Highly-multiplexed imaging analysis of intra-tumour heterogeneity and response to immune checkpoint inhibitors in colorectal cancer

Bortolomeazzi, Michele

Awarding institution:
King's College London

Highly-multiplexed imaging analysis of intra-tumour heterogeneity and response to immune checkpoint inhibitors in colorectal cancer

Michele Bortolomeazzi

Thesis submitted for the Doctor of Philosophy degree in Bioinformatics

March 2022

King’s College London

Abstract

Immune checkpoint inhibitors (ICIs) have been adopted to treat multiple cancer types, including colorectal cancer (CRC). Two anti-Programmed Cell Death 1 (PD1) antibodies, pembrolizumab and nivolumab, have proven effective in patients with metastatic mismatch repair-deficient CRC with microsatellite instability (dMMR-MSI). However, anti-PDI ICIs provide no benefit for a significant fraction of dMMR-MSI CRC patients. These patients resist anti-PD1 treatment from the beginning or develop acquired resistance, leading to disease progression. Inter- and intra-tumour heterogeneity (ITH) are leading causes of resistance through antigen escape and immunosuppression. Thus, further investigation of these factors in CRC and their impact on the tumour microenvironment (TME) is required to identify new biomarkers.

Through three projects, I investigate the impact of ITH on response to ant-PD1 agents in CRC. First, I developed a tool for the analysis of highly-multiplexed images. There is still a lack of software tools encompassing all the workflow steps and allowing flexible, scalable and reproducible analysis. For this reason, I developed the Single-cell Identification from MultiPLexed Images (SIMPLI) pipeline. SIMPLI includes raw image processing and spatially resolved cell- and pixel-level analyses. Every step of SIMPLI's workflow is highly customisable and imaging technology agnostic, thus making it highly flexible. SIMPLI can run on both desktop computers and high-performance computing environments with minimum configuration and automatically manage computational resources allocation. These features make SIMPLI a portable, scalable and reproducible software for analysing highly-multiplexed images.

The second project is the application of SIMPLI to the analysis of ITH in the TME in CRC patients treated with pembrolizumab or nivolumab in a multiregional and multiomic study. Multiple tumour regions with different levels of T cell infiltration underwent DNA, RNA and T-cell receptor sequencing, imaging mass cytometry and multiplex immunofluorescence. These analyses showed that response to anti-PD1 ICI in CRCs did not correlate with tumour mutational burden. Instead, response was linked to the clonality of immunogenic mutations and T cell receptors, dysregulation of the WNT signalling, interferon-gamma and antigen presentation pathways. PDL1$^+$ antigen-presenting macrophages enrichment segregated with response and formed high-density clusters rich in cytotoxic and proliferating PD1$^+$CD8 $^+$T cells.

The third project analysed the inter-tumour heterogeneity of cancer drivers conducted through the Network of Cancer Genes. This repository of manually annotated genes includes 3355 drivers of cancer and non-cancer clonal expansion in 122 cancer types and 12 non-cancer tissues and their system-level properties. These include gene duplicability, essentiality, evolutionary origin, miRNA, and protein interactions. This investigation showed that inter-tumour heterogeneity in cancer drivers across cancer types is caused by the intrinsic features of each tumour type and not just due to differences in their detection. The annotations produced in this project proved valuable for interpreting the genomic and transcriptomic data and the design of the antibody panels for the imaging analysis of the TME in CRC.

In summary, this thesis provides a new software for analysing highly-multiplexed images, a new set of TME features linked to anti-PD1 immunotherapy

response in CRC, and an analysis of the heterogeneity of cancer drivers across multiple cancer types.

List of published research articles and reviews

Bortolomeazzi M, Keddar MR, Ciccarelli FD, Benedetti L: Identification of non-cancer cells from cancer transcriptomic data. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 2020, 1863(6):194445.

Bortolomeazzi M, Montorsi L, Temelkovski D, Keddar MR, Acha-Sagredo A, Pitcher MJ, Basso G, Laghi L, Rodriguez-Justo M, Spencer J et al.: A SIMPLI (Single-cell Identification from MultiPLexed Images) approach for spatially-resolved tissue phenotyping at single-cell resolution. Nature Communications 2022, 13(1):781.

Bortolomeazzi M, Keddar MR, Montorsi L, Acha-Sagredo A, Benedetti L, Temelkovski D, Choi S, Petrov N, Todd K, Wai P et al: Immunogenomics of Colorectal Cancer Response to Checkpoint Blockade: Analysis of the KEYNOTE 177 Trial and Validation Cohorts. Gastroenterology 2021, 161(4):1179-1193.

Dressler L, Bortolomeazzi M, Keddar MR, Misetic H, Sartini G, Acha-Sagredo A, Montorsi L, Wijewardhane N, Repana D, Nulsen J et al: Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the Network of Cancer Genes (NCG) resource. Genome Biology 2022, 23(1):35.

Repana D, Nulsen J, Dressler L, Bortolomeazzi M, Venkata SK, Tourna A, Yakovleva A, Palmieri T, Ciccarelli FD: The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biology 2019, 20(1):1.

Table of Contents

List of figures

List of tables

List of Supplementary tables

Abbreviations

A-FRET	Amplified Förster Resonance Energy Transfer
APC	Adenomatous Polyposis Coli
AKT	v-Akt Murine Thymoma Viral Oncogene
B2M	β2-Microglobulin
BATF	Basic Leucine Zipper ATF-Like Transcription Factor
BRAF	v-Raf Murine Sarcoma Viral Oncogene Homolog B
CAF	Cancer-Associated Fibroblasts
CAR	CD19-Targeting Chimeric Antigen Receptor
CCL4	C-C Motif Chemokine Ligand 4
CDX2	Caudal Type Homeobox 2
CIMP	CpG Island Methylator Phenotype
CIN	Chromosomal Instability
CLR	Crohn’s-like reaction
CMS	Consensus Molecular Subtypes
CODEX	Co-Detection by Indexing
COSMIC	Catalogue Of Somatic Mutations In Cancer
CpG	Cytosine-phosphate-Guanine
CPPIPE	CellProfiler pipeline
CRC	Colorectal Carcinoma
CSV	Comma Separate Values
CTL	Cytotoxic Lymphocytes
CTLA4	Cytotoxic T-Lymphocyte Associated Protein 4
CXCL9	Chemokine (C-X-C motif) ligand 9

CXCL10	Chemokine (C-X-C motif) ligand 10
CyCIF	Cyclic Immunofluorescence
DAPI	4’,6-diamidino-2-phenylindole
DB	Durable Benefit
DBSCAN	Density-based spatial clustering of applications with noise
DC	Dendritic Cell
DII	Diffuse inflammatory infiltration
ECM	Extracellular Matrix
EGFR	Epidermal Growth Factor Receptor
EMT	Epithelial-to-Mesenchymal Transition
ERBB4	Erb-B2 Receptor Tyrosine Kinase 4
FDA	Food and Drug Administration
FDR	False Discovery Rate
FOLR2	Folate Receptor Beta
FOV	Field Of View
FOXP3	Forkhead Box P3
FFPE	Formalin-fixed paraffin-embedded
GzB	Granzyme B
GUI	Graphical User Interface
HE	Hematoxylin and Eosin stain
HLA	Human Leukocyte Antigen
HPC	High Performance Computing
HRP	Horseradish peroxidase
ICGC	International Cancer Genome Consortium

ICIs	Immune Checkpoint Inhibitors
IFNα	Interferon α
IFN γ	Interferon γ
IL	Interleukin
IMC	Imaging Mass Cytometry
ITH	Intra-Tumour Heterogeneity
JAK1	Janus Kinase 1
JAK2	Janus Kinase 2
KRAS	Kirsten Rat Sarcoma Virus Protein
LAG3	Lymphocyte-Activation Gene 3
MCD	Mass Cytometry Data
MDSC	Myeloid-Derived Suppressor Cells
MHC	Major Histocompatibility Complex
MIBI	Multiplexed Ion Beam Imaging
mIF	Multiplex Immunofluorescence
mIHC	Multiplex Immunohistochemistry
MLH1	*MutL* Homolog 1
MMR	Mismatch Repair
MSI	Microsatellite Instability
MSS	Microsatellite Stable
MUC1	Mucin 1
MYC	V-Myc Avian Myelocytomatosis Viral Oncogene Homolog
NCG	Network of Cancer Genes
NCG^{HD}	Network of Cancer Genes and Healthy Drivers

nDB	non-Durable Benefit
NF-κB	Nuclear Factor kappa B
NMS	Non-Maximum Suppression
NRAS	Neuroblastoma RAS Viral Oncogene Homolog
PCAWG	Pan-Cancer Analysis of Whole Genomes
PD1	Programmed Cell Death Protein 1
PDL1	Programmed Death-Ligand 1
PDL2	Programmed Death-Ligand 2
PDAC	Pancreatic Ductal Adenocarcinoma
PDF	Portable Document Format
PIK3CA	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit α
PI3K	Phosphoinositide 3-Kinases
RAS	Rat Sarcoma Virus Protein
ROI	Region Of Interest
SHP1	Src Homology 2 Domain-Containing Protein Tyrosine Phosphatase 1
SHP2	Src Homology 2 Domain-Containing Protein Tyrosine Phosphatase 2
SIMPLI	Single-cell Identification from MultiPLexed Images
TAF	Tumour Associated Fibroblasts
TCF7	Transcription Factor 7
TCGA	The Cancer Genome Atlas
TCIA	The Cancer Imaging Archive

TCR	T-cell Receptor
TGFα	Transforming Growth Factor α
TGFβ	Transforming Growth Factor β
Th1	T helper type 1
Th2	T helper type 2
Th17	T helper type 17
TIFF	Tag Image File Format
TIGIT	T cell Immunoreceptor with Ig and ITIM Domains
TIM3	T cell Immunoglobulin and Mucin-Domain Containing-3
TLS	Tertiary Lymphoid Structures
TMA	Tumour Microarray
TME	Tumour Microenvironment
TNFα	Tumour Necrosis Factor
TNM	The TNM Classification of Malignant Tumours
TP53	Tumour Protein P53
Treg	regulatory T cells
tSNE	t-Distributed Stochastic Neighbour Embedding
TXT	Text
UMAP	Uniform Manifold Approximation and Projection
UV	ultraviolet
VISTA	V-Domain Ig Suppressor of T cell Activation
WES	Whole Exome Sequencing
WNT	Wingless/integrated

Chapter 1. Introduction

1.1 Tumour heterogeneity

Cancer is a highly heterogeneous disease, as even tumours of the same type and histological subtype can present a high level of phenotypic diversity across patients. In addition to this inter-tumour heterogeneity, there is also a large amount of phenotype variability within single tumours. This intra-tumour heterogeneity (ITH) is due to the multiple clones of cancer cells forming their tumour and their interactions with the surrounding tissues, which constitute their tumour microenvironment (TME)[1]. Both inter- and intra-tumour heterogeneity involve molecular, cellular and tissue-level features. These include genetic, epigenetic and transcriptional alterations, as well as variability in cell morphology, organisation and migration, such as different propensities to undergo epithelial-to-mesenchymal transition. This heterogeneity has a strong impact on clinical outcomes by contributing to tumour growth, immune evasion, metastasis and therapeutic resistance[2].

Tumour heterogeneity has been extensively studied through data collected and analysed in several projects, including The Cancer Genome Atlas (TCGA)[3], the International Cancer Genome Consortium (ICGC)[4] Pan-Cancer Analysis of Whole Genomes (PCAWG)[5], the Catalogue Of Somatic Mutations In Cancer (COSMIC)[6], the Network of Cancer Genes (NCG)[7] and The Cancer Imaging Archive (TCIA)[8]. However, further analysis in this field is required to characterise the mechanisms by which tumour heterogeneity drives cancer pathogenesis and to develop more effective diagnosis and treatment strategies[1].

1.1.1 Inter-tumour heterogeneity

Genetic and phenotypic variability across tumours of different patients are defined as inter-tumour heterogeneity. Even between patients with the same cancer type and molecular or histological subtype, there can be significant differences in response to therapy and clinical outcomes. This remained the case even after the development of precision oncology approaches and the application of DNA-seq and RNA-seq technologies to the analysis of tumour heterogeneity[9]. While genetic variation is the underlying cause of inter-tumour heterogeneity, inter-tumour heterogeneity can also manifest at the cell and tissue level, with different cancers having different TMEs.

1.1.1.1 Genetic inter-tumour heterogeneity

Genetic inter-tumour heterogeneity is a direct consequence of cancer evolution. Cancers acquire somatic alterations stochastically during their development, and these mutations are then subject to evolutionary pressures due to random drift and selection pressures. This combination of stochastic and deterministic processes shaping cancer evolution causes tumours to have different driver alterations. Additionally, cancer genomes are subject to various mutational processes, which result in specific frequencies of different types of alterations constituting the tumour’s mutational signatures[10]. Cancers have different mixtures of such signatures, which can vary within cancer types and molecular subtypes because of differences in environmental exposures or other intrinsic and extrinsic factors.

Evaluating the impact of genetic inter-tumour heterogeneity on clinical outcomes is challenging due to the relatively low frequency of individual mutations[9]. This issue can be partially addressed by performing analyses at the pathway level to identify gene sets whose inter-tumour heterogeneity is associated with clinical features. Additionally, this approach has been recently expanded through the integration of transcriptomic and epigenetic data; however, a lot more studies are required to fully characterise the clinical impact of inter-tumour heterogeneity[9].

Genetic heterogeneity is directly reflected in epigenetic and transcriptional diversity. Thus, genetic inter-tumour heterogeneity can thus drive inter-patient variability at the tissue level by directly affecting the biological capabilities of the tumour to interact with the surrounding tissue. These interactions are part of the hallmarks of cancer and include the deregulation of the immune microenvironment to avoid immune-mediated killing and induce inflammation and angiogenesis[11].

1.1.1.2 Inter-tumour microenvironment heterogeneity

Due to underlying genomic heterogeneity and the intrinsic characteristic of the affected tissue and organ, distinct tumours can present different TME statuses. These depend on the immune system’s ability to recognise the tumour and respond to it, a process influenced by several factors. These include somatic mutations, antigen expression, signal pathway activity, availability of immune cell populations for recruitment and T helper type 1 (Th1) skewing[12]. Tumour cell recognition by the immune system depends on the presentation of tumour self-antigens to the adaptive immune system. These antigens include aberrantly expressed differentiation, development, or lineage-specific antigens (like cancer-testis antigens) [13], or tumour cell-specific neoantigens derived from the translation of genes harbouring somatic mutations[14]. However, the production of tumour self-antigens is not sufficient to initiate an antigenic T cell response, and the antigens also need to be presented to the tumour cells in an immunostimulatory context. Antigen presentation requires cleaving the neoantigen into peptides that can be presented via the Major histocompatibility complex (MCH) I molecules on the tumour cell surface. Then, the specific T cell receptor (TCR) recognises the antigens bound to MHCI on naïve $CD8^+$ T cells. This antigen presentation generally occurs in the lymph nodes and spleen but can also happen in Tertiary Lymphoid Structures (TLS) at the periphery of tumours that have been subject to long exposures to inflammatory cytokines[13]. After activation, the resulting short-term memory T cells proliferate and differentiate into effector T cells, effector memory T cells, central memory T cells, and cytotoxic lymphocytes (CTLs) [14]. The resulting anticancer immunity can then develop into one of three main TME phenotypes: the immune-desert, the immune–excluded, and the inflamed phenotype [14].

The immune-desert TME is characterised by minimal inflammation, and though myeloid cells may be present, there is little or no $CD8^+$ T cell infiltration. This low immune activity is probably due to immunological ignorance, the induction of tolerance or a lack of appropriate T cell priming or activation. All these factors prevent the generation of tumour-specific T cells at a rate sufficient to mount an effective response[15].

The immune-excluded TME phenotype instead is characterised by abundant immune cells in the outlying stroma and unable to infiltrate the tumour[14]. Immune cell infiltration is inhibited or blocked through various mechanisms, including inhibitory chemokines, fibrotic nests, vascular factors or barriers, or specific stromal-based inhibition. Immune excluded TMEs can be found in different epithelial cancers, such as colorectal carcinoma (CRC), melanoma and pancreatic ductal adenocarcinoma (PDAC). Immune-excluded TMEs contain CTLs with low expression of the activation markers Granzyme B (GzB) and Interferon γ (IFNγ) compared to tumours with a more inflamed phenotype [15].

Inflamed tumours have TMEs with high levels of infiltration by active CTLs with high expression of Programmed cell death protein 1 (PD1), GzB, and IFNγ. These TMEs are also characterised by high levels of proinflammatory cytokines potentially driving T cell activation and expansion, such as type I and II IFNs, Tumour Necrosis Factor α (TNFα), Interleukin 2 (IL2), IL12, IL23, IL1β [15]. Additionally, many other immune cell types can be present, including B cells regulatory T cells (Treg), myeloid-derived suppressor cells (MDSC), and cancer-associated fibroblasts (CAF). These can express inhibitory factors and checkpoint molecules such as Programmed death-ligand 1 (PDL1) that inhibit CD8$^+$ T cell response and cause their exhaustion. Finally, a subset of inflamed tumours can develop TLS, lymphoid aggregates with a cell composition similar to lymph nodes. They are generally located in the stroma or the tumour invasive margin and can act as additional centres of antigen presentation and lymphocyte activation. TLSs often correlate with a favourable prognosis[16].

1.1.1.3 Inter-tumour heterogeneity of CRC and MSI-CRC

CRC classification initially relied on clinicopathological characteristics. However, tumours with the same histologic features and tumour stage could still have different prognoses and responses to therapy[17]. More recently, the mutation status of genes such as *Rat sarcoma virus protein* (*RAS), v-Raf murine sarcoma viral oncogene homolog B (BRAF), Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit* α *(PIK3CA)* and other initiating molecular events have been employed as biomarkers for the informed clinical management of CRC, together with microsatellite instability (MSI). According to this last feature, CRCs are generally divided into MSI tumours and Microsatellite stable (MSS) tumours[18].

MSI tumours constitute about 12-15% of all CRCs and arise from deficient DNA mismatch repair (MMR), which can be due to germline mutations in MMR genes, or more commonly to *MutL Homolog 1 MLH1* gene inactivation from CpG island methylation[19]. This type of sporadic CRC is also known as the CpG island methylator phenotype (CIMP) and often presents somatic mutations in *BRAF*. MSI tumours are generally located in the proximal and right colon and often present hypermutation. These tumours are commonly poorly differentiated and rich in tumour-infiltrating lymphocytes.

MSS tumours that arise in 65–85% of CRC patients are instead characterised by chromosomal instability (CIN), which does not lead to a hypermutated phenotype[19]. The exact mechanisms underlying CIN in MSS CRC are not well understood as large structural variations copy number alterations, aneuploidy, mutations in *TP53*, and other tumour suppressors can play a role[19].

Genomic instability induced by replication stress, telomere erosion, and promoter hypomethylation can also drive CIN in CRC. In contrast to MSI patients, CIN MSS patients have a relatively unfavourable prognosis, especially in the early stages[18].

MSI status is used to guide clinical management together with the Classification of Malignant Tumors (TNM) staging, and the status of *RAS* and *BRAF* help guide clinical management[17]. These factors inform the administration of adjuvant therapy, and the mutational status of *Kirsten rat sarcoma virus protein (KRAS)* and *Neuroblastoma RAS Viral Oncogene Homolog* (*NRAS)* guides the application of anti- Epidermal growth factor receptor (EGFR) drugs to metastatic CRC[17]. Finally, as described in section 1.2, these features are also biomarkers for response to anti-PD1 immunotherapy. However, these molecular markers do not fully explain the substantial differences in outcome and therapy response observed across CRC patients, even when other pathways such as Wingless/integrated (*WNT)* are considered[19]. Thus, several approaches relying on transcriptomic and epigenetic analysis and somatic mutations have been proposed in recent years. These efforts were harmonised into a more comprehensive classification system known as the consensus molecular subtypes (CMS) of CRC[20]. The CMS system relies on transcriptomic data to classify CRCs into four molecular subtypes: CMS1, CMS2, CMS3 and CMS4 (Figure 1.1).

The CMS1, also known as MSI immune subtype, represents about 14% of total CRCs[19]. CMS1 tumours have MMR defects with an MSI hypermutated phenotype and CIMP, often with *BRAF* mutations and a low level of CIN. The presence of abundant neoepitopes due to the hypermutated phenotype often results in high CD8$^+$ T cell infiltration and high Th1 cell activity[14] (Figure 1.1a). However, immunosuppressive cells with pro-tumour cytokines and immune-checkpoint protein expression often inhibit this immune response. This immunosuppressive environment is thought to be the reason why CMS1 tumours generally respond better to checkpoint-blockade treatment than CMS2–4 tumours[18].

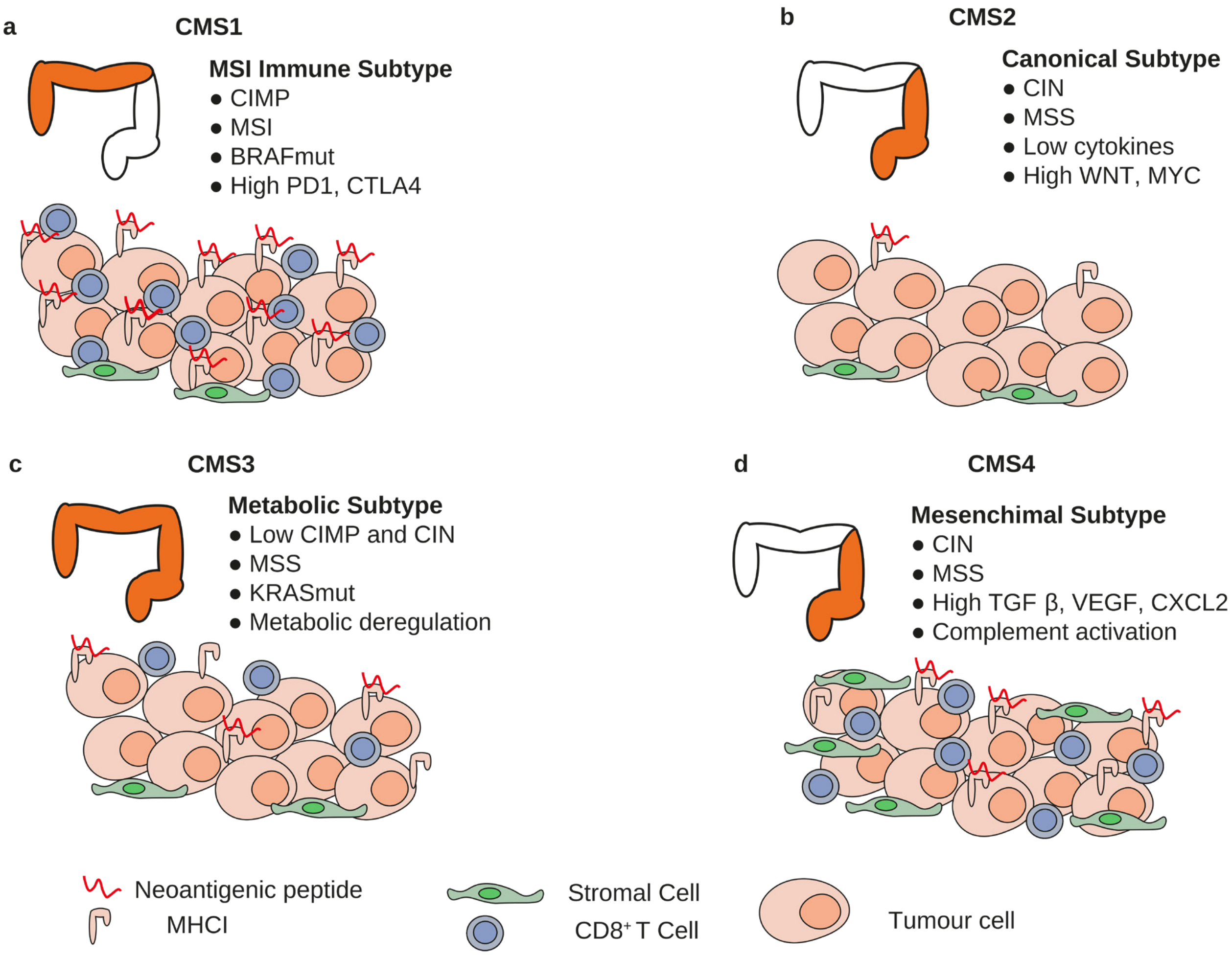
CMS2 or canonical subtype accounts for 37% of all CRCs and is characterised by an MSS phenotype with high CIN[19]. These tumours have high levels of CIN and *TP53* mutations. These tumours have WNT and V-Myc Avian Myelocytomatosis Viral Oncogene Homolog (MYC) pathway activation and elevated EGFR activity (Figure 1.1b).

