS and end (TES) sites. Correlation analyses confirmed the clustering of the same histone marks for different patients and showed the clear separation between the branches relating to the repressive marker H3K27me3, the elongation marker H3K36me3, and the block of histone marks defining active regulatory regions (H3K4me3, H3K27ac, and H3K4me1) (**Figure 24 B**). To capture the epigenomic layer of CRC complexity in a systematic manner rather than based on a single epigenomic feature, we implemented machine learning approaches to perform *de novo* chromatin state characterization on the complete ChIP-seq data for our PDOs, including additional ChIP-seq data for five normal colon tissue, six primary colon tumors and two CRC cell lines (**Errore. L'origine riferimento non è stata trovata.**). Using ChromHMM[174], we explored the combinatorial patterns of the five histone marks in an 8-state model and predicted specific genomic features with high resolution and robustness across our samples. **Figure 24 C** reports the histone marks emission probability heatmap which represents the frequency in which different histone modifications are co-present in the same genomic region. The annotation term for each state was chosen according the Roadmap Epigenomics Consortium nomenclature[186]. In detail, two states were annotated as promoter states ("Flanking Active TSS - FlnkActTSS" and "Active TSS - ActTSS") since they showed enrichment for both H3K4me3 and H3K27ac. The two states with a strong enrichment of H3K4me1 and H3K27ac were defined as "Flanking Active Enhancers - FlnkActEnh" and "Active Enhancers - ActEnh". The state characterized by the presence of H3K4me1 alone was defined as "Weak Enhancers - WkEnh". The "Elongation – Elong" and "Repression - Repr" were characterized by the presence of H3K36me3 and H3K27me3, respectively. "Quiescence" state marks regions without any significant enrichment of histone marks. Interestingly, the promoter and enhancer states were 11% and 30%, respectively, of the total chromatin states identified while the repressed states represented only the 7% (**Figure 24 D**). To further confirm the robustness of our results we verified that the proportion of each chromatin state was comparable across PDOs (**Figure 24 E**). We then compared our ChromHMM data with chromatin accessibility levels for colon adenocarcinoma using ATAC-seq (Assay for Transposase Accessible Chromatin with high-throughput sequencing) datasets obtained from The Cancer Genome Atlas (TCGA). The chromatin states identified in PDOs remarkably concur with chromatin accessibility, with active states displaying the highest and more inactive regions the lowest ATAC-seq signals, respectively (**Figure 24 F**). This provides further support that PDOs preserve the regulatory networks of primary tumors and thus represent a faithful resource to investigate the epigenetic landscape of CRC. Importantly, the ChromHMM-defined chromatin states of CRCs constitute a precise atlas of genome-wide regulatory elements that enables the functional interpretation of ATAC-seq-defined open chromatin regions.

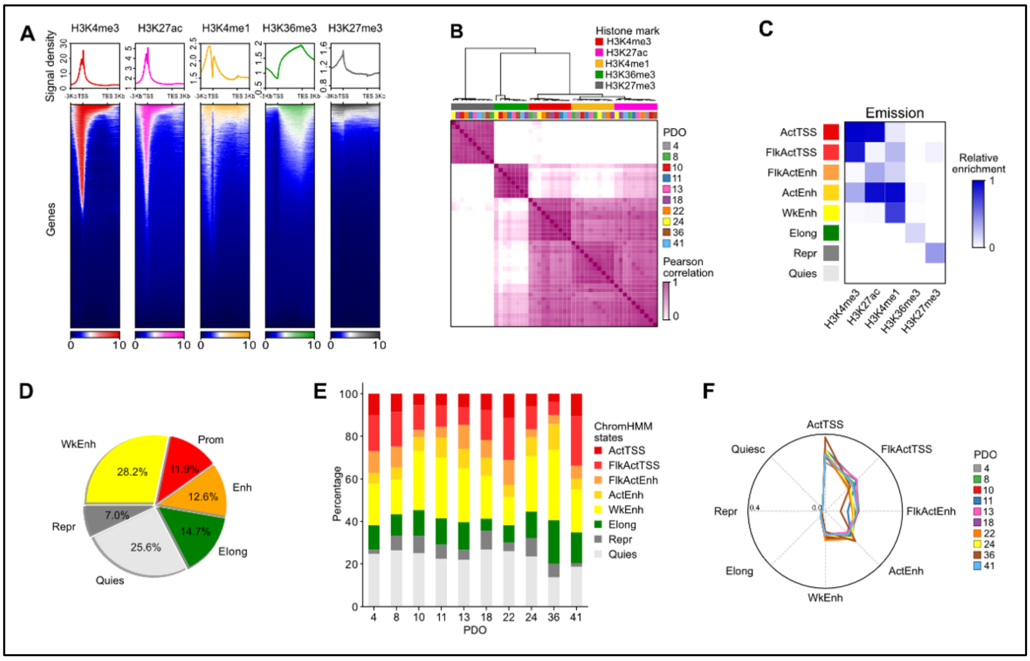


Figure 24 | Epigenomic landscape of CRC using *de novo* chromatin state discovery. (A) Representative density plots (top) of average intensity and corresponding heatmaps (bottom) displaying the relative distribution of H3K4me3 (red), H3K27ac (pink), H3K4me1 (yellow), H3K36me3 (green), and H3K27me3 (grey) signals at regions surrounding +/- 3kb of the gene body for all the genes present in GENCODEv25. **(B)** Pearson correlation heatmap of ChIP-seq data for the complete set of five histone modifications across all PDOs. **(C)** Combinatorial pattern of histone marks in an 8-state model using ChromHMM. The heatmaps show the frequency of the histone modifications found in each state (Emission). **(D)** Average proportion of each chromatin state over all PDOs. The chromatin segments for active/flanking TSS and active/flanking enhancer states are merged into the promoter and enhancer functional elements, respectively. **(E)** Distribution of the eight ChromHMM states for each PDOs. FlkActTSS: Flanking Active TSS, ActTSS: Active TSS, FlkActEnh: Flanking Active Enhancers, ActEnh: Active Enhancers, WkEnh: Weak Enhancers, Elong: Elongation, Repr: Repression, Quies: Quiescence. (F) Spider plot showing the probability of each ChromHMM-defined chromatin state overlapping ATAC-seq regions for TCGA colon adenocarcinoma samples. Probabilities for each PDOs are represented by different colors.

The histone modification pattern of FABP1, a marker of enterocytic differentiation, is an example of an open and active chromatin profile that favours gene expression (**Figure 25 A**). As shown in the immunofluorescence analysis (**Figure 22 C**, second lane), the expression of the FABP1 protein is abundant in all PDOs consistent with the increased RNA-seq levels across all samples, regardless of their molecular subtype. This conservation is also reflected in the epigenetic level as shown by the presence of active histone marks (H3K4me3/H3K4me1/H3K27ac) at the promoter and flanking region, and of H3K36me3 in the gene body. Regulatory variability across PDOs was observed in the gene encoding for laminin subunit α-5 (LAMA5) (**Figure 25 B**), a marker of cell adhesion and migration reported to be involved in metastasis[187]. Active transcription in PDO11 is indicated by a ChromHMM profile that associates with active states around the TSS and with an elongation state along the gene body. On the contrary, LAMA5 is actively silenced in PDO18 evident by the loss of H3K36me3 and the accumulation of the H3K27me3 repressive mark at the promoter and throughout the gene body. Interestingly, the effect of opposite epigenetic profiles at LAMA5, and also at the MUC2 gene, was confirmed at the protein level by 3D immunofluorescence analysis in PDO11 and PDO18 (**Figure 25 C**). PDO11 showed an active chromatin profile at both loci supporting the transcription of the MUC2 and LAMA5 genes; at the protein level we can clearly see that both proteins are present. The abundant expression of the goblet cell-specific marker MUC2 in PDO11 also suggests a mucinous phenotype that is consistent with the MSI status of this tumor[188]. On the contrary, the LAMA5 and MUC2 proteins are not detected in PDO18 consistent with the repressed chromatin profile of their genomic loci. To encourage and facilitate the interrogation of the identified regulatory elements, we created a web browser (available at http://hepic.homic.eu) that enables the visualization of our comprehensive epigenetic resource of human CRC.

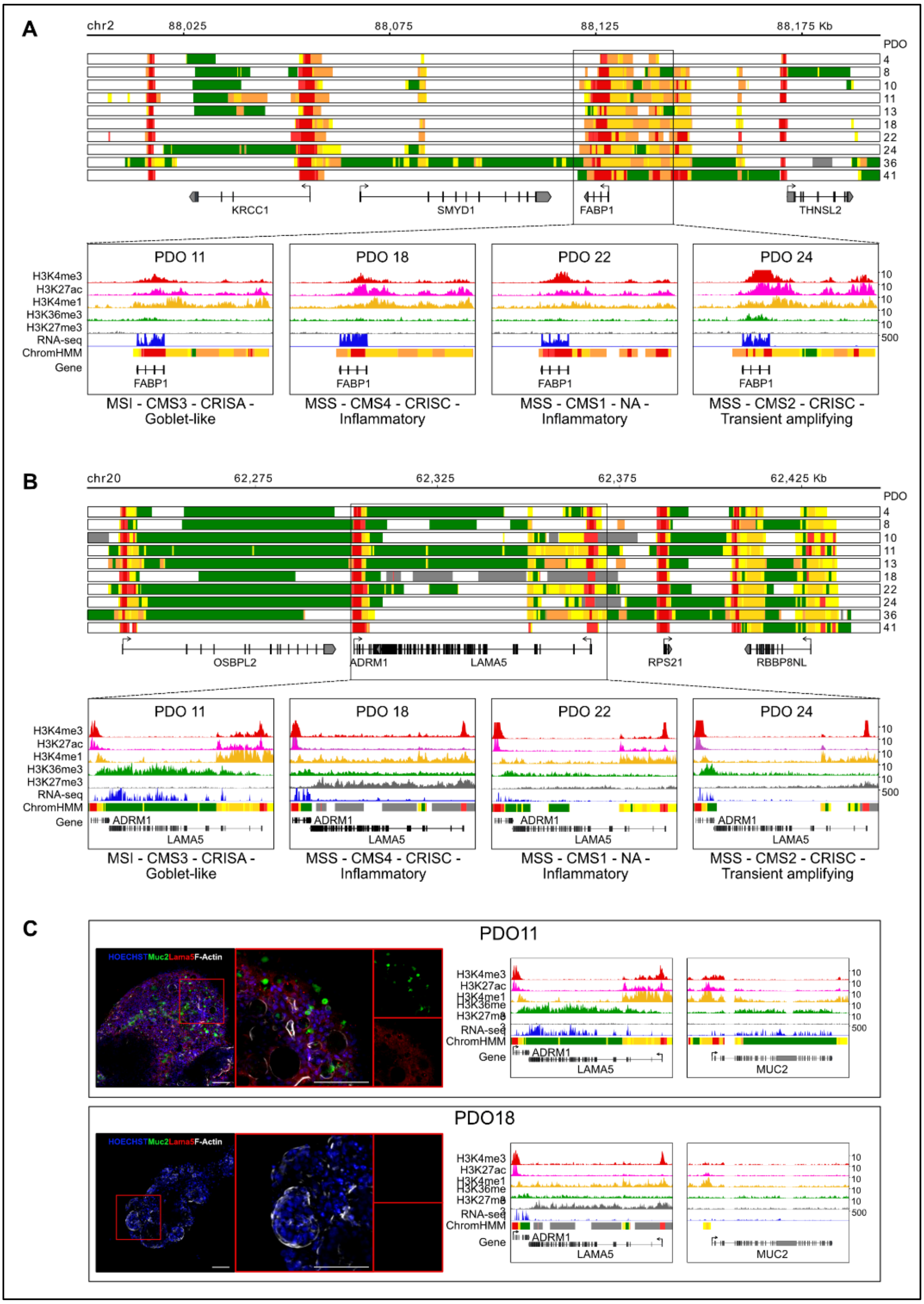


Figure 25 | Genomic overview of two representative genes with open chromatin states and repressed states. Representative tracks of ChromHMM states for the FABP1 **(A)** and LAMA5 **(B)** genomic loci in all PDOs. The expanded regions show H3K4me3, H3K27ac, H3K4me1, H3K36me3 and H3K27me3 ChIP-seq profiles, along with RNA-seq signal and ChromHMM states for PDOs of different molecular subtypes. **(C)** The epigenetic and transcriptional profiles of the LAMA5 and MUC2 genes in PDO11 and PDO18 (right panel) are concordant with their protein expression levels as shown by confocal images of 3D immunofluorescence whole mount analysis on CRC PDOs 11 and 18 stained for MUC2 (green), LAMA5 (red), and F-Actin (white) (left panel).

4.3 Definition of human CRC enhancerome

Following the reconstruction of the epigenomic landscape of colon cancer, we sought to gain further insights into the human colon cancer enhancerome. Starting from our *de novo* chromatin discovery, we used the “Active Enhancers” and “Flanking Active Enhancers” ChromHMM states (**Figure 24 C**), characterized by the co-presence of H3K27Ac and H3K4me1, to select active distal enhancer regions for PDOs and normal colon tissues. We identified a total number of 33,131 enhancers identified in at least two PDOs and/or normal tissues and located 5 kb away from TSS. The collection of all the active enhancer regions of human CRC covered the 3% of the human genome. Unsupervised clustering based on the H3K27ac signals of the ~33K enhancers showed a clear distinction between the enhancerome of PDOs and normal colon tissues (**Figure 26 A**). To discriminate between tumoral versus normal colon active enhancers, we performed a differential activation analysis and we identified 7,828 enhancers that were differentially enriched (gained or lost) in H3K27ac of PDOs compared to normal colon mucosa (adjusted P-value < 0.01 and $|\log2FC| > 2$) (**Figure 26 B**). Of those, 2,419 enhancers were specifically activated in PDOs whereas 5,409 active regulatory regions were upregulated in normal colon mucosa. The different number of detected enhancers was likely related to the low heterogeneity of normal colon tissues, consistent with their transcriptional profile (**Figure 23 A**, group of N tissues). To seek for common epigenetic blueprints across our human CRC library, we looked at the distribution of tumor enriched enhancers in the CRC PDOs. Notably, 20% of the identified gained enhancers was conserved in 8 to 10 CRC PDOs (n=486), and half of the total activated tumor-specific enhancers were shared in 5 out of 10 patients (**Figure 26 C**), independently on their original molecular and histological features (**Table 1**). The remaining enhancers, discovered in 1 to 4 PDOs, were likely related to patients’ heterogeneity and specific molecular features. The relatively even distribution of the non-shared enhancers across PDOs (**Figure 26 D**) along with the high number of conserved enhancers, indicates that there is no bias in the discovery of enhancers across PDOs.

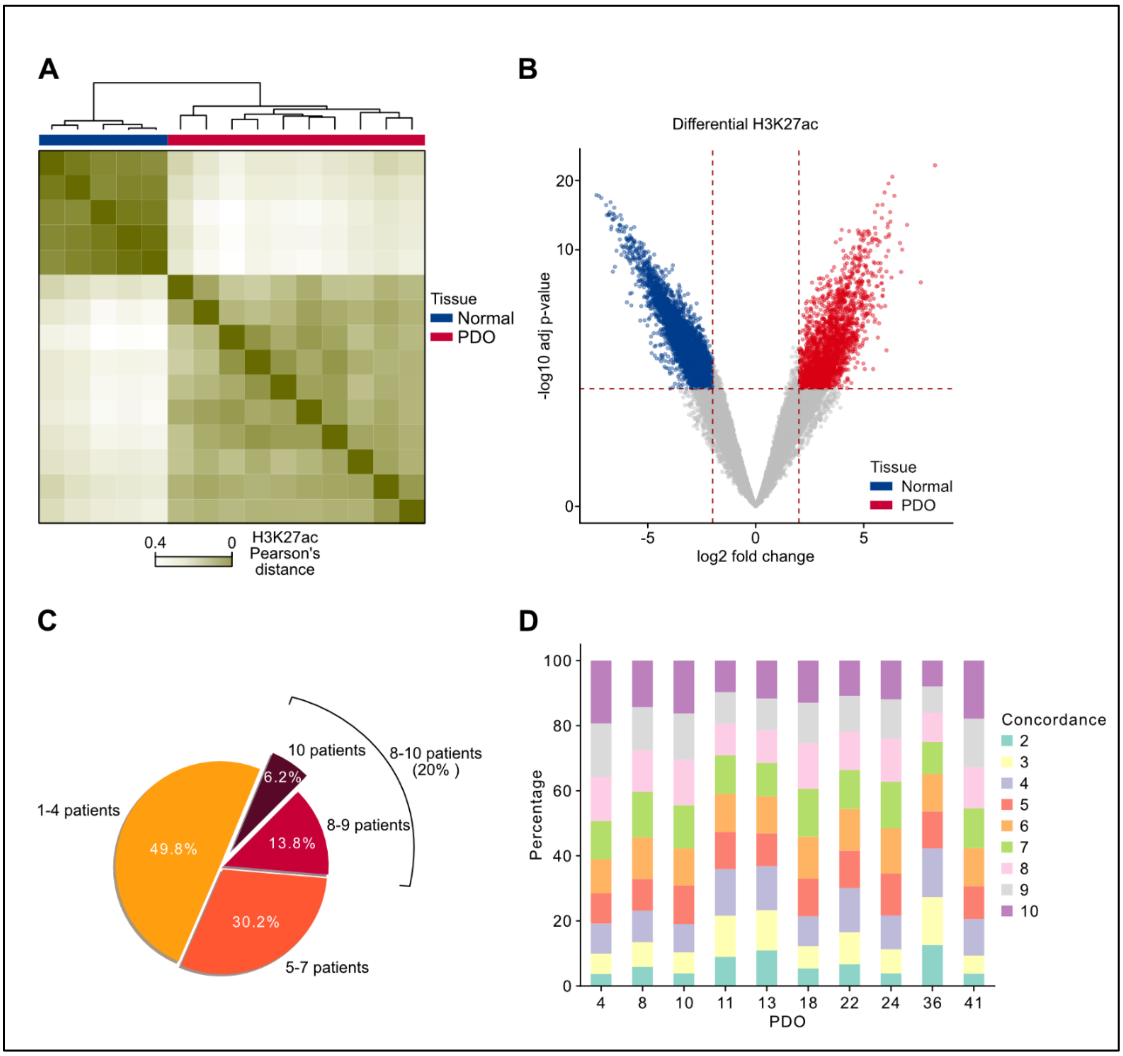


Figure 26 | Human CRC enhancerome definition. (A) Unsupervised clustering and Pearson correlation heatmap of H3K27ac ChIP-seq data for the 33,131 ChromHMM-defined enhancers clearly divides PDOs and normal colon tissues. **(B)** Volcano plot of differentially enriched enhancer regions between PDOs and normal colon mucosa. Dotted lines indicate thresholds for FDR < 0.01 and |log2 fold-change| > 2 **(C)** Pie chart reporting the percentage of differentially gained enhancers in PDOs that were shared across different patients. **(D)** Stacked bar plot reporting the distribution of gained enhancers across PDOs.

We report here the genomic overview of PHLDA1 (**Figure 27 A**, red tracks), a gene upregulated in colon cancers[189,190] and involved in tumor cell proliferation and migration[191]. In line with previous reports[126], PHLDA1 displays a conserved epigenetic signature of regulatory regions located downstream the gene, with high levels of H3K27 acetylation shared among all CRC PDOs compared to the normal colon tissue reference track (**Figure 27 A**, blue tracks). Conversely, a common feature of tumors can be the loss of specific colon enhancer activity related to differentiation programs[62] as underlined by the lack of H3K27 acetylation across and upstream the MUC4 gene region in our PDOs (**Figure 27 B)**. Overall, by exploiting the *de novo* chromatin states reconstruction in PDOs we were able to characterize the CRC enhancerome, identify tumor-specific active enhancers and reveal a novel layer of conserved regulation (consisting of approximately 500 active enhancers) that is independent of tumor diversity.

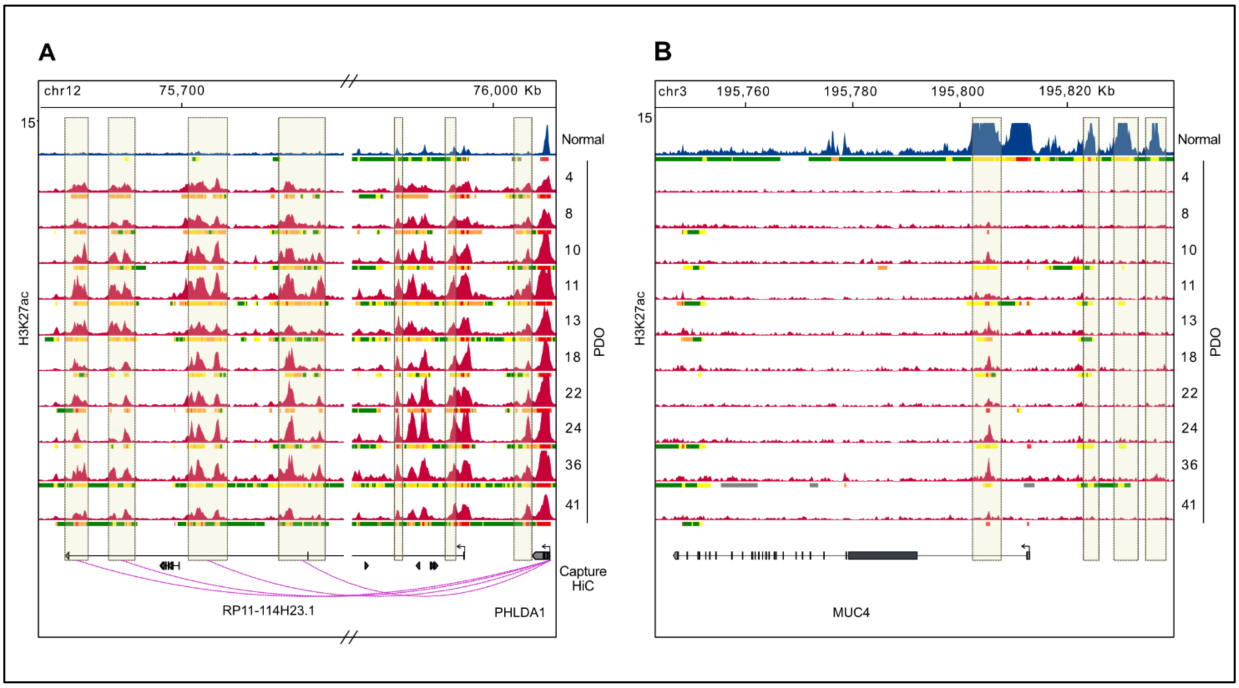


Figure 27 | Genomic overview of gained and lost active enhancer regions in PDOs. (A-B) Representative tracks of H3K27ac and ChromHMM profiles, illustrating examples of a gained **(A)** and a lost **(B)** enhancer region in PDOs compared to normal colon mucosa. Shaded boxes indicate the presence or absence of H3K27ac peaks. The Capture Hi-C track highlights the promoter-enhancer interactions within each genomic region.