CMS3, also known as metabolic CMS, constitutes 13% of all CRCs[19]. CMS3 has mixed genomic and epigenomic features with low MSI and CIMP status with frequent *KRAS* and *PIK3CA* mutations. The expression profile of CMS3 tumours is characterised by high dysregulation of multiple metabolic signatures, active WNT and MYC signalling (at lower levels than CMS2 tumours) and IGBP3 overexpression (Figure 1.1c). The TME of CMS2 and CMS3 CRCs is immune cold with low immune and inflammatory activity and mostly PDL1 negative, as expected from antigenically cold tumours[18].

CMS4 tumours constitute about 23% of all CRCs and have a mesenchymal phenotype due to the upregulation of genes involved in epithelial-to-mesenchymal transition, remodelling of the extracellular matrix, and angiogenesis[5]. Additionally, the TME of these tumours is skewed towards immunosuppression as evidenced by the high expression of pro tumour genes, including the *Transforming Growth Factor β* (TGFβ) pathway, and gene signatures associated with T helper type 17 (Th17) cells and monocytes/macrophages (Figure 1.1d). This immunosuppression results in lower CTL infiltration and a worse prognosis compared to CMS1 CRCs[14]. Finally, the remaining 13% of CRCs fall into a mixed group whose expression signatures do not allow its classification in any of the four previous CMS[19].

The clinical applicability of the CMS system is, however, limited by intra-tumour heterogeneity, which can result in significantly different prognoses even across tumours with the same phenotype[21]. This issue is further compounded in a metastatic setting, in which the metastasis could have a phenotype vastly different from the primary tumour[21]. Additionally, there can be phenotype variability across tumour regions and between metastatic lesions from the same patient both in space and time. Thus, extensive intra-tumour heterogeneity may significantly hinder the prediction of a tumour prognosis and therapy susceptibility when only bulk data from a single tumour region is available[17].

Figure 1.1 The four CMS of CRC



The CMS classification of CRC[20]. For each CMS (a-d) are reported the more frequent primary site, the most common microsatellite status, main driver mutations and altered pathways. CIMP, CpG island methylator phenotype; CIN, chromosomal Instability; MSI, microsatellite instability; MSS microsatellite stable; BRAF, v-Raf murine sarcoma viral oncogene homolog B; PD1, programmed death protein 1; CTLA4, cytotoxic T-lymphocyte-associated protein 4; WNT, wingless/integrated; MYC, V-Myc Avian Myelocytomatosis Viral Oncogene Homolog; KRAS, kirsten rat sarcoma virus protein; TGFβ, transforming growth factor β; VEGF, vascular endothelial growth factor; CXCL2, Chemokine (C-X-C-motif) ligand 2.

1.1.2 Intra-tumour heterogeneity

ITH comprises all the differences in phenotypes between different regions of the same tumour. Multiregional DNA-seq and RNA-seq analyses first and later studies with single-cell technologies have shown significant phenotype variability across a single tumour[9]. While genetic variation is the main cause of ITH, this variability can then manifest at the epigenetic and cell phenotype levels, resulting in different TMEs in different regions of the same tumours.

1.1.2.1 Genetic intra-tumour heterogeneity

The stochastic acquisition of somatic mutations after the transformation of the first tumour cell is the leading cause of genetic ITH. These new somatic mutations can be drivers that increase the fitness of their host cells or have a negative or null effect on fitness. This latter case is the more common, and these so-called passenger mutations constitute the vast majority of mutations picked up by cancer sequencing screens[9]. During the life history of the tumour, malignant cells and their genomes undergo evolutions according to drift and selection[22]. Drift is the variation of allele frequencies in the population due to random external factors impacting the survival and reproduction of different lineages independently of their genotypes. Selection is non-random and depends on the lineage having a genotype, which confers a survival or reproductive advantage over other lineages[22]. The genotypes of the lineage with the higher fitness will thus have a higher frequency in the following generations[22].

As a result of these evolutionary forces, by the time that tumours present clinically, they consist of multiple distinct subpopulations or clones. While clonal mutations are shared by all cancer cells, subclonal mutations are present only in a subset of one or more clones. Subclonal mutations may affect fitness when the TME changes because of increased hypoxia, immune response, therapeutic agents, or after metastatisation[2]. Thus, genetic ITH significantly impacts the tumour prognosis and response to treatment.

The origin of genetic ITH is still being investigated. While point mutations are well understood, they might not be the leading cause of ITH. CIN is a feature of most human cancers, often linked to whole-genome duplication and aneuploidy[23]. These large-scale genomic rearrangements cause much faster rates of genomic mutation, thus increasing ITH[23]. The impact on the phenotype of these large genomic changes is, however, difficult to interpret, as they generally involve several genes and thus impact whole regulatory networks[24]. The few exceptions are cases of complete allele loss or massive focal amplification of an individual or few genes[23].

Genetic ITH can quantify from DNA-seq data by calculating the cancer cell fraction of each somatic mutation from its mutation allele frequencies after adjusting for local copy number and sample purity[9]. The subsequent clustering of these mutations according to their cancer cell fractions allows the estimation of the number of distinct subclones and their relative sizes[9]. This approach is limited by the very high sequencing depth required to distinguish mutations with a low allele frequency from random noise[22]. Additionally, single-sample analyses can underestimate ITH as the sample is not fully representative of the whole tumour, and clonal mutations in one area can be subclonal in another[25]. Multi-region sequencing explores ITH in multiple regions, but deriving a unified clonal landscape from multiple regions is a complex computational problem[9].

Further understanding of ITH and tumour evolution has been achieved through the application of single-cell sequencing technologies, including single-cell DNA and RNA sequencing[25]. These new approaches allow the investigation of ITH at single-cell resolution but are limited by high rates of missing values due to allele drop-out and other technical factors[9].

ITH is hard to study with genetically engineered mouse models, as most available lines rely on powerful combinations of driver mutations within a single cell[2]. These models provide convenient and reproducible experimental systems to study molecular mechanisms, but they rarely display the degree of subclonal and cellular genetic heterogeneity seen in spontaneous cancers[2]. This problem has been partially overcome through the combined use of low-penetrance oncogenes drivers with a source of genetic diversification, such as mutagens or transposons[2]. The use of such models will enable the assessment of the functional relevance of subclonal interactions in tumour evolution in higher detail[2].

While most studies focus on genomic alterations, epigenetic diversity is an essential component of ITH. Epigenetic heterogeneity is involved in the acquisition of traits linked to metastatisation and drug resistance[2, 26]. However, epigenetic changes can be highly plastic: while some changes like the DNA hypermethylation of promoters are stable, others can last only a few cell divisions[2]. Thus, studies on epigenetic ITH mainly focused on DNA methylation[24]. Advances in single-cell sequencing technologies have enabled the single-cell multiomic profiling of tumours, including both genomic and epigenomic analyses for DNA methylation and chromatin accessibility[2].

ITH is not limited to differences in subclones with different genetic and epigenetic statuses, as ITH also involves differences in the TME[26]. This variability in the TME is due to clones with specific epigenetic and genetic makeups interacting with the local environment in different ways, thus giving rise to TME-related ITH[24, 25].

1.1.2.2 Intra-tumour microenvironment heterogeneity

As tumour cells evolve and diversify during tumour evolution, so does their microenvironment due to the dynamic interactions between the TME and tumour cells[27]. In turn, the TME can shape tumour evolution through immune response, cell signalling and other cell-extrinsic factors like pH, oxygen and nutrient availability[2]. The subsequent alteration of the tissue environment can then result in abnormal juxtacrine and paracrine signalling, which contributes to selection and further diversifies cancer cell phenotypes[2]. Finally, the tissue organisation is wholly lost in invasive and metastatic tumours, resulting in highly altered and heterogeneous TMEs and significantly affecting tumour progression and treatment outcomes. These outcomes are shaped by both structural and immunological aspects of TME heterogeneity[27].

Structural ITH is caused by the non-uniform remodelling of several tissue features, including blood and lymphatic vascularisation, as well as the extracellular matrix (ECM) and cell migration[2]. While most tumour cells are separated from stroma-derived ECM by more than one cell layer, some are still in contact with an ECM whose composition is altered compared to normal tissues[2]. Additionally, the distribution of blood and lymphatic vasculature in tumours is generally disorganised and together with the layout of nearby epithelia. These alterations cause significant spatial and temporal variability in oxygenation, availability of nutrients and growth factors, and pH[28]. The combination of gradients of all these factors generates a wide range of microenvironments and thus leads to phenotypic ITH. For example, a decrease in pH at the tumour periphery has been observed to drive the epithelial-to-mesenchymal transition, mediating tissue invasion[28]. Finally, distinct microenvironments can act as evolutionary niches by causing different selective pressures and thus increasing spatial genetic ITH, which has been correlated with poor prognosis[24].

Immunological TME ITH is also strictly linked to genetic ITH. The main way the immune system influences the evolution of ITH is through the three stages: elimination, equilibrium and escape[24]. At the beginning of tumour development, distinct tumour clones produce different sets of neoepitopes from their specific set of somatic mutations. The clones harbouring the strongest neoepitopes are more immunogenic and are thus negatively selected. In turn, the T cells populations with TCRs targeting these strong neoepitopes expand, and the less tumorigenic tumour clones grow[24]. This process can play a significant role in shaping ITH as subclonal neoepitopes have been observed to have a higher immunogenic potential than clonal ones[29]. Immune elimination processes lead to an equilibrium state, in which the immune system maintains net tumour growth to a minimum}[2]. Finally, in the escape phase, the tumour acquires the ability to escape the immune system and progresses to unrestrained growth[24].

Immune escape can be achieved through multiple mechanisms. These include reduced antigen presentation either by loss of heterozygosity at the human leukocyte antigen (HLA) locus or epigenetic repression of neoantigen express, and generation of an immunosuppressive microenvironment leading to immune exclusion or immune exhaustion[24]. The impact of these processes was observed in a multi-region study showing that high densities of CD8$^+$ and CD4$^+$ T cells were associated with high TCR diversity, but these factors did not correlate with genetic heterogeneity[29]. The same study also observed immune-cell-excluded and inflammatory microenvironments in multiple metastases of the same ovarian cancer patient[29].

Other cell populations also contribute to diversity, including antigen-presenting cells and tumour associated fibroblasts (TAF). The latter often comprise multiple diverse populations, which can have different tumour promoting functions like driving a migratory phenotype and Epithelial-to-Mesenchymal Transition (EMT) in tumour cells[24].

Thus, ITH is determined by a combination of genetic, epigenetic, and TME-related factors[27]. The resulting phenotypic ITH then directly impacts response and resistance to therapies[2, 26].

1.1.2.3 Intra-tumour heterogeneity of CRC

CRC is a highly heterogeneous disease, and survival rates vary highly across patients sharing the same TNM stage[30]. This heterogeneity is due to multiple genetic and non-genetic factors[31]. Genetic ITH in CRC is observed in most patients[32] as various multiregional sequencing studies have identified subclonal mutations for many of the main driver genes of CRC, including *KRAS* and *NRAS*[33] as well as *Adenomatous Polyposis Coli (APC)*, *TP53*, and *Erb-B2 Receptor Tyrosine Kinase 4* (*ERBB4)* mutations[34]. Finally, the variant allele frequency of *KRAS*, *NRAS*, *PIK3CA*, or *BRAF* mutations was observed to be highly variable across samples from a single patient[35].

Generally, there is a very high concordance in MSI status and clonal driver mutations between the primary tumour and associated metastases[31]. However, differences in *KRAS* and *TP53* mutations and 18q loss have been reported[36]. This high similarity between primary tumours and the associated metastasis has also been observed at the gene expression level[31].

Histopathological examination of phenotypic ITH in CRC has shown that more than 50% of MSI tumours present different growth patterns with glandular, mucinous, or medullar morphologies[37]. While mixed morphologies were only observed in 10% of MSS tumours, differences in immune infiltration and inflammation status were also observed within one tumour. These variations are highly correlated with the morphology of the tumour, and thus higher ITH was observed in MSI compared to MSS tumours[37].

ITH in CRC is directly reflected in patient survival and response to therapy[30]. CRC patients with higher ITH have shorter survival: the three years overall and progression-free survival are 66% and 23% for patients with low metastatic ITH, but only 18% and 5% and for high ITH patients[38]. High ITH was also associated with a higher incidence of liver metastasis[39] and treatment resistance[30]. Interactions between subclones can also lead to stronger resistance to treatment. For instance, *KRAS* wildtype CRCs generally respond to cetuximab, while *KRAS* mutant CRCs are often resistant. However, *KRAS* mutant subclones can produce *Transforming Growth Factor TGFα* and amphiregulin, which can induce *KRAS* wildtype cells to grow continuously despite treatment[40].

1.2 Cancer Immunotherapy

Several immunotherapeutic agents have been developed over the last decades, and many saw application in the clinic. The mechanisms targeted by cancer immunotherapy differ significantly from those behind chemotherapy or oncogene-targeted therapies. Immunotherapy leverages the patient's immune system to produce a dynamic anticancer response not limited to a single oncogenic target or the high proliferation rates of cancer cells[41].

Immunotherapy agents include immune checkpoint inhibitors (ICIs) and autologous T cells engineered to express a CD19-targeting chimeric antigen receptor (CAR) [42]. The applications of other types of agents are also being investigated, including immunostimulatory monoclonal antibodies, small molecules immunosuppression inhibitors, as well as therapeutic vaccines[42]. This section will focus on ICIs as they are the immunotherapy agents currently used in CRC[43].

1.2.1 Immune checkpoint Inhibitors

As described in section 1.1.2.2, the tumour can escape elimination by the immune system through several mechanisms, including the activation of negative regulatory pathways to suppress the immune response. Most of these pathways depend on immune checkpoints cell surface receptors, which regulate the activation and function of T cells[15]. The physiological role of immune checkpoints is the maintenance of self-tolerance, but tumours frequently exploit them to suppress the immune response[42]. So far, ICIs against three targets have been approved for clinical use by the United States Food and Drug Administration (FDA)[44]. The first was ipilimumab, an antibody against Cytotoxic T-Lymphocyte Associated Protein 4 (CTLA4)[41]. CTLA4 negatively regulates T cell activation through competition for binding with the shared ligands CD80 and CD86. Ipilimumab was approved first for the treatment of advanced-stage melanomas and then of various other cancer types[41].

The second ICIs are also antibodies and target the PD1 receptor. These drugs, named pembrolizumab and nivolumab, were approved first in melanoma and then also for several other cancer types[44]. Finally, the third target of ICI antibodies is PDL1. Three antibodies have been approved as anti-PDL1 agents by the FDA: atezolizumab, durvalumab, and avelumab, which are applied mostly to urothelial carcinoma, non-small-cell lung cancer, and Merkel cell carcinoma[41].

Anti-PD1 and anti-PDL1 antibodies have become more commonly employed than anti-CTLA4 agents because of the lower side effects and higher efficacy[44]. The following section will focus on anti-PD1 agents. In addition to the currently approved therapies, many antibodies and small molecules are in active clinical development. These new ICIs target different checkpoints, including Lymphocyte-activation gene 3 (LAG3), T cell immunoreceptor with Ig and ITIM domains (TIGIT), T cell immunoglobulin and mucin-domain containing-3 (TIM3), CD276, CD39, CD73, and CD47[44].

Despite the great research development efforts behind ICI therapies, most patients do not achieve a durable clinical benefit. For instance, 60%-70% of melanoma patients treated with anti-PD1 therapy do not respond[45], and out of all responders 20–30% have a later relapse[45]. Cases with no response to therapy are classified as primary resistance, while cases where the response is not maintained are categorised as acquired resistance[46]. The understanding of the mechanisms of both types of resistance is still incomplete despite considerable research efforts in this direction, as it will be vital to unlocking the full potential of anti-PD1 agents and other ICIs[46].

1.2.2 Anti-PD1 immunotherapy

Anti-PD1 and anti-PDL1 agents operate by blocking the interactions between PD1 and its ligand that activate this negative regulation pathway. PD1 is expressed by activated T cells, natural killer cells, B cells, macrophages, monocytes, and Dendritic Cells (DCs), but its role is best known in tumour-specific T cells where it is expressed at the highest levels[47]. A broader range of cell types expresses its ligands PDL1 and Programmed death-ligand 2 (PDL2). PDL1 is expressed by several immune cell types, tumour cells, endothelial cells and epithelial cells. PDL2 instead is expressed mainly by activated DCs and macrophages[47].

The binding between PD1 and its ligand on tumour-specific T cells causes the phosphorylation of multiple tyrosine residues in the cytoplasmic region of PD1[48]. The phosphorylation of these residues enables the recruitment of multiple phosphatases, including Src homology 2 domain-containing protein tyrosine phosphatase (SHP1) and (SHP2)[49]. These phosphatases inhibit the stimulatory signals produced by the TCR through MHCI and CD28 interactions[48]. For instance, CD28-mediated signalling is inhibited through interactions with CD80 or CD86, while other downregulated downstream signalling pathways include the PI3K–AKT and the RAS signalling pathways[49]. This inhibition causes the decreased activation of several transcription factors, including Nuclear Factor kappa B (NF-κB) [48] and prevents the activation of pathways necessary for maintaining T cell activation, proliferation, and effector functions, including cytotoxicity and survival[49].

Another mechanism of T Cell inhibition by PD1 is the increase of expression of another series of transcription factors, including Basic Leucine Zipper ATF-Like Transcription Factor (BATF), which further suppress T cell function[49]. This downregulation causes also reduced production of TNFα, IFNγ, and IL2, further inhibiting the anti-tumour immune response[48]. Finally, PD1 is expressed at high levels also on Treg cells. PD1 signalling can stimulate the proliferation of Tregs, which in turn contribute to the creation of an immune-suppressed microenvironment[15].

Several tumour features have been investigated as predictors of response to anti-PD1 ICIs. First, tumours with an immune-inflamed phenotype generally have higher chances to respond[48]. PDL1 expression is also employed as a biomarker because its presence is necessary for PD1-PDL1 interactions to occur[48]. Additionally, PDL1 expression is increased in response to IFNγ signalling and thus reflects the activity status of the immune response, linking its expression to CD8$^+$ T cell responses and antigen presentation[42]. Other markers of active immune response correlate with PDL1 expression, including granzymes and Chemokine (C-X-C motif) ligand 9 and 10 (CXCL9 and CXCL10)[15]. However, tumours with low levels of PDL1 can still occasionally respond to anti-PD1 therapy, and high PDL1 expression is not sufficient to predict response to anti-PD1 therapy[15].

TMB is also an emerging predictor of anti-PD1 response in multiple cancer types[46]. High TMB is associated with a higher response rate and longer survival in patients treated with ICI. This is likely due to higher neoantigen production, which increases the likelihood strongly immunogenic neoantigens will elicit a strong immune response when PD1 inhibition is released by ICI therapy[48].

Prediction of response to ICI is also hindered by the large variety of mechanisms, which can lead to primary or acquired resistance[46]. For instance, mutations in genes involved in antigen processing and presentation can cause ICI resistance. Loss or downregulation of MHCI components such as *β2-Microglobulin* (*B2M)* can result in impaired antigen presentation to cytotoxic T cells[46]. Alternatively, oncogenic signalling can alter TME composition and reduce T cell infiltration for alterations in β-catenin/WNT can reduce immune response through lower production of C-C Motif Chemokine Ligand 4 (CCL4) leading to lower levels of DC infiltration[46]. Additionally, higher expression of other immune checkpoint molecules such as CTLA4, TIM3, LAG3, and V-domain Ig suppressor of T cell activation (VISTA) correlates with anti-PD1 resistance[46]. Another cause of resistance is the presence of different populations of PD1$^+$CD8$^+$ T Cells, not all of which can respond to anti-PD1 treatment[49]. Finally, other immune cells can also influence the TME in ways that affect the anti-PD1 response, including Tregs, MDSCs, T helper type 2 (Th2) T cells, and M2 Tumour Associated Macrophages (TAM)[41]. These cell types can generate an immune-suppressive TME that prevents an effective anti-tumour immune response even after the release of PD1-mediated inhibition on effector T cells[48]. A more comprehensive understanding of all these factors is required to increase the clinical effectiveness of anti-PD1 ICI[46].

1.2.3 Anti-PD1 immunotherapy in CRC

As described in section 1.1.1.3, CRCs can be divided in two phenotypes MSI and MSS CRC. The latter constitute the vast majority of CRC patients and generally do not respond to anti-PD1 immunotherapy[50]. This is thought to be likely caused by their lower mutational burden, leading to fewer neoantigens and lower levels of T Cell infiltration. These tumours generally belong to CMS2, 3 and 4, MSI tumours instead typically belong to CMS1 and are rich in tumour-infiltrating lymphocytes[19]. Thus, MSI tumours are the main targets of anti-PD1 ICI agents in CRC[43].

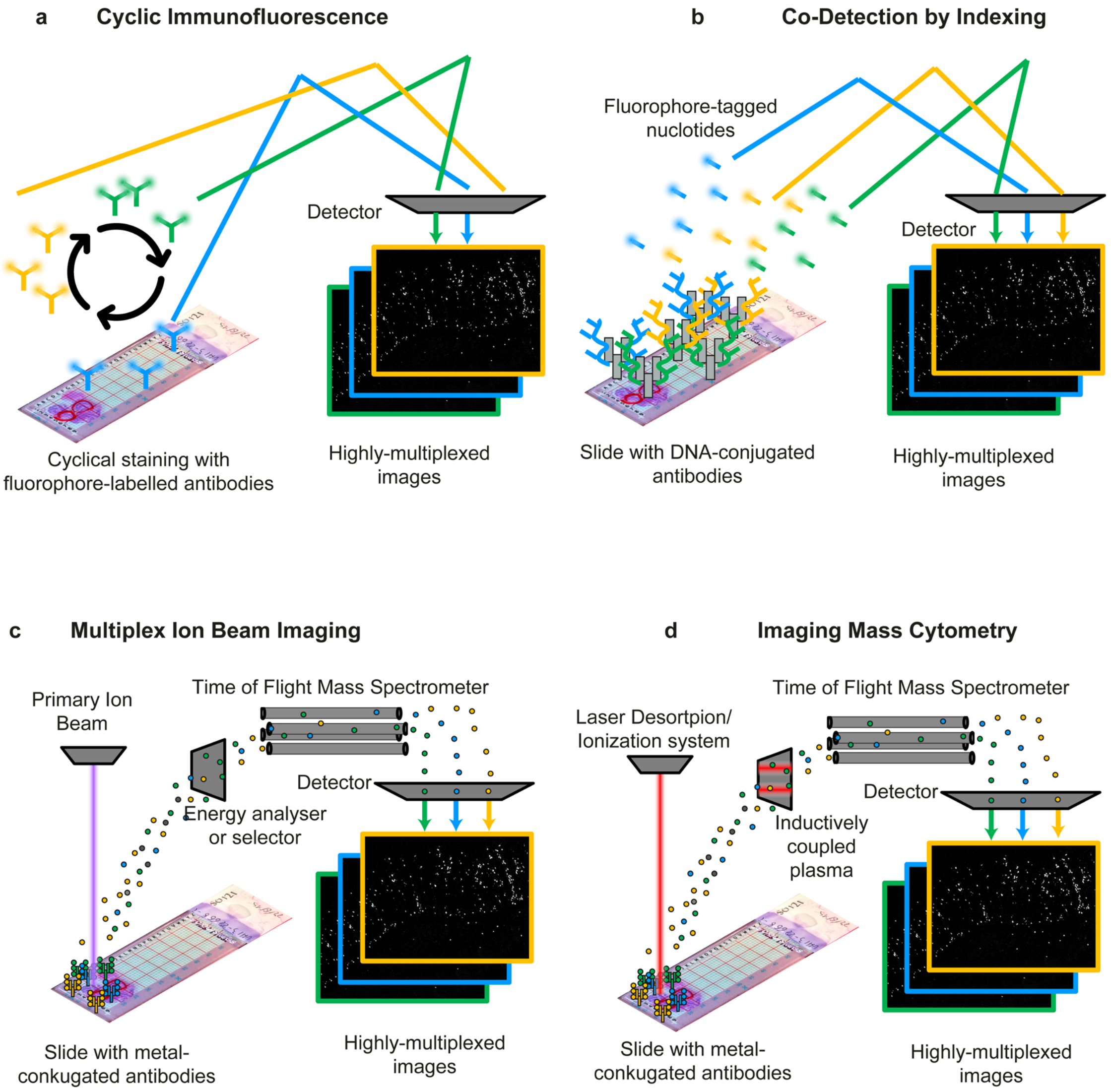
The two anti-PD1 ICI drugs, pembrolizumab and nivolumab, have been approved by the FDA for patients with metastatic MSI CRC. Additionally, the anti-CTLA4 agent ipilimumab has been approved by the FDA for use in combination with nivolumab in metastatic MSI CRC patients previously treated with chemotherapy[51]. The effectiveness of anti-PD1 ICI has been evaluated in multiple clinical studies. For instance, Keynote-177[52] is a phase III trial of first-line pembrolizumab in stage IV MSI CRC which had a 24 months progression free survival rate of 48.3% compared to the 18.6% achieved through chemotherapy[52]l. Additionally, the phase II trial Checkmate 142 trial[53] on combined nivolumab and low-dose ipilimumab therapy in untreated stage 4 dMMR–MSI CRC patients had an objective response rate and a disease control rate of 60% and 84% respectively, with complete response in 7% of patients[53]. The 12 months progression free and overall survival rates were 77% and 83%[53]. Despite these very positive results, there is still a large fraction of MSI CRC patients who do not respond to anti-PD1 therapies. Additionally, these trials evidenced the presence of patients who develop acquired resistances. Altogether, these factors suggest that factors other than the tumour mutational burden are determining for response to anti-PD1 therapies[43].

In addition to the mutational burden, another predictor of immunotherapy response could be the level of tumour-infiltrating $CD3^+CD8^+$ lymphocytes. This can be quantified through the assignment of an immunoscore calculated from density these T cells in the tumour core and invasive margin[54]. Other studies employed gene expression signatures to quantify intra-tumoural cytotoxic T cell infiltration, but the predictive potential of these signatures has not been fully explored in CRC[50].

PDL1 expression measured by immunohistochemical staining could represent a predictive biomarker of anti PD1 response in some tumour types including non-small cell lung cancer and gastric and gastroesophageal junction tumours[10]. However, PDL1 expression was not a predictor of response or survival in CRC junction[10]. Additionally, acquired mutations in Janus Kinase 1 and 2 (*JAK1* and *JAK2)*, which have been identified as markers of anti-PD1 resistance in other cancers, have a still uncertain role in CRC[43]. Truncating mutations in *B2M* were also observed in CRCs that developed resistance to pembrolizumab[50]. Finally, the sporadic or Lynch syndrome associated aetiology of MSI CRCs is not predictive of anti-PD1 therapy response[43].

Further research into the mechanisms underlying response to PD1 beyond the tumour mutational burden is needed. An in-depth investigation of the TME in metastatic MSI CRC is required to identify novel biomarkers of response and clinical strategies[43]. For this reason, tissue-level analyses of the TME in CRC and other cancer types are being conducted with the techniques and approaches described in the following sections.

Figure 1.2 Examples of highly-multiplexed imaging technologies


a
Cyclic Immunofluorescence
Detector
Cyclical staining with fluorophore-labelled antibodies
Highly-multiplexed images
b
Co-Detection by Indexing
Fluorophore-tagged nuclotides
Detector
Slide with DNA-conjugated antibodies
Highly-multiplexed images
c
Multiplex Ion Beam Imaging
Primary Ion Beam
Time of Flight Mass Spectrometer
Detector
Energy analyser or selector
Slide with metal-conkugated antibodies
Highly-multiplexed images
d
Imaging Mass Cytometry
Laser Desortpion/ Ionization system
Time of Flight Mass Spectrometer
Detector
Inductively coupled plasma
Slide with metal-conkugated antibodies
Highly-multiplexed images

Schematic representation of four of the highly-multiplexed imaging technologies reported in table 1 (a-d). For each imaging technology are reported the type of antibody conjugates employed in the staining process, and the detection system needed.

1.3 Tissue-level analysis

Animal cells are always acting as part of the higher-order organisations represented by tissues, formed of cells of multiple types, extracellular matrix and signalling molecules. These form a microenvironment whose composition, structure, and interaction are responsible for the functions of the tissue within the organ. Tissue-level interactions underlie several disease conditions, including cancer, and determine the outcome of therapies acting directly on tissue environment, such as cancer immunotherapy[14]. For these reasons, investigating the dynamic organisation of tissues is of paramount importance for the understanding of most physiological and pathological functions[55]. In the case of the tumour microenvironment and its relationship with immunotherapy, there are still several open questions that need to be addressed at the tissue level[56]. For example, which intrinsic or extrinsic factors enable immune evasion? How do vascular endothelial cells and cancer cells interact to promote growth and metastasis? Does the spatial organisation of cell-cell interactions influence the intra- and inter-clonal heterogeneity in solid tumours? These and other research questions are being addressed with several techniques[56], which will be discussed in the following paragraphs.

1.3.1 Transcriptomics for tissue-level analysis

Since different cell types have different expression profiles, tissue composition can be characterised at the transcriptomic level. The first approaches were based on the bulk RNA-seq analysis of pooled, heterogeneous mixtures of cells from tissue samples. However, these measurements relied on an average quantification of gene expression influenced by the differences in state, phenotype and transcriptional profiles from cells of the same type, which can obscure proportional and subpopulation or state-specific differences[57].

Then, the development of several single-cell sequencing technologies enabled the direct quantification of tissue composition and heterogeneity at the single-cell resolution. These approaches, often combining other single-cell omics targeting genome sequences, protein expression, DNA methylation, and chromatin accessibility, enabled the investigation of several aspects of tissue heterogeneity in cancer[58].

The analysis of cancer tissue composition from expression data is presented in a review[59] I wrote with Mohamed Reda Keddar, Francesca D. Ciccarelli and Lorena Benedetti.

1.3.1.1 Identification of non-cancer cells from cancer transcriptomic data

BBA - Gene Regulatory Mechanisms 1863 (2020) 194445

Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

journal homepage: www.elsevier.com/locate/bbagrm

Review

Identification of non-cancer cells from cancer transcriptomic data[☆]

Michele Bortolomeazzi[1], Mohamed Reda Keddar[1], Francesca D. Ciccarelli[*], Lorena Benedetti[*]

Cancer Systems Biology Laboratory, The Francis Crick Institute, London NW1 1AT, UK
School of Cancer and Pharmaceutical Sciences, King's College London, London SE11UL, UK

ARTICLE INFO

Keywords:
Tumour microenvironment
RNA-seq
Gene expression profiles
Cell-specific signatures
GSEA
Deconvolution

ABSTRACT

Interactions between cancer cells and non-cancer cells composing the tumour microenvironment play a primary role in determining cancer progression and shaping the response to therapy. The qualitative and quantitative characterisation of the different cell populations in the tumour microenvironment is therefore crucial to understand its role in cancer. In recent years, many experimental and computational approaches have been developed to identify the cell populations composing heterogeneous tissue samples, such as cancer. In this review, we describe the state-of-the-art approaches for the quantification of non-cancer cells from bulk and single-cell cancer transcriptomic data, with a focus on immune cells. We illustrate the main features of these approaches and highlight their applications for the analysis of the tumour microenvironment in solid cancers. We also discuss techniques that are complementary and alternative to RNA sequencing, particularly focusing on approaches that can provide spatial information on the distribution of the cells within the tumour in addition to their qualitative and quantitative measurements.

This article is part of a Special Issue entitled: Transcriptional Profiles and Regulatory Gene Networks edited by Dr. Federico Manuel Giorgi and Dr. Shaun Mahony.

1. Introduction

Cancers arising from epithelial cells account for 80–90% of all solid cancers [1]. However, cancer cells do not grow in isolation. The malignant epithelium is in fact surrounded by stromal cells including fibroblasts, immune and endothelial cells which altogether form the tumour microenvironment (TME). Stromal cells in the TME sustain and regulate tumour growth, immune evasion and drug resistance mechanisms [2]. In the past decade, the interest of the cancer research community in the TME has progressively grown because of its role in new therapies that target the host immune system [2–4]. In particular, T cells are able to recognise and eliminate tumour cells. However, tumours develop resistance mechanisms preventing T cell activation. Immunotherapies currently used in the clinic have two main mechanisms of action. They either boost the immune response by activating T cells or they restore the immune response that has been inactivated during tumour growth. Anticancer vaccines and chimeric antigen receptor T cells represent successful attempts to activate the anticancer immune response [3]. Immune checkpoint inhibitors release the brakes imposed by tumour cells on T cells, restoring the host antitumour immune response. These drugs are already successfully applied to treat a variety of tumours, including melanoma, lymphoma, lung, renal cell, head and neck squamous, bladder, liver and gastro-oesophageal cancers [3,5]. However, despite their encouraging success, still many patients do not respond to immunotherapy or develop resistance over time. Understanding TME complexity is therefore essential to predict which patients would benefit from immunotherapy, in full agreement with a personalised approach to cancer therapy.

Tumour infiltrating immune cells can be either beneficial or detrimental for cancer development depending on their localisation, abundance and function. For instance, the presence of $CD8^+$ T cells and T helper cells is usually associated with good prognosis [2,6] while myeloid derived suppressor cells are predictive of bad outcome [7]. Therefore, the detailed characterisation of immune infiltrates is being progressively incorporated into the clinical practice [6,8]. A method widely used in the clinic to estimate the abundance of tumour infiltrating immune cells is the haematoxylin and eosin (H&E) staining. Although this staining is not specific for any particular cell type, it has proven to be clinically relevant for several cancer types [6]. For instance, high levels of lymphocyte infiltration estimated from H&E

[☆] This article is part of a Special Issue entitled: Transcriptional Profiles and Regulatory Gene Networks edited by Dr. Federico Manuel Giorgi and Dr. Shaun Mahony.

[*] Corresponding authors.

E-mail addresses: francesca.ciccarelli@crick.ac.uk (F.D. Ciccarelli), lorena.benedetti@crick.ac.uk (L. Benedetti).

[1] Equal contribution.

https://doi.org/10.1016/j.bbagrm.2019.194445
Received 21 June 2019; Received in revised form 20 September 2019; Accepted 7 October 2019
Available online 22 October 2019

staining are predictive of better prognosis in non–small cell lung cancer [9]. More cell-specific methods for the clinical quantification of immune cells include the combination of up to five antibodies to detect the presence of different immune cell populations using immunohistochemistry (IHC) or immunofluorescence (IF) [6]. These chromogenic or fluorescent labelling-based approaches also provide some spatial information on how epithelial and stromal cells are distributed within the tumour. This analysis is however restricted to the small portion of the tumour that can be sliced from a formalin-fixed paraffin embedded (FFPE) cancer block. It therefore may not be representative of the whole tumour mass. Moreover, the number of cell populations that can be identified is limited due to the small number of markers that can be tested. In this respect, serial IF constitutes a major improvement allowing several rounds of sequential staining of the same sections using up to 12 antibodies [10]. Similarly, high-parameter flow cytometry can profile up to 27 markers in disaggregated cells from several centimetres of tumour mass [11]. These approaches are still being developed and are not yet part of the clinical practice.

Similarly, approaches based on the quantification of protein expression with mass-spectrometry can also reveal detailed profiles of the tumour immune infiltrates. Imaging mass cytometry (IMC) [12] and multiplex ion beam imaging (MIBI) [13] allow the simultaneous identification of up to 40 markers in about 1mm^2 of tissue area. IMC and MIBI provide spatial information on the distribution of cells within the tissue, which adds additional layers of relevant information. Other methods rely on mRNA quantification either using fluorescent probes, like the NanoString nCounter [14], or next-generation sequencing (NGS). NanoString nCounter can be applied to slices of FFPE or fresh frozen (FF) tissues leading to the quantification of up to 800 markers. NGS-based approaches like RNA sequencing (RNA-seq) can be applied to bulk cancer samples or to previously isolated single cells. Despite not providing any spatial information, RNA-seq enables a comprehensive and unbiased characterisation of tumour infiltrating immune cells [15,16]. Moreover, the latest advancements in the field of transcriptomics are beginning to provide spatial resolution ranging from a few cells to subcellular levels [17,18].

In this review, we describe the main methods currently used to quantify tumour-infiltrating cell populations, with a particular focus on those based on bulk and single cell RNA sequencing (scRNA-seq). We also comment on alternative and complementary approaches that are emerging for TME characterisation.

2. RNA sequencing of cancer samples

RNA-seq allows the quantification of gene expression and enables the profiling of a number of genes far greater than other approaches based on probes or antibodies. In the context of cancer biology, RNA-seq is a useful tool for tumour classification, patient stratification and for studying response to therapy, [19,20].

2.1. Bulk RNA sequencing

Bulk RNA-seq refers to the sequencing of RNA from the bulk cancer mass and it consists of four steps (Fig. 1A).

The first step is the extraction of RNA from either FF or FFPE cancer samples. FF samples yield higher quantity and better quality RNA and are thus preferentially used in large scale sequencing projects such as The Cancer Genome Atlas (TCGA). However, the vast majority of samples archived in hospital cancer biobanks are FFPE tissue blocks [21]. Paraffin embedding and long-term storage are known to cause the fragmentation of nucleic acids, while crosslinking is a direct consequence of formalin fixation. This usually leads to low quantity and bad quality RNA [21]. A de-modification step in which the RNA is heated in amine-rich or organocatalytic buffers can be performed to revert formaldehyde linkages and improve RNA quality [21]. Independently of the sample source, the quality of the extracted RNA is a key factor for all downstream analysis and should be carefully evaluated. The main RNA quality metric is the 28 s:18 s rRNA ratio, generally expressed as a RNA integrity number (RIN), with a higher value indicating more intact RNA. While there is no consensus on the RIN value to be used as a quality threshold, generally RIN values below 5 can negatively impact the library preparation and sequencing steps [22].

The second step is the depletion of rRNAs that usually constitute > 80% of the total RNA. There are several approaches for rRNA depletion, depending on RNA quality [23]. In one of them, mRNA is enriched through poly-A enrichment using oligo-dT beads. This method generates high quality expression data that strongly correlates with measurements from independent techniques such as microarrays [24]. However, it requires high quality and intact input RNA, because the capture is done with a poly-T primer against the 3′ end of the transcript. Thus, it is not always suitable for FFPE samples [23]. mRNA enrichment can also be achieved through exon capture probes after cDNA synthesis. In a comparative study with matched FF and FFPE tissue, the best correlation between FF and FFPE expression data was obtained with exon capture RNA [23]. However, the coverage is mostly limited to the captured sequences. This, is because the RNA is partially fragmented so the exonic probes will pull down small fragments containing the target sequences. This approach allows to recover RNA fractions of > 98% of the exome [25]. Alternatively, rRNAs can be removed with techniques based on hybridisation, duplex digestion, or not-so-random RT-PCR priming [23].

In the third step the RNA is fragmented, generally by heat digestion with divalent cations. Finally, in the fourth step the fragmented RNA is converted into cDNA and ligated to adapters to generate the library for sequencing. The most commonly used NGS platforms for RNA-seq are HiSeq and MiSeq Illumina.

2.2. Single-cell RNA sequencing

The recent development of high-throughput scRNA-seq technologies allows to profile the transcriptome of thousands of individual cells per sample [26] (Fig. 1B). These approaches mostly differ in the techniques used for single-cell isolation. Cells can be isolated by fluorescence-activated cell sorting (FACS), as in the MARS-Seq method, which performs scRNA-seq on thousands of cells previously sorted into 384-well plates [27]. Alternatively, single cells can be separated using microfluidic chambers. This is achieved through micron-scale well arrays (as in Seq-Well [28]) or by separating cells in aqueous microdroplets forming an emulsion with an oil phase (as in Chromium [29], Drop-seq [30] and inDrop [31]). Microfluidic-based methods require lower reaction volumes and enable the screening of up to hundreds of thousand cells at lower costs [26]. Cell shape and stickiness (for example of fibroblasts or cancer cells) can affect the efficiency of these methods, biasing single cell capture towards certain cell types over others [32]. Due to the high number of cells, sequencing depth is limited to around 50,000 reads/cell, which is sufficient for clustering and identifying different populations [33].

FFPE samples represent a major challenge for scRNA-seq because the tissue cannot be disaggregated to obtain single cells. However, single cells can be isolated using a computer-guided laser capture microdissection (LCM) system. Although this approach has a throughput of hundreds of cells only, it offers the advantage that each sequenced cell can be mapped back to its original location in the tissue [34].

In the case of FACS or microfluidic-based methods, cells are barcoded during the cDNA synthesis step using beads bound to primers containing a cell-specific barcode, a poly-T capture sequence, and a Unique Molecular Identifier (UMI). While cell-specific barcodes are identical within each cell-containing droplet or well, UMI sequences are different and allow the counting of individual mRNA molecules. This reduces the effects of duplicates that can be generated during cDNA amplification [35]. Since a poly-T capture sequence is used, only the 3′

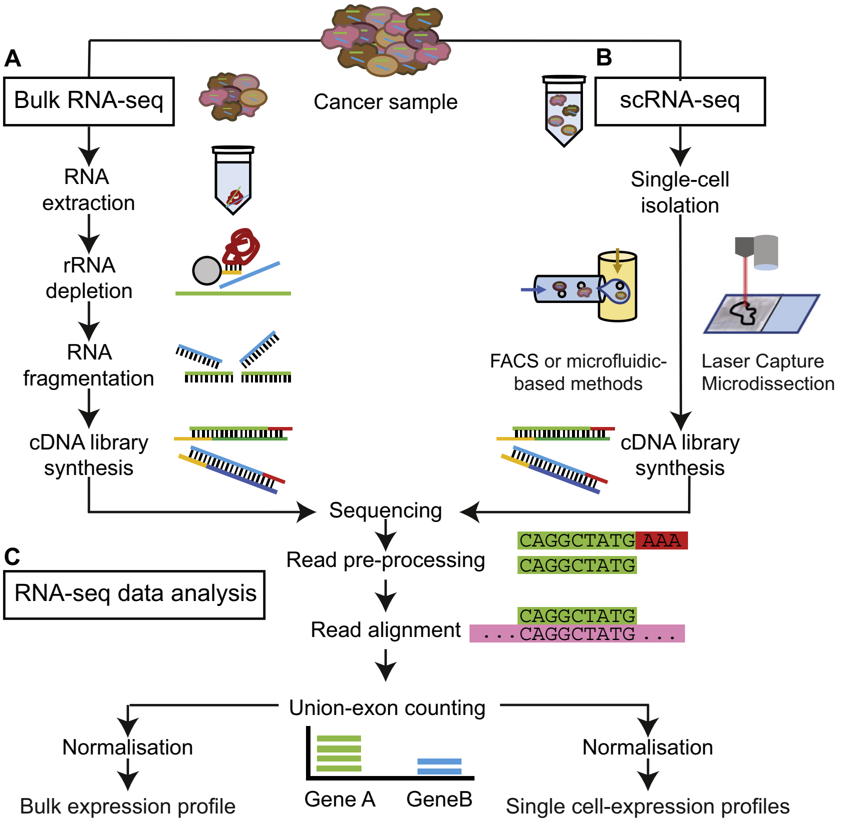


Fig. 1. Workflows of bulk and scRNA-seq experiments. (A) Bulk RNA-seq of solid tumours is based on four steps: RNA extraction from the cancer tissue, rRNA depletion, RNA fragmentation, and cDNA library synthesis for sequencing. (B) scRNA-seq from solid tumour samples requires single cell isolation either through FACS or microfluidics-based methods or laser capture microdissection. cDNA libraries from individual cells are then synthesised and sequenced. (C) Analytical approaches for the quantification of gene expression for bulk RNA-seq and scRNA-seq data. After pre-processing, the reads are aligned to the reference transcriptome or genome. Reads mapping to the exons are counted and normalised to generate gene expression profiles. FACS = fluorescence-activated cell sorting.

RNAs ends are sequenced [26]. This achieves a throughput of thousands to hundreds of thousands of cells, despite the limitations imposed by sequencing cost and capacity. The main drawbacks of poly-T capture beads are that low abundance transcripts may be lost mainly due to limited capture efficiency [36] and it is not suitable for detecting mutations or splicing variants [33].

In the case of LCM, isolated cells are generally sequenced at lower throughput using full-length scRNA-seq protocols, like Smart-seq2 [37]. These methods use Switching Mechanism at 5′End of RNA Template (SMART) chemistry. In this respect the recently developed Smart-3SEQ protocol is particularly suited for FFPE samples [38].

2.3. RNA-seq data analysis

Conceptually similar analytical approaches can be applied to the quantification of gene expression from either bulk RNA-seq or scRNA-seq data [36] (Fig. 1C). In fact, although scRNA-seq-specific methods have started to be developed [36], bulk RNA-seq analysis tools are still successfully applied to scRNA-seq data [39].

The input data for the quantification of gene expression are the raw sequencing reads, which undergo pre-processing to remove adaptor sequences, trim poor-quality bases, and discard low-quality reads, usually derived from poor quality RNA. Libraries with a high number of low-quality reads have lower complexity. This affects the detection of lowly expressed genes and can negatively bias the quantification of gene expression [22].

Pre-processed reads are then aligned to either the reference transcriptome or genome. Aligned reads may undergo post-mapping quality control to evaluate sequence overrepresentation and fragment-size biases. Finally, reads mapping to the exons are counted using union-exon counting methods. In the case of bulk RNA-seq, read counts are normalised to account for gene length and library size and obtain the sample gene expression profile. Different types of gene expression measures can be used, including Reads Per Kilobase of transcript per Million mapped reads (RPKM), Fragments Per Kilobase of transcript per Million mapped reads (FPKM), or Transcripts Per Million (TPM). In the case of scRNA-seq a direct molecule counting based on UMIs can be performed providing an absolute measure of gene expression [35]. If UMIs are not used, scRNA-seq specific normalisation tools can be applied [39]. Moreover, several quality control metrics are usually used to exclude cells with too few or degraded RNA or cell doublets accidentally captured in the same reaction chamber. For instance, bad quality reads or a large percentage of unmapped reads in scRNA-seq can be an index of RNA degradation. Also, high mitochondrial-to-nuclear gene mapping ratio or low mRNA abundance are linked to apoptotic or damaged cells which have lost most of their cytoplasmic mRNAs [39].

The resulting gene expression data from either bulk or scRNA-seq can be used as input for determining the cell-type composition in the sample. Furthermore, scRNA-seq data can be used to define refined cell-type specific expression profiles. These applications of bulk and scRNA-seq are described in detail in the next sections.

3. Quantification of non-cancer cells from bulk transcriptomic data

The bulk transcriptomic profile of a cancer sample is an admixture of transcripts from cancer and non-cancer cells. It therefore offers a qualitative and quantitative representation of the diverse cell types that are present in the sample.

In recent years, many computational approaches have been designed to estimate the abundance of the various cell populations of the TME from bulk transcriptomic data [15] (Table 1). Such approaches leverage on reference signatures consisting of either marker genes and/or expression profile matrices that are specific for a given cell population. Therefore, to quantify the non-cancer component of the TME from cancer expression data, it is paramount to derive robust marker genes or profile matrices. These are generated using a signature derivation pipeline that consists of three main steps (Fig. 2). In the first step, gene

Table 1
Examples of approaches for the quantification of tumour-infiltrating cells from bulk transcriptomic data. For each approach, we report the underlying mathematical method, the type of expression data used to derive the signatures, the total number of marker genes included in the signatures and the final number of non-cancer cell populations considered. Only methods that implement their own reference signatures and that have been applied to the analysis of cancer samples are reported. ssGSEA = single sample gene set enrichment analysis, GSVA = gene set variation analysis.

Approach	Computational method	Source of expression data	Marker genes (n)	Cell populations (n)
Angelova et al. [40]	ssGSEA	Microarray	812	31
Charoentong et al. [41]	ssGSEA	Microarray	782	28
ConsensusTME [42]	ssGSEA	Microarray, bulk RNA-seq	Cancer type specific	18
xCell [43]	ssGSEA and spillover compensation	Microarray, bulk RNA-seq	10,808	64
Tamborero et al. [44]	Scoring (GSVA)	Microarray, bulk RNA-seq	401	16
MCP-counter [45]	Log-transformed geometric mean of expression	Microarray	522	10
Danaher et al. [46]	Log-transformed geometric mean of expression	Microarray, bulk RNA-seq	60	14
ImSig [47]	Arithmetic mean of expression	Microarray, bulk RNA-seq	318	7
CIBERSORT [48]	Deconvolution, nu support vector regression	Microarray	547	22
TIMER [49]	Deconvolution, constrained least square fitting	Microarray	Cancer type specific	6
EPIC [50]	Deconvolution, constrained least square fitting	scRNA-seq	118	10
quanTIseq [51]	Deconvolution, constrained least square fitting	Bulk RNA-seq	153	10

expression data of the cell populations of interest are collected from gene expression databases (e.g. GEO, IRIS, ArrayExpress) and/or from the literature (Fig. 2A). In the second step, these expression data are curated and normalised to allow their comparative analysis (Fig. 2B). Finally, cell type-specific markers (Fig. 2C) or reference expression profile matrices (Fig. 2D) are derived from the normalised transcriptional profiles of cell populations.