4.4 Identification of transcription factors involved in the CRC enhancerome regulation

One of the essential steps to untangle the complex regulatory network orchestrating the CRC enhancerome is the study of transcription factor occupancy at the identified enhancer regions.

To this extent, we performed motif enrichment analysis within the accessible regions of the conserved gained enhancers. Interestingly, one of the enriched motifs in PDOs was that of the TEAD family transcription factors, suggesting a role for the YAP/TAZ transcriptional coactivators as putative transcriptional regulators of the conserved CRC enhancerome. The identification of YAP/TAZ as putative regulators of the CRC enhancerome is further supported by the identification of motifs for AP-1 factors (*i.e.*, Jun and Fos family members) as highly enriched in the conserved CRC enhancers. Indeed, AP-1 has been recently established in various publications as an intimate partner of YAP/TAZ, co-occupying disproportionally and pervasively *cis*-regulatory regions also bound by YAP/TAZ and TEAD [192].

To further validate the strength of this findings, we sought to determine which of the genes annotated to gained enhancers were upregulated in the majority of PDOs compared to normal tissues. To properly assign each of the 2,419 gained enhancers to a putative target gene, we integrated our ChIP-seq data with capture Hi-C data on human colon cancer[169] and the remaining differentially activated enhancers were annotated using the nearest protein-coding gene overlapping a ChromHMM-defined active promoter state (see methods section). This analysis annotated our 2,419 CRC-specific gained enhancers to 1,932 genes. Then, we selected those genes that were differentially expressed in CRC PDOs versus normal tissue based on RNA-seq analysis (n=495, padj <0.05) (**Figure 28 A**). Next, we investigated the biological function of the tumor-specific enhancerome target genes by performing a functional enrichment analysis. The results, plotted in **Figure 28**, show the most significantly enriched pathways, associated to our CRC-specific enhancers. Overall, the biological relevance of all the detected pathways (**Figure 28 B, C**) confirm the tumor-specific nature of the identified CRC enhancerome. Interestingly, the presence of the Hippo signalling pathway as the first biological term suggests an involvement of the mechano-transducers YAP/TAZ in the activation of this tumor-specific epigenetic program. In summary, through integrative analyses of multi-omics datasets we characterized the CRC enhancerome highlighting a role for the YAP/TAZ co-activators in regulating the cancer-cell intrinsic active enhancers.

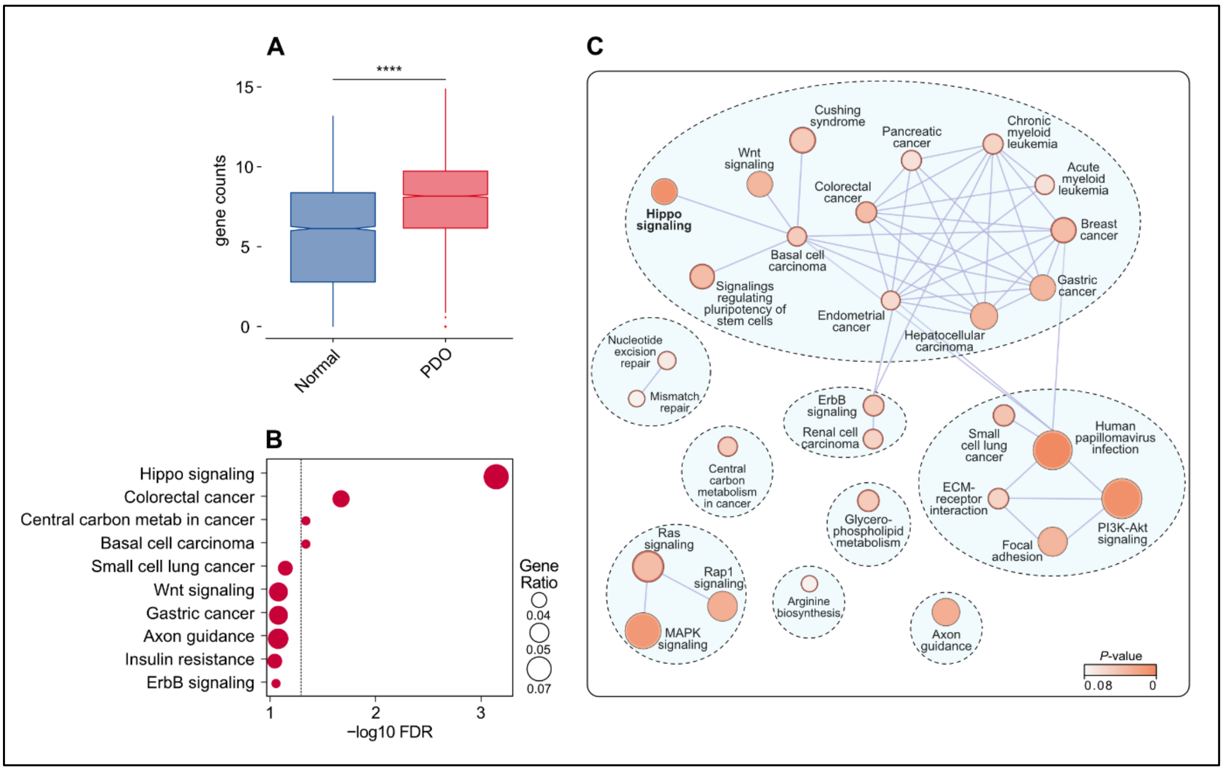


Figure 28 | Functional enrichment analysis of tumor-specific enhancerome regulated genes. (A) Normalized gene count distribution of the gained-enhancer associated genes in normal adjacent tissue and PDOs are reported in box plot. **** *P* < 0.0001 (Wilcoxon rank sum test). **(B)** Significantly enriched pathways related to gained enhancers in PDOs (g:SCS threshold < 0.05). The size of the circles corresponds to the number of gained-enhancer associated genes present in the geneset of a particular pathway (Gene Ratio). The dotted line indicates the threshold for significantly enriched pathways (false discovery rate < 0.05). **(C)** Network constructed considering the overlap between pathways enriched for tumor-specific genes regulated by gained enhancers. Circles represent pathway terms and the size of each circle is proportional to the number of genes present in the pathways' genesets. The circles are coloured according to the enrichment P-value.

4.5 YAP/TAZ as key regulators of the conserved CRC enhancerome

The Hippo pathway transducers YAP/TAZ are stably activated in CRC and other types of cancer[193] due to different functional[194] and mechanical stimuli[195]. In order to decipher the role of YAP/TAZ in the regulation of CRC enhancer regions, we first inspected the expression level of the Hippo signalling transducers YAP and TAZ. The transcriptional profile of both YAP and TAZ showed an enrichment in the tumor counterpart, including both PDOs and tumor tissues, compared with the normal adjacent tissues (**Figure 29 A**). Interestingly, TAZ showed the largest difference in gene expression levels compared to the normal counterpart (log2FC > 7). This result was also confirmed at the protein level using immunohistochemistry analysis. YAP/TAZ were not detected in the nucleus of normal tissue samples but, consistent with literature data[194,196], they were instead mostly localized in the nucleus of primary tumors and PDOs (**Figure 29 B**, second and third rows), confirming the hyperactivation of this complex in CRC. We next sought to investigate the role of the transcriptional activators YAP/TAZ in the epigenetic regulation of CRC PDOs. During tumor progression, disturbed tissue architecture increases compression forces and alters the stiffening and composition of the extracellular matrix (ECM)[195] favouring a stable activation of YAP/TAZ in a large fraction of cancer cells. Nevertheless, YAP/TAZ constantly shuttle between cytoplasm and nucleus, where they can bind to the DNA through the interaction with TEAD and other co-factors[197]. The continuous exchange between cytoplasm and nucleus, the sensitivity to the tissue stiffness and the dependency to TEAD family for the DNA binding[198] make YAP/TAZ extremely difficult to fix and immunoprecipitate. For this reason, to successfully perform a ChIP-seq experiment on YAP/TAZ a large quantity of material is required which is often difficult to obtain. To overcome this challenge, we set up and optimized a protocol to efficiently perform YAP/TAZ immunoprecipitation starting from a low amount of cells using the ChIPmentation protocol (see methods section). We generated a genomic map of TAZ recruitment to the chromatin showing the distribution of TAZ signals across ChromHMM-defined active enhancers (n=33,131) and all annotated promoters (**Figure 29 C**). We identified 14,878 statistically significant TAZ peaks, comparable with previous reports[192]. **Figure 29 D** shows the enrichment of TAZ at the promoter level of Hippo signalling canonical target genes.

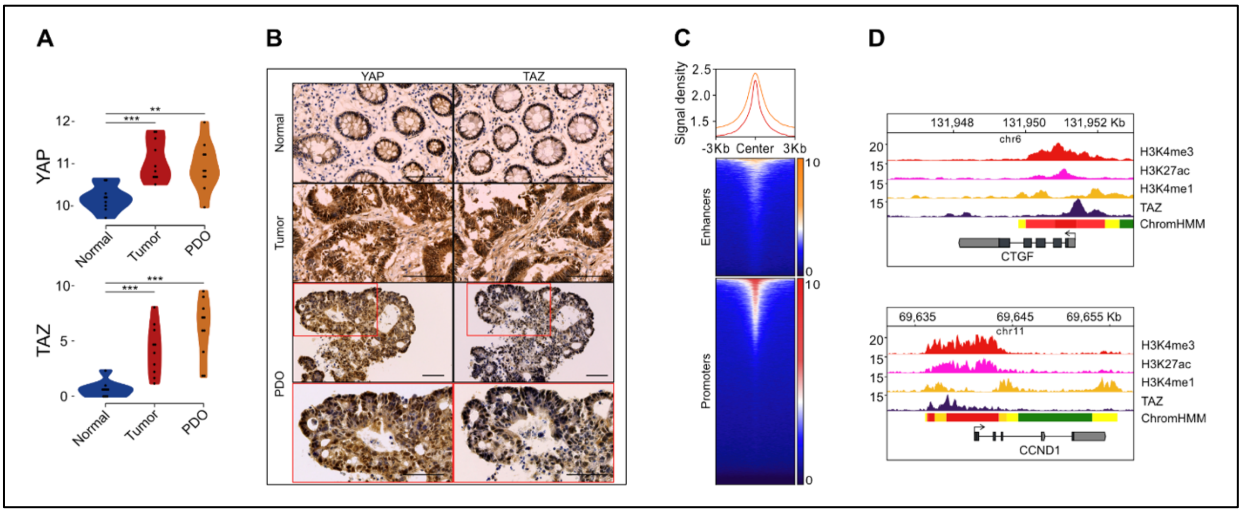


Figure 29 | YAP/TAZ transcripts and proteins are enriched in CRC tumors. (A) Violin plot shows an enrichment of YAP1 and WWTR1 normalized gene counts distribution in primary tumors and PDOs compared to normal colon tissues. ** $P < 0.01$, *** $P < 0.001$ (Wilcoxon rank sum test). **(B)** Representative immunohistochemistry images of a normal tissue, a primary tumor, and an organoid line stained for YAP1 and TAZ. The fourth row provides a magnification of the boxed area in the third row. Scale bars, 50 µm. **(C)** Signal density plot (top) and corresponding heatmaps (bottom) displaying the relative distribution of TAZ peaks around enhancer and promoter regions. **(D)** Genomic overview of the YAP1/TAZ canonical targets CTGF (top) and CCND1 (bottom). Profiles for H3K4me3, H3K27ac, and H3K4me1, ChromHMM states and TAZ ChiP-seq signals are reported.

Furthermore, the genomic profile of TAZ enrichment is also mirrored by the YAP immunoprecipitation profile (**Figure 30 A)**.

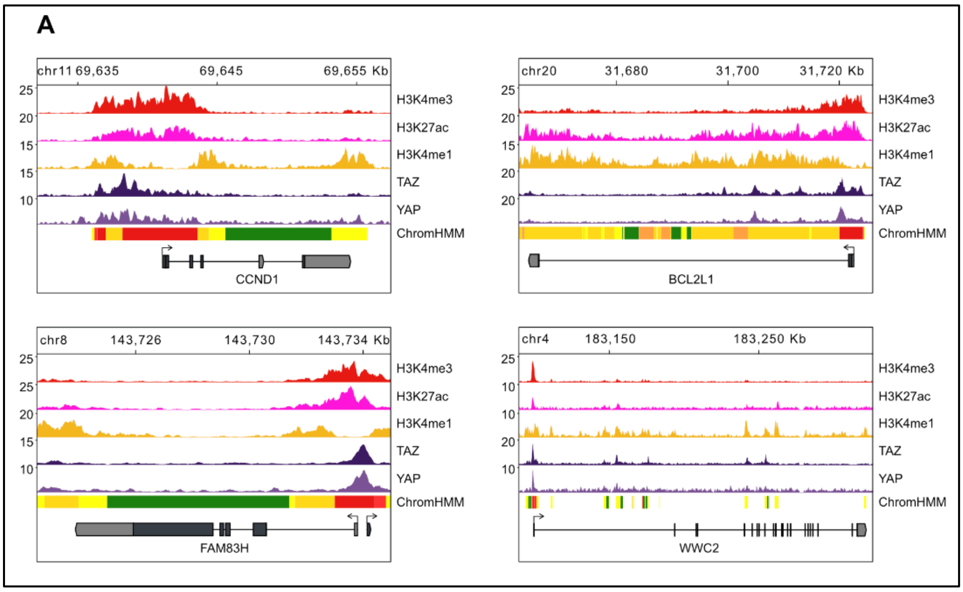


Figure 30 | Genomic overview of YAP/TAZ canonical targets. **(A)** Genomic overview of CCND1 (top left), BCL2L1 (top right), FAM83H (bottom left), and WWC2 (bottom right) showing the distribution of H3K4me3, H3K27ac, H3K4me1, TAZ, and YAP ChIP-seq signals along with ChromHMM states.

Since TEAD family proteins are necessary for the binding of TAZ to the DNA, we confirmed their cooperative interaction in human CRC[199] by searching for transcription factor binding motifs encompassing the summit of TAZ peaks and we found the TEAD family binding motif as the most enriched (**Figure 31 A**). To further confirm the active role of this transcriptional activator in CRC, we combined the CRC-specific ChromHMM states and the ChIP-seq data for TAZ to characterize its genomic occupancy. Interestingly, the majority of TAZ peaks were located at active regulatory regions with the 95% of peaks equally distributed across promoters and enhancers (**Figure 31 B**). Focusing on the gained CRC-enhancerome, we assessed the overlap of TAZ peaks with the total number of differentially activated enhancers (n=2,419, **Figure 26 B**) as well as those conserved in at least 50% or 80% of PDOs (**Figure 26 C**). Notably, TAZ enrichment increased with the level of enhancer conservation across PDOs (**Figure 31 C**), suggesting a role for TAZ in regulating the shared CRC enhancerome independent of patient-driven tumor diversity. TAZ bound 40% of highly conserved gained enhancers (n=195) compared to less than 20% of all gained enhancers. This result indicates a key role of YAP/TAZ in the regulation of the CRC-conserved enhancers. The core 195 enhancers shared by at least eight PDOs were assigned to their interacting promoter, through capture Hi-C data or the nearest active TSS, generating a CRC enhancerome signature of 211 TAZ-bound genes.

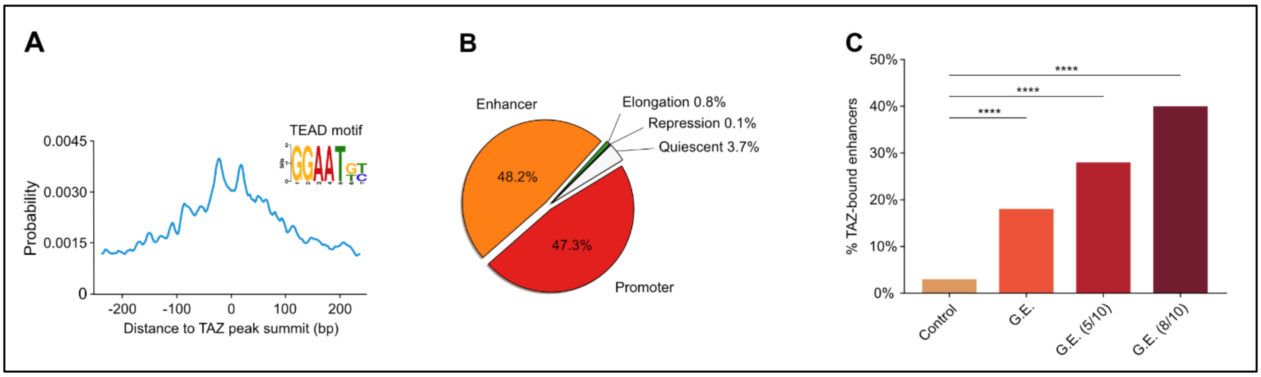


Figure 31 | YAP/TAZ are key regulators of the conserved CRC enhancerome. (A) Enrichment of the TEAD binding motif around the summit of TAZ peaks. **(B)** Distribution of TAZ peaks across functional elements for active and inactive genomic regulatory regions defined in our ChromHMM de novo discovery analysis (pie chart). **(C)** Enrichment of TAZ in CRC-specific enhancers. **** *P* < 0.0001 (Fisher's exact test). G.E.: Gained enhancers.

Interestingly, TAZ itself was one of the signature genes shared by all PDOs. As shown by capture Hi-C data, its expression was regulated by a TAZ-bound intronic enhancer, marked by an "Active Enhancer" chromatin state and located almost 100 kb downstream the TSS (**Figure 32 A**, boxed area). The enrichment of TAZ at its promoter and the reinforcement performed by the possible TAZ-regulated enhancer suggests previously unreported feedback loop driving its transcriptional modulation. The CRC enhancerome signature also included the gene that encodes for epiregulin (EREG), a known ligand of the epidermal growth factor (EGF) receptor whose expression is increased in numerous human cancers[200]. Furthermore, EREG is a target of the Hippo signalling pathway and is involved in intestinal regeneration and YAP function[196]. The regulation of EREG is mediated by two long distant downstream enhancers detected thought capture Hi-C data (**Figure 32 B**). These TAZ-bound enhancers were enriched for H3K4me1, H3K27ac and H3K4me3. The enrichment of H3K4me3 and the high level of RNA-seq peaks further underlines the hyperactivation of these regulatory regions.

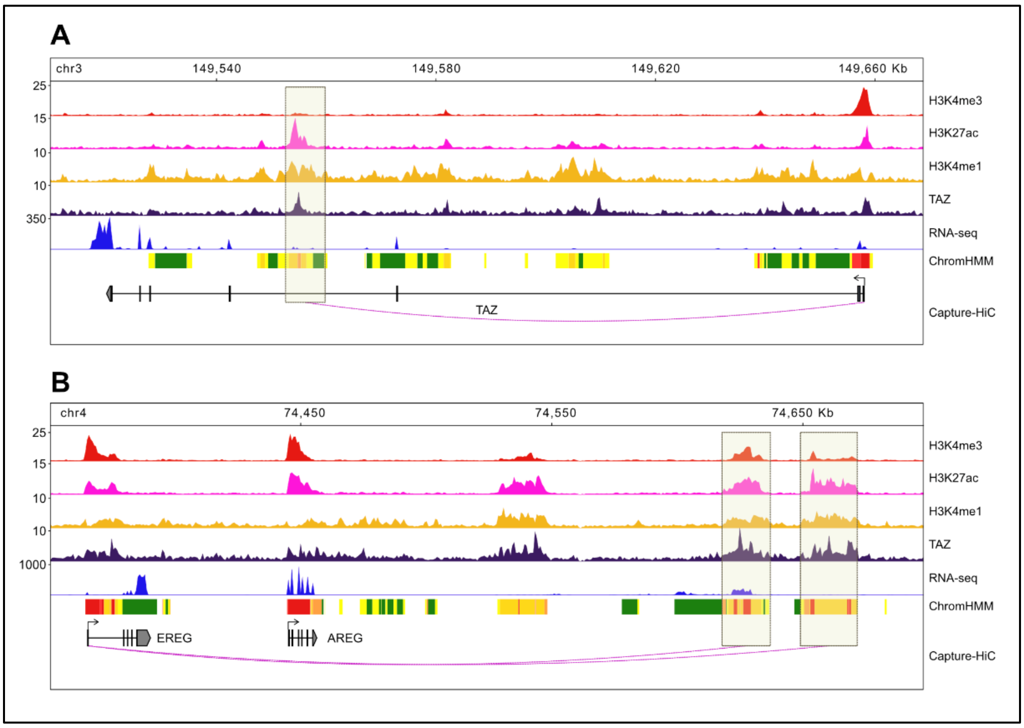


Figure 32 | Genomic overview of genes regulated by TAZ-bound conserved CRC enhancers. (A-B) Representative genomic regions of TAZ target genes displaying Capture Hi-C interactions between promoters and conserved CRC enhancers. The tracks show H3K4me3, H3K27ac, H3K4me1 and TAZ ChiP-seq signals along with RNA-seq and ChromHMM profiles for the TAZ **(A)** and EREG **(B)** genomic loci.