3.1. Cell type-specific signatures based on marker genes

A marker gene signature consists of a set of genes that should be expressed specifically by the cell population represented by that signature. The first approaches that were developed to build cell type-specific marker gene signatures used microarray data of purified cell populations (Table 1). One of the first large-scale efforts to build such signatures used microarray-derived expression profiles of immune cells sorted from different tissues, including peripheral blood and bone marrow [40]. Differentially expressed marker genes across cell populations were then identified using ANOVA and further refined by applying a fold-change threshold based on their median expression. Furthermore, as the marker genes of a cell population are expected to be co-expressed, only those with an average correlation coefficient between all other markers of the same population of at least 0.6 were kept. Following this approach, a final set of 812 immune-related marker genes was obtained. The signatures derived from these markers were then used to estimate the abundance of 31 colorectal cancer-infiltrating immune cell populations [40]. The same pipeline was later applied to build signatures for 28 immune cell populations used to characterise the TME of TCGA tumours [41].

Another approach based on signatures derived from microarray data of purified stromal populations is MCP-counter [45]. In this case, however, the area under the curve (AUC) and the signal-to-noise ratio were used in addition to the expression fold-change threshold to select the marker genes. In addition, the signatures were derived taking the hierarchical classification of immune cells into account. This allowed the generation of robust signatures for both parental populations (e.g. all T cells) and subpopulations (e.g. CD8$^+$ T cells). In total, 522 marker genes were derived to define ten stromal cell populations. MCP-counter was applied to estimate the abundance of these populations in a large dataset of non-hematopoietic human tumours [45]. A recent

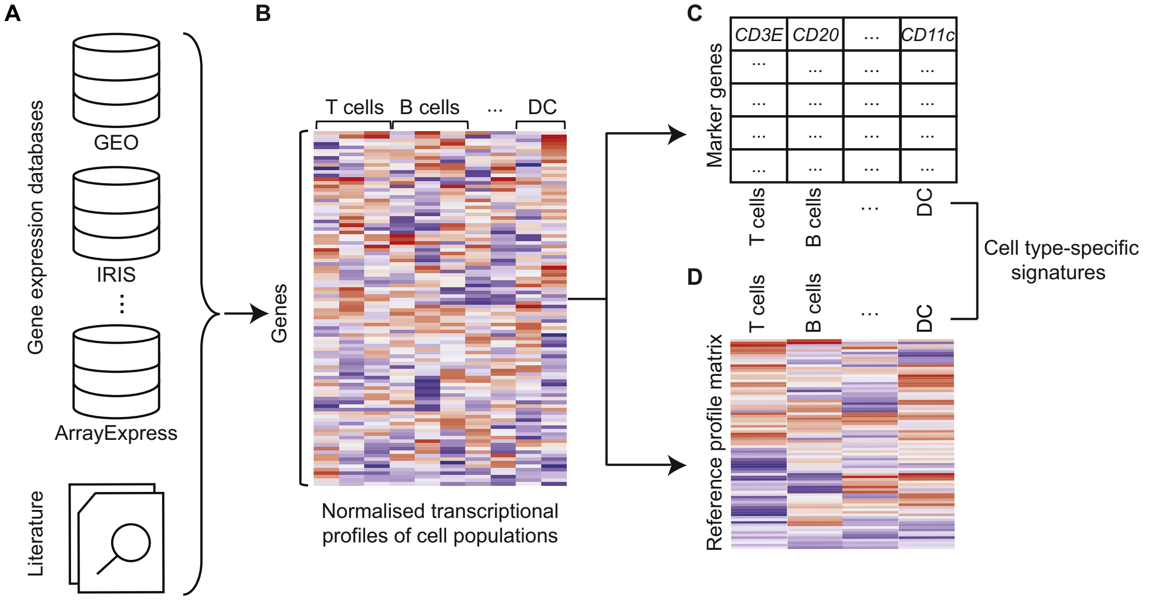


Fig. 2. Computational framework to derive reference signatures. (A) Gene expression data of purified cell populations and marker genes are collected from gene expression databases and/or the literature. (B) They are then normalised to derive cell type-specific transcriptional profiles. (C) Profiles are used to derive cell type-specific reference marker genes through differential expression and correlation analyses. (D) Alternatively, the transcriptional profiles can be aggregated to generate reference expression profile matrices. GEO = Gene Expression Omnibus database, IRIS = Immune Response In Silico database, DC = dendritic cells.

benchmark study found MCP-counter particularly reliable for the comparison of immune infiltrates across samples due to the robustness of its signatures [52]. It performed particularly well in the quantification of B cells, CD8$^+$ T cells, macrophages, natural killer (NK) cells and cancer-associated fibroblasts (CAFs).

More recent approaches started to use cancer RNA-seq expression data to derive marker gene signatures. For example, xCell [43] employs signatures derived from RNA-seq, microarray and Cap Analysis of Gene Expression datasets of tumours and normal tissues from different sources. Unlike other methods, xCell uses more than one signature for each considered population. Signatures were first derived from each data source individually based on marker gene overexpression analysis with different thresholds. Then, for each data source, the top three signatures were kept based on the t-statistic of their enrichment scores (ES) between the cell population they define and all the others. A total of 489 signatures were obtained to define 64 cell populations, making xCell the broadest and most comprehensive quantification approach to date. xCell was applied to characterise infiltrates in TCGA and TARGET data [43]. In the comparative study cited above [52], xCell resulted particularly suitable to estimate the abundance of CD4$^+$ T cells, T regulatory cells and endothelial cells.

In addition to deriving *ex novo* signatures, cancer RNA-seq data has also been used to refine pre-existing signatures to make them more specific for the quantification of infiltrates in tumour samples. Danaher et al. [46] were the first to derive signatures from an initial compendium of 14 previously published immune cell signatures. Using bulk RNA-seq data from 24 TCGA cancer types, the authors measured the co-expression patterns of markers associated with a given signature using a pairwise similarity metric. Then, they built a pairwise similarity matrix for each cancer type and applied hierarchical clustering using the average similarity values across the 24 cancer types. They only considered as final markers for a specific cell type the genes with the highest co-expression patterns across tumours. By using bulk RNA-seq data from the TME, the differences between intratumoral and purified immune cell expression patterns are accounted for [46].

A very similar RNA-seq dataset from TCGA was used to select the most representative signatures from an initial list of marker gene sets obtained from three literature sources [44]. The specificity of the initial signatures was assessed through a correlation analysis using the signature ESs instead of marker gene expression as in other approaches. For each literature source, a pairwise correlation matrix was computed for all the ES of the signatures across the TCGA samples. Sources were discarded when the overall correlation picture of their signatures poorly agreed with biological knowledge. For instance, sources with signatures from cell populations known to be highly co-infiltrated, but that resulted to be negatively correlated, were discarded. Compared to Danaher et al., this approach is less susceptible to the quality of gene expression data, since the correlations are done on the ES values. This strategy yielded a curated set of 16 immune signatures defined by 401 marker genes that were then used to characterise the immune infiltrates in the same TCGA cohort [44].

ConsensusTME [42] is a more inclusive approach as compared to the others because it integrated pre-existing signatures instead of refining them separately. For each cell population, a new set of markers was obtained combining previously defined sets. Additionally, genes whose expression showed a correlation coefficient higher than -0.2 with tumour purity scores derived from 32 TCGA cancer types were filtered out. This step was justified because the correlation of gene expression with tumour purity is indicative of the fact that cancer cells may express these marker genes thus invalidating their specificity for a particular stromal population [42].

In addition to using expression profiles from purified cell populations or refining previous signatures, gene sets can also be derived *ex novo* from bulk transcriptomic data. For instance, ImSig [47] relies on a collection of immune signatures derived from microarray datasets of healthy and disease human samples. For each dataset, a gene correlation network was computed and subsequent clustering was performed to identify modules of co-expressed genes. These modules were then manually annotated to identify those corresponding to immune cell types and extract 318 associated marker genes defining seven immune cell populations. ImSig was applied to characterise the immune infiltrates in TCGA samples [47].

3.2. Cell type-specific signatures based on profile matrices

Instead of sets of marker genes, cell type-specific signatures can also consist of reference expression profile matrices of marker genes in a particular cell population. CIBERSORT [48] was the first tool to use a curated signature matrix of reference expression profiles to estimate the proportion of 22 immune cell populations. Marker genes were first selected from microarray expression data of isolated immune cells using differential expression analysis and fold-change ranking. The expression value of each marker gene and immune cell population in the reference matrix was defined as the median expression of that gene across all transcriptome profiles for that population [48]. TIMER [49] uses a different expression profile matrix for each one of 23 TCGA cancer types to estimate the abundance of six immune cell populations. In this case, marker genes were collected from the Immune Response In Silico database [53] and filtered out if positively correlated with TCGA tumour purity. Expression profiles of isolated immune cells were then obtained from the Human Primary Cell Atlas [54]. For each immune cell type, the reference profile was taken as the median expression of the filtered marker genes across corresponding transcriptome profiles. Unlike the profile matrices of the above methods, EPIC [50] was the first to use a profile matrix derived from scRNA-seq data of primary and non-lymphoid metastatic melanoma samples. Marker genes were identified by differential expression analysis and the resulting profile of a cell type was taken as the average expression of corresponding markers. Out of the considered stromal populations, EPIC was recommended for the deconvolution of B cells, CD4$^+$ and CD8$^+$ T cells, NK cells, CAFs and endothelial cells [52]. quanTIseq [51] was the first method to derive its signature matrix entirely from bulk RNA-seq data of purified cell populations. Marker genes were selected based on their differential expression between cell types and filtered out if highly expressed in tumour cells. The reference profile of each cell population was computed as the median expression over corresponding RNA-seq purified profiles. The approach was found to be particularly suitable for the deconvolution of regulatory and CD8$^+$ T cells [52]. Notably, quanTIseq implements a whole RNA-seq data processing pipeline, from read pre-processing to TME cell type quantification. This avoids technical differences between the bulk tumour sample and the reference profile matrix.

3.3. Computational methods for the quantification of tumour-infiltrates

The cell type-specific signatures derived from either marker genes or profile matrices can then be used to quantify non-cancer cells of the TME. Computational approaches developed so far can be broadly divided into gene set scoring approaches and deconvolution approaches.

Gene set scoring approaches leverage on marker gene signatures to provide relative abundance scores indicative of how enriched a cell population of interest is in the bulk tumour sample. Most of these approaches implement Gene Set Enrichment Analysis (GSEA) methods to quantify cell populations defined by their corresponding marker gene set in each individual sample. In these GSEA-based methods, genes from bulk transcriptomic data are first ranked in decreasing order of their expression. Cell populations are then considered to be enriched or depleted if their marker genes are among the top or bottom expressed genes, respectively. An example of GSEA-based methods is single-sample GSEA (ssGSEA) [55] that computes an ES in each sample by ranking the genes according to their absolute expression value. ESs are calculated for every pair of sample and marker gene set. This is achieved by integrating the difference between the empirical cumulative distribution of the rank-normalised gene expression inside and outside the gene set [55]. ssGSEA was directly used for the characterisation of the TME in several cancer types [40–42]. xCell uses ssGSEA for the calculation of the raw enrichment score of a cell population, which is then adjusted through a spillover technique to correct for cell type collinearity [43]. xCell is therefore less prone to background predictions, i.e. the artificial abundance estimation of cell types that are actually absent. For this reason, it was recommended for use when the main interest is to identify the presence of a particular cell population in the sample [52]. Unlike ssGSEA, Gene Set Variation Analysis (GSVA) [56] still applies GSEA but accounts for expression variability across large and heterogeneous datasets. It uses a non-parametric estimation of the cumulative density function of the expression profile of each gene. GSVA has been used to quantify tumour-immune infiltrates and characterise the immunophenotypes of TCGA samples [44].

Other gene set scoring methods that are not based on GSEA use the log-transformed geometric [45,46] or arithmetic [47] mean of the normalised marker gene expression values in the tumour sample (Table 1). Although these methods are more dependent on the quality of gene expression data than GSEA-based methods, they provide abundance scores that are directly proportional to marker gene expression [46]. This facilitates their interpretation. For instance, if marker genes associated to a particular cell population are twice as expressed in sample A than in sample B, one can infer that this cell population is twice as abundant in A than in B (assuming the absence of aberrant expression of any of those markers by some tumour cells). This fold change would not be reflected by GSEA-based approaches as they provide scores computed from gene ranks.

Deconvolution approaches estimate the fraction of each cell population in the sample from transcriptomic data using both marker gene sets and expression profile matrices. These methods consider the expression profile of a heterogeneous tissue sample as the sum of the expression profiles of the composing cell populations weighted by their relative fractions [57]. Deconvolution can be partial to find only the fraction of each cell population, or complete to derive also the associated expression profiles [57]. Partial deconvolution requires a reference expression profile matrix containing an aggregate of the expression profile of each marker gene. It is usually based on least square regression to minimise the differences between the bulk expression values and the product of the reference expression profiles with the estimated fractions [57]. Tools implementing least square regression include PERT [58], DeconRNASeq [59], TIMER [60], EPIC [50], and quanTIseq [51]. Machine learning based on nu-support vector regression (nu-SVR) has also been applied in the context of partial deconvolution, such as CIBERSORT [48] and Mysort [61]. Although nu-SVR was a first step towards handling outlier gene expression values, the recently proposed FARDEEP [62] was the first approach to directly address this issue. FARDEEP uses an adaptive least trimmed square model to detect and remove outliers prior to cell fraction estimation and thereby increase estimation robustness. All these partial deconvolution methods rely on a linear model of gene expression that considers the total bulk mRNA as the sum of the mRNAs of the composing cell populations. However, solving deconvolution equations on the linear scale is not always efficient [63]. This is because RNA-seq data generally have a skewed asymmetric distribution with a longer right tail of highly expressed genes. To account for this skewedness in gene expression data, dtangle [63] implements a multivariate logistic model that solves the linearly-modelled deconvolution problem on the logarithmic scale.

Complete deconvolution approaches, also known as unsupervised methods, estimate both cell fractions and their expression profiles [57]. Most of these methods are based on non-negative matrix factorisation that factorises the bulk expression profiles as the product of non-negative cell fractions and cell-specific profiles. Examples of tools implementing non-negative matrix factorisation include deconf [64] and a semi-supervised algorithm that incorporates prior knowledge of cell type-specific markers [65]. Other approaches that also use cell type-specific markers are based on quadratic programming [66] or on maximum likelihood estimation [67]. Recently, DeMixT [68] has been developed to de-convolute bulk RNA-seq cancer data into tumour and stromal components. DeMixT considers the input data as a linear additive model of tumour and stroma. Then, their relative proportions and corresponding expression profiles are estimated using the iterated conditional modes algorithm and a gene-set-based component merging approach [68].

3.4. Limitations of TME quantification from bulk transcriptomic data

Both gene set scoring- and deconvolution-based approaches present several limitations when characterising the TME from bulk tumour data.

First, as mentioned above, the scores derived from gene set scoring approaches cannot be interpreted as cell type proportions within the sample. One of the reasons for this is that the sizes of the marker gene sets can be highly variable, biasing the scoring towards larger sets. Thus, gene set scoring approaches do not allow intra-sample comparisons of different cell populations. This is partially solved in deconvolution-based approaches as they provide cellular fractions that can be related to cell population abundances both within and across samples.

Second, most cell type-specific signatures are derived from expression data of cell populations that were isolated from non-cancer tissues, generally peripheral blood. This is likely to affect the abundance estimation in bulk tumour samples for at least two reasons. First, the immune cell composition varies across cancers [15]. Second, some marker genes can be expressed also by tumour cells [43]. Some approaches reduce these biases by incorporating tumour-specific expression profiles when constructing cell type-specific signatures.

Third, most partial deconvolution approaches rely on static cell type-specific signature matrices that assume constant expression profiles of the cell populations across samples. This assumption neglects sample-specific variations in time and space [57]. Moreover, given the variability and diversity of the TME, it is likely that not all cell populations are accounted for by the existing quantification approaches. In addition, not all cell populations considered by these approaches are necessarily present in the cancer samples (referred to as background predictions [52]). As a result, partial deconvolution approaches may produce under- or over-estimated cell fractions [57]. To address this, some methods avoid restricting the cell fraction estimation to the populations under consideration [50,51,69]. Instead, they estimate the fraction of uncharacterised cells within the tumour bulk to provide more accurate estimations.

Fourth, mRNA abundances across cell types are often neglected by partial deconvolution methods when estimating TME cell fractions. Only EPIC [50] and quanTIseq [51] correct for this by normalising each estimated cell type abundance by a corresponding scaling factor representing the mRNA content of that cell type. Therefore, these methods allow a more reliable comparison of cell population abundances as they can be interpreted as actual cell fractions. Both methods were recommended for immuno-oncology applications as their fractions are comparable both across and within samples [52]. An alternative approach, ABIS [70], used a reference profile matrix normalised for cell type-specific mRNA abundance instead of correcting estimated abundances by a scaling factor. However, ABIS was derived from and applied to blood-derived expression data, and has not been applied to cancer transcriptomic data yet.

Finally, often only a small set of cell populations is used to benchmark quantification approaches. This is because experimental techniques to derive ground-truth quantifications (such as flow cytometry) allow the simultaneous profiling of a limited number of cell types [43,48]. Recently, five deconvolution-based and two scoring-based approaches were systematically compared by assessing their performance on estimating nine stromal populations [52]. Four metrics

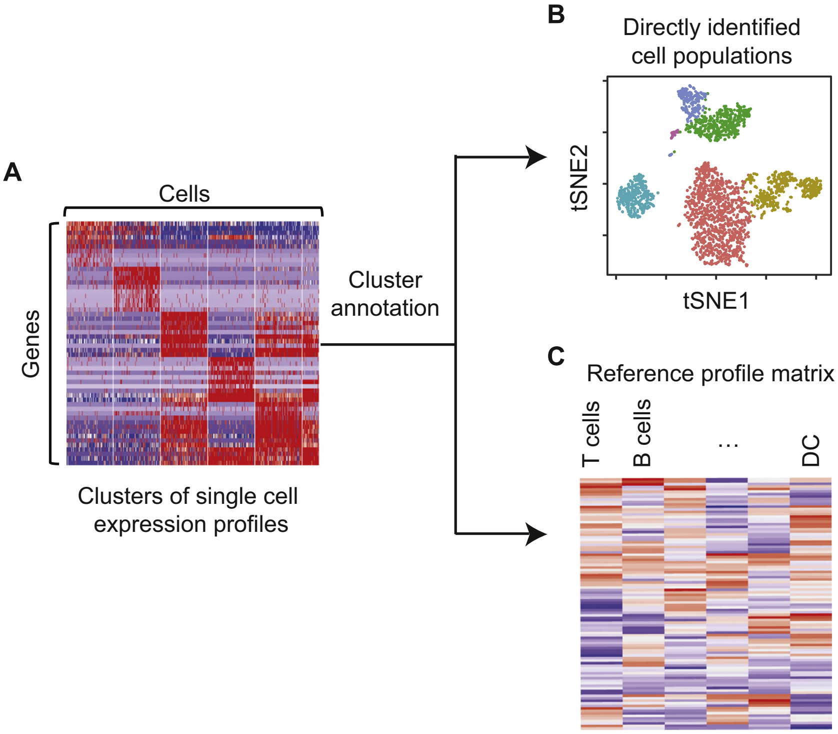


Fig. 3. Single-cell RNA-seq for the identification of TME cell populations. (A) Clustered scRNA-seq profiles of cancer samples are annotated according to the expression of known marker genes. (B) Cell populations can be directly identified from the annotated clusters, and visualised after dimensionality reduction. (C) Alternatively, the annotated clusters can also be used to derive high-resolution reference profile matrices. DC = dendritic cells, tSNE = t-distributed Stochastic Neighbour Embedding.

assessed each quantification approach: predictive performance, minimal detection fraction, background predictions, and spillover effect on both real and simulated bulk RNA-seq datasets. Spillover effect measures the over-estimation of a cell population due to the inaccurate estimation of others. Interestingly, the performance of the tested approaches varied across cell types, with poor performances on CD4$^+$ T cells and dendritic cells, overall. Deconvolution-based approaches were found to be more likely to estimate minimal immune cell fractions even when these were absent (i.e. background predictions).

4. Quantification of non-cancer cells from single-cell transcriptomic data

The last five years have seen an increasing number of studies applying scRNA-seq to characterise the TME across different cancer types [19]. scRNA-seq data can be used to identify the sequenced cells directly or to generate reference expression profile matrices to de-convolute bulk RNA-seq data (Fig. 3).

TME cell populations can be directly quantified from scRNA-seq data by clustering and annotating the resulting clusters according to the expression of cell type-specific marker genes (Fig. 3A). This allows to assign the clusters to specific stromal cell populations (Fig. 3B). This approach has been used to profile the tumour infiltrates in melanoma [71], hepatocellular carcinoma [72], breast [73–75], colorectal [76] and lung cancer [77].

Cell type-specific reference expression profile matrices can also be derived from scRNA-seq data (Fig. 3C). Deconvolution-based quantification methods can then use these matrices to estimate the abundance of different cell populations from bulk tumour transcriptomic data. For example, EPIC [50] used a reference matrix derived from expression profiles of a melanoma scRNA-seq dataset [71]. This approach was later extended to two other scRNA-seq datasets(normal blood [29] and ovarian cancer [78]) to build five reference expression profile matrices [78]. Each matrix was obtained using a different strategy to average gene expression across and within single cell datasets and cell types. Then, CIBERSORT was applied with each reference matrix [48] and reference marker genes from independent sources [48,71,78]. The best deconvolution results for ten stromal and two cancer cell types, on both simulated and real bulk RNA-seq data were obtained by averaging expression within both cell types and datasets [78]. This highlights the strong dependence of deconvolution methods on the quality of reference profile matrices. The newest version of CIBERSORT, CIBERSORTx [79], accounts for this dependency by allowing the use of reference signatures obtained from single cells or from bulk expression data. CIBERSORTx also uses nu support vector regression to estimate cell type proportions. However, before deconvolution, it performs normalisation and batch correction of platform-specific variations between the reference signatures and the bulk RNA-seq data [79].

5. Other approaches for the quantification of tumour infiltrating cells

Alternative approaches that are not based on tumour gene expression profiles can also be used to characterise the TME. They are usually based on the multidimensional analysis of proteins or RNAs from thousands of cells detected either from solid tissue sections or disaggregated tissues (Table 2). These methods can be categorised in four main groups according to the detection technology, namely chromogenic or fluorescent labelling, mass-spectrometry, and DNA probes coupled with bulk- or single-cell sequencing.

5.1. *Chromogenic or fluorescent labelling methods*

The latest development in IHC and IF allow multiplexed assays of tens of markers through repeated rounds of staining. They also permit the analysis of large regions of interest at high spatial resolution. For instance, the IHC-based approach SIMPLE [88] was used to quantify the association between mono-myelocytic and exhausted T-cell density and the response to GVAX vaccination in pancreatic ductal adenocarcinomas [89]. Similarly, the MultiOmyx IF platform [90] was applied to investigate resistance to rituximab–CHOP in diffuse large B-Cell lymphoma, pointing to high PD-1 expressing CD8$^+$ T cells and PD-L1 expressing macrophages as mediators of resistance.

Flow cytometry employs a fluidic system and fluorescent labelling

Table 2
Non-transcriptomic approaches for the quantification of tumour-infiltrating cells. For each method, we report its detection technology and compatibility with FFPE samples, the type and the maximum number of measurable markers, and its throughput per run. For techniques providing spatial information, we also report their spatial resolution. FFPE = formalin-fixed paraffin-embedded, IHC = immunohistochemistry, IF = immunofluorescence, cyTOF = cytometry by time-of-flight, IMC = imaging mass cytometry, MIBI = multiplexed ion beam imaging, DSP = digital spatial profiling.

Method	Technology	FFPE	Markers	Throughput per run	Spatial resolution
Multiplex IHC [80]	Chromogenic-antibodies	Y	< 12 proteins	~500 mm^2/run	< 1 µm
Multiplex IF [81]	Fluorescent-antibodies	Y	< 50 proteins	~500 mm^2	< 1 µm
Flow Cytometry [82]	Fluorescent-antibodies	Y	< 28 proteins	~10^7 cells	N
cyTOF [83]	Mass spectrometry	Y	< 40 proteins	~10^7 cells	N
IMC [12]	Mass spectrometry	Y	< 40 proteins	~ 1 mm^2	1 µm
MIBI [13]	Mass spectrometry	Y	< 50 proteins	~ 1 mm^2	0.2 µm
Spatial transcriptomics [17]	DNA probes Bulk DNA-seq	N	> 1500 genes	1007 spots/slide	200 µm
NanoString DSP [84]	DNA probes Bulk DNA-seq	Y	< 40 proteins < 90 genes	600 µm^2	10 µm
REAP-seq [85]	DNA probes scDNA-seq	N	~1500 genes 82 surface markers	~4000 cells	N
Abseq [86]	DNA probes scDNA-seq	N	> 600 genes 30 surface markers	> 10,000 cells	N
CITE-seq [87]	DNA probes scDNA-seq	N	~1500 genes > 20 surface markers	~10,000 cells	N

to isolate and characterise cells according to the expression of 5–15 markers. Since flow cytometry is not destructive, the cells can be used for further analyses. Due to its intrinsic robustness, flow cytometry is often used to validate computational methods that quantify TME populations [50]. Recently developed high dimensional flow-cytometry techniques allow the quantification of up to 28 markers in single cells [82]. Routine flow cytometry analyses cannot be applied to this technique as it requires specialised computational tools for the unbiased identification of cell populations from this larger number of markers [82].

5.2. Mass spectrometry

Mass cytometry, also known as cytometry by time-of-flight (cyTOF) [83], similarly to flow cytometry, uses a fluidic system to isolate cells. However, cyTOF marker detection is based on time-of-flight (TOF) mass spectrometry instead of fluorescence. Cells are first labelled with heavy metal-tagged antibodies, which are then distinguished according to the atomic mass of the associated metal ions. Ion counts are acquired across the mass spectra, and combined to form events as in flow-cytometry experiments. These events are then thresholded according to signal intensity across all channels to discard events caused by debris. After this filtering the event data is exported in the standard FCS format used also for flow cytometry [91]. cyTOF was recently used to characterise the TME of breast cancer leading to the identification of TME features that can be used for patient stratification [92]. Because cyTOF does not rely on fluorophores, the detection specificity is not reduced by spectral overlap and autofluorescence. This increases the number of markers that can be quantified in a single experiment. However, the limiting factor is the number of pure heavy metal isotopes available (Table 2). Despite higher specificity, cyTOF has still lower throughput than flow-cytometry (< 1000 cells/s compared to about 10,000 cells/s). Both cyTOF and high-throughput flow cytometry data differ from scRNA-seq data in two main aspects: the number of analysed cells is much higher ~10^7, and the possibility to quantify only up to about 40 markers (Table 2). After gating to remove doublets and select only intact single cells, a multidimensional analysis of the single cell data can be performed. First unsupervised clustering is employed to group cells into different subpopulations. Then, differential cell population abundance and/or differential marker expression across different conditions can be analysed. Finally, the different cell populations and the expression of markers of interest can be visualised using dimensionality reduction approaches [91].

Other approaches leverage mass spectrometry with heavy metal ion-tagged antibodies for the imaging of FF or FFPE tissues. The best-known examples are IMC [12] and MIBI [13], which differ mainly in the way the heavy metal ions are separated from the tissue slide. IMC uses a UV laser to ablate pre-selected areas of the tissue slide and the resulting gas is then ionised with inductively coupled plasma before TOF mass spectrometry is applied [12]. MIBI instead relies on a primary ion beam to liberate the heavy metals chelated to the antibodies as secondary ions. These ions are then analysed with sector field [13] or TOF mass analysers. The ion counts obtained from rasterising the slide with the laser or the ion beam are finally used to reconstruct a multidimensional image composed of one layer per ion/marker. The tissue areas scanned in both IMC and MIBI are much smaller than those acquired with multiplex IHC and IF. However, they provide greater sensitivity, with at least five orders of magnitude of linear dynamic range, and can use a higher number of markers. MIBI can reach a resolution higher than IMC (of < 500 nm as compared to about 1 µm). In contrast, IMC has faster scan sampling times which makes it suitable for the ablation of larger areas and has been further adapted to quantify mRNAs from FFPE tissues [93]. After removing background noise, IMC and MIBI images can be used to identify single cells with image segmentation techniques. Then, for each of these cells, the expression values of each marker can be extracted to obtain a matrix similar to those derived from cyTOF or high dimensional flow cytometry. This matrix, generally containing a much lower number of cells, generally in the order of 10^3 per image, can be analysed with unsupervised clustering. Finally, the spatial information contained in the images can be leveraged to identify significant cell-cell interactions through neighbourhood analysis, or by investigating the localisation of specific cell population in the tissue.

Both IMC and MIBI have been applied to TME characterisation. For example, MIBI revealed a positive correlation between the expression of immunoregulatory proteins and the tumour-immune composition and organisation in triple negative breast cancer [94]. IMC enabled the analysis of the relationship between CD8$^+$ T cell infiltration, the extracellular domain of HER2, and response to trastuzumab in breast cancer [95].

5.3. DNA probes coupled with bulk sequencing

Two recently developed techniques, spatial transcriptomics [17] and NanoString digital spatial profiling (DSP) [84], can quantify gene expression in specific areas of tissue samples. Both methods rely on DNA or DNA-RNA probes coupled with fluorescent labelling to retain spatial information of gene expression.

In spatial transcriptomics, this is achieved through an array of 100 µm-large spots of spatially barcoded oligo-dT probes. After placing the tissue on the array, mRNAs can be reverse-transcribed directly in situ and then sequenced. The spatial barcode sequences from the array probes are retained in the RNA-seq reads and this allows to trace them back to the original spot in the tissue. Spatial transcriptomics requires intact RNA (therefore it cannot currently be applied to FFPE samples) and cannot reach single cell resolution. However, it has a higher throughput than, for example, multiplexed sequential FISH techniques (about 1000 spots per sample able to detect > 1500 genes per spot). Spatial transcriptomics also provides greater flexibility than other in situ sequencing approaches, as it does not require customised instruments. Spatial transcriptomic data consists of an expression matrix where each row corresponds to a gene and each column corresponds to a spot coordinate. An integration of scRNA-seq and spatial transcriptomics was recently applied to study the spatial composition of the TME in pancreatic ductal adenocarcinoma [96].

NanoString DSP [84] relies on the NanoString nCounter platform [14] to quantify antibody-bound proteins or hybridised transcripts using specific photocleavable DNA probes. The probes are then hybridised with complementary fluorescent-labelled RNA probes. To obtain spatial information three consecutive slides are used, one for IHC or in situ RNA hybridisation to visualise the tissue and select the area of interest and the other two for protein and RNA quantification. After the tissue area is selected, the photocleavable probes are released with UV light and collected by microcapillary aspiration for quantification with the NanoString nCounter platform. Area selection in NanoString DSP is flexible ranging from simple to complex shapes associated with tissue compartments or single cells. This approach can be used in both FF and FFPE samples, but the number of markers is limited to about 40 proteins and 90 transcripts [97]. NanoString DSP read counts are normalised using spike-in probes to account for capture and amplification efficiency. Moreover, since ROIs differ in size both within and across samples, area normalisation is also applied. Additionally, ROI background is corrected through the addition of negative RNA probes and isotype antibodies; while transcripts and antibodies against cellular proteins address differences in cellularity across ROIs. The output data consists of a matrix with the normalised intensities of the protein and mRNA markers in each ROI [97]. NanoString DSP has been recently applied to quantify 32 proteins and 82 transcripts in tumour and stromal regions of non-small cell lung cancer [97].

The characterisation of tissue regions from marker expression is achieved by processing the expression matrices with dimensionality reduction followed by hierarchical clustering. The clustered features can then be placed back on the tissue images to relate them with tissue architecture [98].

5.4. DNA probes coupled with single-cell sequencing

In the past three years single-cell approaches that integrate transcriptomics and cell-surface protein quantification have emerged [20]. These approaches quantify protein expression in single cells through DNA-tagged antibodies. In parallel they allow RNA expression profiling in the same cell using microdroplet- or microwell- based scRNA-seq [99]. The most widely used technologies implementing this approach are REAP-seq [85], Abseq [86] and CITE-seq [87]. REAP-seq is based on the Chromium sequencing platform [29] and, while it has a relatively low throughput (about 4000 cells per run), it allows the quantification of up to 82 different proteins. Abseq relies on the BD Rhapsody sequencing platform [100] and can quantify of up to 600 genes and 30 proteins in > 10,000 cells [86]. Finally, CITE-seq [87] uses either Drop-seq [30] or other microdroplet-based technologies to measure about 1500 genes and 20 proteins in > 10,000 cells [101]. These high throughput scRNA-seq methods allow to perform multimodal RNA-protein analyses on large single-cell datasets. For example, CITE-seq has been used to characterise rare immune cell phenotypes by splitting scRNA-seq derived clusters into subsets with high and low expression of specific surface markers [87].

6. Conclusion

The success of cancer immunotherapy has led to an increased interest in the fine characterisation of TME composition. This is indeed the first step to understand how the TME influences response to therapy [2]. In addition to a better knowledge of the interactions between cancer and non-cancer cells, TME characterisation can also be exploited as biomarker for patient stratification and prognosis. For example, the quantification of tumour infiltrating CD3$^+$ and CD8$^+$ T cells using digital pathology from IHC slides has a validated prognostic value for predicting colorectal cancer recurrence [8]. This measure, called Immunoscore, represents the first step towards the adoption of standardised immune-based assays for colorectal cancer classification. Other similar efforts are extending this approach to a broader set of cancer types [6].

Despite their undoubted utility, the incorporation of technically sophisticated methods that allow a thorough analysis of the TME in the clinical setting is still challenging. Indeed, these techniques are usually expensive, highly sensitive to the quality of the input material and require specialised expertise for their analysis. This is particularly the case for the more recent approaches such as high throughput scRNA-seq and mass spectrometry-based imaging. Moreover, the turnaround time is often not compatible with the decision-making process of the clinical practice. Further efforts are needed to harmonise the depth and specificity of the TME analysis achieved in the research setting to the requirements of time and cost-effective clinical assays.

Future developments in the characterisation of the TME should incorporate spatial information and integrate different types of omic data. Emerging approaches have already started to link the expression of marker genes to their localisation within the tissue enabling a deeper understanding of the tumour-TME interactions. However, these approaches are currently limited to tiny regions of the tumour that may not be representative of the whole tumour mass. Increasing the tissue area that can be analysed will also increase the robustness of the results. Similarly, new technologies enabling the simultaneous multi-omic analyses of the TME are being developed, particularly in the single cell setting. They combine scRNA-seq, scDNA-seq, single cell T and B cell receptor sequencing, single cell epigenomics and small and non-coding scRNA-seq [101]. In combination with functional studies, these techniques will enable a further in-depth description of all the cell populations constituting the TME and their interactions [20].

We are at the beginning of an exciting era where technological innovations can effectively contribute to improve not only our understanding of cancer biology, but also the way we treat cancer patients.

Author contributions

All authors contributed to write the manuscript.

Funding

This study has received funding from Cancer Research UK [C43634/A25487], the Cancer Research UK King's Health Partners Centre at King’s College London [C604/A25135 and C604/A25189], the CRUK City of London Centre Award [C7893/A26233], and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No CONTRA-766030.

Transparency document

The transparency document associated with this article can be found in the online version.

Declaration of competing interest

The authors declare that no potential conflicts of interest were disclosed.

References

[1] S. Frank, Dynamics of Cancer: Incidence, Inheritance, and Evolution, Princeton University Press, Place Published, 2007.

[2] M. Binnewies, E.W. Roberts, K. Kersten, V. Chan, D.F. Fearon, M. Merad, L.M. Coussens, D.I. Gabrilovich, S. Ostrand-Rosenberg, C.C. Hedrick, R.H. Vonderheide, M.J. Pittet, R.K. Jain, W. Zou, T.K. Howcroft, E.C. Woodhouse, R.A. Weinberg, M.F. Krummel, Understanding the tumor immune microenvironment (TIME) for effective therapy, Nat. Med. 24 (2018) 541–550.

[3] M.F. Sanmamed, L. Chen, A paradigm shift in cancer immunotherapy: from enhancement to normalization, Cell 175 (2018) 313–326.

[4] A. Ribas, J.D. Wolchok, Cancer immunotherapy using checkpoint blockade, Science 359 (2018) 1350–1355.

[5] J.J. Havel, D. Chowell, T.A. Chan, The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy, Nat. Rev. Cancer 19 (2019) 133–150.

[6] S. Hendry, R. Salgado, T. Gevaert, P.A. Russell, T. John, B. Thapa, M. Christie, K. van de Vijver, M.V. Estrada, P.I. Gonzalez-Ericsson, M. Sanders, B. Solomon, C. Solinas, G. Van den Eynden, Y. Allory, M. Preusser, J. Hainfellner, G. Pruneri, A. Vingiani, S. Demaria, F. Symmans, P. Nuciforo, L. Comerma, E.A. Thompson, S. Lakhani, S.R. Kim, S. Schnitt, C. Colpaert, C. Sotiriou, S.J. Scherer, M. Ignatiadis, S. Badve, R.H. Pierce, G. Viale, N. Sirtaine, F. Penault-Llorca, T. Sugie, S. Fineberg, S. Paik, A. Srinivasan, A. Richardson, Y. Wang, E. Chmielik, J. Brock, D.B. Johnson, J. Balko, S. Wienert, V. Bossuyt, S. Michiels, N. Ternes, N. Burchardi, S.J. Luen, P. Savas, F. Klauschen, P.H. Watson, B.H. Nelson, C. Criscitiello, S. O'Toole, D. Larsimont, R. de Wind, G. Curigliano, F. Andre, M. Lacroix-Triki, M. van de Vijver, F. Rojo, G. Floris, S. Bedri, J. Sparano, D. Rimm, T. Nielsen, Z. Kos, S. Hewitt, B. Singh, G. Farshid, S. Loibl, K.H. Allison, N. Tung, S. Adams, K. Willard-Gallo, H.M. Horlings, L. Gandhi, A. Moreira, F. Hirsch, M.V. Dieci, M. Urbanowicz, I. Brcic, K. Korski, F. Gaire, H. Koeppen, A. Lo, J. Giltnane, M.C. Rebelatto, K.E. Steele, J. Zha, K. Emancipator, J.W. Juco, C. Denkert, J. Reis-Filho, S. Loi, S.B. Fox, Assessing tumor-infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the international immuno-oncology biomarkers working group: part 2: TILs in melanoma, gastrointestinal tract carcinomas, non-small cell lung carcinoma and mesothelioma, endometrial and ovarian carcinomas, squamous cell carcinoma of the head and neck, genitourinary carcinomas, and primary brain tumors, Adv. Anat. Pathol. 24 (2017) 311–335.

[7] B. Weide, A. Martens, H. Zelba, C. Stutz, E. Derhovanessian, A.M. Di Giacomo, M. Maio, A. Sucker, B. Schilling, D. Schadendorf, P. Büttner, C. Garbe, G. Pawelec, Myeloid-derived suppressor cells predict survival of patients with advanced melanoma: comparison with regulatory T cells and NY-ESO-1- or melan-A–specific T cells, Clin. Cancer Res. 20 (2014) 1601.

[8] F. Pagès, B. Mlecnik, F. Marliot, G. Bindea, F.-S. Ou, C. Bifulco, A. Lugli, I. Zlobec, T.T. Rau, M.D. Berger, I.D. Nagtegaal, E. Vink-Börger, A. Hartmann, C. Geppert, J. Kolwelter, S. Merkel, R. Grützmann, M. Van den Eynde, A. Jouret-Mourin, A. Kartheuser, D. Léonard, C. Remue, J.Y. Wang, P. Bavi, M.H.A. Roehrl, P.S. Ohashi, L.T. Nguyen, S. Han, H.L. MacGregor, S. Hafezi-Bakhtiari, B.G. Wouters, G.V. Masucci, E.K. Andersson, E. Zavadova, M. Vocka, J. Spacek, L. Petruzelka, B. Konopasek, P. Dundr, H. Skalova, K. Nemejcova, G. Botti, F. Tatangelo, P. Delrio, G. Ciliberto, M. Maio, L. Laghi, F. Grizzi, T. Fredriksen, B. Buttard, M. Angelova, A. Vasaturo, P. Maby, S.E. Church, H.K. Angell, L. Lafontaine, D. Bruni, C. El Sissy, N. Haicheur, A. Kirilovsky, A. Berger, C. Lagorce, J.P. Meyers, C. Paustian, Z. Feng, C. Ballesteros-Merino, J. Dijkstra, C. van de Water, S. van Lent-van Vliet, N. Knijn, A.-M. Mușină, D.-V. Scripcariu, B. Popivanova, M. Xu, T. Fujita, S. Hazama, N. Suzuki, H. Nagano, K. Okuno, T. Torigoe, N. Sato, T. Furuhata, I. Takemasa, K. Itoh, P.S. Patel, H.H. Vora, B. Shah, J.B. Patel, K.N. Rajvik, S.J. Pandya, S.N. Shukla, Y. Wang, G. Zhang, Y. Kawakami, F.M. Marincola, P.A. Ascierto, D.J. Sargent, B.A. Fox, J. Galon, International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study, Lancet 391 (2018) 2128–2139.

[9] M. Rakaee, T.K. Kilvaer, S.M. Dalen, E. Richardsen, E.-E. Paulsen, S.M. Hald, S. Al-Saad, S. Andersen, T. Donnem, R.M. Bremnes, L.-T. Busund, Evaluation of tumor-infiltrating lymphocytes using routine H&E slides predicts patient survival in resected non–small cell lung cancer, Hum. Pathol. 79 (2018) 188–198.

[10] R.E. Parra, A. Francisco-Cruz, I.I. Wistuba, State-of-the-art of profiling immune contexture in the era of multiplexed staining and digital analysis to study paraffin tumor tissues, Cancers, 11, (2019).

[11] C.M. Mousset, W. Hobo, R. Woestenenk, F. Preijers, H. Dolstra, A.B. van der Waart, Comprehensive phenotyping of T cells using flow cytometry, Cytometry A, 2019.

[12] C. Giesen, H.A.O. Wang, D. Schapiro, N. Zivanovic, A. Jacobs, B. Hattendorf, P.J. Schüffler, D. Grolimund, J.M. Buhmann, S. Brandt, Z. Varga, P.J. Wild, D. Günther, B. Bodenmiller, Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry, Nat. Methods 11 (2014) 417.

[13] M. Angelo, S.C. Bendall, R. Finck, M.B. Hale, C. Hitzman, A.D. Borowsky, R.M. Levenson, J.B. Lowe, S.D. Liu, S. Zhao, Y. Natkunam, G.P. Nolan, Multiplexed ion beam imaging of human breast tumors, Nat. Med. 20 (2014) 436.

[14] G.K. Geiss, R.E. Bumgarner, B. Birditt, T. Dahl, N. Dowidar, D.L. Dunaway, H.P. Fell, S. Ferree, R.D. George, T. Grogan, J.J. James, M. Maysuria, J.D. Mitton, P. Oliveri, J.L. Osborn, T. Peng, A.L. Ratcliffe, P.J. Webster, E.H. Davidson, L. Hood, K. Dimitrov, Direct multiplexed measurement of gene expression with color-coded probe pairs, Nat. Biotechnol. 26 (2008) 317.

[15] F. Finotello, Z. Trajanoski, Quantifying tumor-infiltrating immune cells from transcriptomics data, Cancer Immunol. Immunother. 67 (2018) 1031–1040.

[16] F. Petitprez, C.-m. Sun, L. Lacroix, C. Sautès-Fridman, A. de Reyniès, W.H. Fridman, Quantitative Analyses of the Tumor Microenvironment Composition and Orientation in the Era of Precision Medicine, Frontiers in Oncology 8 (2018) 390.

[17] F. Salmén, P.L. Ståhl, A. Mollbrink, J.F. Navarro, S. Vickovic, J. Frisén, J. Lundeberg, Barcoded solid-phase RNA capture for spatial transcriptomics profiling in mammalian tissue sections, Nat. Protoc. 13 (2018) 2501–2534.

[18] J.H. Lee, E.R. Daugharthy, J. Scheiman, R. Kalhor, T.C. Ferrante, R. Terry, B.M. Turczyk, J.L. Yang, H.S. Lee, J. Aach, K. Zhang, G.M. Church, Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues, Nat. Protoc. 10 (2015) 442–458.

[19] F. Valdes-Mora, K. Handler, A.M.K. Law, R. Salomon, S.R. Oakes, C.J. Ormandy, D. Gallego-Ortega, Single-cell transcriptomics in cancer immunobiology: the future of precision oncology, Front. Immunol. 9 (2018) 2582.

[20] F. Finotello, F. Eduati, Multi-omics profiling of the tumor microenvironment: paving the way to precision immuno-oncology, Front. Oncol. 8 (2018) 430.

[21] L.C. Wehmas, C.E. Wood, R. Gagne, A. Williams, C. Yauk, M.M. Gosink, D. Dalmas, R. Hao, R. O'Lone, S. Hester, Demodifying RNA for transcriptomic analyses of archival formalin-fixed paraffin-embedded samples, Toxicol. Sci. 162 (2017) 535–547.

[22] Q. Sheng, K. Vickers, S. Zhao, J. Wang, D.C. Samuels, O. Koues, Y. Shyr, Y. Guo, Multi-perspective quality control of Illumina RNA sequencing data analysis, Briefings in Functional Genomics 16 (2016) 194–204.

[23] J. Li, C. Fu, T.P. Speed, W. Wang, W.F. Symmans, Accurate RNA sequencing from formalin-fixed cancer tissue to represent high-quality transcriptome from frozen tissue, JCO precision oncology 2 (2018) 1–9.

[24] Y. Guo, Q. Sheng, J. Li, F. Ye, D.C. Samuels, Y. Shyr, Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data, PLoS One 8 (2013) e71462.

[25] I, Illumina, Illumina, Exon Capture Protocol, 2017.

[26] X. Zhang, T. Li, F. Liu, Y. Chen, J. Yao, Z. Li, Y. Huang, J. Wang, Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-Seq systems, Molecular Cell, 73 e135 (2019) 130–142.

[27] D.A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, I. Amit, Massively parallel single-cell RNA-Seq for marker-free decomposition of tissues into cell types, Science 343 (2014) 776.

[28] T.M. Gierahn, M.H. Wadsworth Ii, T.K. Hughes, B.D. Bryson, A. Butler, R. Satija, S. Fortune, J.C. Love, A.K. Shalek, Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput, Nat. Methods 14 (2017) 395.

[29] G.X.Y. Zheng, J.M. Terry, P. Belgrader, P. Ryvkin, Z.W. Bent, R. Wilson, S.B. Ziraldo, T.D. Wheeler, G.P. McDermott, J. Zhu, M.T. Gregory, J. Shuga, L. Montesclaros, J.G. Underwood, D.A. Masquelier, S.Y. Nishimura, M. Schnall-Levin, P.W. Wyatt, C.M. Hindson, R. Bharadwaj, A. Wong, K.D. Ness, L.W. Beppu, H.J. Deeg, C. McFarland, K.R. Loeb, W.J. Valente, N.G. Ericson, E.A. Stevens, J.P. Radich, T.S. Mikkelsen, B.J. Hindson, J.H. Bielas, Massively parallel digital transcriptional profiling of single cells, Nat. Commun. 8 (2017) 14049.

[30] Z. Evan, A. Macosko, R. Basu, J. Satija, K. Nemesh, M. Shekhar, I.T. Goldman, Allison R. Bialas, N. Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, A. Regev, Steven A. McCarroll, Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets, Cell 161 (2015) 1202–1214.

[31] A.M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D.A. Weitz, M.W. Kirschner, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells, Cell 161 (2015) 1187–1201.

[32] A.A. Kolodziejczyk, J.K. Kim, V. Svensson, J.C. Marioni, S.A. Teichmann, The technology and biology of single-cell RNA sequencing, Mol. Cell 58 (2015) 610–620.

[33] Q.H. Nguyen, N. Pervolarakis, K. Nee, K. Kessenbrock, Experimental considerations for single-cell RNA sequencing approaches, Frontiers in Cell and Developmental Biology 6 (2018) 108.

[34] S. Nichterwitz, G. Chen, J. Aguila Benitez, M. Yilmaz, H. Storvall, M. Cao, R. Sandberg, Q. Deng, E. Hedlund, Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling, Nat. Commun. 7 (2016) 12139.

[35] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, S. Linnarsson, Quantitative single-cell RNA-seq with unique molecular identifiers, Nat. Methods 11 (2013) 163.

[36] A.A. AlJanahi, M. Danielsen, C.E. Dunbar, An introduction to the analysis of single-cell RNA-sequencing data, Molecular Therapy - Methods & Clinical Development 10 (2018) 189–196.

[37] S. Picelli, Å. Björklund, O.R. Faridani, S. Sagasser, G. Winberg, R. Sandberg, Smart-seq2 for sensitive full-length transcriptome profiling in single cells, Nat. Methods 10 (2013) 1096.

[38] J.W. Foley, C. Zhu, P. Jolivet, S.X. Zhu, P. Lu, M.J. Meaney, R.B. West, Gene-expression profiling of single cells from archival tissue with laser-capture microdissection and Smart-3SEQ, bioRxiv, (2019) 207340.

[39] R. Rostom, V. Svensson, S.A. Teichmann, G. Kar, Computational approaches for interpreting scRNA-seq data, FEBS Lett. 591 (2017) 2213–2225.

[40] M. Angelova, P. Charoentong, H. Hackl, M.L. Fischer, R. Snajder, A.M. Krogsdam, M.J. Waldner, G. Bindea, B. Mlecnik, J. Galon, Z. Trajanoski, Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy, Genome Biol. 16 (2015) 64.