Another key gene that was annotated to the TAZ-regulated conserved CRC enhancers was Forkhead box Q1(FOXQ1), which is involved in cell cycle regulation, cell signalling and tumorigenesis[201]. The enrichment of YAP/TAZ at the promoter and a highly active enhancer upstream of FOXQ1, in conjunction with an increased expression level of this gene, pinpoints FOXQ1 as a new possible YAP/TAZ target gene (**Figure 33 A**). To verify this finding, we performed *in situ* hybridization on normal and tumor tissues confirming that FOXQ1 gene expression is restricted to the CRC sections that express YAP in the nucleus (**Figure 33 B**). Taken together, these results revealed a conserved CRC-enhancerome core that is regulated by the Hippo signalling pathway effector TAZ independently of patient-to-patient tumor molecular diversity. This shared-enhancerome core offers new insights into CRC epigenetics by highlighting enhancer regions that are involved in tumor transcriptional deregulation and controlled by the Hippo transducer TAZ.

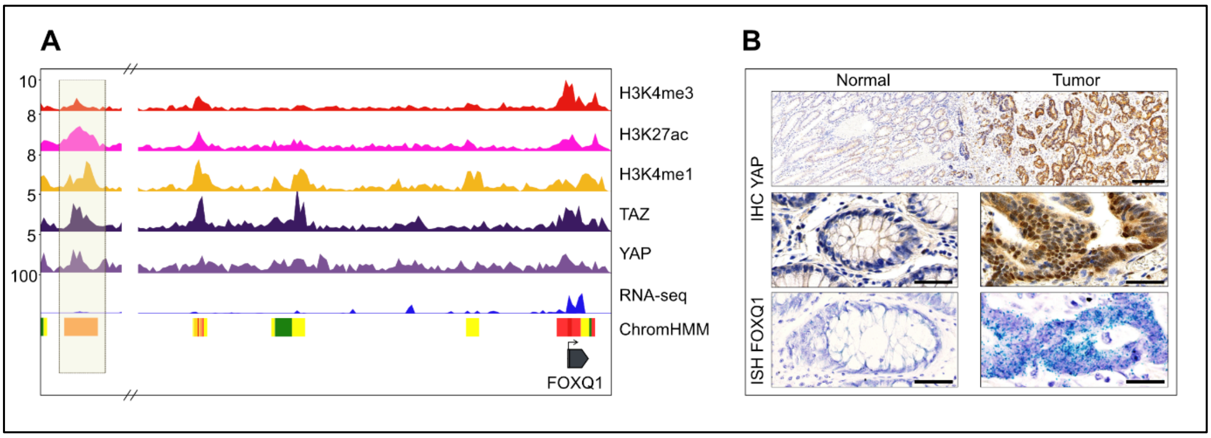


Figure 33 | FOXQ1 as a potentially new YAP/TAZ target gene. (A) Genomic overview of FOXQ1 showing the distribution of H3K4me3, H3K27ac, H3K4me1, TAZ, and YAP ChIP-seq signals along with RNA-seq levels and ChromHMM states. **(B)** FOXQ1 expression in the same tissues expressing YAP. Images of a human CRC (right) and nearby healthy colon mucosa (left) tissue within the same section. The graphs show: immunohistochemical (IHC) staining for YAP (upper and middle panels; scale bars, 250 and 50 mm, respectively) and RNA *in situ* hybridization (ISH) for FOXQ1 (bottom panel; scale bars, 50 mm). Nuclei were counterstained with hematoxylin.

4.6 TAZ-regulated CRC enhancer blueprint is shared by various types of cancer

After the identification of the YAP/TAZ-regulated CRC conserved enhancerome, we asked whether this core of enhancers was also shared in other cancer types. To investigate the relevance of the YAP/TAZ-regulated CRC enhancerome in human cancer pathology, we assessed the chromatin accessibility levels of the 195 core CRC-enhancers in 23 diverse cancer types using ATAC-seq data obtained from TCGA[116] (**Figure 34 A**). The 195 CRC enhancer regions displayed a strong chromatin accessibility profile across all colon adenocarcinoma (COAD) TCGA samples (**Figure 34 A**, left part of the heatmap) validating their regulatory role and underlining the CRC-specific nature of the TAZ-regulated conserved enhancerome detected in the PDOs library. Notably, 46 out of the 195 active regulatory elements (23%) were highly accessible in all cancer types (**Figure 34**, enhancer cluster in blue). We refer to these accessible and shared regulatory elements as “ultra-conserved” providing a core of pan-cancer enhancers, with a possible involvement in the molecular mechanisms at the basis of tumor biology and maintenance. To validate this finding, we analysed H3K27ac occupancy from several primary tumor and normal adjacent tissues (**Table 3**). Interestingly, we observed that there was an enrichment of H3K27ac signal at the pan-cancer enhancers in all primary tumors compared to normal tissues (**** *P* <

0.0001, Wilcoxon rank sum test, **Figure 34 B**), confirming the epigenetic activation of these enhancer regions in diverse cancer types.

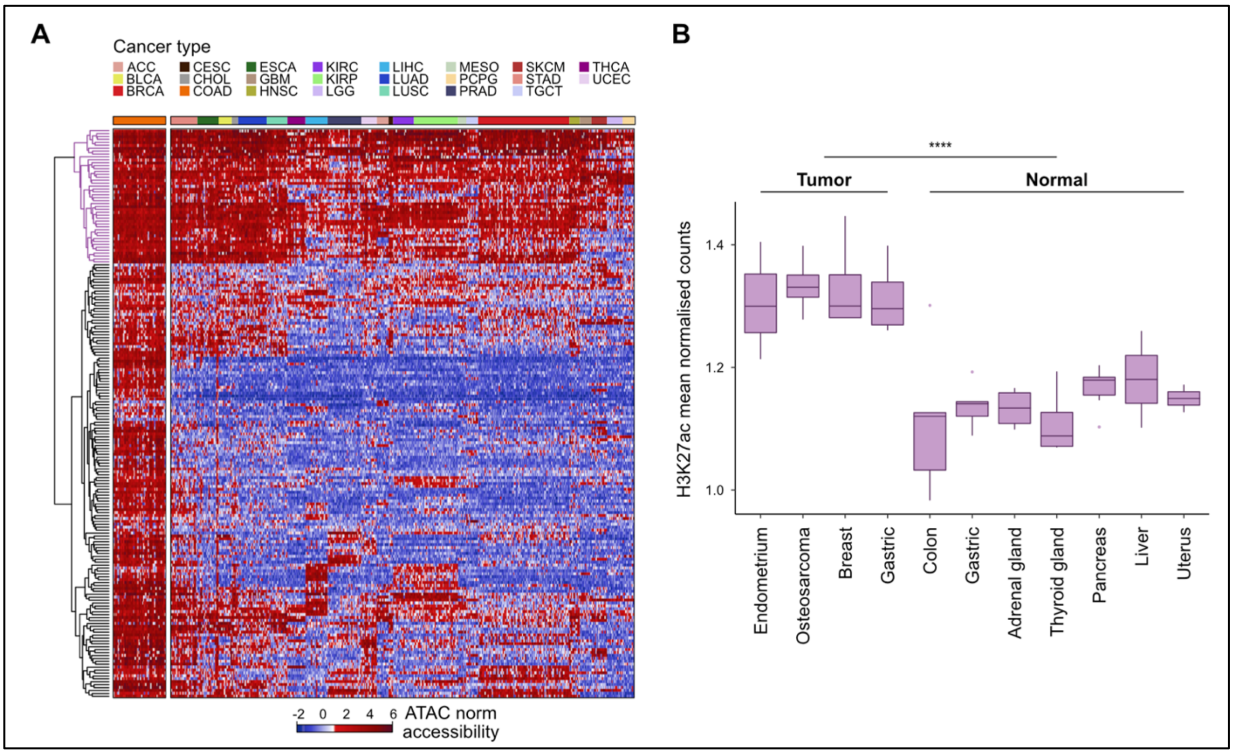


Figure 34 | Conserved CRC enhancers are shared by other types of cancer. (A) Chromatin accessibility profiles of the 195 conserved gained enhancers in 23 diverse primary human cancer types reveals a signature of 46 pan-cancer enhancers with highly conserved accessibility profiles across cancer types. The heatmap represents log2 normalized insertion counts of ATAC-seq data derived from TCGA. Colon adenocarcinoma (COAD) samples are the first cancer type reported on the left of the heatmap. ACC, Adrenocortical carcinoma; BLCA, Bladder Urothelial Carcinoma; BRCA, Breast invasive carcinoma; CESC, Cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, Cholangiocarcinoma; COAD, Colon adenocarcinoma; ESCA, Esophageal carcinoma; GBM, Glioblastoma multiforme; HNSC, Head and Neck squamous cell carcinoma; KIRC, Kidney renal clear cell carcinoma; KIRP, Kidney renal papillary cell carcinoma; LGG, Brain Lower Grade Glioma; LIHC, Liver hepatocellular carcinoma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma; MESO, Mesothelioma; PCPG, Pheochromocytoma and Paraganglioma; PRAD, Prostate adenocarcinoma; SKCM, Skin Cutaneous Melanoma; STAD, Stomach adenocarcinoma; TGCT, Testicular Germ Cell Tumors; THCA, Thyroid carcinoma; UCEC, Uterine Corpus Endometrial Carcinoma **(B)** H3K27ac signal intensities for the pan-cancer enhancers in primary tumors compared to normal tissues. ****$P <$ 0.0001, Wilcoxon rank sum test.

Among the genes annotated to the ultra-conserved enhancers, some are involved in cancer (*e.g.* MYC) and/or are known target genes of the Hippo signalling pathway (*e.g.* EREG, PHLDA1, FJX1)[194], whereas for some genes (*e.g.* UBE2H) there are no previous reports of their role as YAP/TAZ target genes (**Figure 35 A**).

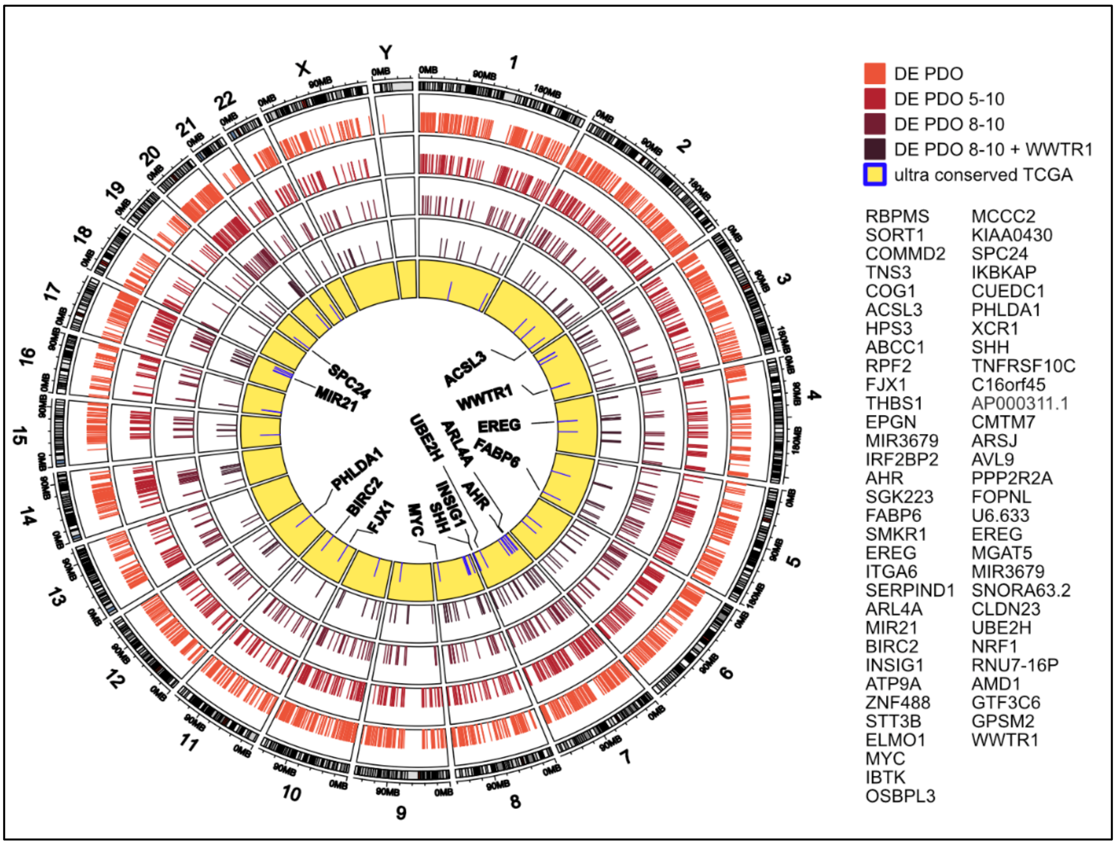


Figure 35 | Circos plot of the CRC gained enhancerome layers. (A) Circos plot showing the genomic distribution of CRC-specific gained enhancers. The outer ring displays the chromosomes. Tracks are described from outside to inside. Track 1: Enhancers differentially activated in PDOs compared to normal colon tissue (n=2,419). Track 2: Gained enhancers shared by at least 5 PDOs (n=1,216). Track 3: Gained enhancers shared by at least 8 PDOs (n=486). Track 4: TAZ-bound gained enhancers shared by at least 8 PDOs (n=195). Track 5: TAZ-bound ultra-conserved pan-cancer enhancer regions (n=46) with high chromatin accessibility profiles in diverse human cancers based on TCGA ATAC-seq data. The inner ring highlights some of the genes annotated to the 46 pan-cancer enhancers. The number of enhancers for each chromosome is balanced. The list of the 46 pan-cancer enhancers is reported on the right.

Interestingly, the intronic enhancer of TAZ was part of the 46 pan-cancer enhancers bound by YAP/TAZ. **Figure 36 A** displays the genomic overview of the TAZ locus considering the epigenetic profile of normal adjacent tissue, PDOs and ATAC-seq data deriving from diverse types of tumors. The normal tissue displays a repressed configuration of the chromatin, lack of gene expression (consistent with **Figure 29 A**) and lack of TAZ enrichment at both enhancer or promoter level. On the contrary, the chromatin profile in PDOs shows a highly active chromatin configuration together with high gene expression and TAZ enrichment. Considering the TCGA ATAC-seq profiles, we can observe the presence of multiple, highly accessible regions in all the tumors. One of the open chromatin regions coincides with the YAP/TAZ-bound CRC-conserved enhancer (**Figure 36 A**, boxed area), suggesting the functional role of this transcriptional activator in diverse human cancer types. Overall, through integrative multi-omics analyses of stroma-free PDOs we identified a YAP/TAZ-regulated CRC active enhancerome that is shared by all TCGA colon adenocarcinoma samples, demonstrating that this enhancer signature captures the epigenetic profile intrinsic to cancer-cells independent of stromal contribution. Extending the relevance of our findings to other cancer pathologies, we revealed a core of "ultra-conserved" pan-cancer enhancers that display an active chromatin profile in diverse solid tumors, suggesting a role for the Hippo signalling pathway effectors in the deregulation that characterizes the human cancer enhancerome.

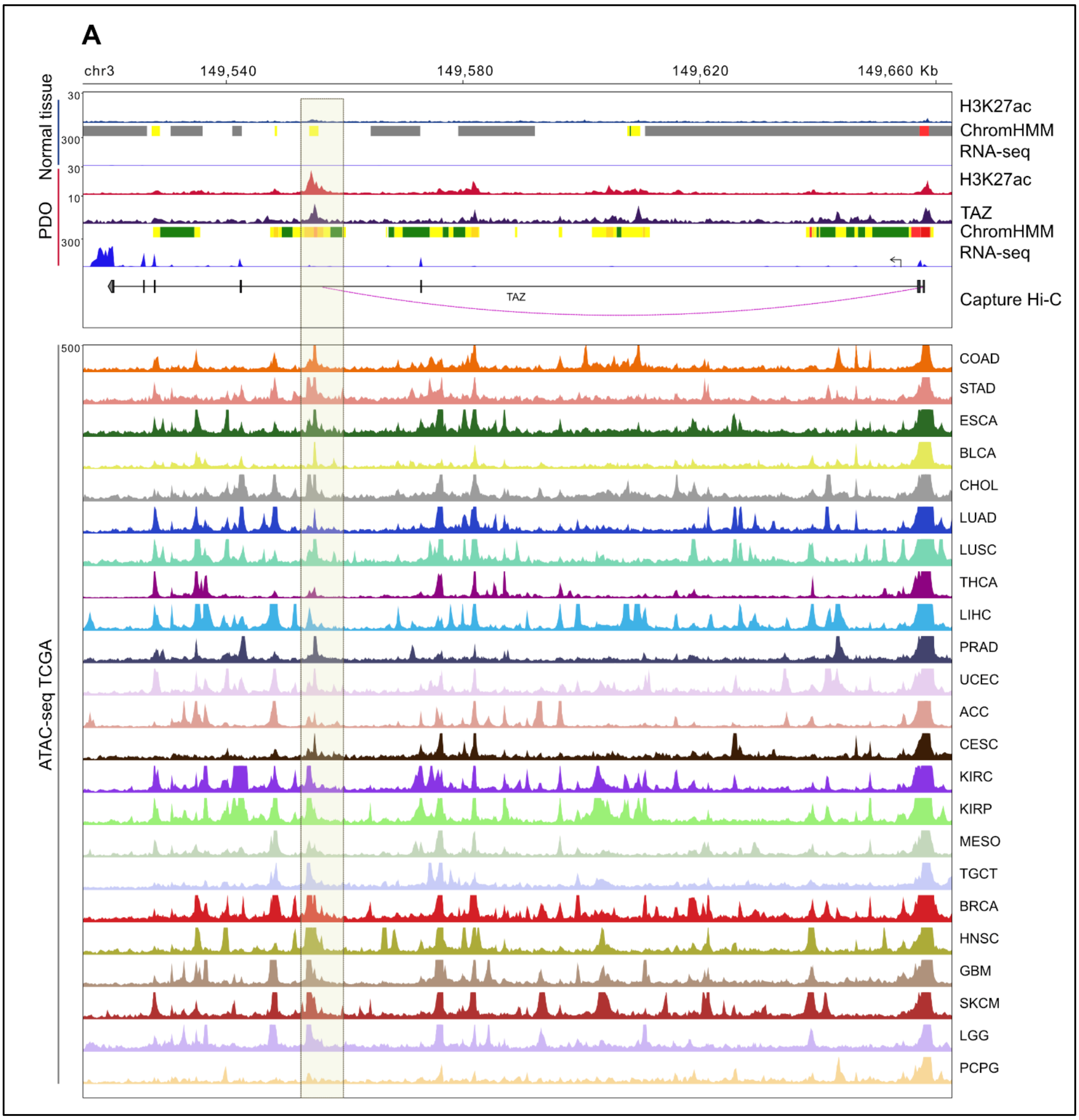


Figure 36 | Genomic overview of the TAZ locus. (A) Genomic overview of the TAZ locus in representative samples of normal tissue and PDOs, and in TCGA cancer types. Upper panel shows: H3K27ac ChIP-seq profiles, ChromHMM states and RNA-seq signals in normal tissue; H3K27ac and TAZ ChIP-seq profiles, ChromHMM states, and RNA-seq signals in PDOs; and CRC capture Hi-C data. Bottom panel displays ATAC-seq profiles for 23 TCGA cancer types.

4.7 Single-cell landscape of the pan-cancer core of enhancers

We next sought to investigate the epigenetic abberration at single cell resolution. To this extent, we investigated the distribution and expression of the signature genes regulated by the YAP/TAZ-bound pan-cancer core of enhancers at single cell resolution by performing single-cell RNA sequencing of a primary CRC tissue. Following quality control analyses to remove contaminants and low quality cells, we analyzed 3044 cells and 19460 genes. Graph-based clustering of the CRC tissue identified eleven clusters. Using previously reported markers[80,170,184,202], we were able to discriminate six normal-like clusters (**Figure 37 A, B**): stem cells (LGR5, SMOC2, RGMB, in lightgreen), early and late progenitors (in grey and orange, respectively) on the basis of decreasing level of stemness and cell-cycle genes and increasing level of terminally differentiated epithelial cells, enterocytes (CA1, KRT20, FABP1, in purple), goblet cells (SPINK4, REG4, TFF3, in darkgreen), and paneth cells (SPIB, CA7, BEST4, in azure). We further confirmed the identity of these clusters by performing geneset enrichment analysis (GSEA) of the ranked differentially expressed genes for each normal-like cluster against reference genesets of differentiated epithelial cells published in a previous study[184] (**Figure 37 C**).

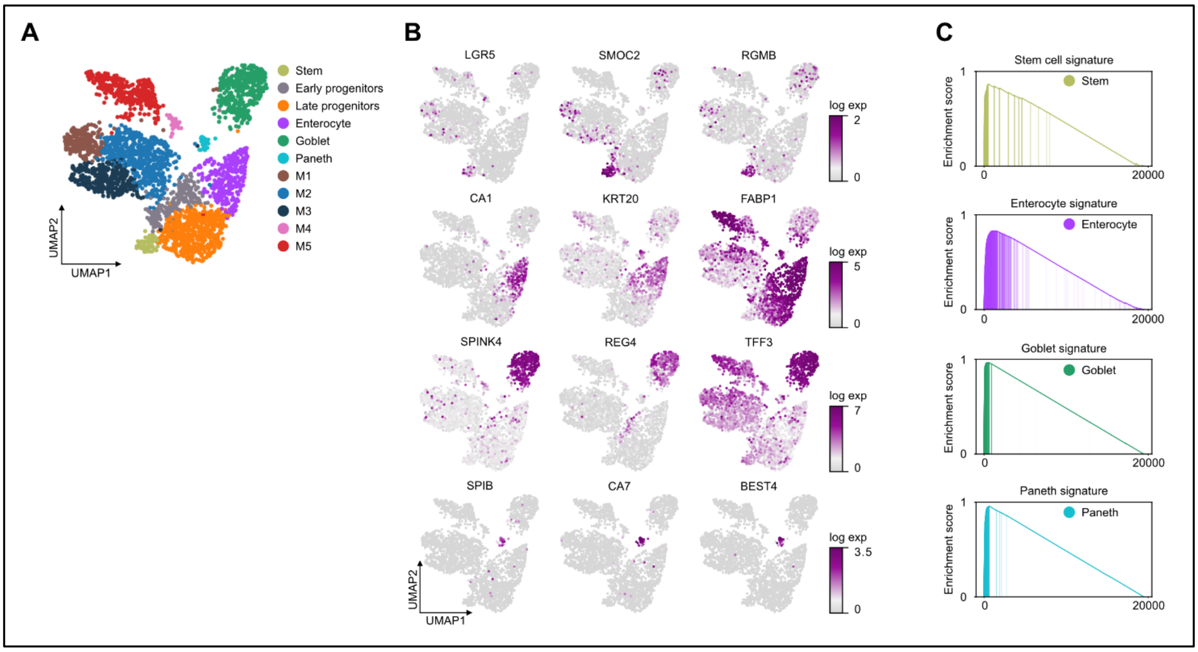


Figure 37 | Identification of known intestinal subpopulations in CRC primary tumor. (A) UMAP visualization reporting the six known populations (Stem cells, early and late progenitors, Enterocytes, Goblet and Paneth cells) and five unknowns clusters (M1-M5) identified using unsupervised clustering of a CRC primary tumor. Each point depicts a single cell, colored according to the cluster it belongs to. **(B)** UMAP coloured by the log expression of known marker genes for differentiated epithelial cells: stem cells (LGR5, SMOC2, RGMB), enterocytes (CA1, KRT20, FABP1), goblet cells (SPINK4, REG4, TFF3), paneth cells (SPIB, CA7, BEST4). **(C)** Geneset enrichment analysis of differentially expressed genes in CRC primary tumor clusters using reference genesets of differentiated epithelial cells published in previous work[184].

We next focused on the five additional clusters with an undefined cell phenotype (M1 to M5) and a substantial deviation from the normal epithelium differentiation programs (**Figure 37 A**). Copy number variation (CNV) inference from gene expression data[185] confirmed the presence of genetic aberrations in these clusters and revealed intratumor heterogeneity (**Figure 38 A**). M1-5 clusters display diverse CNV patterns compared to normal-like cell populations, suggesting the existence of two distinct genetic clones (clone A and B) which are represented by clusters M1-M3 and M4-M5, respectively (**Figure 38 B**). Despite their undefined phenotypes, the five malignant clusters were characterized by heterogeneous transcriptional states related to cancer; stemness in M1, ribosomal biogenesis in M2-M3, proliferation in M4, and hypoxia in M5 (**Figure 38 C**).

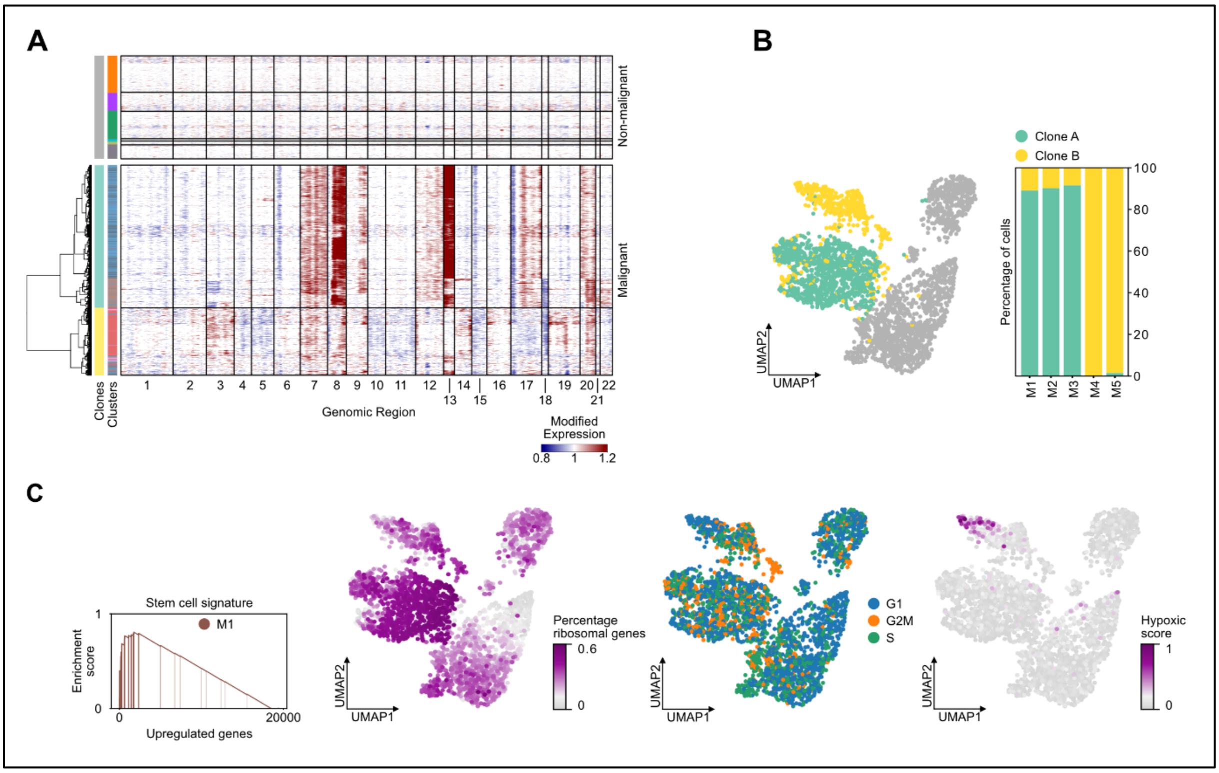


Figure 38 | Characterization of the heterogeneous transcriptional state of the malignant clusters. (A) Heatmap of scRNAseq-inferred large-scale chromosomal CNVs for individual cells based on average expression intensity across chromosomal segments. CNVs amplifications and deletions are denoted in red and blue, respectively. Comparison to the reference normal-like cells of the primary CRC tissue reveals two main patterns of CNVs, indicative of different genetic clones. **(B)** UMAP plot depicting clone A and B cells based on inferred CNV patterns (left) and barplot showing the percentage of clone A and B cells within each malignant population (right). **(C)** Transcriptional states in malignant clusters M1 to M5. GSEA of M1 enrichment in stem signature (first column). UMAP plots depicting increased expression of ribosomal genes in M2 and M3 (second column), cell-cycle genes in M4 (third column), and a hypoxic score (CA9, SLC2A3, SLC2A1, HIF1A, VEGFA, PFKP, HK2, BNIP3, PDK1) in M5 (fourth column).

We then performed a pseudo-time analysis and identified a differentiation trajectory originating from the stem compartment, progressing through transient amplifying cells, and finally culminating in enterocytes, goblet cells and the malignant cluster M5 (**Figure 39 A, B**). Analysis using PArtition-based Graph Abstraction (PAGA) confirmed the existence of two main branches of differentiation within the primary CRC tissue (**Figure 39 C**). With stem cells at the root, malignant clusters mark a trajectory of tumoral states from M1 to M5 that is divergent from the normal differentiation paths of non-malignant cells toward enterocytes, paneth or goblet cells.

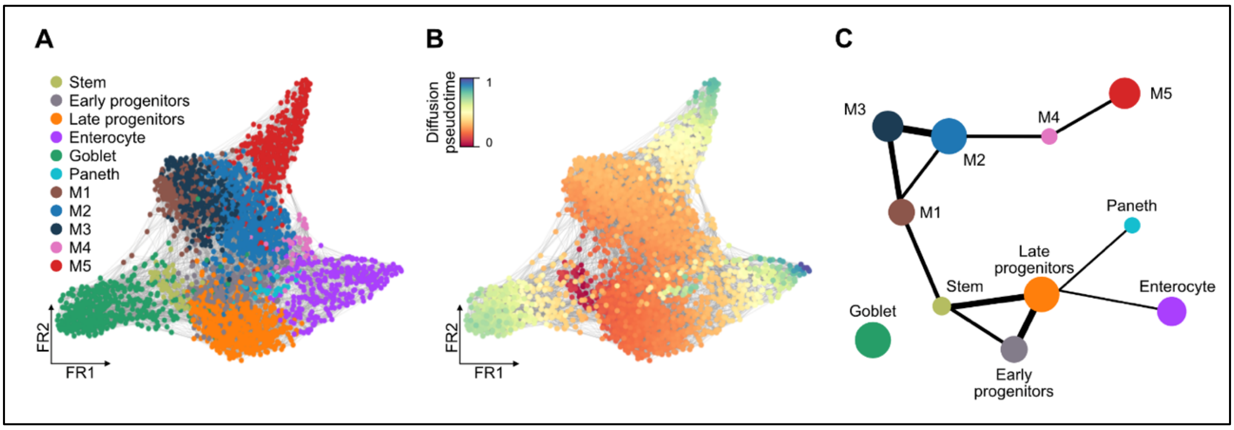


Figure 39 | Trajectory analysis of the primary CRC tissue. (A) Force-directed graph depicting the malignant and non-malignant clusters of a primary CRC tissue. **(B)** Pseudo-temporal ordering of all clusters with stem cells (depicted in red) as root cells. **(C)** Trajectory analysis based on cell connectivity (see Methods session, Data visualization and trajectory analysis, Partition based graph abstraction) distinguishes two main branches of non-malignant and malignant subpopulations. Nodes correspond to the distinct cell clusters with the node size being proportional to the number of cells in the cluster. The thickness of the edges denotes the strength of connectivity between two clusters.

After the extensive characterization of the cell populations present in the CRC primary tumor, we investigated the cancer regulatory blueprint defined by genes annotated to the YAP/TAZ regulated pan-cancer core of enhancers (**Figure 34 A**). Hierarchical clustering analysis using the genes annotated to the YAP/TAZ-regulated pan-cancer enhancers (n=46) showed that the gene signature of this blueprint can distinguish between the malignant and non-malignant cell populations (**Figure 40 A, B**).

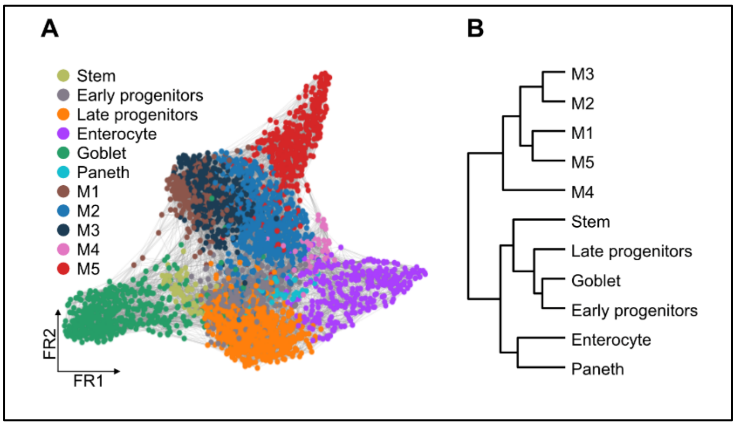


Figure 40 | Cancer regulatory blueprint distinguishes between malignant and non-malignant clusters. (A) Force-directed graph depicting the malignant and non-malignant clusters of a primary CRC tissue. **(B)** Hierarchical clustering analysis and resulting dendrogram of cell populations based on genes associated to the CRC conserved YAP/TAZ-bound enhancers distinguishes non-malignant from malignant cell clusters.

To evaluate the signal distribution and intensity of the gene signature we exploited a score based on the co-expression of genes within cells and their level of expression, hereinafter referred to as cancer regulatory blueprint score (CRB score, see methods for more details). We observed that the CRB score is mostly absent from non-malignant populations but highly enriched in the malignant clusters (**Figure 41 A, B**). Interestingly, the CRB score is equally enriched in both genetic clones regardless of the genetic and transcriptional differences that define them. Moreover, M1 and M5 showed the highest enrichment for the CRB score (**Figure 41 B**). M1 portraits as the most immature part of the tumor, enriched in stem markers LGR5+, SMOC2+, RGMB+ and it is the starting point of the cancer-related transcriptional trajectory. Instead, M5 displays high expression of differentiated colon epithelium markers (*e.g.* KRT20+ and FABP1+) and represents the end point of the tumor developmental branch. This widespread cellular distribution of the cancer regulatory blueprint across diverse parts of the tumor irrespective of the intratumoral genetic and transcriptional heterogeneity suggests the involvement of the YAP/TAZ-regulated pan-cancer enhancers in both cancer initiation and maintenance. Previous studies[185,196] reported the role of YAP/TAZ in the promotion of stem-like properties, amongst a diverse array of downstream effects. We thus asked whether the epigenetically-driven deregulation could relate to stemness. Interestingly, we found that not all malignant cells with an active cancer regulatory blueprint displayed stem-like properties. For instance, M1 and M5, the tumoral clusters with the highest CRB score (**Figure 41 B, D, E, F**), had moderate and low stemness scores, respectively (**Figure 41 C, E, F**). Conversely, the normal-like stem compartment displayed low CRB scores (**Figure 41 A, B, E, F**), consistent with the specificity of the blueprint in the malignant cell populations. This suggests that the blueprint is a feature of cancer that relates to YAP/TAZ-driven effects on tumoral functional states that reach beyond the acquisition of stemness. Collectively, we show that the cancer regulatory blueprint is enriched in the malignant cell populations despite their genetic and transcriptional heterogeneity, is not related to stemness per se, and is associated to an aberrant YAP/TAZ activation that is required for both tumorigenesis and maintenance of the cancer cell state.

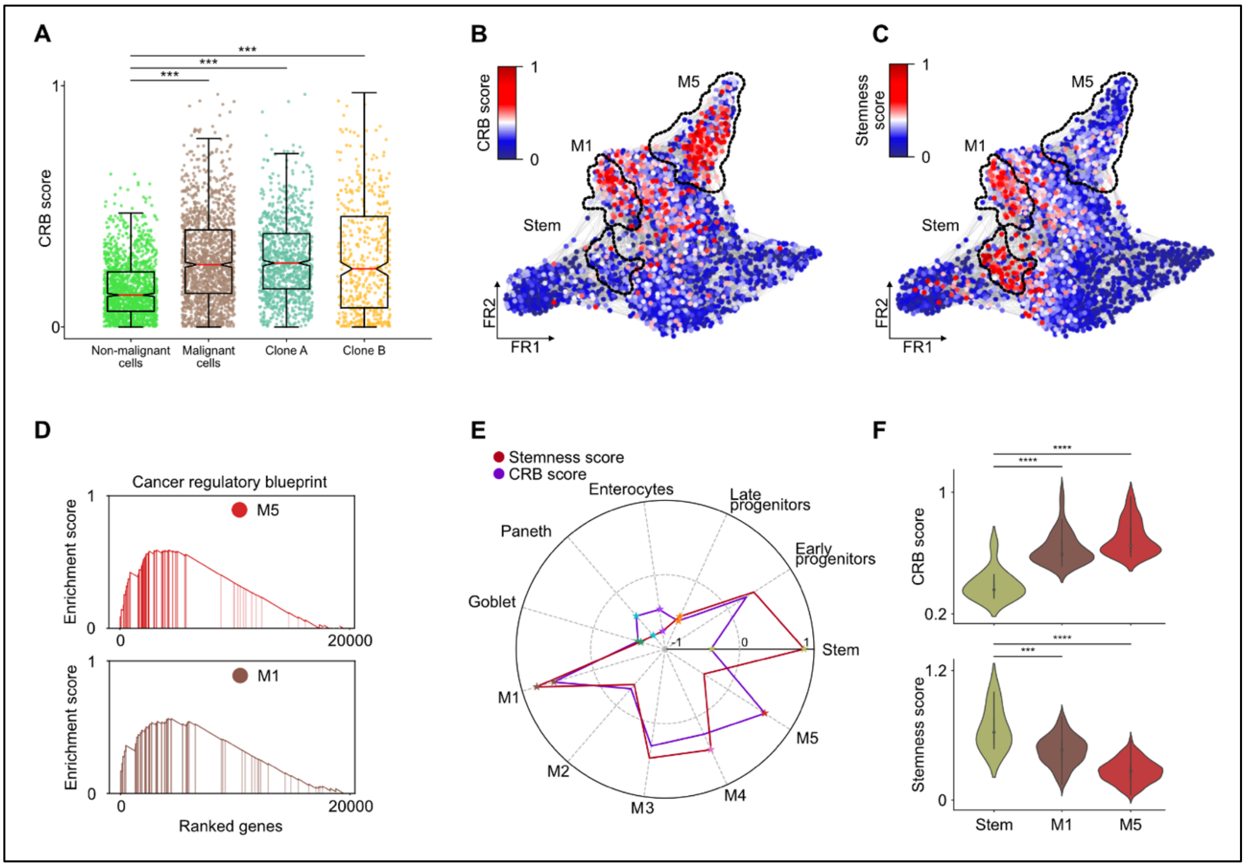


Figure 41 | Malignant cells display significantly higher CRB scores. (A) Boxplot reporting the distribution of the CRB score in all non-malignant (green) and malignant (brown) clusters, and separately in malignant cells of clone A (sapphire) and clone B (yellow). Boxplots describe the median and interquartile range with whiskers denoting the 1.5 x interquartile range. **(B-C)** UMAP representation of the epigenetic **(B)** and stemness **(C)** score across all cell clusters. Contour lines denote cells of the stem, M1 and M5 cell populations. **(D)** GSEA plots showing the enrichment of the cancer regulatory blueprint in the M5 (top) and M1 (bottom) malignant cell populations. **(E)** Enrichment of the CRB and stemness scores across the malignant and non-malignant clusters of the CRC primary tissue. Stars on the edges denote a statistically significant enrichment (FDR < 0.05). **(F)** Violin plots of the top quantile of the CRB score in the stem, M1 and M5 clusters (top), and distribution of the stemness score calculated in the same cells (bottom). *** $P < 0.001$, **** $P < 0.0001$, Wilcoxon rank sum test.

4.8 Pan-cancer enhancers core signature is shared among diverse types of cancer at single cell resolution

Considering the great heterogeneity of cancers, we next asked whether our findings can be confirmed in an independent CRC dataset or a different tumor type. To this end, we took advantage of single cell RNA sequencing data of whole CRC[170] and adenocarcinoma (LUAD)[171] tissues, including normal tissue counterparts. Considering the extensive work for cell population identification and labelling, we decided to maintain the original cell annotation as published in the two studies.

We re-analysed a total of 63,689 cells from normal and tumor tissues of 23 Korean patients affected by CRC[170]. Since the single cell experiment was performed on the whole tissues, it was possible to identify not only epithelial cells but also B cells, Mast cells, Myeloid, Stromal cells and T cells from both normal and tumor counterpart (**Figure 42 A, B**) Furthermore, the 23 collected samples represented the heterogeneity of CRC; in fact the samples were previously classified according to CMS classification and the number of samples belonging to each of the CMS subtypes 1 to 3 was well-balanced. CMS4 samples consisted of only 11 epithelial cells, probably due to their highly mesenchymal nature. We found that the CRB score was largely absent in the non-epithelial cells but was highly enriched in epithelial cells (**Figure 42 C, D**). Focusing on the epithelial compartment we observed not only the specificity of the CRB score for cancer cells (**Figure 42 E**) but also an enrichment of the CRB score in the malignant cell populations despite the molecular and phenotypic differences due to the presence of different CMS groups (**Figure 42 F**).

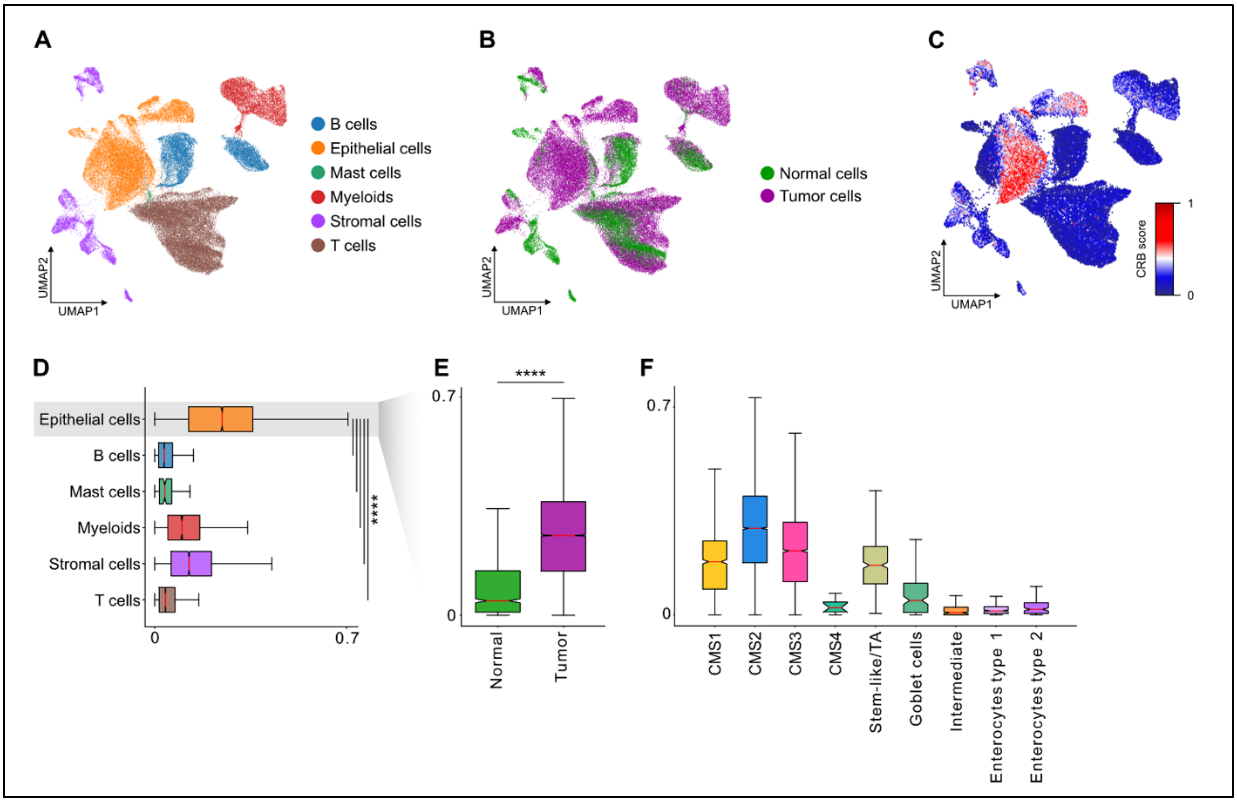


Figure 42 | CRB-score is enriched in tumor epithelial cells of CRC from 23 patients. (A) UMAP visualization reporting the major cell populations identified in the analysis of 23 primary CRC tumor and normal tissues. **(B)** UMAP visualization showing the tissue of origin (Normal or Tumor) for each cell **(C)** UMAP representation of the epigenetic score **(D)** Boxplot reporting the distribution of the CRB score in the identified cell populations, Epithelial cells are reported in orange. Boxplots describe the median and interquartile range with whiskers denoting the 1.5 x interquartile range. **** P < 0.0001, Wilcoxon rank sum test. **(E)** Boxplot reporting the distribution of the CRB score in normal (darkgreen) and tumor (purple) epithelial cells**** P < 0.0001, Wilcoxon rank sum test. **(F)** Boxplot reporting the distribution of the CRB score in normal cell sub-populations (*i.e.* Stem-like TA, Goblet cells, Intermediate, Enterocytes type 1, Enterocytes type 2) and tumor (CM1, CM2, CM3, CM4) epithelial cells. All pairwise comparisons between each CMS and the other sub-populations were statistically significant (**** P < 0.0001, Wilcoxon rank sum test). The only non-significant was the comparison between CMS1 and Stem-like/TA. CMS4 was not considered in this evaluation.