[41] P. Charoentong, F. Finotello, M. Angelova, C. Mayer, M. Efremova, D. Rieder, H. Hackl, Z. Trajanoski, Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade, Cell Rep. 18 (2017) 248–262.

[42] A. Jiménez-Sánchez, O. Cast, M. Miller, Comprehensive benchmarking and integration of tumour microenvironment cell estimation methods, bioRxiv, (2018) 437533.

[43] D. Aran, Z. Hu, A.J. Butte, xCell: digitally portraying the tissue cellular heterogeneity landscape, Genome Biol. 18 (2017) 220.

[44] D. Tamborero, C. Rubio-Perez, F. Muiños, R. Sabarinathan, J.M. Piulats, A. Muntasell, R. Dienstmann, N. Lopez-Bigas, A. Gonzalez-Perez, A pan-cancer landscape of interactions between solid tumors and infiltrating immune cell populations, Clin. Cancer Res. 24 (2018) 3717–3728.

[45] E. Becht, N.A. Giraldo, L. Lacroix, B. Buttard, N. Elarouci, F. Petitprez, J. Selves, P. Laurent-Puig, C. Sautès-Fridman, W.H. Fridman, A. de Reyniès, Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression, Genome Biol. 17 (2016) 218.

[46] P. Danaher, S. Warren, L. Dennis, L. D'Amico, A. White, M.L. Disis, M.A. Geller, K. Odunsi, J. Beechem, S.P. Fling, Gene expression markers of tumor infiltrating leukocytes, Journal for ImmunoTherapy of Cancer 5 (2017) 1–15.

[47] A.J. Nirmal, T. Regan, B.B. Shih, D.A. Hume, A.H. Sims, T.C. Freeman, Immune cell gene signatures for profiling the microenvironment of solid tumors, Cancer Immunology Research, 44 (2018) canimm.0342.2018.

[48] A.M. Newman, C.L. Liu, M.R. Green, A.J. Gentles, W. Feng, Y. Xu, C.D. Hoang, M. Diehn, A.A. Alizadeh, Robust enumeration of cell subsets from tissue expression profiles, Nat. Methods 12 (2015) 453–457.

[49] B. Li, E. Severson, J.C. Pignon, H. Zhao, T. Li, J. Novak, P. Jiang, H. Shen, J.C. Aster, S. Rodig, S. Signoretti, J.S. Liu, X.S. Liu, Comprehensive analyses of tumor immunity: implications for cancer immunotherapy, Genome Biol. 17 (2016) 1–16.

[50] J. Racle, K. de Jonge, P. Baumgaertner, D.E. Speiser, D. Gfeller, Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data, eLife 6 (2017).

[51] F. Finotello, C. Mayer, C. Plattner, G. Laschober, D. Rieder, H. Hackl, A. Krogsdam, W. Posch, D. Wilflingseder, S. Sopper, D. Johnson, Y. Xu, Y. Wang, M.E. Sanders, M.V. Estrada, P. Ericsson-gonzalez, J. Balko, N.D. Miranda, Z. Trajanoski, quanTIseq: Quantifying Immune Contexture of Human Tumors, 2017.

[52] G. Sturm, F. Finotello, F. Petitprez, J.D. Zhang, J. Baumbach, W.H. Fridman, M. List, T. Aneichyk, Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology, Bioinformatics 35 (2019) i436–i445.

[53] A.R. Abbas, D. Baldwin, Y. Ma, W. Ouyang, A. Gurney, F. Martin, S. Fong, M. van Lookeren Campagne, P. Godowski, P.M. Williams, A.C. Chan, H.F. Clark, Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data, Genes & Immunity 6 (2005) 319–331.

[54] N.A. Mabbott, J. Baillie, H. Brown, T.C. Freeman, D.A. Hume, An expression atlas of human primary cells: inference of gene function from coexpression networks, BMC Genomics 14 (2013) 632.

[55] D.A. Barbie, P. Tamayo, J.S. Boehm, S.Y. Kim, S.E. Moody, I.F. Dunn, A.C. Schinzel, P. Sandy, E. Meylan, C. Scholl, S. Fröhling, E.M. Chan, M.L. Sos, K. Michel, C. Mermel, S.J. Silver, B.A. Weir, J.H. Reiling, Q. Sheng, P.B. Gupta, R.C. Wadlow, H. Le, S. Hoersch, B.S. Wittner, S. Ramaswamy, D.M. Livingston, D.M. Sabatini, M. Meyerson, R.K. Thomas, E.S. Lander, J.P. Mesirov, D.E. Root, D.G. Gilliland, T. Jacks, W.C. Hahn, Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1, Nature 462 (2009) 108–112.

[56] S. Hänzelmann, R. Castelo, J. Guinney, GSVA: gene set variation analysis for microarray and RNA-Seq data, BMC Bioinformatics 14 (2013) 7.

[57] F. Avila Cobos, J. Vandesompele, P. Mestdagh, K. De Preter, Computational deconvolution of transcriptomics data from mixed cell populations, Oxford University Press, Bioinformatics, 2018, pp. 1969–1979.

[58] W. Qiao, G. Quon, E. Csaszar, M. Yu, Q. Morris, P.W. Zandstra, PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions, PLoS Comput. Biol. 8 (2012) e1002838.

[59] T. Gong, J.D. Szustakowski, DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data, Bioinformatics 29 (2013) 1083–1085.

[60] T. Li, J. Fan, B. Wang, N. Traugh, Q. Chen, J.S. Liu, B. Li, X.S. Liu, TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells, Cancer Res. 77 (2017) e108–e110.

[61] H.H.-S. Lu, S.-Y. Su, W.-Y. Kuo, W.-C. Chung, S.-H. Chen, J.-M. Ho, C.-Y. Lin, A gene profiling deconvolution approach to estimating immune cell composition from complex tissues, BMC Bioinformatics 19 (2018) 154.

[62] Y. Hao, M. Yan, B.R. Heath, Y.L. Lei, Y. Xie, Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares, PLoS Comput. Biol. 15 (2019) e1006976.

[63] G.J. Hunt, S. Freytag, M. Bahlo, J.A. Gagnon-Bartsch, dtangle: accurate and robust cell type deconvolution, Bioinformatics 35 (12) (2018) 2093–2099.

[64] D. Repsilber, S. Kern, A. Telaar, G. Walzl, G.F. Black, J. Selbig, S.K. Parida, S.H. Kaufmann, M. Jacobsen, Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach, BMC Bioinformatics 11 (2010) 27–42.

[65] R. Gaujoux, C. Seoighe, Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study, Infect. Genet. Evol. 12 (2012) 913–921.

[66] Y. Zhong, Y.-W. Wan, K. Pang, L.M. Chow, Z. Liu, Digital sorting of complex tissues for cell type-specific gene expression profiles, BMC Bioinformatics 14 (2013) 89.

[67] D.A. Liebner, K. Huang, J.D. Parvin, MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples, Bioinformatics 30 (2014) 682–689.

[68] Z. Wang, S. Cao, J.S. Morris, J. Ahn, R. Liu, S. Tyekucheva, F. Gao, B. Li, W. Lu, X. Tang, I.I. Wistuba, M. Bowden, L. Mucci, M. Loda, G. Parmigiani, C.C. Holmes, W. Wang, Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration, iScience 9 (2018) 451–460.

[69] C. Maurer, S.R. Holmstrom, J. He, P. Laise, T. Su, A. Ahmed, H. Hibshoosh, J.A. Chabot, P.E. Oberstein, A.R. Sepulveda, J.M. Genkinger, J. Zhang, A.C. Iuga, M. Bansal, A. Califano, K.P. Olive, Experimental microdissection enables functional harmonisation of pancreatic cancer subtypes, Gut 68 (2019) 1034.

[70] G. Monaco, B. Lee, W. Xu, S. Mustafah, Y.Y. Hwang, C. Carré, N. Burdin, L. Visan, M. Ceccarelli, M. Poidinger, A. Zippelius, J. Pedro de Magalhães, A. Larbi, RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types, Cell Reports, 26 e1627 (2019) 1627–1640.

[71] I. Tirosh, B. Izar, S.M. Prakadan, M.H. Wadsworth, D. Treacy, J.J. Trombetta, A. Rotem, C. Rodman, C. Lian, G. Murphy, M. Fallahi-Sichani, K. Dutton-Regester, J.-R. Lin, O. Cohen, P. Shah, D. Lu, A.S. Genshaft, T.K. Hughes, C.G.K. Ziegler, S.W. Kazer, A. Gaillard, K.E. Kolb, A.-C. Villani, C.M. Johannessen, A.Y. Andreev, E.M. Van Allen, M. Bertagnolli, P.K. Sorger, R.J. Sullivan, K.T. Flaherty, D.T. Frederick, J. Jané-Valbuena, C.H. Yoon, O. Rozenblatt-Rosen, A.K. Shalek, A. Regev, L.A. Garraway, Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq, Science 352 (2016) 189–196.

[72] C. Zheng, L. Zheng, J.K. Yoo, H. Guo, Y. Zhang, X. Guo, B. Kang, R. Hu, J.Y. Huang, Q. Zhang, Z. Liu, M. Dong, X. Hu, W. Ouyang, J. Peng, Z. Zhang, Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing, Cell 169 (2017) 1342–1356 (e1316).

[73] P. Savas, B. Virassamy, C. Ye, A. Salim, C.P. Mintoff, F. Caramia, R. Salgado, D.J. Byrne, Z.L. Teo, S. Dushyanthen, A. Byrne, L. Wein, S.J. Luen, C. Poliness, S.S. Nightingale, A.S. Skandarajah, D.E. Gyorki, C.M. Thornton, P.A. Beavis, S.B. Fox, C. Kathleen Cuningham, Foundation Consortium for Research into Familial Breast, P.K. Darcy, T.P. Speed, L.K. Mackay, P.J. Neeson, S. Loi, Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis, Nat. Med. 24 (2018) 986–993.

[74] W. Chung, H.H. Eum, H.O. Lee, K.M. Lee, H.B. Lee, K.T. Kim, H.S. Ryu, S. Kim, J.E. Lee, Y.H. Park, Z. Kan, W. Han, W.Y. Park, Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer, Nat. Commun. 8 (2017) 15081.

[75] E. Azizi, A.J. Carr, G. Plitas, A.E. Cornish, C. Konopacki, S. Prabhakaran, J. Nainys, K. Wu, V. Kiseliovas, M. Setty, K. Choi, R.M. Fromme, P. Dao, P.T. McKenney, R.C. Wasti, K. Kadaveru, L. Mazutis, A.Y. Rudensky, D. Pe'er, Single-cell map of diverse immune phenotypes in the breast tumor microenvironment, Cell, 174 e1236 (2018) 1293–1308.

[76] H. Li, E.T. Courtois, D. Sengupta, Y. Tan, K.H. Chen, J.J.L. Goh, S.L. Kong, C. Chua, L.K. Hon, W.S. Tan, M. Wong, P.J. Choi, L.J.K. Wee, A.M. Hillmer, I.B. Tan, P. Robson, S. Prabhakar, Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors, Nat. Genet. 49 (2017) 708.

[77] X. Guo, Y. Zhang, L. Zheng, C. Zheng, J. Song, Q. Zhang, B. Kang, Z. Liu, L. Jin, R. Xing, R. Gao, L. Zhang, M. Dong, X. Hu, X. Ren, D. Kirchhoff, H.G. Roider, T. Yan, Z. Zhang, Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing, Nat. Med. 24 (2018) 978–985.

[78] M. Schelker, S. Feau, J. Du, N. Ranu, E. Klipp, G. MacBeath, B. Schoeberl, A. Raue, Estimation of immune cell content in tumour tissue using single-cell RNA-seq data, Nat. Commun. 8 (2017) 2032.

[79] A.M. Newman, C.B. Steen, C.L. Liu, A.J. Gentles, A.A. Chaudhuri, F. Scherer, M.S. Khodadoust, M.S. Esfahani, B.A. Luca, D. Steiner, M. Diehn, A.A. Alizadeh, Determining cell type abundance and expression from bulk tissues with digital cytometry, Nat. Biotechnol. 37 (7) (2019) 773–782.

[80] E.C. Stack, C. Wang, K.A. Roman, C.C. Hoyt, Multiplexed immunohistochemistry, imaging, and quantitation: a review, with an assessment of Tyramide signal amplification, multispectral imaging and multiplex analysis, Methods 70 (2014) 46–58.

[81] J.-R. Lin, M. Fallahi-Sichani, J.-Y. Chen, P.K. Sorger, Cyclic immunofluorescence (CycIF), a highly multiplexed method for single-cell imaging, Current Protocols in Chemical Biology 8 (2016) 251–264.

[82] K.M. McKinnon, Flow cytometry: an overview, Curr. Protoc. Immunol. 120 (2018) 5.1.1–5.1.11.

[83] D.R. Bandura, V.I. Baranov, O.I. Ornatsky, A. Antonov, R. Kinach, X. Lou, S. Pavlov, S. Vorobiev, J.E. Dick, S.D. Tanner, Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry, Anal. Chem. 81 (2009) 6813–6822.

[84] A. White, N. Elliott, Y. Liang, S.E. Warren, J.M. Beechem, Spatially resolved, multiplexed digital characterization of protein distribution and abundance in formalin-fixed, paraffin-embedded (FFPE) diffuse large B cell lymphoma tissue sections based on the Nanostring® digital spatial profiling (DSP) technology, Blood 130 (2017) 1196.

[85] V.M. Peterson, K.X. Zhang, N. Kumar, J. Wong, L. Li, D.C. Wilson, R. Moore, T.K. McClanahan, S. Sadekova, J.A. Klappenbach, Multiplexed quantification of proteins and transcripts in single cells, Nat. Biotechnol. 35 (2017) 936.

[86] P. Shahi, S.C. Kim, J.R. Haliburton, Z.J. Gartner, A.R. Abate, Abseq: ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding, Sci. Rep. 7 (2017) 44447.

[87] M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P.K. Chattopadhyay, H. Swerdlow, R. Satija, P. Smibert, Simultaneous epitope and transcriptome measurement in single cells, Nat. Methods 14 (2017) 865.

[88] G. Glass, J.A. Papin, J.W. Mandell, Simple: a sequential immunoperoxidase labeling and erasing method, Journal of Histochemistry & Cytochemistry 57 (2009) 899–905.

[89] T. Tsujikawa, S. Kumar, R.N. Borkar, V. Azimi, G. Thibault, Y.H. Chang, A. Balter, R. Kawashima, G. Choe, D. Sauer, E. El Rassi, D.R. Clayburgh, M.F. Kulesz-Martin,

E.R. Lutz, L. Zheng, E.M. Jaffee, P. Leyshock, A.A. Margolin, M. Mori, J.W. Gray, P.W. Flint, L.M. Coussens, Quantitative multiplex immunohistochemistry reveals myeloid-inflamed tumor-immune complexity associated with poor prognosis, Cell Rep. 19 (2017) 203–217.

[90] M.J. Gerdes, C.J. Sevinsky, A. Sood, S. Adak, M.O. Bello, A. Bordwell, A. Can, A. Corwin, S. Dinn, R.J. Filkins, D. Hollman, V. Kamath, S. Kaanumalle, K. Kenny, M. Larsen, M. Lazare, Q. Li, C. Lowes, C.C. McCulloch, E. McDonough, M.C. Montalto, Z. Pang, J. Rittscher, A. Santamaria-Pang, B.D. Sarachan, M.L. Seel, A. Seppo, K. Shaikh, Y. Sui, J. Zhang, F. Ginty, Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue, Proc. Natl. Acad. Sci. 110 (2013) 11982.

[91] L.R. Olsen, M.D. Leipold, C.B. Pedersen, H.T. Maecker, The anatomy of single cell mass cytometry data, Cytometry Part A 95 (2019) 156–172.

[92] J. Wagner, M.A. Rapsomaniki, S. Chevrier, T. Anzeneder, C. Langwieder, A. Dykgers, M. Rees, A. Ramaswamy, S. Muenst, S.D. Soysal, A. Jacobs, J. Windhager, K. Silina, M. van den Broek, K.J. Dedes, M. Rodríguez Martínez, W.P. Weber, B. Bodenmiller, A single-cell atlas of the tumor and immune ecosystem of human breast cancer, Cell, 177 e1318 (2019) 1330–1345.

[93] D. Schulz, V.R.T. Zanotelli, J.R. Fischer, D. Schapiro, S. Engler, X.-K. Lun, H.W. Jackson, B. Bodenmiller, Simultaneous multiplexed imaging of mRNA and proteins with subcellular resolution in breast cancer tissue samples by mass cytometry, Cell Systems, 6 e25 (2018) 25–36.

[94] L. Keren, M. Bosse, D. Marquez, R. Angoshtari, S. Jain, S. Varma, S.-R. Yang, A. Kurian, D. Van Valen, R. West, S.C. Bendall, M. Angelo, A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging, Cell, 174 e1319 (2018) 1373–1387.

[95] D.E. Carvajal-Hausdorf, J. Patsenker, K.P. Stanton, F. Villarroel-Espindola, A. Esch, R.R. Montgomery, A. Psyrri, K.T. Kalogeras, V. Kotoula, G. Foutzilas, K.A. Schalper, Y. Kluger, D.L. Rimm, Multiplexed (18-Plex) measurement of signaling targets and cytotoxic T cells in trastuzumab-treated patients using imaging mass cytometry, Clin. Cancer Res. 25 (10) (2019) 3054–3062.

[96] R. Moncada, F. Wagner, M. Chiodin, J.C. Devlin, M. Baron, C.H. Hajdu, D.M. Simeone, I. Yanai, Integrating single-cell RNA-Seq with spatial transcriptomics in pancreatic ductal adenocarcinoma using multimodal intersection analysis, bioRxiv, (2019) 254375.

[97] J. Decalf, M.L. Albert, J. Ziai, New tools for pathology: a user's review of a highly multiplexed method for in situ analysis of protein and RNA expression in tissue, J. Pathol. 247 (2019) 650–661.

[98] P.L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J.F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J.O. Westholm, M. Huss, A. Mollbrink, S. Linnarsson, S. Codeluppi, Å. Borg, F. Pontén, P.I. Costea, P. Sahlén, J. Mulder, O. Bergmann, J. Lundeberg, J. Frisén, Visualization and analysis of gene expression in tissue sections by spatial transcriptomics, Science 353 (2016) 78.

[99] Y. Simoni, M.H.Y. Chng, S. Li, M. Fehlings, E.W. Newell, Mass cytometry: a powerful tool for dissecting the immune landscape, Curr. Opin. Immunol. 51 (2018) 187–196.

[100] E.Y. Shum, E.M. Walczak, C. Chang, H.C. Fan, Quantitation of mRNA transcripts and proteins using the BD Rhapsody^{TM} single-cell analysis system, Springer, Place Published, Single Molecule and Single Cell Sequencing, 2019, pp. 63–79.

[101] Y. Hu, Q. An, K. Sheu, B. Trejo, S. Fan, Y. Guo, Single cell multi-omics technology: methodology and application, Frontiers in Cell and Developmental Biology 6 (2018) 28.

1.3.2 Highly-multiplexed imaging for tissue-level analysis

Several highly-multiplexed methods have been developed for imaging large numbers of markers from a single tissue section (Table 1). These methods differ mainly in the technology they rely on for marker detection, which then determines its sensitivity, spatial resolution, throughput and maximum imaged area[60]. The primary detection technologies are chromogens, fluorescence, DNA sequencing, and mass spectrometry, all relying on different probes linked to antibodies specific to the markers of interest.

1.3.2.1 Multiplexed immunohistochemistry methods

The first highly-multiplexed imaging approaches to be developed relied on antibodies bound to different chromogens for marker detection. These multiplexed immunohistochemistry (mIHC) based methods (Table 1) were derived from conventional immunohistochemistry through sequential staining and detection or by using multiple chromogens with different colours and multispectral imaging. The first approach is employed by sequential immunohistochemistry, which relies on several cycles of labelling, stripping, and imaging[61]. The imaging is performed on a regular bright-field microscope, and the single-marker images are then coregistered to obtain the final multi-channel image. The use of standard bright-field microscopes, in addition to the higher light levels in bright field microscopy, leads to lower costs and shorter image acquisition times relative to immunofluorescence (IF) microscopy[62]. Also, the already widespread use of conventional IHC approaches can leverage an extensive knowledge base in pathology associated with bright field techniques. More recently, the use of multispectral images and different chromogenic dies allowed the imaging of the slide in a single step[63]. However, approaches relying on multispectral imaging are generally limited in the number of detectable markers because of the overlaps between the absorption spectra of the chromogens[63]. Finally, a limitation of both types of mIHC methods is the limited range where optical densities linearly correlate with marker levels[64].

1.3.2.2 Multiplex immunofluorescence methods

Multiplex immunofluorescence (mIF) methods (Table 1) have the advantage of being quantitative in a much wider dynamic range compared to mIHC methods. Fluorescent emissions have a linear and additive nature and relatively well-defined emission spectra. Fluorescence-based platforms are multiplexed through spectral unmixing or sequential staining. The first approach is employed by the Vectra Polaris (Akoya Biosciences) to measure up to nine markers simultaneously in Formalin-fixed paraffin-embedded (FFPE) or fresh tissue sections. Higher levels of multiplexing can be achieved at the cost of more extended preparation and acquisition times with sequential protocols such as Cyclic immunofluorescence (CyCIF)[65] (Figure 1.2 a). An improved version of this approach t-CycIF[66] increases throughput by imaging up to four channels simultaneously in each cycle. Finally, t-CycIF allows the use of fluorophores conjugated to secondary antibodies in the last cycle to increase sensitivity for a specific subset of markers.

1.3.2.3 Methods based on DNA-probes

Methods based on DNA-probes (Table 1) use antibodies conjugated to unique oligonucleotide sequences to reach high levels of multiplexing, comparable to sequential mIF methods while maintaining a high throughput. These high throughputs are achieved by staining with all antibodies in a single round and then using complementary probes or single nucleotides labelled with fluorophores for detection. InSituPlex employs complementary probes[67] to detect up to 15 markers from FFPE tissue slides[62]. More recently, the co-detection by indexing (CODEX)[68] method (Figure 1.2 b) enabled the imaging of up to 60 markers using fluorophore labelled nucleotides. A fully automated fluidics system performs several rounds of labelling by adding single labelled nucleotides and imaging. The sequence of all probes is derived simultaneously by reading the fluorescence signal associated with each nucleotide with commonly available three-coloured fluorescence microscopes [69]. This staining and detection technology enables CODEX to produce staining patterns comparable to IHC and IF methods.

1.3.2.4 Mass spectrometry-based methods

Mass spectrometry-based methods (Table 1) leverage antibodies bound to heavy metal cations, mostly lanthanides, to image tissues with up to 50 markers[60]. The use of mass spectrometry for detection enables the use of large panels by completely removing the need to account for the spectral overlap between different chromogens or fluorophores. Additionally, these technologies are not affected by the natural autofluorescence of biological tissues. However, these approaches can have issues with steric hindrance between labelled antibodies on the tissue and non-specific binding between antibodies and the tissue, thus requiring extensive panel optimisation.

The two main methods relying on this technology are Multiplexed Ion Beam Imaging (MIBI)[70] (Figure 1.2 c) and Imaging Mass Cytometry (IMC)[71] (Figure 1.2 d). MIBI employs an ionic beam, while IMC relies on a ultraviolet (UV) laser to ionise the sample by sequentially ablating small square spots. A magnetic sector mass analyser in MIBI and a time-of-flight analyser in IMC separate the heavy metal cations. The separated ions are quantified in each ablated spot on the tissue, and each ablated spot will constitute a pixel in the final image with a channel for each metal and intensity equal to that metal's measured ion current. Image acquisition is slow and expensive because of the need to scan the slide and ablate every single spot separately. For this reason, both methods are generally not applied to image the whole slide but to regions of interest (ROIs) of about 1mm^2. MIBI can potentially image a higher number of markers than IMC and reach a higher resolution of 0.25µm compared to the 1µm of IMC at the expense of longer acquisition times.