To investigate the CRB in a different cancer type, we re-analysed scRNA-seq data of 208,506 cells from 44 patients affected with LUAD. Besides epithelial cells, the LUAD dataset also contains B cells, Endothelial cells, Fibroblasts, MAST cells, Myeloid cells, NK cells, Oligodendrocytes, and T lymphocytes. In addition, it includes pleural fluids (PE), and lymph node (mLN) or (mBrain) brain metastases, as well as distant normal lymph nodes (nLN). Notably, we noticed a strong enrichment of the cancer regulatory blueprint in epithelial cells compared to non-epithelial cells deriving from all collected districts. Collectively, these findings suggest that the CRB score is specifically enriched in epithelial cells of solid tumors (**Figure 43 A, B, C, D**). Considering the variegated origin of the samples, which included also lymph nodes and metastasis, we decided to focus our attention on the epithelial cells deriving from normal and tumor counterpart of the lung tissue (**Figure 43 B**, cells coloured in green and magenta). As in CRC, we confirmed also in LUAD the specificity of the cancer regulatory blueprint for malignant cells deriving from tumor tissue compared to epithelial cells from normal tissue. (**Figure 43 E**).

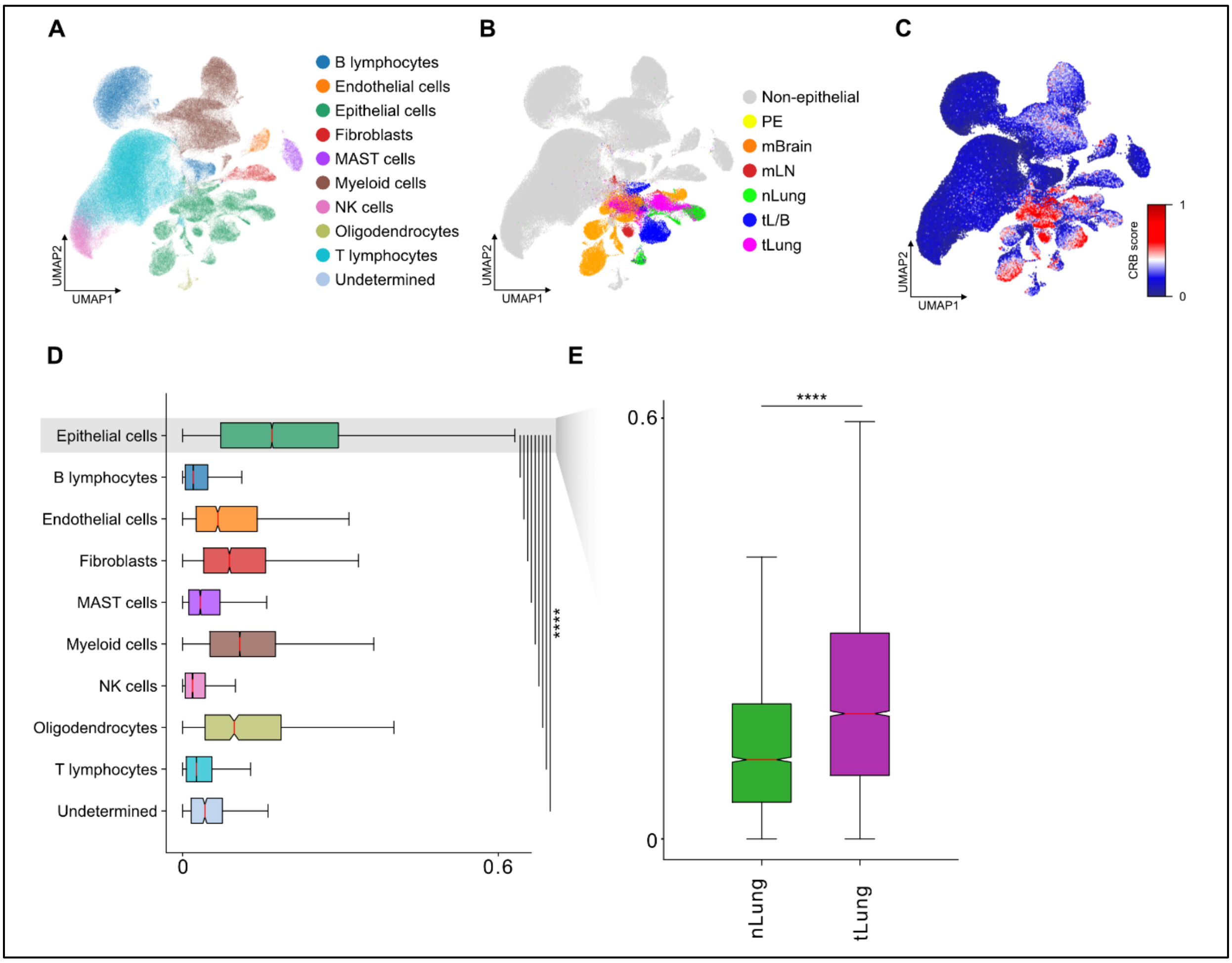


Figure 43 | CRB-score is enriched in diverse types of cancer. (A) UMAP visualization reporting the major cell populations identified in the analysis of 44 patients affected by LUAD. **(B)** UMAP visualization showing non-epithelial cells (in grey) and epithelial cells coloured according to tissues of origin: Cancer tissue-derived whole cells from primary sites (tLung and tL/B), pleural fluids (PE), lymph node (mLN), and brain metastases (mBrain), as well as normal tissues from lungs (nLung) **(C)** UMAP representation of the epigenetic score **(D)** Boxplot reporting the distribution of the CRB score in all cell populations identified. Epithelial cells are reported in dark green. Boxplots describe the median and interquartile range with whiskers denoting the 1.5 x interquartile range. **(E)** Boxplot reporting the distribution of the CRB score in normal (nLung, in dark-green) and tumor (tLung, in purple) epithelial cells.

5. Discussion and conclusion

Our study provides the epigenetic landscape of human CRC, unveiling the existence of an aberrant pan-CRC enhancerome core regulated by the transcriptional coactivators YAP/TAZ and active in more than 20 types of human malignancies. It is increasingly recognized that the deregulated epigenome is a universal feature of cancer, challenging the previously prevailing paradigm that cancer is a genetic disease[5,6]. Although mutations in epigenetic regulators are widespread among different cancer types[117,118], the transcriptional and epigenetic changes that occur in tumors cannot be attributed solely to the spectrum of oncogenic mutations[120]. In fact, the dysfunctional epigenome in cancer is often preceded by a continuum of epigenetic alterations in healthy tissues as a direct consequence of age and tissue damage[203,204]. The current surge of interest in epigenomics is further fueled by the failures in translating genetic findings into novel therapeutics that work for the majority of patients. In this context, can tumor deconvolution at the epigenomic level lead to insights that may drive more effective cancer therapies? Motivated by these outstanding issues of key relevance to cancer biology, we sought to decipher the epigenetic landscape of human CRC using a heterogeneous library of patient-derived tumor organoids.

Previous efforts to identify CRC-specific enhancers have focused on characterizing chromatin accessibility using ATAC-seq data[116] or single epigenetic features predictive of specific functional elements[126,205,206]. Going beyond previous epigenetic analyses, our work employs a systematic *de novo* strategy to discover biologically-informative chromatin states, providing a global overview of all functional genomic elements[207]. By combining more than 60 chromatin maps we identified 8 different chromatin states, providing a comprehensive set of genome-wide regulatory regions including promoter and enhancer, as well as elongating and repressed genomic regions. This data generates an extensive functional annotation of the human genome in CRC allowing the interrogation of diverse modes of epigenetic regulation, including that of repressed promoters. Our dataset is composed by a large set of ChIP-seq and RNA-seq data from normal adjacent tissue and tumor samples deriving from the same patients, providing a comprehensive view of correlated activity patterns in human CRC and an essential resource for exploring not only the specific epigenetic programs that drive patient heterogeneity but also common epigenetic blueprints. To facilitate the use of these data we created a web application tool (available at http://hepic.homic.eu) that offers users the opportunity to browse the genome-wide maps of chromatin states and individual histone marks across the heterogeneous library of CRC PDOs.

In the last few years, there was an increasing interest in the study of enhancers and their fundamental role in tumor development, progression and metastasis[122,123]. To this extent, we exploited a machine learning approach to provide a more robust characterization of different classes of enhancers compared to regulatory elements predicted by individual histone marks[174,186,208]. The combinatorial pattern of histone modifications allowed to precisely discriminate active enhancers from other genomic elements within chromatin accessible regions. For instance, the epigenetic profile of TAZ gene locus (**Figure 36 A**), clearly shows the presence of multiple open chromatin regions across diverse tumor types corresponding to a small set of *de novo* active enhancers states.

Human CRCs are characterized by ostensibly endless combinations of oncogenic lesions resulting in a high degree of intratumoral and intertumoral genetic heterogeneity[13]. Is the epigenetic level similarly complex or, rather, does it represent a much-simplified layer of integration of genetic and microenvironmental inputs into a restricted, shared set of transcriptional states? Based on ChromHMM-defined chromatin states, we found two main groups of enhancers that are differentially active in PDOs compared to normal mucosa. While half of these enhancers displayed low levels of conservation across PDOs, the remaining half was conserved in at least 50% of the tumor organoids, including those displaying microsatellite instability (**Figure 26 C**). This comes in striking contrast with the reported recurrence of mutated genes in CRC; with the exception of few driver genes the vast majority of the recurrently mutated genes are shared by less than 10% of tumors. Thus, our findings indicate that despite the profound genetic heterogeneity, CRC is characterized by a common aberrant enhancerome.

To provide an in-depth characterization of the shared CRC enhancerome, we sought to understand which transcription factors orchestrate the activation of these *cis*-regulatory elements. Motif discovery and functional enrichment analyses highlighted AP1 and TEAD families along with the Hippo pathway, pinpointing the YAP/TAZ transcriptional coactivators as major regulators of the human CRC enhancerome. YAP/TAZ are induced in the majority of solid tumors[209], triggering several hallmarks of cancer such as proliferation, phenotypic plasticity, drug resistance and metastasis (**Figure 44 A**)[193]. These functions are exerted in the nucleus (**Figure 44 B**) through interaction with other DNA-binding partners (primary TEAD) and occupancy of distal cis-regulatory elements, that are in touch with their cognate promoters through chromatin loops[98,192,210,211]. AP-1 family motifs are often found in close proximity to TEAD elements (**Figure 44 C**), enabling the formation of a YAP/TAZ/TEAD/AP1 complex that is able to transcriptionally cooperate to regulate tumor cell proliferation and motility[192,210,212,213]. One of the most known mechanism of YAP/TAZ regulation is via Hippo signalling pathway[197]. However, it is becoming increasingly evident that YAP/TAZ activity is not only regulated by the Hippo cascade, but is also involved in a plethora of mechanisms, including cell-cell adhesions, epithelial cell polarity, microenvironment stiffness, metabolic pathways and extracellular growth factors (**Figure 44 D**)[197,214].

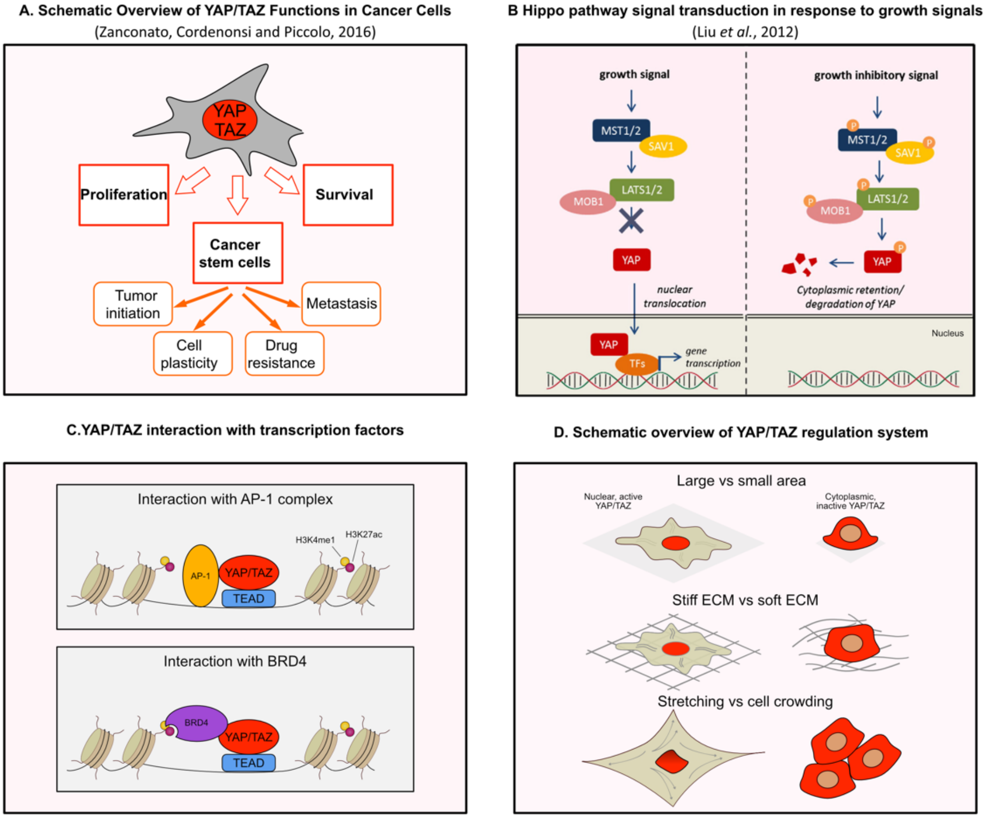

A. Schematic Overview of YAP/TAZ Functions in Cancer Cells
(Zanconato, Cordenonsi and Piccolo, 2016)
YAP
TAZ
Proliferation
Survival
Cancer stem cells
Tumor initiation
Metastasis
Cell plasticity
Drug resistance
B Hippo pathway signal transduction in response to growth signals
(Liu et al., 2012)
growth signal
MST1/2
SAV1
MOB1
LATS1/2
YAP
nuclear translocation
YAP
TFs
gene transcription
growth inhibitory signal
P
MST1/2
SAV1
MOB1
LATS1/2
YAP
Cytoplasmic retention/ degradation of YAP
Nucleus
C.YAP/TAZ interaction with transcription factors
Interaction with AP-1 complex
H3K4me1
H3K27ac
AP-1
YAP/TAZ
TEAD
Interaction with BRD4
BRD4
YAP/TAZ
TEAD
D. Schematic overview of YAP/TAZ regulation system
Large vs small area
Nuclear, active YAP/TAZ
Cytoplasmic, inactive YAP/TAZ
Stiff ECM vs soft ECM
Stretching vs cell crowding

Figure 44 | (A) Schematic representation of YAP/TAZ functions in tumors (from Zanconato *et al.*[193]). **(B)** Hippo pathway signal cascade with all the factors involved in this pathway (from Liu *et al.* [215]). **(C)** Example of two of the most studied co-factors of YAP/TAZ **(D)** Overview of the new mechanism emerging in YAP/TAZ regulation.

In line with their pervasive activation in human epithelial tumors[193], we confirmed the transcriptional upregulation and nuclear translocation of YAP/TAZ, and further unveiled the YAP/TAZ chromatin recruitment at distal enhancers in human CRC. In addition, we found an enrichment of TEAD and AP1 motifs at YAP/TAZ-bound genomic elements, supporting the involvement of these TFs as YAP/TAZ partners also in this type of cancer, as suggested by previous studies[192,210,212,213].

Strikingly, YAP/TAZ were most enriched in the highly conserved gained enhancers (**Figure 31 C**), highlighting these transcription factors as driving forces of the CRC deregulated enhancerome. The relevance of this epigenetic signature was extended to diverse malignancies of epithelial cells, suggesting a previously undescribed universal role for YAP/TAZ as master regulators of tumor-associated epigenetic shifts. In the light of the recently reported YAP/TAZ-dependent transcriptional addiction in cancer[98], we speculate that the core of 46 pan-cancer gained enhancers identified in our study could be at the roots of the cancer transcriptional addiction.

Therefore, they might represent a unique epigenetic “fil rouge” that can be exploited for potential therapeutic targets. It remains to be evaluated whether a subset of these highly conserved gained enhancers is active in metastatic tumors and/or premalignant lesions.

Collectively, different aspects of our study design have enabled the identification of the pan-cancer YAP/TAZ-driven enhancerome. First, we interrogated chromatin states that define active enhancers with high resolution, as opposed to open chromatin regions that are indicative of diverse active regulatory elements. Second, by exploiting primary tissue-derived tumor organoids we were able to dissect the cancer cell-intrinsic epigenetic alterations devoid of the influence of stromal cells[31,32,216] and reveal that this active enhancerome is shared amongst diverse cancer types. Finally, organoid cultures preserve the mechanical forces and stress of a 3D cellular architecture and organization, which is essential for maintaining YAP/TAZ activation in *ex vivo* models[217,218].

Recent scRNA-seq studies have revealed the complex transcriptional mosaic of tumors, characterized by a continuum of differentiation and the presence of diverse transcriptional states among cancer cell[219,220]. This heterogeneity is only partially determined by genetic events indicating that epigenetic programs may influence the emergence of cancer cell states. In line with these notions, we show that the YAP/TAZ-orchestrated deregulation, which is largely absent in normal cells, specifically underlies the cellular states of tumor cells regardless of their clonality or functional status. Moreover, we confirmed our findings considering two scRNA-seq datasets on primary CRC[170] and LUAD[171] demonstrating that the identified cancer regulatory blueprint is a specific feature of malignant cells across diverse types of cancer.

Still, there are unaddressed questions regarding the regulation of YAP/TAZ. Depending on the molecularly distinct checkpoints to overcome, different tumor types can exploit intrinsic (oncogenic lesions, Hippo pathway alterations) or biomechanical cues to activate YAP/TAZ[193]. In addition, YAP/TAZ can employ self-sustaining positive loops to maintain their functions[212,221]. Interestingly, we found a tumor-specific epigenetic mechanism of *TAZ* regulation: a positive feedback loop between an intronic YAP/TAZ-bound active enhancer and TAZ promoter itself, shared by all CRC PDOs. This intronic enhancer was also observed in the TCGA panel of 23 tumor types, suggesting that the TAZ self-regulation is relevant to a wide range of cancers (**Figure 36**). This transcriptional feedback mechanism combined with the inhibition and persistent activation of the Hippo and Wnt pathways, respectively[222], may provide a constant fuel of YAP/TAZ for uncontrolled proliferation of tumor cells, sustaining the recently reported YAP/TAZ-dependent transcriptional addiction in cancer[98].

The accumulative evidence pinpointing YAP/TAZ as key mediators of epigenetic reprogramming and transcriptional addiction in cancer, makes them appealing targets for therapeutic intervention. This is corroborated by the association of YAP/TAZ activation with poor prognosis in human epithelial tumors[193], and resistance to both chemotherapeutic drugs[155,223] and molecularly targeted therapies[224,225]. Deeper understanding of the mechanisms by which YAP/TAZ exert their nuclear function can link the inhibition of these coactivators to an arsenal of potent epigenetic agents. For instance, YAP/TAZ-mediated transcriptional addiction is achieved through interaction with the bromodomain and extraterminal domain (BET) coactivator BRD4, demonstrating the rational use of BET inhibitors in impairing expression of YAP/TAZ-regulated genes and YAP/TAZ-induced oncogenic functions and drug resistance. Nuclear inhibitors of YAP/TAZ that act as competitors for TEAD binding have also been described as a valid strategy to constrain YAP/TAZ functions[193]. A different approach to control YAP/TAZ activity could arise from the epigenetic modulation of TAZ gene expression. The dispensability of YAP/TAZ for normal tissue homeostasis[194,196] provides a further argument in favor of exploiting YAP/TAZ as master regulator of pan-cancer enhancerome to design therapeutic strategies that are of clinical relevance to a significant number of patients and in cancer pathologies beyond CRC.

As previously discussed, organoids provide a powerful and versatile tool for the development of new therapeutic approaches targeting not only the cancer-intrinsic features but also the tumor microenvironment. Indeed, the long-term goals arising from this project are i) the validation of promoter-enhancer interaction using capture-HiC technique[226,227] on different PDOs to confirm the correct target genes regulated by the shared core of YAP/TAZ bound enhancers, ii) the epigenetic editing of the established promoter-enhancer interactions by CRISPR epigenetic editing[228–230] (*e.g.,* the generation of a specific dCas9-KRAB repressor complex) in order to shut down the function of single or multiple regulatory elements and evaluate how these perturbations affect the expression of their target genes and the functional state in malignant cells, and iii) the generation of a co-culture system[69–71,231] between tumor infiltrating (TI) CD4$^+$ Treg and PDOs to study the interplay and the mutual effects of TI-Treg and malignant cells. Together, these approaches provide the opportunity to untangle the complex regulatory network orchestrating human malignancy leading toward a new era of more precise and efficient therapeutic strategies.