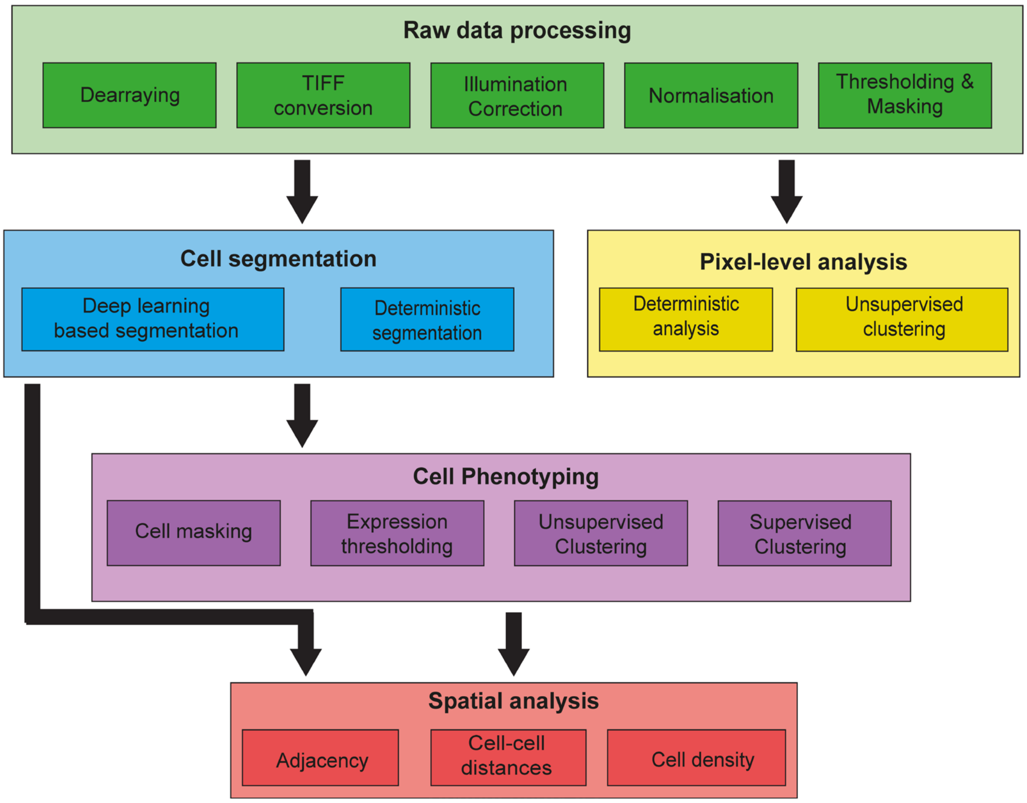
The following section will examine the information content of highly-multiplexed images, outline the main steps of their analysis to investigate tissue structure, composition and function, and provide an overview of the software tools currently available for the analysis.

Table 1: highly-multiplexed imaging technologies

Method	Technology	Markers	Imaged Area	Resolution
Sequential mIHC[61]	Chromogenes	< 12 proteins	Whole Slide	>0.2µm
Discovery ultra[63]	Chromogenes	6 proteins	Whole Slide	>0.2µm
Vectra Polaris (Akoya Biosciences)	Fluorescence	< 9 proteins	Whole Slide	>0.2µm
CyCIF[65]	Fluorescence	<60 proteins or RNA	Whole Slide	>0.2µm
t-CycIF[66]	Fluorescence	<60 proteins or RNA	Whole Slide	>0.2µm
InSituPlex[67]	DNA probes	8 proteins	Whole Slide	>0.2µm
CODEX[68]	DNA probes	<60 proteins	Whole Slide	>0.2µm
MIBI[70]	Mass spectrometry	<100 proteins	Above 1 mm^2 high costs and run times	>0.2µm
IMC[71]	Mass spectrometry	<50 proteins or RNA	Above 1 mm^2 high costs and run times	1µm

For each method, are reported its core technology, the maximum number and type of imaged markers, the maximum imaged area, and its spatial resolution.

Figure 1.3 Highly-multiplexed image information and analysis



Analysis of the three levels of information in highly multiplexed images. Flowchart showing the five steps of the workflow for the analysis of highly multiplexed images: raw data processing, pixel-level analysis, cell segmentation, cell phenotyping and spatial analysis. Inside each step are reported the main processes, which can be performed at that analysis stage.

1.4 Analysis of highly-multiplexed image data

Highly-multiplexed images of tissues contain three levels of information: the marker level or pixel level, the cell level and the spatial level[55]. The analysis of information from all three levels is necessary for the comprehensive characterisation of the tissue.

1.4.1 Information obtainable from highly-multiplexed image data

The first level of information contained in highly-multiplexed image data consists of the intensity values of each marker in every pixel, which is directly related to the expression of these markers in the tissue. While an analysis at the pixel level allows only an indirect quantification of the cell composition of the tissue, at the same time, it is not impacted by the biases or artefacts, which can be introduced by cell segmentation.

The second level of information is the cell-level data acquired by performing cell segmentation on the images. These data consist of the expression values of all markers in the pixels belonging to each cell, generally reduced to their summary statistics and morphological features. The most commonly measured morphological features include area, perimeter, eccentricity and solidity, and the cell's centroid coordinates in the image. This information can then be used to classify cells according to their lineage or functional phenotypes. After cells have been phenotyped, the cell composition of the tissue or its compartment can then be quantified, and the proportions of cells associated with each phenotype can be compared across samples and experimental conditions.

Finally, the third level of information is spatial information, the distribution of cells within the tissue and each of its compartments. The spatial distribution of a single cell phenotype or lineage at a time can be investigated by analysing the density distribution of the cells within the image. These features can then be compared within and across samples to investigate the heterogeneity of cellular patterning in the tissue. The possibility of leveraging this third level of information constitutes the main advantage of imaging-based technologies over single-cell sequencing technologies for performing biological analysis at the tissue level.

1.4.2 Highly-multiplexed image analysis workflow

The workflow for analysing highly multiplexed images can be divided into five main steps: raw data processing, pixel-level analysis, cell segmentation, cell phenotyping and spatial analysis (Figure 1.3). These steps are highly variable and are often skipped or adapted according to the tissue of interest, imaging modality, and research questions.

1.4.2.1 Raw data processing

The raw data processing step consists of processing the imaging data produced by the instrument to generate images, which can be used as input for the downstream steps of the analysis (Figure 1.3). In most cases, the images need to be converted from the file formats output by the instrument to TIFF, the most commonly employed by highly-multiplexed image analysis tools. TIFF images can be single-channel when the intensities of each marker are stored in a separate file or multi-channel when a single file is used for all markers.

Other operations commonly performed at this stage are tumour microarray (TMA) dearraying, stitching and registration, and illumination correction. TMA dearraying consists of identifying each spot in the whole slide image of the TMA and saving it as a separate TIFF image. Illumination correction is often required for images from digital microscopy modalities and removes non-homogeneous illumination across the image field. Stitching and registration are required when the imaged area is larger than the microscope field of view (FOV) to derive a single image by aligning and merging the acquisitions from multiple overlapping FOVs.

After the necessary operations have been completed, the resulting TIFF images can then undergo normalisation and further processing to remove background artefacts or to produce binary masks to be used as input for the following analysis steps. Normalisation is often performed because it allows the use of similar thresholds across images both for preprocessing and cell phenotyping. However, normalisation can also introduce artefacts when staining is not uniform within the same tissue or when images from samples with different properties issue-specific properties (tissue age, length of time in fixation solution) are analysed together. This latter case can be mitigated by performing the normalisation independently for each sample and channel.

Binary masks can be derived from the thresholding of the images of one or more markers and their combination through Boolean operators. Alternatively, these masks can be derived from a small set of user annotations by performing pixel classification with random-forest classifiers or other machine-learning algorithms. These masks can then be applied to the raw or the normalised images to remove background signals or artefacts or used in the downstream analysis to identify specific tissue compartments or cell types.

1.4.2.2 Pixel-level analysis

The pixel-level analysis requires as input the intensities of all markers in the sample, generally as a single- or multi-channel TIFF image, and optionally binary masks defining specific tissue compartments or cell types. The intensity values of all markers in each pixel are then analysed with a deterministic approach or by unsupervised clustering (Figure 1.3). The deterministic approach measures the pixel intensities for selected markers in the image. Alternatively, if binary masks were derived for the markers of interest, the positive areas of these masks can be measured by counting the number of non-zero pixels in the mask. These measurements can then be related to the presence and prevalence of specific cell types and biological processes in the tissue.

The unsupervised approach consists of the unsupervised clustering of all pixels according to their marker intensity values. Then the median intensity values of all markers in each cluster can be employed to identify the cell types, tissue structures, or biological processes associated with each cluster.

The pixel-level analysis relies only on the first level of information in highly-multiplexed biological images and is fully cell-agnostic. Its independence from cellular features makes it a powerful approach for analysing tissues or cell types, in which cell segmentation is problematic. Finally, the pixel-level analysis can be used as further validation of the results derived from analysis performed at the cell-level[55].

1.4.2.3 Cell segmentation

The cell segmentation step identifies cells in an image by determining which pixels constitute a cell. This step enables the extraction of the single-cell data from the image, which constitute the second level of information in highly multiplexed images. There are two main approaches for this step: deterministic cell segmentation and deep-learning-based segmentation (Figure 1.3).

Deterministic segmentation generally relies on marker-controlled watershed algorithms[72]. With these approaches, an intensity gradient transformation is applied to the input image, then local maxima are identified in the input image with an h-maxima transform or other approaches. Finally, the watershed algorithm is applied to the gradient image and the local maxima are used as seeds; only areas containing at least one seed are considered. Multiple variations of this basic approach were developed to improve the selection of seeds and avoid over-or under- segmentation[72]. Deterministic segmentation approaches do not require manual training or large labelled datasets like the deep-learning methods. However, to obtain an accurate segmentation, fairly extensive parameter tuning and empirical testing is required[72].

More recently, deep-learning methods have been successfully applied to cell segmentation. These approaches consist of training a model using ground truth datasets of cell images acquired from the same or similar issues with the same imaging modality and manually annotated by experts. These models can then be applied to the segmentation of new images. Several deep-learning frameworks and architectures have been applied to cell segmentation including: Stardist[73], NucleAIzer[74], CellPose[75], or Mesmer[76].

Machine learning algorithms generally outperform deterministic segmentation approaches; however, model training and evaluation are labour intensive and require computational expertise, which is unavailable to all users[77].

Nuclei are commonly used as a target for segmentation because most cells have one nucleus and nuclear stains, DNA intercalators, or antibodies targeting histones are available for most imaging modalities with high signal-to-noise ratios. However, membrane markers are often used to increase segmentation accuracy and include the cytoplasm and membrane in the segmented objects.

The output of the cell segmentation step is the single-cell data, which consists of the expression values of all markers in the pixels belonging to each cell and can additionally include morphological features[55]. The pixels belonging to each cell are also uniquely labelled by producing a cell mask in which the pixels of each cell are assigned a unique value, generally stored in 16-bit integer format.

1.4.2.4 Cell phenotyping

After single-cell data has been acquired, cells can be assigned to specific populations, compartments or functional phenotypes. The most straightforward approaches for single-cell phenotyping consist of classifying cells according to threshold on the expression of user-selected markers or to the overlap with masks defining specific tissue compartments or populations. The masks used in this step need to be produced in the raw data processing step and can effectively classify cells belonging to different tissue compartments (Figure 1.3). Expression thresholding is applied in the same way as it is performed for techniques measuring fewer markers and non-imaging-based techniques like flow-cytometry. These hypotheses driven approaches can accurately identify previously known cell populations and phenotypes. However, these methods become impractical for tissue wide phenotyping because of the need to identify appropriate thresholds and masks for many markers. This limitation can be overcome by using supervised clustering methods, which can automatically classify cells into pre-selected phenotypes without the need for the user to select a threshold for each marker.

More recently, unsupervised clustering approaches have been applied to cell phenotyping. These discovery-driven methods require only the selection of the marker expression and morphological features to be used as input with no previous knowledge of the populations to be identified. Phenotyping by unsupervised clustering can be performed with multiple software applying different clustering algorithms like Phenograph[78], FlowSOM[79], Seurat[80].

The unsupervised clustering results then need to be annotated to enable their interpretation in the context of the physiology, pathology, and morphology of the tissue. This interpretation is generally assisted by graphical visualisations of the clustered cells and the marker expression that characterises them. These visualisations can be heatmaps showing the median expression values of each marker in every cluster or t-distributed stochastic neighbour embedding (tSNE)[81] and Uniform Manifold Approximation and Projection (UMAP)[82] plots.

1.4.2.5 Spatial analysis

After cells have been assigned to different lineages or phenotypes, the third level of information in the image can be leveraged to study the cell interactions that define the organisation and function of the tissue[55] (Figure 1.3). The spatial distribution of cells in the tissue can be compared across experimental conditions or against a random distribution obtained by computing cell adjacency frequencies during multiple rounds of permutation of the cell population or phenotype annotations. This analysis enables the identification of cell populations or phenotypes, which interact preferentially or are separated in the tissue organisation.

The spatial distribution of cells in the tissues can be calculated differently. The simplest approach quantifies the frequencies at which each cell population or phenotype is directly adjacent to every cell population or phenotype. These frequencies can be calculated from the cell mask produced in the cell-segmentation step. This approach can be extended by considering, instead of only directly adjacent cells, all cells within a user-specified number of pixels. The resulting set of cell-cell interactions can then be analysed with tools like imcRtools[83].

An alternative way to measure the distribution of two cell populations or phenotypes within the tissue is to calculate all the minimum distances between cells of two different populations or phenotypes of interest. The cell-cell distances are calculated as the Euclidean distances between the centroids of the cells of interest.

Finally, the distribution of a cell type or population can be measured as cell density (Figure 1.3). This approach identifies regions in the image or tissue where the density of cells of a phenotype or population is above a user-defined threshold. The cell densities across the tissue can be measured with a sliding window like in CytoMAP[84] or with distance-based clustering algorithms like Density-based spatial clustering of applications with noise (DBSCAN)[85].

1.4.3 Software for highly-multiplexed image analysis

Several software tools implemented the highly multiplexed image analysis workflow or some of its steps (Table 2). These differ in terms of the supported imaging technologies, the image analysis steps they cover (Figure 1.3), and the algorithms, tools, and libraries they employ for each step. These tools can be divided into two main categories: analysis pipelines and interactive tools.

1.4.3.1 Highly-multiplexed image analysis pipelines

Analysis pipelines are non-interactive software in which every step of the analysis is configured before running the analysis, which then proceeds without user interaction. This software generally does not have a graphical user interface (GUI), and the user needs to rely on other tools like cytomapper[86] to visualise the results of the analysis and any intermediate output. Analysis pipelines are generally developed for imaging modalities with a high number of markers like IMC[71] and CODEX[68].

Examples of such software are imcyto[87] and ImcSegmentationPipeline[88] for IMC[71] and CODEX Toolkit[68] for CODEX[68] (Table 2). A recently released analysis pipeline designed for the analysis of multiple modalities of highly multiplexed images is MCMICRO[89]. MCIMICRO[89] enables the technology-agnostic processing of highly-multiplexed images for pixel-level and cell-level analyses. A software with a Graphical User Interface (GUI) allowing the user to build analysis pipelines that are then executed non-interactively is CellProfiler4[90]. This tool can operate with images from all imaging modalities, but the GUI is inconvenient for configuring workflows with multiple markers, as most processes need to be configured independently for each marker. While plugins have been developed for simplifying the application of CellProfiler4[90] to IMC[71] derived data. Pipeline configuration with CellProfiler4[90] can still be time-consuming and labour intensive when dealing with tens of markers.

1.4.3.2 Interactive software for highly-multiplexed image analysis

Interactive analysis software allows the user to actively select, configure and run different analysis steps interactively through a GUI. The GUI enables both the configuration of the analysis and the visualization of results, making this software generally more user friendly than analysis pipelines. This is particularly the case for analyses that require the manual annotation of ROIs in whole slide images or the manual classification of pixels and cells for the training of deep-learning models.

An example of software focused on these applications (Table 2) is Ilastik[91], which provides a user-friendly interface for the training of deep-learning models for pixel and cell classification and cell segmentation. These models can then be applied directly from the GUI or run in headless mode without the GUI. Another software providing the ability to run the analysis also in headless mode is QuPath[92], which also provides an extensive scripting framework. The previous examples of tools were designed for the analysis of images with a lower number of markers and can make the processing of MIBI[70], IMC[71] or CODEX[68] images cumbersome. histoCAT++[93] is an interactive tool for the analysis of IMC[71] data, but is however less flexible as it does not provide any options to extend its capability through plugins or scripting.

1.4.3.3 Main features of highly-multiplexed image analysis software

Essential features to consider in the design or selection of software for highly-multiplexed image analysis are reproducibility, portability, scalability and flexibility[94].

Generally, image analysis pipelines offer a higher standard of reproducibility than Interactive analysis software[95]. This higher reproducibility derives from the greater ease of saving and sharing the pipeline configuration without the need for macros or project files as those used by software like QuPath[92] and Ilastik[91]. The reproducibility of analysis pipelines is further increased by the use of automatic workflow managers like Nextflow[96] and Galaxy[97] in the software implementation[95, 98]. Examples of pipelines implemented with this workflow manager are imcyto[87] and MCMICRO[89], which also has a Galaxy[97] implementation.

The portability of the analysis software (i.e., the ability to run it on different computing platforms) is strictly linked to both the reproducibility and the scalability analysis [98]. This is particularly the case for High-Performance Computing (HPC) environments, which are required for experiments involving tens or hundreds of samples that would likely be impossible or too time-consuming on regular desktop hardware. Interactive tools often need to be preconfigured on a different system and then run in headless mode in the HPC environment, with the user manually managing the available cores and memory. Instead, pipelines relying on workflow management tools like Nextflow[96] have a significant advantage of being able to be seamlessly deployed on most HPC platforms and automatically allocate memory and cores to each process in the pipeline[95].

Finally, image analysis software needs to be flexible enough to analyse a large variety of tissues to answer a large variety of possible research questions. Interactive software tends to be highly flexible and often has extensive collections of publicly available plugins, like CellProfiler4[90]. Other tools are further extendable by providing a scripting interface, like QuPath[92], which supports scripting in Java. Currently available highly-multiplexed imaging analysis pipelines, while being scalable and reproducible, generally lack flexibility. This is caused by multiple factors, including the possibility to select only one tool or algorithm for most steps of the pipeline. Additionally, these pipelines often do not allow the user to skip processes or start and terminate the pipeline at any point of the analysis workflow. This is the case for imcyto[87], and MCMICRO[89] with the latter only allowing the user to select between alternative tools to use for some steps of the analysis.

Table 2: Software for the analysis of highly-multiplexed imaging data

Software	Raw data processing	Pixel-level analysis	Cell segmentation	Cell phenotyping	Spatial analysis	Interactive	GUI	Parallelisation	Imaging technologies
CODEX Toolkit[68]	✓	✘	✓	✓	✓	✘	✘	✘	3
ImcSegmentationPipeline[88]	✓	✘	✓	✘	**Partial**	✘	✘	✓	1
Imcyto[87]	✓	✘	✓	✘	✘	✘	✘	✓	1
CellProfiler4[90]	✓	✓	✓	✓	**Partial**	✘	✓	✓	1-6
HistoCAT++[93]	✓	✘	✓	✓	✓	✓	✓	✘	1,2,4,5
QuPath[92]	✓	✘	✓	✓	✓	✓	✓	✓	1-6
Ilastik[91]	✘	✓	✓	✓	✘	✓	✓	✓	1-6
MCMICRO[89]	✓	✓	✓	✓	✘	✘	✘	✓	1-6

For each software, are reported the parts of the highly-multiplexed image analysis workflow it covers, if it can be parallelised, if it is interactive, if it has a Graphical User Interface (GUI), and the multiplexed imaging platform it can be applied to (1: Imaging Mass Cytometry; 2: multiplexed immunofluorescence; 3: co-detection by indexing; 4: Multiplexed Ion Beam Imaging; 5: multiplexed immunohistochemistry; 6: spatial transcriptomic visualisation). A method was considered compatible with a given imaging platform in the original publication or other studies.

1.5 Aims

This thesis aimed to perform a comprehensive investigation of the genetic and immunological factors shaping the response to ICI in CRC. In particular, this investigation focused on identifying the differences in the composition and spatial organisation of the TME of patients who had a durable benefit from anti-PD1 treatment and those who did not.

To achieve this result, I developed a new software tool for analysing highly multiplexed images: Single-cell Identification from MultiPLexed Images (SIMPLI). I then tested SIMPLI on data generated from different imaging technologies and human tissues. I described this software in a recently published article of which I am the first author [99] (Chapter three).

I then applied SIMPLI to analyse IMC images from CRC patients treated with anti-PD1 CRC. I performed this analysis of the TME in the context of a multiregional and multiomic study of samples from CRC patients treated with pembrolizumab (KEYNOTE 177 clinical trial[52]) or nivolumab. This work aimed to characterise the cellular and molecular determinants of response to anti-PD1 ICI. We reported the results of this study in an article of which I have co-first authorship (Chapter four).

Throughout all these studies, I have relied on the repertoire of CRC cancer driver genes stored in the NCG database[7]. NCG is a web resource maintained by our group, which provides annotations on cancer genes and their systems-level properties. In this study, we employed the gene annotations in the database to interpret somatic mutation and gene expression data and the design of all the IMC and mIF panels.

During the 2019 update of the NCG database, we studied the heterogeneity of drivers across primary sites and cancer types[100]. A description of this resource, together with the results of the accompanying analysis of cancer drivers, was published in a 2019 paper, of which I am a co-first author[100] (Chapter five).

In the 2022 iteration of the database the Network of Cancer Genes and Healthy Drivers NCG^{HD} [7], we expanded our analysis of inter-tumour heterogeneity to include both genes whose driver role depends on mutations in their noncoding sequences and genes identified as drivers of non-malignant clonal expansion in non-cancer tissues[7]. We also collected the systems-level properties of these two additional gene categories and evaluated their heterogeneity across cancer types and primary sites. This database expansion and relative analysis is the subject of a 2022 paper of which I share the co-first authorship (Chapter six).

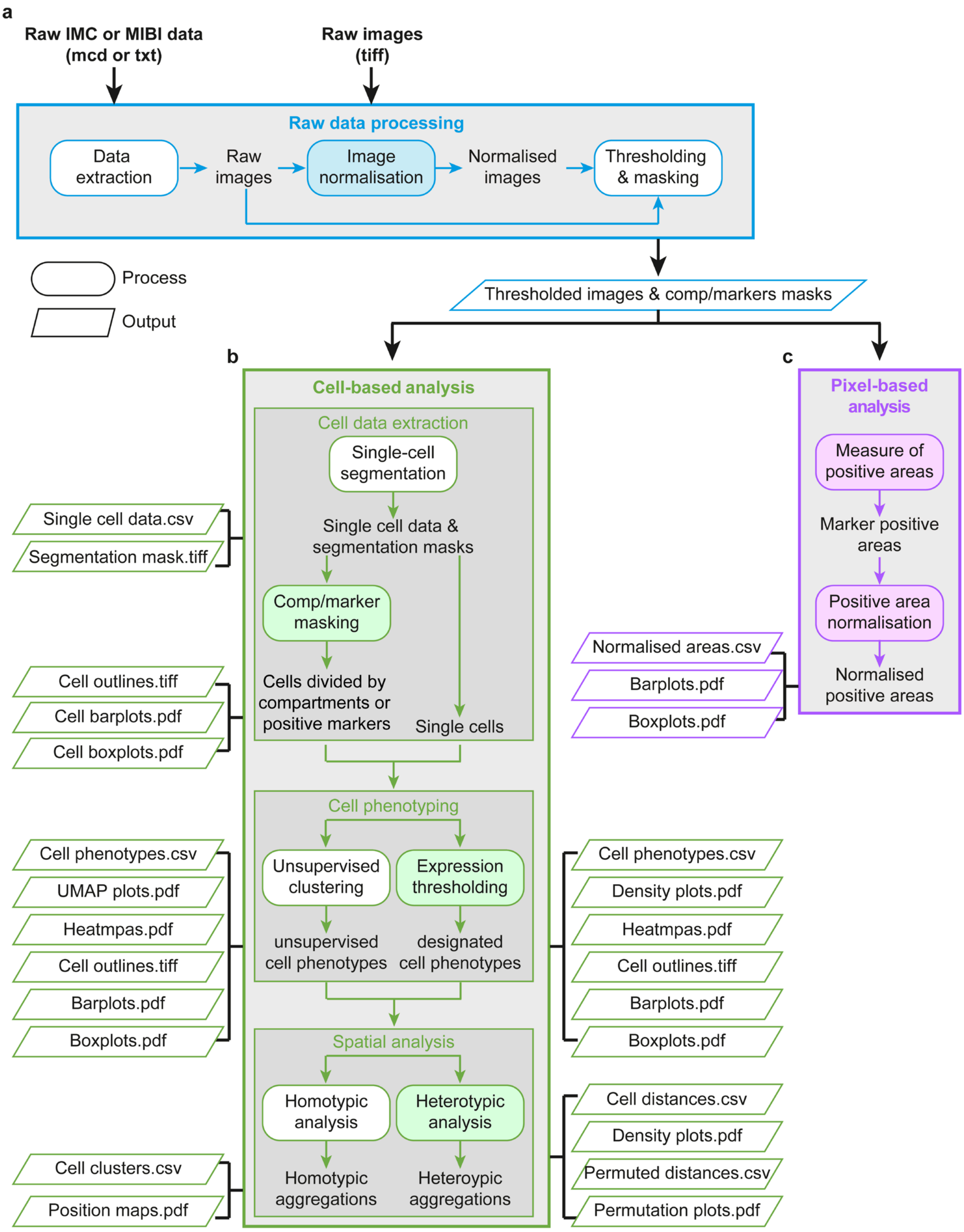
Chapter 2. Materials and Methods

2.1 SIMPLI's algorithm and implementation

SIMPLI is implemented as a Nextflow[96] pipeline, which manages the execution of several R, Python and Bash scripts. These scripts are either custom-coded or wrappers around different software tools. These tools and their dependencies are managed through three Singularity containers[101] hosted at Sylabs.io. SIMPLI's workflow is divided into three main steps: raw data processing, cell-based analysis, pixel-based analysis. Each step is composed of multiple independent processes. These processes produce Tagged Image File Format (TIFF) images and masks or tables in Comma Separate Values (CSV) format.