6. Bibliography

1. Hassanpour SH, Dehghani M. Review of cancer from perspective of molecular. *J Cancer Res Pract*. Published online 2017. doi:10.1016/j.jcrpr.2017.07.001

2. Takeshima H, Ushijima T. Accumulation of genetic and epigenetic alterations in normal cells and cancer risk. *npj Precis Oncol*. 2019;3(1):1-8. doi:10.1038/s41698-019-0079-0

3. Liu J, Dang H, Wang XW. The significance of intertumor and intratumor heterogeneity in liver cancer. *Nat Publ Gr*. 2018;50:416. doi:10.1038/emm.2017.165

4. Meacham CE, Morrison SJ. Tumour heterogeneity and cancer cell plasticity. *Nature*. 2013;501(7467):328-337. doi:10.1038/nature12624

5. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell*. 2000;100(1):57-70. doi:10.1016/S0092-8674(00)81683-9

6. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. *Cell*. 2011;144(5):646-674. doi:10.1016/j.cell.2011.02.013

7. Lazebnik Y. What are the hallmarks of cancer? *Nat Rev Cancer*. 2010;10(4):232-233. doi:10.1038/nrc2827

8. Hanahan D, Robert A W. Biological Hallmarks of cancer. *Holland-Frei cancer Med*. 2017;01(April):1-10. doi:10.1002/9781119000822.hfcm002

9. Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB. Colorectal cancer. *Lancet*. 2019;394(10207):1467-1480. doi:10.1016/S0140-6736(19)32319-0

10. Al-Tassan N, Chmiel NH, Maynard J, et al. Inherited variants of MYH associated with somatic G:C→T:A mutations in colorectal tumors. *Nat Genet*. 2002;30(2):227-232. doi:10.1038/ng828

11. Kashfi SMH, Golmohammadi M, Behboudi F, Nazemalhosseini-Mojarad E, Zali MR. MUTYH the base excision repair gene family member associated with polyposis colorectal cancer. *Gastroenterol Hepatol from Bed to Bench*. Published online 2013. doi:10.22037/ghfbb.v6i0.458

12. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell*. 1990;61(5):759-767. doi:10.1016/0092-8674(90)90186-I

13. Muzny DM, Bainbridge MN, Chang K, et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. Published online 2012. doi:10.1038/nature11252

14. Fearon ER. Molecular genetics of colorectal cancer. *Annu Rev Pathol Mech Dis*. Published online 2011. doi:10.1146/annurev-pathol-011110-130235

15. Walther A, Johnstone E, Swanton C, Midgley R, Tomlinson I, Kerr D. Genetic prognostic and predictive markers in colorectal cancer. *Nat Rev Cancer*. Published online 2009. doi:10.1038/nrc2645

16. Blank A, Roberts DE, Dawson H, Zlobec I, Lugli A. Tumor heterogeneity in primary colorectal cancer and corresponding metastases. Does the apple fall far from the tree? *Front Med*. 2018;5(AUG). doi:10.3389/fmed.2018.00234

17. Califano A, Alvarez MJ. The recurrent architecture of tumour initiation, progression and drug sensitivity. *Nat Rev Cancer*. Published online 2017. doi:10.1038/nrc.2016.124

18. Ma S, Ogino S, Parsana P, et al. Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis. *Genome Biol*. Published online 2018. doi:10.1186/s13059-018-1511-4

19. Wang W, Kandimalla R, Huang H, et al. Molecular subtyping of colorectal cancer: Recent progress, new challenges and emerging opportunities. *Semin Cancer Biol*. 2019;55:37-52. doi:10.1016/j.semcancer.2018.05.002

20. Nitsche U, Zimmermann A, Späth C, et al. Mucinous and Signet-Ring Cell Colorectal Cancers Differ from Classical Adenocarcinomas in Tumor Biology and Prognosis. *Ann Surg*. 2013;258(5):775-783. doi:10.1097/SLA.0b013e3182a69f7e

21. Tong GJ, Zhang GY, Liu J, et al. Comparison of the eighth version of the American joint committee on cancer manual to the seventh version for colorectal cancer: A retrospective review of our data. *World J Clin Oncol*. Published online 2018. doi:10.5306/wjco.v9.i7.148

22. Amin MB, Greene FL, Edge SB, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin*. Published online 2017. doi:10.3322/caac.21388

23. Loeb LA. Cancer cells exhibit a mutator phenotype. *Adv Cancer Res*. 1997;72:25-56. doi:10.1016/s0065-230x(08)60699-5

24. Ho AS, Turcan S, Chan TA. Epigenetic therapy: Use of agents targeting deacetylation and methylation in cancer management. *Onco Targets Ther*. 2013;6:223-232. doi:10.2147/OTT.S34680

25. Arvelo F, Sojo F, Cotte C. Biology of colorectal cancer. *Ecancermedicalscience*. 2015;9.

doi:10.3332/ecancer.2015.520

26. Sadanandam A, Lyssiotis CA, Homicsko K, et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat Med*. 2013;19(5):619-625. doi:10.1038/nm.3175

27. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. *Nat Med*. 2015;21(11):1350-1356. doi:10.1038/nm.3967

28. Linnekamp JF, Van Hooff SR, Prasetyanti PR, et al. Consensus molecular subtypes of colorectal cancer are recapitulated in in vitro and in vivo models. *Cell Death Differ*. Published online 2018. doi:10.1038/s41418-017-0011-5

29. Wang J, Mouradov D, Wang X, et al. Colorectal Cancer Cell Line Proteomes Are Representative of Primary Tumors and Predict Drug Sensitivity. *Gastroenterology*. Published online 2017. doi:10.1053/j.gastro.2017.06.008

30. Schütte M, Risch T, Abdavi-Azar N, et al. Molecular dissection of colorectal cancer in pre-clinical models identifies biomarkers predicting sensitivity to EGFR inhibitors. *Nat Commun*. Published online 2017. doi:10.1038/ncomms14262

31. Isella C, Brundu F, Bellomo SE, et al. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. *Nat Commun*. 2017;8:15107. doi:10.1038/ncomms15107

32. Calon A, Lonardo E, Berenguer-Llergo A, et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat Genet*. Published online 2015. doi:10.1038/ng.3225

33. Fujii M, Shimokawa M, Date S, et al. A Colorectal Tumor Organoid Library Demonstrates Progressive Loss of Niche Factor Requirements during Tumorigenesis. *Cell Stem Cell*. Published online 2016. doi:10.1016/j.stem.2016.04.003

34. Gonçalves E, Segura-Cabrera A, Pacini C, et al. Drug mechanism-of-action discovery through the integration of pharmacological and CRISPR screens. *Mol Syst Biol*. Published online 2020. doi:10.1101/2020.01.14.905729

35. Iorio F, Knijnenburg TA, Vis DJ, et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*. Published online 2016. doi:10.1016/j.cell.2016.06.017

36. Raju KL, Augustine D, Rao RS, et al. Biomarkers in tumorigenesis using cancer cell lines: A systematic review. *Asian Pacific J Cancer Prev*. Published online 2017. doi:10.22034/APJCP.2017.18.9.2329

37. Torsvik A, Stieber D, Enger PO, et al. U-251 revisited: Genetic drift and phenotypic consequences of long-term cultures of glioblastoma cells. *Cancer Med*. Published online 2014. doi:10.1002/cam4.219

38. John T, Kohler D, Pintilie M, et al. The ability to form primary tumor xenografts is predictive of increased risk of disease recurrence in early-stage non-small cell lung cancer. *Clin Cancer Res*. Published online 2011. doi:10.1158/1078-0432.CCR-10-2224

39. Fan H, Demirci U, Chen P. Emerging organoid models: Leaping forward in cancer research. *J Hematol Oncol*. Published online 2019. doi:10.1186/s13045-019-0832-4

40. Schutgens F, Clevers H. Human Organoids: Tools for Understanding Biology and Treating Diseases. Published online 2019. doi:10.1146/annurev-pathmechdis

41. SMITH E, COCHRANE WJ. Cystic organoid teratoma; report of a case. *Can Med Assoc J*. Published online 1946.

42. Dutta D, Heo I, Clevers H. Disease Modeling in Stem Cell-Derived 3D Organoid Systems. *Trends Mol Med*. Published online 2017. doi:10.1016/j.molmed.2017.02.007

43. Crespo M, Vilar E, Tsai SY, et al. Colonic organoids derived from human induced pluripotent stem cells for modeling colorectal cancer and drug testing. *Nat Med*. Published online 2017. doi:10.1038/nm.4355

44. Greggio C, De Franceschi F, Figueiredo-Larsen M, et al. Artificial three-dimensional niches deconstruct pancreas development in vitro. *Dev*. Published online 2013. doi:10.1242/dev.096628

45. McCracken KW, Catá EM, Crawford CM, et al. Modelling human development and disease in pluripotent stem-cell-derived gastric organoids. *Nature*. Published online 2014. doi:10.1038/nature13863

46. Sato T, Stange DE, Ferrante M, et al. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology*. 2011;141(5):1762-1772. doi:10.1053/j.gastro.2011.07.050

47. Wang K, Yuen ST, Xu J, et al. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat Genet*. Published online 2014. doi:10.1038/ng.2983

48. Huch M, Dorrell C, Boj SF, et al. In vitro expansion of single Lgr5 + liver stem cells induced by Wnt-driven regeneration. *Nature*. Published online 2013. doi:10.1038/nature11826

49. Huch M, Gehart H, Van Boxtel R, et al. Long-term culture of genome-stable bipotent stem cells from adult human liver. *Cell*. Published online 2015. doi:10.1016/j.cell.2014.11.050

50. Huch M, Bonfanti P, Boj SF, et al. Unlimited in vitro expansion of adult bi-potent pancreas progenitors through the Lgr5/R-spondin axis. *EMBO J*. Published online 2013. doi:10.1038/emboj.2013.204

51. Karthaus WR, Iaquinta PJ, Drost J, et al. Identification of multipotent luminal progenitor cells in human prostate organoid cultures. *Cell*. Published online 2014. doi:10.1016/j.cell.2014.08.017

52. McCauley HA, Wells JM. Pluripotent stem cell-derived organoids: Using principles of developmental biology to grow human tissues in a dish. *Dev*. Published online 2017. doi:10.1242/dev.140731

53. Matano M, Date S, Shimokawa M, et al. Modeling colorectal cancer using CRISPR-Cas9-mediated engineering of human intestinal organoids. *Nat Med*. Published online 2015. doi:10.1038/nm.3802

54. Nakayama M, Sakai E, Echizen K, et al. Intestinal cancer progression by mutant p53 through the acquisition of invasiveness associated with complex glandular formation. *Oncogene*. Published online 2017. doi:10.1038/onc.2017.194

55. Schell MJ, Yang M, Teer JK, et al. A multigene mutation classification of 468 colorectal cancers reveals a prognostic role for APC. *Nat Commun*. Published online 2016. doi:10.1038/ncomms11743

56. Cheung KJ, Gabrielson E, Werb Z, Ewald AJ. Collective invasion in breast cancer requires a conserved basal epithelial program. *Cell*. Published online 2013. doi:10.1016/j.cell.2013.11.029

57. Wu JS, Li ZF, Wang HF, et al. Cathepsin B defines leader cells during the collective invasion of salivary adenoid cystic carcinoma. *Int J Oncol*. Published online 2019. doi:10.3892/ijo.2019.4722

58. Libanje F, Raingeaud J, Luan R, et al. ROCK 2 inhibition triggers the collective invasion of colorectal adenocarcinomas . *EMBO J*. Published online 2019. doi:10.15252/embj.201899299

59. Björk JK, Åkerfelt M, Joutsen J, et al. Heat-shock factor 2 is a suppressor of prostate cancer invasion. *Oncogene*. Published online 2016. doi:10.1038/onc.2015.241

60. Vellinga TT, Den Uil S, Rinkes IHB, et al. Collagen-rich stroma in aggressive colon tumors induces mesenchymal gene expression and tumor cell invasion. *Oncogene*. Published online 2016. doi:10.1038/onc.2016.60

61. Gao H, Chakraborty G, Zhang Z, et al. Multi-organ Site Metastatic Reactivation Mediated by Non-canonical Discoidin Domain Receptor 1 Signaling. *Cell*. Published online 2016. doi:10.1016/j.cell.2016.06.009

62. van de Wetering M, Francies HE, Francis JM, et al. Prospective Derivation of a Living Organoid Biobank of Colorectal Cancer Patients. *Cell*. 2015;161(4):933-945. doi:10.1016/j.cell.2015.03.053

63. Nuciforo S, Fofana I, Matter MS, et al. Organoid Models of Human Liver Cancers Derived from Tumor Needle Biopsies. *Cell Rep*. Published online 2018. doi:10.1016/j.celrep.2018.07.001

64. Sachs N, de Ligt J, Kopper O, et al. A Living Biobank of Breast Cancer Organoids Captures Disease Heterogeneity. *Cell*. Published online 2018. doi:10.1016/j.cell.2017.11.010

65. Saito Y, Muramatsu T, Kanai Y, et al. Establishment of Patient-Derived Organoids and Drug Screening for Biliary Tract Carcinoma. *Cell Rep*. Published online 2019. doi:10.1016/j.celrep.2019.03.088

66. Yao Y, Xu X, Yang L, et al. Patient-Derived Organoids Predict Chemoradiation Responses of Locally Advanced Rectal Cancer. *Cell Stem Cell*. Published online 2020. doi:10.1016/j.stem.2019.10.010

67. Nakamura H, Sugano M, Miyashita T, et al. Organoid culture containing cancer cells and stromal cells reveals that podoplanin-positive cancer-associated fibroblasts enhance proliferation of lung cancer cells. *Lung Cancer*. Published online 2019. doi:10.1016/j.lungcan.2019.04.007

68. Tsai S, McOlash L, Palen K, et al. Development of primary human pancreatic cancer organoids, matched stromal and immune cells and 3D tumor microenvironment models. *BMC Cancer*. Published online 2018. doi:10.1186/s12885-018-4238-4

69. Dijkstra KK, Cattaneo CM, Weeber F, et al. Generation of Tumor-Reactive T Cells by Co-culture of Peripheral Blood Lymphocytes and Tumor Organoids. *Cell*. Published online 2018. doi:10.1016/j.cell.2018.07.009

70. Finnberg NK, Gokare P, Lev A, et al. Application of 3D tumoroid systems to define

immune and cytotoxic therapeutic responses based on tumoroid and tissue slice culture molecular signatures. *Oncotarget*. Published online 2017. doi:10.18632/oncotarget.19965

71. Neal JT, Li X, Zhu J, et al. Organoid Modeling of the Tumor Immune Microenvironment. *Cell*. Published online 2018. doi:10.1016/j.cell.2018.11.021

72. Vlachogiannis G, Hedayat S, Vatsiou A, et al. Patient-derived organoids model treatment response of metastatic gastrointestinal cancers. *Science (80-)*. Published online 2018. doi:10.1126/science.aao2774

73. Gehart H, Clevers H. Tales from the crypt: new insights into intestinal stem cells. *Nat Rev Gastroenterol Hepatol*. Published online 2019. doi:10.1038/s41575-018-0081-y

74. Peterson LW, Artis D. Intestinal epithelial cells: Regulators of barrier function and immune homeostasis. *Nat Rev Immunol*. Published online 2014. doi:10.1038/nri3608

75. Allaire JM, Crowley SM, Law HT, Chang SY, Ko HJ, Vallance BA. The Intestinal Epithelium: Central Coordinator of Mucosal Immunity. *Trends Immunol*. Published online 2018. doi:10.1016/j.it.2018.04.002

76. McDole JR, Wheeler LW, McDonald KG, et al. Goblet cells deliver luminal antigen to CD103 + dendritic cells in the small intestine. *Nature*. Published online 2012. doi:10.1038/nature10863

77. Rothenberg ME, Nusse Y, Kalisky T, et al. Identification of a cKit+ colonic crypt base secretory cell that supports Lgr5+ stem cells in mice. *Gastroenterology*. 2012;142(5):1195-1205.e6. doi:10.1053/j.gastro.2012.02.006

78. Sasaki N, Sachs N, Wiebrands K, et al. Reg4+ deep crypt secretory cells function as epithelial niche for Lgr5+ stem cells in colon. *Proc Natl Acad Sci U S A*. Published online 2016. doi:10.1073/pnas.1607327113

79. Gribble FM, Reimann F. Enteroendocrine Cells: Chemosensors in the Intestinal Epithelium. *Annu Rev Physiol*. Published online 2016. doi:10.1146/annurev-physiol-021115-105439

80. Haber AL, Biton M, Rogel N, et al. A single-cell survey of the small intestinal epithelium. *Nature*. Published online 2017. doi:10.1038/nature24489

81. Sato T, Vries RG, Snippert HJ, et al. Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature*. Published online 2009. doi:10.1038/nature07935

82. Korinek V, Barker N, Moerer P, et al. Depletion of epithelial stem-cell compartments in the small intestine of mice lacking Tcf-4. *Nat Genet*. Published online 1998. doi:10.1038/1270

83. Barker N, Van Es JH, Kuipers J, et al. Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature*. Published online 2007. doi:10.1038/nature06196

84. Kim KA, Kakitani M, Zhao J, et al. Medicine: Mitogenic influence of human R-spondin1 on the intestinal epithelium. *Science (80-)*. Published online 2005. doi:10.1126/science.1112521

85. Ohta Y, Sato T. Intestinal tumor in a dish. *Front Med*. Published online 2014. doi:10.3389/fmed.2014.00014

86. Holmberg FE, Seidelin JB, Yin X, et al. Culturing human intestinal stem cells for regenerative applications in the treatment of inflammatory bowel disease. *EMBO Mol Med*. Published online 2017. doi:10.15252/emmm.201607260

87. Reynolds A, Wharton N, Parris A, et al. Canonical Wnt signals combined with suppressed TGFβ/BMP pathways promote renewal of the native human colonic epithelium. *Gut*. Published online 2014. doi:10.1136/gutjnl-2012-304067

88. Sato T, Van Es JH, Snippert HJ, et al. Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature*. Published online 2011. doi:10.1038/nature09637

89. Hong SN, Dunn JCY, Stelzner M, Martín MG. Concise Review: The Potential Use of Intestinal Stem Cells to Treat Patients with Intestinal Failure. *Stem Cells Transl Med*. Published online 2017. doi:10.5966/sctm.2016-0153

90. WADDINGTON CH. Canalization of Development and the Inheritance of Acquired Characters. *Nature*. 1942;150(3811):563-565. doi:10.1038/150563a0

91. Goldberg AD, Allis CD, Bernstein E. Epigenetics: A Landscape Takes Shape. *Cell*. 2007;128(4):635-638. doi:10.1016/j.cell.2007.02.006

92. Flavahan WA, Gaskell E, Bernstein BE. Epigenetic plasticity and the hallmarks of cancer. doi:10.1126/science.aal2380

93. Cutter AR, Hayes JJ. A brief review of nucleosome structure. *FEBS Lett*. 2015;589(20PartA):2914-2922. doi:10.1016/j.febslet.2015.05.016

94. Morales V, Richard-foy N. *Role of Histone N-Terminal Tails and Their Acetylation in Nucleosome Dynamics*. Vol 20.; 2000.

95. Biswas S, Rao CM. Epigenetic tools (The Writers, The Readers and The Erasers) and their implications in cancer therapy. *Eur J Pharmacol*. 2018;837:8-24. doi:10.1016/j.ejphar.2018.08.021

96. Barnes CE, English DM, Cowley SM. Acetylation and Co: An expanding repertoire of histone acylations regulates chromatin and transcription. *Essays Biochem*. 2019;63(1):97-107. doi:10.1042/EBC20180061

97. Sun XJ, Man N, Tan Y, Nimer SD, Wang L. The role of histone acetyltransferases in normal and malignant hematopoiesis. *Front Oncol*. 2015;5(MAY). doi:10.3389/fonc.2015.00108

98. Zanconato F, Battilana G, Forcato M, et al. Transcriptional addiction in cancer cells is mediated by YAP/TAZ through BRD4. *Nat Med*. 2018;24(10):1599-1610. doi:10.1038/s41591-018-0158-8

99. Alqahtani A, Choucair K, Ashraf M, et al. Bromodomain and extra-terminal motif inhibitors: A review of preclinical and clinical advances in cancer therapy. *Futur Sci OA*. 2019;5(3). doi:10.4155/fsoa-2018-0115

100. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. *Cell Res*. 2011;21(3):381-395. doi:10.1038/cr.2011.22

101. Michalak EM, Burr ML, Bannister AJ, Dawson MA. The roles of DNA, RNA and histone methylation in ageing and cancer. *Nat Rev Mol Cell Biol*. 2019;20(10):573-589. doi:10.1038/s41580-019-0143-1

102. Hyun K, Jeon J, Park K, Kim J. Writing, erasing and reading histone lysine methylations. *Exp Mol Med*. 2017;49(4):e324-e324. doi:10.1038/emm.2017.11

103. Zhang T, Cooper S, Brockdorff N. The interplay of histone modifications - writers that read. *EMBO Rep*. 2015;16(11):1467-1481. doi:10.15252/embr.201540945

104. Ordoñez R, Martínez-Calle N, Agirre X, Prosper F. DNA methylation of enhancer elements in myeloid neoplasms: Think outside the promoters? *Cancers (Basel)*. 2019;11(10):1424. doi:10.3390/cancers11101424

105. Li J, Duns G, Westers H, Sijmons R, van den Berg A, Kok K. SETD2: an epigenetic modifier with tumor suppressor functionality. *Oncotarget*. 2016;7(31):50719-50734. doi:10.18632/oncotarget.9368

106. Laugesen A, Højfeldt JW, Helin K. Molecular Mechanisms Directing PRC2 Recruitment and H3K27 Methylation. *Mol Cell*. 2019;74(1):8-18. doi:10.1016/j.molcel.2019.03.011

107. Yadon AN, Van de Mark D, Basom R, Delrow J, Whitehouse I, Tsukiyama T. Chromatin Remodeling around Nucleosome-Free Regions Leads to Repression of Noncoding RNA Transcription. *Mol Cell Biol*. 2010;30(21):5110-5122. doi:10.1128/mcb.00602-10

108. Maehara K, Ohkawa Y. Exploration of nucleosome positioning patterns in transcription factor function. *Sci Rep*. 2016;6(1):1-11. doi:10.1038/srep19620

109. Gasperini M, Tome JM, Shendure J. Towards a comprehensive catalogue of validated and target-linked human enhancers. doi:10.1038/s41576-019-0209-0

110. Kempfer R, Pombo A. Methods for mapping 3D chromosome architecture. *Nat Rev Genet*. Published online December 17, 2019. doi:10.1038/s41576-019-0195-2

111. Schoenfelder S, Fraser P. Long-range enhancer–promoter contacts in gene expression control. *Nat Rev Genet*. 2019;20(8):437-455. doi:10.1038/s41576-019-0128-0

112. ENCODE Project Consortium EP, Encode Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74. doi:10.1038/nature11247

113. Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al. The NIH roadmap epigenomics mapping consortium. *Nat Biotechnol*. 2010;28(10):1045-1048. doi:10.1038/nbt1010-1045

114. Chadwick LH. The NIH Roadmap Epigenomics Program data resource. *Epigenomics*. 2012;4(3):317-324. doi:10.2217/epi.12.18

115. Martens JHA, Stunnenberg HG. BLUEPRINT: Mapping human blood cell epigenomes. *Haematologica*. 2013;98(10):1487-1489. doi:10.3324/haematol.2013.094243

116. Corces MR, Granja JM, Shams S, et al. The chromatin accessibility landscape of primary human cancers. *Science (80-)*. 2018;362(6413):eaav1898. doi:10.1126/science.aav1898

117. Shen H, Laird PW. Interplay between the cancer genome and epigenome. *Cell*. Published online 2013. doi:10.1016/j.cell.2013.03.008

118. You JS, Jones PA. Cancer Genetics and Epigenetics: Two Sides of the Same Coin? *Cancer Cell*. 2012;22(1):9-20. doi:10.1016/j.ccr.2012.06.008

119. Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol*. Published online 2015. doi:10.1038/nrm3949

120. Kron KJ, Bailey SD, Lupien M. Enhancer alterations in cancer: A source for a cell identity crisis. *Genome Med*. Published online 2014. doi:10.1186/s13073-014-0077-3

121. Roe L, Normand C, Wren MA, Browne J, O'Halloran AM. The impact of frailty on

healthcare utilisation in Ireland: Evidence from the Irish longitudinal study on ageing. *BMC Geriatr*. Published online 2017. doi:10.1186/s12877-017-0579-0

122. Sur I, Taipale J. The role of enhancers in cancer. *Nat Rev Cancer*. Published online 2016. doi:10.1038/nrc.2016.62

123. Roe JS, Hwang C Il, Somerville TDD, et al. Enhancer Reprogramming Promotes Pancreatic Cancer Metastasis. *Cell*. Published online 2017. doi:10.1016/j.cell.2017.07.007

124. Lin CY, Erkek S, Tong Y, et al. Active medulloblastoma enhancers reveal subgroup-specific cellular origins. *Nature*. Published online 2016. doi:10.1038/nature16546

125. Patten DK, Corleone G, Győrffy B, et al. Enhancer mapping uncovers phenotypic heterogeneity and evolution in patients with luminal breast cancer. *Nat Med*. Published online 2018. doi:10.1038/s41591-018-0091-x

126. Cohen AJ, Saiakhova A, Corradin O, et al. Hotspots of aberrant enhancer activity punctuate the colorectal cancer epigenome. *Nat Commun*. 2017;8:1-13. doi:10.1038/ncomms14400

127. Ohno S. Major sex-determining genes. *Monogr Endocrinol*. Published online 1978.

128. Lambert SA, Jolma A, Campitelli LF, et al. The Human Transcription Factors. *Cell*. 2018;172:650-665. doi:10.1016/j.cell.2018.01.029

129. Lambert M, Jambon S, Depauw S, David-Cordonnier MH. Targeting transcription factors for cancer treatment. *Molecules*. 2018;23(6). doi:10.3390/molecules23061479

130. Bradner JE, Hnisz D, Young RA. Transcriptional Addiction in Cancer. *Cell*. 2017;168(4):629-643. doi:10.1016/j.cell.2016.12.013

131. Piccolo S. Linking cancer transcriptional addictions by CDK7 to YAP/TAZ. *Genes Dev*. 2020;34(1-2):4-6. doi:10.1101/gad.335562.119

132. Maxam AM, Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A*. Published online 1977. doi:10.1073/pnas.74.2.560

133. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. Published online 1977. doi:10.1073/pnas.74.12.5463

134. Clark TA, Sugnet CW, Ares M. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science (80-)*. Published online 2002. doi:10.1126/science.1069415

135. Nagalakshmi U, Wang Z, Waern K, et al. The transcriptional landscape of the yeast

genome defined by RNA sequencing. *Science (80-)*. Published online 2008. doi:10.1126/science.1158441

136. Kivioja T, Vähärautio A, Karlsson K, et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods*. Published online 2012. doi:10.1038/nmeth.1778

137. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Rep*. Published online 2012. doi:10.1016/j.celrep.2012.08.003

138. Ranzani V, Rossetti G, Panzeri I, et al. The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4. *Nat Immunol*. 2015;16(3):318-325. doi:10.1038/ni.3093

139. Barbieri I, Kouzarides T. Role of RNA modifications in cancer. *Nat Rev Cancer*. Published online April 16, 2020:1-20. doi:10.1038/s41568-020-0253-2

140. Frye M, Jaffrey SR, Pan T, Rechavi G, Suzuki T. RNA modifications: What have we learned and where are we headed? *Nat Rev Genet*. Published online 2016. doi:10.1038/nrg.2016.47

141. Hoadley KA, Yau C, Wolf DM, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*. Published online 2014. doi:10.1016/j.cell.2014.06.049

142. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods*. Published online 2013. doi:10.1038/nmeth.2639

143. Macosko EZ, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*. Published online 2015. doi:10.1016/j.cell.2015.05.002

144. Zheng GXY, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. Published online 2017. doi:10.1038/ncomms14049

145. Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat Commun*. Published online 2019. doi:10.1038/s41467-019-12266-7

146. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol*. Published online 2019. doi:10.15252/msb.20188746

147. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet*. 2012;13(12):840-852. doi:10.1038/nrg3306

148. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol*. Published online 2015. doi:10.1002/0471142727.mb2129s109

149. Schmidl C, Rendeiro AF, Sheffield NC, Bock C. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat Methods*. 2015;12(10):963-965. doi:10.1038/nmeth.3542

150. Schwartzman O, Tanay A. Single-cell epigenomics: Techniques and emerging applications. *Nat Rev Genet*. Published online 2015. doi:10.1038/nrg3980

151. Kawakami H, Zaanan A, Sinicrope FA. Microsatellite Instability Testing and Its Role in the Management of Colorectal Cancer. *Curr Treat Options Oncol*. 2015;16(7):30. doi:10.1007/s11864-015-0348-2

152. Fujii M, Matano M, Nanki K, Sato T. Efficient genetic engineering of human intestinal organoids using electroporation. *Nat Protoc*. Published online 2015. doi:10.1038/nprot.2015.088

153. Mahe MM, Aihara E, Schumacher MA, et al. Establishment of Gastrointestinal Epithelial Organoids. *Curr Protoc Mouse Biol*. 2013;3(4):217-240. doi:10.1002/9780470942390.mo130179

154. Schindelin J, Arganda-Carreras I, Frise E, et al. Fiji: an open-source platform for biological-image analysis. *Nat Methods*. 2012;9(7):676-682. doi:10.1038/nmeth.2019

155. Cordenonsi M, Zanconato F, Azzolin L, et al. The hippo transducer TAZ confers cancer stem cell-related traits on breast cancer cells. *Cell*. Published online 2011. doi:10.1016/j.cell.2011.09.048

156. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol*. 2017;35(4):316-319. doi:10.1038/nbt.3820

157. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. Published online 2013. doi:10.1093/bioinformatics/bts635

158. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923-930.

doi:10.1093/bioinformatics/btt656

159. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

160. R Foundation for Statistical Computing. *R: A Language and Environment for Statistical Computing.*; 2018.

161. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102

162. Ramírez F, Ryan DP, Grüning B, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*. 2016;44(W1):W160-5. doi:10.1093/nar/gkw257

163. Eide PW, Bruun J, Lothe RA, Sveen A. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Sci Rep*. 2017;7(1):16618. doi:10.1038/s41598-017-16747-x

164. Andrews S. FastQC: A quality control tool for high throughput sequence data. Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/. doi:citeulike-article-id:11583827

165. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25. doi:10.1186/gb-2009-10-3-r25

166. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324

167. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):R137. doi:10.1186/gb-2008-9-9-r137

168. Ramírez F, Bhardwaj V, Arrigoni L, et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. doi:10.1038/s41467-017-02525-w

169. Orlando G, Law PJ, Cornish AJ, et al. Promoter capture Hi-C-based identification of recurrent noncoding mutations in colorectal cancer. *Nat Genet*. 2018;50(10):1375-1380. doi:10.1038/s41588-018-0211-z

170. Lee HO, Hong Y, Etlioglu HE, et al. Lineage-dependent gene expression programs

influence the immune landscape of colorectal cancer. *Nat Genet*. Published online 2020. doi:10.1038/s41588-020-0636-z

171. Kim N, Kim HK, Lee K, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun*. Published online 2020. doi:10.1038/s41467-020-16164-1

172. Wolf FA, Angerer P, Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol*. Published online 2018. doi:10.1186/s13059-017-1382-0

173. Ross-Innes C, Stark R, Teschendorff A, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*. 2012;481(7381):389-393. http://discovery.ucl.ac.uk/1335413/

174. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*. 2012;9(3):215-216. doi:10.1038/nmeth.1906

175. Love MI, Anders S, Huber W. *Differential Analysis of Count Data - the DESeq2 Package*. Vol 15.; 2014. doi:110.1186/s13059-014-0550-8

176. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842. doi:10.1093/bioinformatics/btq033

177. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*. 2010;38(4):576-589. doi:10.1016/j.molcel.2010.05.004

178. Machanick P, Bailey TL. MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics*. Published online 2011. doi:10.1093/bioinformatics/btr189

179. Zhu LJ, Gazin C, Lawson ND, et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*. 2010;11(1):237. doi:10.1186/1471-2105-11-237

180. Shannon P, Markiel A, Ozier O, et al. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-2504. doi:10.1101/gr.1239303

181. Leek JT, Storey JD. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet*. 2007;3(9):e161. doi:10.1371/journal.pgen.0030161

182. McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*.

Published online 2017. doi:10.1093/bioinformatics/btw777

183. Kowalczyk MS, Tirosh I, Heckl D, et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res*. Published online 2015. doi:10.1101/gr.192237.115

184. Wang Y, Song W, Wang J, et al. Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine. *J Exp Med*. Published online 2020. doi:10.1084/jem.20191130

185. Patel AP, Tirosh I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (80-)*. Published online 2014. doi:10.1126/science.1254257

186. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317-330. doi:10.1038/nature14248

187. Bartolini A, Cardaci S, Lamba S, et al. BCAM and LAMA5 mediate the recognition between tumor cells and the endothelium in the metastatic spreading of KRAS-mutant colorectal cancer. *Clin Cancer Res*. Published online 2016. doi:10.1158/1078-0432.CCR-15-2664

188. Jass JR, Do KA, Simms LA, et al. Morphology of sporadic colorectal cancer with DNA replication errors. *Gut*. Published online 1998. doi:10.1136/gut.42.5.673

189. Van der Flier LG, Sabates-Bellver J, Oving I, et al. The Intestinal Wnt/TCF Signature. *Gastroenterology*. Published online 2007. doi:10.1053/j.gastro.2006.08.039

190. Gaspar C, Cardoso J, Franken P, et al. Cross-species comparison of human and mouse intestinal polyps reveals conserved mechanisms in adenomatous polyposis coli (APC)-driven tumorigenesis. *Am J Pathol*. Published online 2008. doi:10.2353/ajpath.2008.070851

191. Sakthianandeswaren A, Christie M, D'Andreti C, et al. PHLDA1 expression marks the putative epithelial stem cells and contributes to intestinal tumorigenesis. *Cancer Res*. Published online 2011. doi:10.1158/0008-5472.CAN-10-2342

192. Zanconato F, Forcato M, Battilana G, et al. Genome-wide association between YAP/TAZ/TEAD and AP-1 at enhancers drives oncogenic growth. *Nat Cell Biol*. 2015;17(9):1218-1227. doi:10.1038/ncb3216

193. Zanconato F, Cordenonsi M, Piccolo S. YAP/TAZ at the Roots of Cancer. *Cancer Cell*. 2016;29(6):783-803. doi:10.1016/j.ccell.2016.05.005

194. Azzolin L, Panciera T, Soligo S, et al. YAP/TAZ incorporation in the β-catenin destruction complex orchestrates the Wnt response. *Cell*. Published online 2014. doi:10.1016/j.cell.2014.06.013

195. Egeblad M, Nakasone ES, Werb Z. Tumors as organs: Complex tissues that interface with the entire organism. *Dev Cell*. Published online 2010. doi:10.1016/j.devcel.2010.05.012

196. Gregorieff A, Liu Y, Inanlou MR, Khomchuk Y, Wrana JL. Yap-dependent reprogramming of Lgr5+ stem cells drives intestinal regeneration and cancer. *Nature*. Published online 2015. doi:10.1038/nature15382

197. Pocaterra A, Romani P, Dupont S. YAP/TAZ functions and their regulation at a glance. *J Cell Sci*. 2020;133(2). doi:10.1242/jcs.230425

198. Totaro A, Panciera T, Piccolo S. YAP/TAZ upstream signals and downstream responses. *Nat Cell Biol*. Published online 2018. doi:10.1038/s41556-018-0142-z

199. Liu X, Li H, Rajurkar M, et al. Tead and AP1 Coordinate Transcription and Motility. *Cell Rep*. Published online 2016. doi:10.1016/j.celrep.2015.12.104

200. Riese DJ, Cullum RL. Epiregulin: Roles in normal physiology and cancer. *Semin Cell Dev Biol*. Published online 2014. doi:10.1016/j.semcdb.2014.03.005

201. Kaneda H, Arao T, Tanaka K, et al. FOXQ1 is overexpressed in colorectal cancer and enhances tumorigenicity and tumor growth. *Cancer Res*. Published online 2010. doi:10.1158/0008-5472.CAN-09-2161

202. Grün D, Lyubimova A, Kester L, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*. Published online 2015. doi:10.1038/nature14966

203. Teschendorff AE, West J, Beck S. Age-associated epigenetic drift: Implications, and a case of epigenetic thrift? *Hum Mol Genet*. Published online 2013. doi:10.1093/hmg/ddt375

204. Vandiver AR, Irizarry RA, Hansen KD, et al. Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin. *Genome Biol*. Published online 2015. doi:10.1186/s13059-015-0644-y

205. Akhtar-Zaidi B, Cowper-Sallari R, Corradin O, et al. Epigenomic enhancer profiling defines a signature of colon cancer. *Science (80-)*. Published online 2012. doi:10.1126/science.1217277

206. Hung S, Saiakhova A, Faber ZJ, et al. Mismatch repair-signature mutations activate gene

enhancers across human colorectal cancer epigenomes. *Elife*. Published online 2019. doi:10.7554/eLife.40760

207. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*. 2010;28(8):817-825. doi:10.1038/nbt.1662

208. Creyghton MP, Cheng AW, Welstead GG, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci*. 2010;107(50):21931-21936. doi:10.1073/pnas.1016071107

209. Wang Y, Xu X, Maglic D, et al. Comprehensive Molecular Characterization of the Hippo Signaling Pathway in Cancer. *Cell Rep*. Published online 2018. doi:10.1016/j.celrep.2018.10.001

210. Stein C, Bardet AF, Roma G, et al. YAP1 Exerts Its Transcriptional Control via TEAD-Mediated Activation of Enhancers. *PLoS Genet*. Published online 2015. doi:10.1371/journal.pgen.1005465

211. Monroe TO, Hill MC, Morikawa Y, et al. YAP Partially Reprograms Chromatin Accessibility to Directly Induce Adult Cardiogenesis In Vivo. *Dev Cell*. Published online 2019. doi:10.1016/j.devcel.2019.01.017

212. Gill MK, Christova T, Zhang YY, et al. A feed forward loop enforces YAP/TAZ signaling during tumorigenesis. *Nat Commun*. Published online 2018. doi:10.1038/s41467-018-05939-2

213. Maglic D, Schlegelmilch K, Dost AF, et al. YAP-TEAD signaling promotes basal cell carcinoma development via a c-JUN/AP1 axis. *EMBO J*. Published online 2018. doi:10.15252/embj.201798642

214. Kim MK, Jang JW, Bae SC. DNA binding partners of YAP/TAZ. *BMB Rep*. 2018;51(3):126-133. doi:10.5483/BMBRep.2018.51.3.015

215. Liu AM, Wong KF, Jiang X, Qiao Y, Luk JM. Regulators of mammalian Hippo pathway in cancer. *Biochim Biophys Acta - Rev Cancer*. 2012;1826(2):357-364. doi:10.1016/j.bbcan.2012.05.006

216. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nat Commun*. Published online 2015. doi:10.1038/ncomms9971

217. Gaspar P, Tapon N. Sensing the local environment: Actin architecture and Hippo signalling. *Curr Opin Cell Biol*. Published online 2014. doi:10.1016/j.ceb.2014.09.003

218. Low BC, Pan CQ, Shivashankar G V., Bershadsky A, Sudol M, Sheetz M. YAP/TAZ as mechanosensors and mechanotransducers in regulating organ size and tumor growth. *FEBS Lett*. Published online 2014. doi:10.1016/j.febslet.2014.04.012

219. Hinohara K, Polyak K. Intratumoral Heterogeneity: More Than Just Mutations. *Trends Cell Biol*. Published online 2019. doi:10.1016/j.tcb.2019.03.003

220. Li H, Courtois ET, Sengupta D, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet*. Published online 2017. doi:10.1038/ng.3818

221. Yuan WC, Pepe-Mooney B, Galli GG, et al. NUAK2 is a critical YAP target in liver cancer. *Nat Commun*. Published online 2018. doi:10.1038/s41467-018-07394-5

222. Yu FX, Meng Z, Plouffe SW, Guan KL. Hippo pathway regulation of gastrointestinal tissues. *Annu Rev Physiol*. Published online 2015. doi:10.1146/annurev-physiol-021014-071733

223. Bartucci M, Dattilo R, Moriconi C, et al. TAZ is required for metastatic activity and chemoresistance of breast cancer stem cells. *Oncogene*. Published online 2015. doi:10.1038/onc.2014.5

224. Kim MH, Kim J, Hong H, et al. Actin remodeling confers BRAF inhibitor resistance to melanoma cells through YAP / TAZ activation . *EMBO J*. Published online 2016. doi:10.15252/embj.201592081

225. Lin YH, Zhen YY, Chien KY, et al. LIMCH1 regulates nonmuscle myosin-II activity and suppresses cell migration. *Mol Biol Cell*. Published online 2017. doi:10.1091/mbc.E15-04-0218

226. Mifsud B, Tavares-Cadete F, Young AN, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet*. Published online 2015. doi:10.1038/ng.3286

227. Cope NF, Fraser P. Chromosome conformation capture. *Cold Spring Harb Protoc*. Published online 2009. doi:10.1101/pdb.prot5137

228. Amabile A, Migliara A, Capasso P, et al. Inheritable Silencing of Endogenous Genes by Hit-and-Run Targeted Epigenetic Editing. *Cell*. Published online 2016. doi:10.1016/j.cell.2016.09.006

229. Li K, Liu Y, Cao H, et al. Interrogation of enhancer function by enhancer-targeting CRISPR epigenetic editing. *Nat Commun*. Published online 2020. doi:10.1038/s41467-

020-14362-5

230. Klann TS, Black JB, Chellappan M, et al. CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat Biotechnol*. Published online 2017. doi:10.1038/nbt.3853

231. Meng Q, Xie S, Kenneth Gray G, et al. Clonal expansion of personalized anti-tumor T cells from circulation using tumor organoid-immune co-cultures † Equal contribution. *bioRxiv*. Published online November 8, 2020:2020.11.06.371807. doi:10.1101/2020.11.06.371807

7. Appendix

7.1 Experimental protocols

7.1.1 Isolation of human primary tissues

MATERIALS

REAGENTS:

- PBS
- Gentamicin (20 µg/ml)
- PBS-EDTA (2.5 mM)

PROCEDURE

Primary colonic normal and tumoral tissues were processed according to a previously published protocol[152].

I. Surgically resected specimens were reduced to a size of 3-5 mm, extensively washed with cold PBS and gentamicin (20 µg/ml) and incubated with PBS-EDTA (2.5 mM) rocking on a wheel for 1 h at 4°C.

II. After PBS-EDTA treatment, tissue samples were washed with cold PBS - 1% FBS to release normal crypts and tumoral counterpart.

III. Cells suspension were collected by centrifuging at 400 g for 5 min at 4°C and used for transcriptomic and epigenomic analyses.

7.1.2 Patient derived colorectal cancer organoids culture

MATERIALS

REAGENTS:

- Matrigel® Growth Factor Reduced Basement Membrane Matrix, Phenol Red-Free (Corning)
- Advanced DMEM/F12 (Life Technologies)
- Penicillin/streptomycin (Euroclone)
- HEPES (Life Technologies)
- GlutaMAX (Life Technologies)
- B27 (Life Technologies)

- N2 (Life Technologies)
- N-Acetyl Cysteine (Sigma-Aldrich)
- Nicotinamide (Sigma-Aldrich)
- Human EGF (Peprotech)
- Human Noggin (Peprotech)
- Human Gastrin (Sigma)
- A83-01 (Tocris)
- SB202190 (Sigma)
- TrypLETM Express Enzyme (12605010, Thermo Fisher)

PROCEDURE

PDOs were established and maintained as previously described[152].

I. Tumor cells suspension isolated from CRC biopsies were embedded in drops of Matrigel® Growth Factor Reduced Basement Membrane Matrix, Phenol Red-Free (Corning) to establish CRC PDOs libraries.

II. Droplets of matrigel containing tumor cells suspension or established organoids were maintained in 24 well plate overlaid by 500 µl of the organoid culture medium (Advanced DMEM/F12 (Life Technologies) supplemented with penicillin/streptomycin (Euroclone), 10 mM HEPES (Life Technologies), 2 mM GlutaMAX (Life Technologies), 1X B27 (Life Technologies), 1X N2 (Life Technologies), 1 mM N-Acetyl Cysteine (Sigma-Aldrich), 10 mM Nicotinamide (Sigma-Aldrich), 50 ng/ml human EGF (Peprotech), 100 ng/ml human Noggin (Peprotech), 10 nM human Gastrin (Sigma), 500 nM A83-01 (Tocris), 10 µM SB202190 (Sigma).

III. The organoids were split once per week by mechanical disruption or enzymatic digestion using TrypLETM Express Enzyme (12605010, Thermo Fisher) and regularly checked for mycoplasma contamination.

7.1.3 Whole mount staining of PDOs

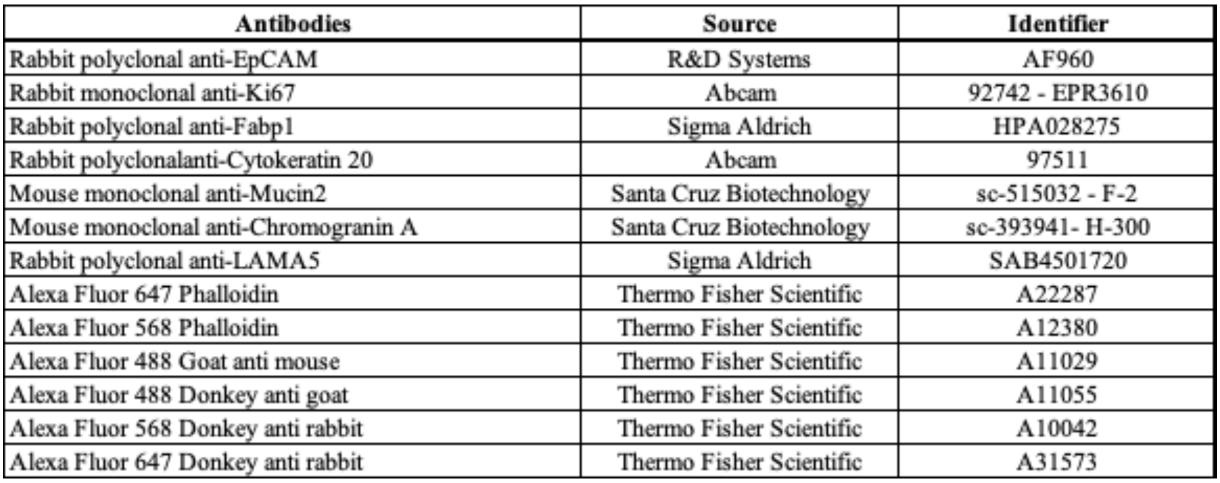
MATERIALS

REAGENTS:

- NH4Cl

- Triton X-100 (Sigma-Aldrich)
- Donkey Serum or Normal Goat Serum (Sigma-Aldrich)
- PBS
- Hoechst 33342
- NaN2

ANTIBODIES:



Antibodies	Source	Identifier
Rabbit polyclonal anti-EpCAM	R&D Systems	AF960
Rabbit monoclonal anti-Ki67	Abcam	92742 - EPR3610
Rabbit polyclonal anti-Fabp1	Sigma Aldrich	HPA028275
Rabbit polyclonalanti-Cytokeratin 20	Abcam	97511
Mouse monoclonal anti-Mucin2	Santa Cruz Biotechnology	sc-515032 - F-2
Mouse monoclonal anti-Chromogranin A	Santa Cruz Biotechnology	sc-393941- H-300
Rabbit polyclonal anti-LAMA5	Sigma Aldrich	SAB4501720
Alexa Fluor 647 Phalloidin	Thermo Fisher Scientific	A22287
Alexa Fluor 568 Phalloidin	Thermo Fisher Scientific	A12380
Alexa Fluor 488 Goat anti mouse	Thermo Fisher Scientific	A11029
Alexa Fluor 488 Donkey anti goat	Thermo Fisher Scientific	A11055
Alexa Fluor 568 Donkey anti rabbit	Thermo Fisher Scientific	A10042
Alexa Fluor 647 Donkey anti rabbit	Thermo Fisher Scientific	A31573

Table 4 | Primary and secondary antibodies/conjugates used for whole mount staining of PDOs.

EQUIPMENT:

- SP5 microscope (Leica Microsystems)
- 10× (NA 0.3) or 20x (NA 0.7) dry objectives (TCS SP5; Leica)

SOFTWARE:

- ImageJ software

PROCEDURES

Isolated organoids embedded in Matrigel in µ-Plate Angiogenesis 96 Well (Ibidi) were fixed in 4% paraforlmaldehyde in PBS for 1 hour, at 4°C. The whole mount staining protocol was performed as previously described[153], with some modifications.

I. After fixation, the auto-fluorescence was quenched with 50 mM NH4Cl for 30 minutes and the organoids were permeabilizated with 0,5% Triton X-100 (Sigma-Aldrich) for 1 h and blocked with 10% Donkey Serum or Normal Goat Serum (Sigma-Aldrich) in PBS with 0.2% Triton X-100 overnight, at 4°C in mild shaking.

II. Primary and secondary antibodies were diluted in 5% of serum and applied respectively ~35 and ~12 hours at 4°C, in mild shaking.

III. Cell nuclei were stained with 20 µg/ml Hoechst 33342 in PBS with 0,2% Triton X-100 for 2 hours and the organoids were stored in PBS with 0,02% NaN2 until the acquisition.

IV. Fluorescence images were captured with confocal laser-scanning SP5 microscope (Leica Microsystems) equipped with eight laser lines and four PMT detectors, using 10× (NA 0.3) or 20x (NA 0.7) dry objectives (TCS SP5; Leica), 5 or 10 µm z-step interval and 1024x1024 or 2048x2048 image format.

V. For each acquired confocal z-stack field, maximum intensity projections (MIP) were generated using ImageJ software (National Institutes of Health)[154].

7.1.4 Chromatin Immunoprecipitation (ChIP) assay and library construction

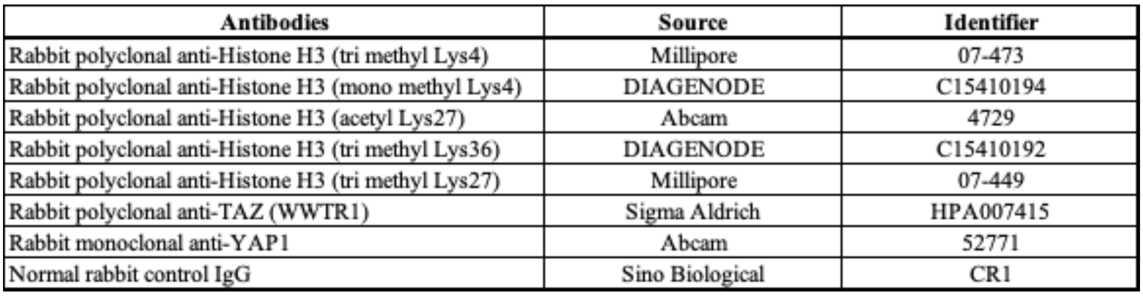
MATERIALS

REAGENTS:

- Cell Recovery Solution (Matrisperse Cell Recovery Solution - Sacco-L004419 CPB40253)
- Formaldheyde (F8775 SIGMA)
- PBS
- Glycine
- Sonication lysis buffer (10 mM Tris pH 8.0, 0.25% SDS, 2 mM EDTA, plus protease inhibitors)
- ProteinG-Dynabeads (Invitrogen)
- RIPA-LS
- RIPA-HS
- RIPA-LiCl
- Tris 10mM pH8
- TE 1x
- Elution buffer (10 mM Tris-HCl pH 8.0, 5 mM EDTA pH 8.0, 300 mM NaCl, 0.4% SDS) Proteinase K
- Qiagen MinElute kit (Qiagen)

- EB buffer

ANTIBODIES:



Antibodies	Source	Identifier
Rabbit polyclonal anti-Histone H3 (tri methyl Lys4)	Millipore	07-473
Rabbit polyclonal anti-Histone H3 (mono methyl Lys4)	DIAGENODE	C15410194
Rabbit polyclonal anti-Histone H3 (acetyl Lys27)	Abcam	4729
Rabbit polyclonal anti-Histone H3 (tri methyl Lys36)	DIAGENODE	C15410192
Rabbit polyclonal anti-Histone H3 (tri methyl Lys27)	Millipore	07-449
Rabbit polyclonal anti-TAZ (WWTR1)	Sigma Aldrich	HPA007415
Rabbit monoclonal anti-YAP1	Abcam	52771
Normal rabbit control IgG	Sino Biological	CR1

Table 5 | Antibodies used for ChIP-seq and ChIPmentation protocols.

EQUIPMENT:

Covaris® M220 focused-ultrasonicator (settings: duty factor 20%, peak incidence power 75 Watt, cycles per burst 200)

PROCEDURES

I. For ChIP experiments, matrigel droplet containing ~0.3 x 10^6 organoid cells/well was dissolved using Cell Recovery Solution (Matrisperse Cell Recovery Solution - Sacco-L004419 CPB40253), following the indicated procedure.

II. PBS-washed organoids pellet was fixed as whole in Formaldheyde (F8775 SIGMA) PBS-solution (final 1%), for 10 min rocking at room temperature and quenched with 0.125 M Glycine for 5 min. P

III. BS-washed organoid pellets were lysed with 500 µl of 1X sonication lysis buffer (10 mM Tris pH 8.0, 0.25% SDS, 2 mM EDTA, plus protease inhibitors) and incubated for at least 10 min at 4°C. Lysed chromatin was sheared at 200–500 bp fragments using Covaris® M220 focused-ultrasonicator (settings: duty factor 20%, peak incidence power 75 Watt, cycles per burst 200, 8-15 minutes).

IV. For organoids and crypts, ~500 ng and ~1000 ng respectively of sonicated chromatin was incubated with antibody (H3K27ac abcam 4729; H3K4me3 Millipore 07-473; H3K4Me1 DIAGENODE C15410194; H3K36me3 DIAGENODE C15410192; H3K27me3 07449 Millipore) overnight at 4 °C on wheel.

V. Antibody/antigen complexes were recovered with blocked ProteinG-Dynabeads (Invitrogen) for 2 h at 4 °C and washed with RIPA-LS (twice), RIPA-HS (twice), RIPA-LiCl (twice), once with Tris 10mM pH8 and once with TE 1x, followed by reverse crosslinking overnight. The washed immunocomplexes were incubated with ChIP elution buffer (10 mM Tris-HCl pH 8.0, 5 mM EDTA pH 8.0, 300 mM NaCl, 0.4% SDS) supplemented with 0.8 mg/ml Proteinase K for 1 h at 55°C and overnight at 65°C, for reverse crosslinking.

VI. The immunoprecipitated DNA was then purified by Qiagen MinElute kit (Qiagen) and eluted in 22 µl EB buffer.

ChIP-seq libraries were constructed with TruSeq ChIP Library Preparation Kit (Illumina), according to the manufacturer's instructions and sequenced on the Illumina HiSeq2500 platform.

7.1.5 ChIPmentation assay and library preparation

REAGENTS:

- Lyses buffer I (50 mM HEPES, pH 7.5, 10 mM NaCl, 1 mM EDTA, 10% Glycerol, 0.5% NP-40, 0.25% Triton X-100, plus protease inhibitors)
- Lyses buffer II (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, plus protease inhibitors)
- Lyses buffer III
- Blocked ProteinG-Dynabeads (Invitrogen)
- Wash buffer
- Salt buffer
- Tris pH8
- Tagment DNA Enzyme from the Nextera DNA Sample Prep Kit (Illumina)
- Nextera DNA Sample Prep Kit (Illumina)
- Proteinase K (NEB)
- SPRI AMPure XP beads

ANTIBODIES:

See **Table 5**.

EQUIPMENT:

Covaris® M220 focused-ultrasonicator (settings: duty factor 10%, peak incidence power 75 Watt, cycles per burst 200)

PROCEDURES

ChIPmentation was carried out as previously described[149] with small modifications in cell lysis and washes after recovering.

I. The crosslinked pellet was lysed in buffer I (50 mM HEPES, pH 7.5, 10 mM NaCl, 1 mM EDTA, 10% Glycerol, 0.5% NP-40, 0.25% Triton X-100, plus protease inhibitors) in ice.

II. The pellet was recovered and lysed with buffer II (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, plus protease inhibitors) at room temperature and sonicated in lyses buffer III using Covaris® M220 focused-ultrasonicator (settings: duty factor 10%, peak incidence power 75 Watt, cycles per burst 200, 15 minutes).

III. Sonicated chromatin was incubated with anti-WWTR1 (Sigma Aldrich, HPA007415) or anti-YAP1 (abcam 52771) overnight at 4 °C on the wheel.

IV. For control libraries, an immunoprecipitation with nonspecific IgG rabbit antibody was used.

V. Antibody/antigen complexes were recovered with blocked ProteinG-Dynabeads (Invitrogen) and washed with low salt wash buffer (twice), high salt buffer (twice) and once with Tris pH8.

VI. Beads were then resuspended and incubated in tagmentation reaction containing Tagment DNA Enzyme from the Nextera DNA Sample Prep Kit (Illumina).

VII. Beads were then washed and incubated with elution buffer plus Proteinase K (NEB) to revert formaldehyde cross-linking.

VIII. Library preparation for ChIPmentation was performed using custom Nextera primers as described for ATAC-seq and enriched libraries were purified using 1.8V of SPRI AMPure XP beads and sequenced with Illumina HiSeq2500.

7.2 Computational protocols

7.2.1 RNA-seq data processing

- RNA-seq reads were sequenced as paired-end reads on Illumina HiSeq2500 and analysed with a custom pipeline built using Nextflow[156].

- FastQC v0.11.769 (http://www.bioinformatics.babraham.ac.uk/projects/) was used to perform the quality control of the sequenced reads:

```
>fastqc sample_name_R1.fastq sample_name_R2.fastq
```

- To remove adapters and low quality reads, reads were trimmed using BBDuk (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/):

```
>bbduk.sh in1=sample_name_R1.fastq in2=sample_name_R1.fastq
out1=tmp1.fastq.gz out2=tmp2.fastq.gz ref=list_of_adapters ktrim=r k=23
mink=11 hdist=1 tpe tbo qin=33
```

```
>bbduk.sh in1=tmp1.fastq.gz in2=tmp2.fastq.gz
out1=sample_name_R1_trim.fastq.gz out2= sample_name_R2_trim.fastq.gz
qtrim=rl trimq=20 minlen=50 qin=33
```

- The reads were aligned to the human reference genome hg38 (GENCODE Release 25 basic gene annotation) using STARv2.5.3a[157]:

```
>STAR --genomeDir STAR_index --readFilesIn sample_name_R1_trim.fastq.gz
sample_name_R2_trim.fastq.gz --readFilesCommand zcat --genomeLoad
LoadAndRemove --outFileNamePrefix sample_name --outReadsUnmapped Fastx
--outSAMtype BAM SortedByCoordinate --alignIntronMax 1000000 --
quantMode GeneCounts --outFilterMismatchNmax 9 --outFilterMultimapNmax
20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --alignMatesGapMax
1000000
```

- FeatureCounts4-Subreadv1.6.2[158] with default parameters was used to perform the quantification of the reads:

```
>featureCounts -p -B -C -t exon -g gene_id -a genome.gtf -o
sample_name_counts.txt -s 2 -T sample_name.bam
```

- For the visualization of RNA-seq tracks, the normalized coverage tracks were generated using the *bamCoverage* function of deeptools[162]. Separate tracks for forward and reverse transcripts were generated for each independent sample.

```
>bamCoverage -b sample_name.bam -o sample_name_forward.bw --
filterRNAstrand forward --normalizeTo1x 3049315783 --minMappingQuality 10
```

```
>bamCoverage -b sample_name.bam -o sample_name_reverse.bw --filterRNAstrand reverse --normalizeTo1x 3049315783 --minMappingQuality 10
```

7.2.2 ChIP-seq data processing

- ChIP-seq reads were sequenced as single-end reads on Illumina HiSeq2500 and analysed with a custom pipeline built using Nextflow[156].

- FastQC v0.11.769 (http://www.bioinformatics.babraham.ac.uk/projects/) was used to perform the quality control of the sequenced reads:

  ```
  >fastqc sample_name_R1.fastq
  ```

- The reads were aligned to the human reference genome hg38 (GENCODE Release 25 basic gene annotation) using Bowtie v1.2.2[165]), sorted using SAMtoolsv1.8[166] and directly converted into binary files (BAM). `>bowtie -S -m 1 --best --strata -v 3 >sbowtie_index | samtools view -bS - | samtools sort -n -T sample_name -O BAM -o sample_name.bam`

- PCR duplicate reads were marked and removed using SAMtoolsv1.8:

  ```
  >samtools fixmate -r -m sample_name.bam -| samtools sort -@ 4 - |
  samtools markdup -r -s - sample_name_no_dup.bam
  ```

- For sharp histone modifications (H3K4me3 and H3K27ac) the peaks were called with MACS2 v2.1.0[167] using matched input DNA as a control with the following command line:

  ```
  >macs2 callpeak -t sample_name_no_dup.bam sample_input_no_dup.bam -f
  BAM -g 3049315783 -n sample_name –nomodel –extsize 200 -B -q 0.01
  ```

- For sharp histone modifications (H3K4me1, H3K36me3 and H3K27me3) the peaks were called with MACS2 v2.1.0[167] using matched input DNA as a control with the following command line:

  ```
  >macs2 callpeak -t sample_name_no_dup.bam sample_input_no_dup.bam --broad
  -f BAM -g 3049315783 -n sample_name –nomodel –extsize 200 -B -q 0.01
  ```

- Peaks overlapping ENCODE blacklisted regions (BL) hg38 (*i.e.* regions in the human genome with signal artefacts in NGS experiments, https://www.encodeproject.org/annotations/ENCSR636HFF/) were removed:

  ```
  >bedtools intersect -a sample_name_peak_file blacklisted_regions -v |
  grep chr > sample_name_peak_file_noBL
  ```

- For the visualization of ChIP-seq tracks, Bedgraph tracks were generated using MACS2 bdgcmp function, converted into bigwig using UCSC bedClip and bedGraphToBigWig functions:

```
>macs2 bdgcmp -t sample_name_pileup -c sample_name_control_lambda -o
sample_name_FE.bdg -m FE
```

```
>LC_COLLATE=C sort -k1,1 -k2,2n sample_name _FE.bdg >
sample_name_FE_sort.bdg
```

```
>bedGraphToBigWig sample_name_FE_sort.bdg chromosome_size_file
sample_name_FE.bw
```

7.2.3 Density and heatmap plot for each histone modification

- Filtered and sorted BAM files were used to generate normalized coverage tracks using the bamCoverage function from deepTools[162] suite:

```
>bamCoverage -v -b sample_name.bam -o sample_name_norm.coverage.bw --
normalizeUsing RPGC –effectiveGenomeSize 3049315783 --extendReads 200 -
-binSize 1
```

- The average signal profile and the heatmap plot along the genebody were calculated using computeMatrix scale-regions with default parameters and GENCODE Release 25 basic gene annotation:

```
>computeMatrix scale-regions --regionsFileName annotation.gtf --
scoreFileName sample_name --outFileName sample_name.genebody.gz --
regionBodyLength 6000 --upstream 3000 --downstream 3000 --
missingDataAsZero
```

```
>plotHeatmap --matrixFile sample_name.genebody.gz --outFileName
sample_name.genebody.heatmap.col.pdf --colorList "#00004c,#0000ff,
white,#F27F7F,#EC3F3F,#E60000" --zMin 0 --zMax 10 --heatmapHeight 20 --
dpi 300 --yMin 0 --yMax 30
```

7.2.4 ChromHMM analysis

- ChromHMM analysis was run with a custom pipeline built using Nextflow[156] using all the samples available for the study and the public available samples as described in "***De novo* chromatin state characterization**".
- The datasets were down-sampled to a maximum depth of 45 million reads (the median read depth over all samples considered in this analysis).

- First, the cellmarkfiletable was created to analyse all the available samples as multiple cell type to be treated concatenating them. The control data was used to adjust the binarization threshold locally. The cellmarkfiletable reports the name of the PDO in the first column, the histone modification type in the second column, the bam files without the duplicates in the third column and the bam file for the associated input in the fourth column. The columns were tab separated and the files were all collected in the same directory:

```
PDO8   H3K4me3     sample_name_no_dup.bam  input_no_dup.bam
PDO8   H3K27Ac     sample_name_no_dup.bam  input_no_dup.bam
PDO8   H3K4me1     sample_name_no_dup.bam  input_no_dup.bam
PDO8   H3K27me3    sample_name_no_dup.bam  input_no_dup.bam
PDO8   H3K36me3    sample_name_no_dup.bam  input_no_dup.bam
PDO10  H3K4me3     sample_name_no_dup.bam  input_no_dup.bam
PDO10  H3K27Ac     sample_name_no_dup.bam  input_no_dup.bam
....
```

- The data binarization in which the genome is fractioned in contiguous bins of 200 bps was executed using the hg38 assembly:

```
>java -Xmx4g -jar ChromHMM.jar BinarizeBam chromosome_length_file_hg38
inputdir cellmarkfiletable output directory
```

- Then, the Model learning was executed using the binarized data localized in the input directory. The *LearnModel* function calculates the probability that certain histone marks are present in the same genomic region and can be co-present with other histone marks. The combination of multiple histone modifications in the same bin over the genome is used to define an *n* number of states, defined by the use, in our case numstates=8,10,12:

```
>java -Xmx4g -jar ChromHMM.jar LearnModel inputdir outputdir numstates
hg38
```

- To compare the models generated with different number of states and evaluate the best one the *CompareModels* function was run:

```
>java -Xmx4g -jar ChromHMM.jar CompareModels
emission_model_8,emission_model_10,emission_model_12
directoy_to_compare outdir
```

The 8-state model was chosen for downstream analysis since it captured the key interaction between histone marks and because it was the model with minimal redundancy.