Additionally, most SIMPLI processes in both the cell-based and pixel-based analysis steps can optionally produce plots to visualise the analysis results. These plots are produced with additional custom R scripts relying on the ggplot2 package and are saved as Portable Document Format (PDF) files. The user can select which processes to execute or skip and from which process to start the analysis. Additionally, the user can disable the plotting of results altogether or enable it only for specific purposes. These settings, together with the main parameters for most processes, can be configured through a Nextflow[96] configuration file or directly through the command line. Finally, if the user chooses not to execute the upstream processes of a given process or step, the paths required input files are specified in a user-supplied CSV metadata file, without the need for a specific folder structure. The full workflow of SIMPLI is described in the following paragraphs.

Figure 2.1 SIMPLI workflow diagram



The workflow of Single-cell Identification from MultiPLexed Images (SIMPLI) consists of three main steps (rectangles): raw data processing (a), cell-based analysis (b) and pixel-based analysis (c). Each step is formed by multiple stand-alone processes (rounded rectangles). These processes are implemented as R, python or bash scripts, which rely on custom code to perform their function (blue, green and pink) or are wrappers for established tools and libraries (white). Each process can produce outputs, including images, tables or plots (parallelograms).

a. Raw data processing. Raw acquisition data in Mass Cytometry Data (MCD) or Text (TXT) format are converted to Tag Image File Format (TIFF) images with the imctools[102] library. The resulting images or user-provided TIFF files can optionally be normalised with a custom R script. The final process of this step is the thresholding and masking of the raw or the normalised images with a user-provided CellProfiler4[90] pipeline to generate masks for each marker and optionally for tissue compartments as well as images for the following steps.

b. Cell-based analysis. This workflow step includes cell data extraction, cell phenotyping, and spatial analysis. First, cell segmentation is performed with CellProfiler4[90], StarDist[73] or both. The resulting single-cell data can be assigned to tissue compartments or populations according to their overlap with masks produced in the previous step or provided by the user. Then, cells can be further phenotyped by performing unsupervised clustering with Seurat[103] v2.3.0 or by applying expression thresholds to one or more markers with an ad hoc R script in the expression thresholding process. Finally, the spatial distribution of cells in the tissue can be analysed to determine the presence of aggregations of one cell type (homotypic analysis process). This analysis is performed with an R script applying the DBSCAN[85] algorithm with the fpc[104] R package. Alternatively, the distribution of the minimum distances between cells of two user-defined cell types can be quantified with the heterotypic analysis process.

c. Pixel-based analysis. The areas positive for a specific marker or user-defined combination of markers are measured from masks provided by the user or derived from the raw data processing step. These measures are then normalised over the area of the whole image or tissue compartments. This step is performed using an ad hoc R script.

This figure was reproduced from figure S1 of Bortolomeazzi et al[99].

2.1.1 Raw image processing

Three processes constitute the raw image processing step: data extraction, normalisation, thresholding, and masking. The data extraction process is required to extract images from the raw acquisition data in mass cytometry data (MCD) or text (TXT) format produced by imaging mass cytometry [71] experiments. In this process, a python script using the imctools[102] v.1.0.7 library is applied to extract images in uncompressed 16-bit TIFF format, which are then used as input for the downstream processes. The user can choose the output images to be single multi-channel images with one channel per marker or single-channel images, one for each marker. Data extraction should be skipped if the input images are already 16-bit TIFF files. The following process is data normalisation. This process performs 99th percentile normalisation of the raw TIFF images generated in the Image extraction process or specified by the user if the image extraction process is skipped. Images are normalised individually by marker and by sample. This process also provides the option to format the output as uncompressed 16-bit multi- or single-channel images. Next is the thresholding and masking process, which employs a user-supplied CellProfiler4[90] pipeline in CellProfiler pipeline (CPPIPE) format. This process is used to perform the image manipulations that will generate the processed images and masks, which can then be used as input for the pixel-based or cell-based analysis. The input images for this process can be derived from images generated in the image normalisation process, images generated in the Image extraction process if the normalisation process is skipped, or user-specified images. The CellProfiler4[90] pipelines employed by SIMPLI need to be built with CellProfiler[90] versions >4.0.7 and configured to run in a non-interactive environment. The pipeline needs to process images from one sample at a time, as parallelisation is achieved by running multiple instances of the pipeline. Finally, the pipeline should be configured to output images as single-channel 16-bit TIFF files. After this project, the pipeline can perform one or both of the next two steps in parallel.

2.1.2 Cell-based analysis

The cell-based analysis consists of extracting single-cell data by cell segmentation and compartment/marker masking. The resulting single-cell data can then be employed for phenotyping user-defined cell populations. Then, the distribution of cells belonging to these populations or phenotypes can be quantified with the spatial analysis processes.

2.1.2.1 Single-cell data extraction

Single-cell data extraction in SIMPLI consists of single-cell segmentation and compartment/marker masking. Segmentation in SIMPLI can be performed with two processes: deterministic segmentation and deep learning segmentation. Both processes can take as input the images and masks produced by the thresholding and masking process, or user-supplied images. These images need to be single-channel TIFF files. Deterministic segmentation is performed with a user-specified CellProfiler4[90] pipeline run independently on each sample. Cells can be identified as primary, secondary, or tertiary objects, or segmentation can also be performed with the watershed module. Deep learning segmentation is performed with StarDist[73] v.0.7.3 using one of the default models bundled with StarDist[73] or a user-provided model. These models need to operate in YXC format, with rows as the first dimension, columns as the second, and all other dimensions being marker intensity values. The probability threshold used for calling cells and the overlap threshold above which Non-Maximum Suppression (NMS) is performed can be derived from the default values included in the model or provided by the user. The output of both segmentation processes consists of cell masks and single-cell data. Cell masks are 16-bit integer TIFF files in which the pixels belonging to each cell are assigned a unique value. Single-cell data consists of a CSV file with a row for each cell, containing the unique identifier mapping the cell to the pixel in the cell mask, the minimum, maximum, and mean intensity values for each marker and the X and Y coordinates of the cell's centroid in the image. Additional columns with spatial features of the cell can also be included. The user can choose to perform segmentation with one or both processes and from which process the single-cell data and cell masks should be taken as input for downstream analyses or provide compatible cell masks and single-cell data.

After segmentation, the pipeline can include the compartment/marker masking process, which allows identifying cells belonging to different populations or tissue compartments according to the overlap of their areas with those of specific masks. The masks and the overlap thresholds to apply are selected by the user through an *ad hoc* metadata file and applied in the order they were specified. The compartment/marker masking process output consists of the single-cell data annotated with the population or compartment to which each cell was assigned. Cells not overlapping any mask by a fraction greater than the threshold are marked as unassigned. The results of this process can be visualised as position maps showing the outlines of every cell coloured according to its population or compartment. Additionally, their proportions are quantified in barplots, and if at least two samples from two different categories were analysed, the proportions are compared across categories and visualised as boxplots.

2.1.2.2 Cell Phenotyping

Cells belonging to user-defined compartments or populations can then be further phenotyped with two independent processes: unsupervised clustering and expression thresholding. The user can run none, one, or both of these processes in parallel. The output of this process consists of the single-cell data with additional columns indicating to which phenotype or cluster the cells were assigned.

The unsupervised clustering process performs unsupervised clustering on cells from one or more populations or compartments with a user-defined set of markers. The clustering is performed with the R package Seurat[103] v2.3.0. The Euclidean distances between the cells to cluster are calculated in a principal component analysis space and used to derive a k-nearest neighbour graph with the cells as nodes. These are then clustered by partitioning the graph with the Louvain algorithm at user-defined levels of resolution, thus generating a set of cell phenotypes for each level of resolution. The unsupervised clustering results can optionally be plotted for each level of resolution. Clusters of cell phenotypes are plotted as scatterplots in UMAP[82] space, coloured by cluster, sample, and by the expression level of each marker employed in the clustering. The UMAP projection is calculated with the R package uwot[105] v.0.1.1. Additionally, the median expression of each marker in every cluster is visualised as a heatmap. Finally, if the analysis includes at least two samples from two different categories, the distributions of the proportion of cells of each cluster in every sample are compared and visualised as boxplots.

The other phenotyping process is expression thresholding, in which cells from user-defined compartments or populations are phenotyped by applying user-defined thresholds to marker expression values. These thresholds can be combined into the Boolean expression to allow the selection of cell phenotypes defined by the expression of more than one marker. The distribution of the expressions values of each marker employed in the thresholding across all cells of interest is optionally visualised as density plots. The median marker expression values in each phenotype identified by thresholding are reported as heatmaps, and the proportions of cells of each phenotype can be shown in boxplots.

2.1.3 Spatial analysis

After cells have been assigned to populations, compartments or phenotypes, two spatial analysis processes can be employed to quantify distances between cells of the same (homotypic aggregations) or different (heterotypic aggregations) types from their centroid coordinates. The homotypic aggregation process is used to identify clusters of cells of the same type within a user-defined distance with DBSCAN[85] as implemented in the fpc[104] v.2.2.9 R package. Homotypic aggregations are optionally visualised as position maps reporting the location of all cells of interest and highlighting cells belonging to high-density clusters as well as the cluster borders.

The heterotypic aggregation process calculates the minimum Euclidean distances between cells of two user-defined cell types in each sample with a custom R script. The distributions of minimum distances produced with the heterotypic aggregations process can be visualised as density plots. The distributions are also plotted by category if the analysis involves at least two samples divided into two categories. Finally, the results of the heterotypic analysis can be compared to those expected from a random distribution of cells through a permutation test. The permutation test results are also represented in the output as density plots.

2.1.4 Pixel-based analysis

The third step in SIMPLI’s workflow is the pixel-based analysis, which can be run as a standalone workflow or directly downstream from the raw image processing step. The user can provide the input masks or they can be derived from the raw image processing step. The pixel-based analysis is independent of the cell-based analysis and can be run in alternative or parallel to it. In this step, the areas of the masks for specific markers or user-defined Boolean combinations of markers are measured with a custom R script relying on the EBImage[106] v.3.10 package.

The images from each marker selected by the user are loaded, and if they are not already binary masks, pixels with an intensity > 0 are set to 1. If the user has selected to measure the area of a combination of masks, the Boolean expression is evaluated, treating the masks as binary matrices. Then the areas of the resulting masks are measured by counting all positive pixels by taking the sum of all pixels in the image.

The areas of the masks or combinations of masks are selected by the user for normalisation are then measured in the same way. These values are then employed to calculate the ratios between the measured areas and the area used for normalisation. These operations are applied iteratively on all samples.

The output of the pixel-based analysis step consists of a CSV file with the normalised areas reported as percentages of the areas of the masks used for normalisation. If the analysis involves at least two samples divided into two categories, the user can choose to plot the comparisons of the normalised areas across the two categories as boxplots.

2.2 Patient and sample description

Two main groups of samples were analysed, samples used for the testing of SIMPLI during and after development, and samples analysed to study the composition of the TME in colorectal cancer patients treated with ICI. The first group includes normal colon mucosa and appendix samples imaged with IMC[71], a publicly available as well as one sample imaged with mIF from a cohort of ICI treated colorectal cancer patients.

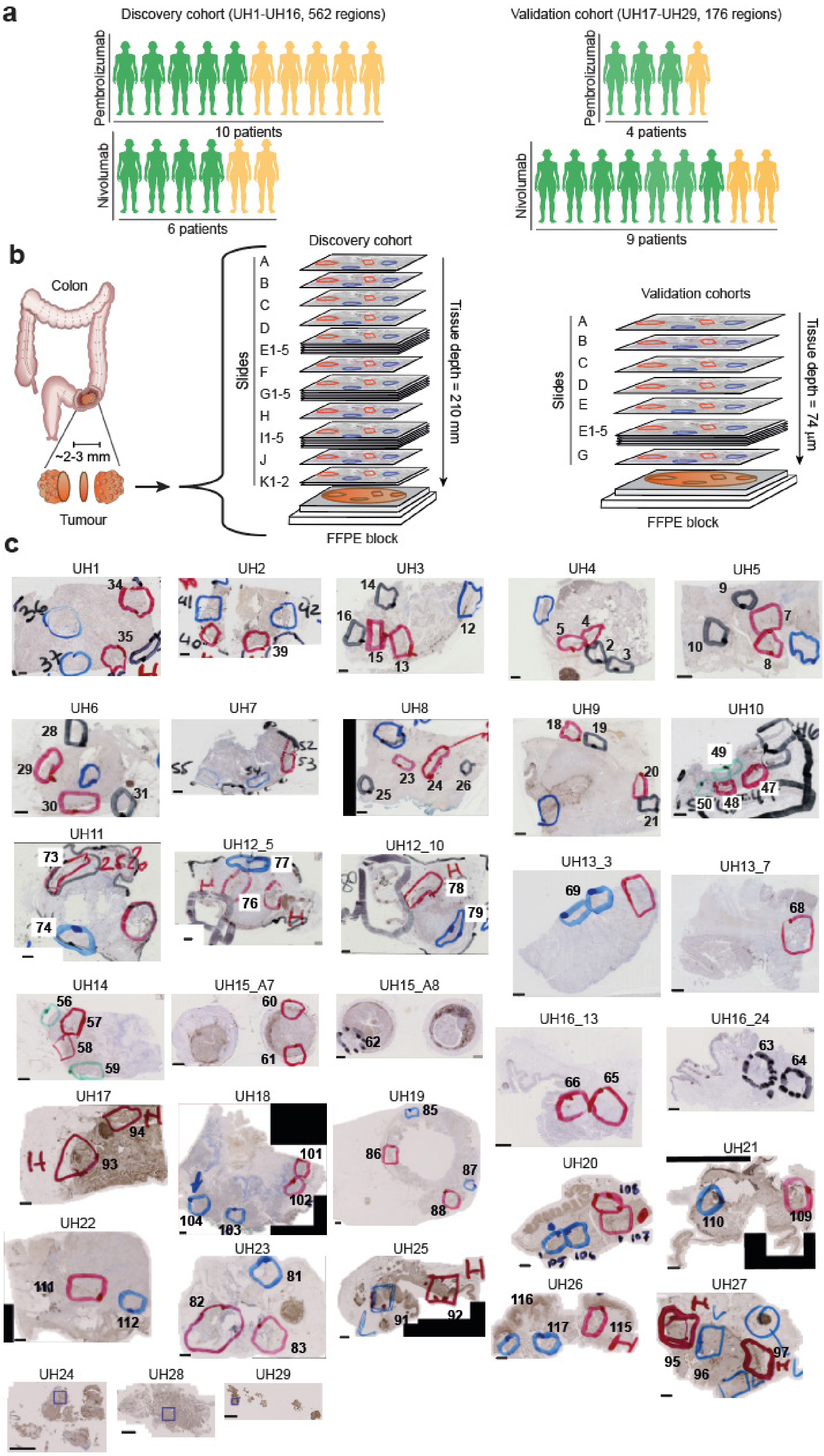
2.2.1 Samples used to test SIMPLI

Non-cancerous colon mucosa was extracted from six individuals during surgery for the recision of colorectal cancers. The extracted samples were then preserved as FFPE blocks (CLN1-CLN6; Supplementary table 1_S), then reviewed by an expert pathologist. All patients provided written informed consent in accordance with approved institutional guidelines (University College London Hospital, REC Reference: 20/YH/0088; Istituto Clinico Humanitas, REC Reference: ICH-25-09).

One FFPE block of normal appendix (APP1; Supplementary table 1_S) was obtained from a patient after surgery and reviewed by an expert pathologist. The study of human appendix was approved by Brighton West Research Ethics Committee, REC reference 10/H1111/014 Biology of T follicular helper cells.

To test SIMPLI's ability to process CODEX[68] derived data, a subset of images from a published dataset of colorectal CODEX[68] images was analysed (Supplementary table 1_S). The images were downloaded from The Cancer Imaging Archive (https://doi.org/10.7937/tcia.2020.fqn0-0326). The dataset consisted of CODEX[68] images from 35 colorectal cancer samples divided into two groups according to their histology: CLR (Crohn's-like reaction) and DII (diffuse inflammatory infiltration) according to the amount of peritumoural inflammation and tertiary lymphoid structures. Each sample has four images corresponding to four 0.6mm spots from two separate 70-core tissue microarrays, stained with 58 antibodies and two DNA markers, and imaged at a 377 nm/pixel resolution. For each of the 35 samples, one representative image was selected for further analysis, manually selecting the core containing both tumour and peritumoural immune infiltrates and having the best focus.

Figure 2.2 ICI response analysis patients and samples



The patient cohorts and design of the study and the regions selected for the analysis in each CRC sample.

a. CRC patient cohorts in the study. The clinical benefit each patient received from treatment was assessed with RECIST 1.1.

b. Schematic of the sequential slides derived from the FPPE blocks from validation and discovery cohorts samples. For samples from the discovery cohort slides A, B, F, H and J were used for CD3 immunohistochemistry, slide C for IMC[71], slide D for mIF, slides E1-5 for WES, slides G1-5 for RNA-seq, slides I1-5 for TCR-seq, and slides K1-2 for A-FRET. For samples UH17-UH23 and UH25-UH27 in the validation cohort, the following sequential slides were used: slides A, E and G for CD3 IHC, slide C for IMC[71] with panel II, slide D for mIF and slides F1-5 RNA-seq. For the three biopsies UH24, UH28, UH29 only IHC, IMC[71] and mIF were performed.

c. Region selection for DNA and RNA extraction ins slide A. Multiple regions with variable CD3 infiltration were identified in slide A by a certified pathologist. These were then projected to all other slides to enable multiregional analysis using a stereo microscope. The projected regions were manually dissected with a needle in slides used for DNA or RNA extractions.

This figure panels were reproduced from figure 1 and figure S1 of Bortolomeazzi et al. [107].

2.2.2 Colorectal cancer patient cohorts for ICI response analysis

Two cohorts of patients were selected for this study: a discovery cohort and a validation cohort (Supplementary table 1_I). For both cohorts, the formal Response Evaluation Criteria in Solid Tumours (RECIST) version 1.1[108] were employed to assess response to therapy. In particular, 12 months without disease progression from the beginning of immunotherapy were required for a patient to be considered as having achieved a durable benefit (DB). Patients showing disease progression before 12 months were instead considered as having had a non-durable benefit (nDB) from immunotherapy. All patients were consented at the UCL Cancer Institute Pathology Biobank - REC reference 15/YH/0311.

The discovery cohort consisted of 16 patients (Supplementary table 1_I): 10 (UH1-UH10) treated with 200mg of Pembrolizumab every three weeks, and six (UH11-UH16) treated with 240mg of Nivolumab every two weeks (Figure 2.2a). Patients UH1-UH10 were part of the KEYNOTE 177 clinical trial (ClinicalTrials.gov, NCT02563002)[52].

The validation cohort was composed of 13 patients (Supplementary table 1_I): three patients (UH17-UH19) enrolled in the KEYNOTE 177 trial (ClinicalTrials.gov, NCT02563002)[52] and treated with 200mg of Pembrolizumab every three weeks; one patient (UH26) treated with 2mg/kg of pembrolizumab every three weeks, seven patients (UH20-UH25 and UH29) treated with 240mg of Nivolumab every two weeks; one patient (UH27) treated with 1mg/kg of Ipilimumab in combination with 3mg/kg of Nivolumab every three weeks for four cycles, then by 240 mg of Nivolumab alone every two weeks; one patient (UH28) treated with 3mg/kg of Nivolumab for two cycles, then a combination of 1mg/kg of Ipilimumab with 3mg/kg of Nivolumab every three weeks for three cycles. All patients treated with Nivolumab (UH20-UH25 and UH27-UH29) belonged to the UK wide Bristol Myers Squibb Individual Patient Supply Request Programme as per Article 5/1 of Article Directive 2001/83/EC.

The tumour content in each FFPE block was assessed by a board-certified surgical pathologist. Each FFPE block was then cut with a microtome into serial sections for CD3 and HE staining, IMC[71], mIF, A-FRET (4mm thick), Whole exome sequencing (WES) and RNA-seq (10mm thick).

The FFPE blocks of samples UH1-UH16 underwent sectioning with a microtome to obtain 24 sequential sections (Figure 2.2b). These then were employed for CD3 staining (slides A, B, F, H and J) (Figure 2.2c), IMC (slide C), mIF Immunofluorescence (mIF, slide D), WES (slides E1-5), RNA-seq (slides G1-5), TCR-seq (Slides I1-5) and A-FRET (slides K1-2) (Supplementary table 2_I).

A total of 11 sequential sections (Figure 2.2b) were derived from samples UH17-UH23 and UH25-UH27 and used for CD3 staining (slides A, E and G) (Figure 2.2c), Hematoxylin and eosin stain (HE) staining (slide B), IMC[71] (slide C), mIF (slide D) and RNA-seq (slides F1-5). RNA-seq could not be performed on the three biopsies UH24, UH28 and UH29, thus only CD3 and HE staining (slides A and B); IMC[71] (slide C) and mIF (Slide D) were performed on these samples (Supplementary table 2_I).

2.3 Imaging mass cytometry staining and ablation

First, the dilution of each antibody in the panel was optimized by staining and ablating FFPE appendix sections. The resulting images were reviewed by a mucosal immunologist to select for each antibody the dilution producing the highest signal to noise ratio.

Then the slides were incubated at 60°C for one hour, dewaxed, and rehydrated before heat-induced antigen retrieval (HIER) with Antigen Retrieval

Reagent-Basic (R&D Systems) in a pressure cooker. Slides were blocked by incubating them in a blocking solution of 10% BSA (Sigma), 0.1% Tween (Sigma), and 2% Kiovig (Shire Pharmaceuticals) Superblock Blocking Buffer (Thermo Fisher) for two hours at room temperature. A primary antibody mix containing each antibody at the selected concentration in blocking solution was added to the slides, which were then incubated overnight at 4°C. Then the slides were washed twice in PBS and PBS-0.1% Tween and treated with the DNA intercalator Cell-ID™ Intercalator-Ir (Fluidigm) (containing the two iridium isotopes 191Ir and 193Ir) 1.25 mM in a PBS solution for 30-minutes. Finally, the slides were washed first in PBS and then in MilliQ water before air-drying.

ROI selection was performed by loading the stained slides in the Hyperion Imaging System (Fluidigm) imaging module to obtain light-contrast high-resolution images of approximately 4mm^2. These images were then examined to select 1mm^2 ROIs in each slide. These were then ablated at a 1 µm/pixel resolution and 200 Hz frequency to obtain the multiplexed images in MCD and TXT format.

2.3.1 Staining and ablation of the samples employed to test SIMPLI

A 26 antibodies-panel covering the main immune populations of the gastrointestinal tract, as well as other stromal and epithelial tissue components, was assembled from 17 already metal-tagged antibodies (Fluidigm) and nine non-pre-tagged ones, which were tagged using the Maxpar X8 metal conjugation kit after testing by immunohistochemistry (Supplementary table 2_S). Samples CLN1-CLN6 and APP1 (Supplementary table 1_S) were cut with a microtome to produce four µm-thick slides, which were stained as described above. Colon mucosa ROIs in CLN1-CLN6 were selected to include the entire thickness of the colon mucosa, with epithelial crypts in longitudinal orientation. For the APP1 ROI, an area containing a lymphoid follicle in its whole depth alongside a portion of lamina propria and epithelium was selected.

2.3.2 Staining and ablation of samples used for the ICI response analysis

A total of 42 antibodies were assembled in three panels targeting the main immune and stromal cell populations of the colonic and rectal TME (IMC panels I, II and III; Supplementary table 5_I). Of these antibodies, 25 were already metal-tagged (Fluidigm), and the remaining 17 were instead tested by immunohistochemistry and tagged using the Maxpar X8 metal conjugation kit (Fluidigm).

In each sample except for UH18, UH24, UH28 and UH29 (Supplementary table 2_I), regions with high and low CD3$^+$ content were selected from CD3$^+$ quantification performed on an adjacent sequential slide. Inside these regions, 1mm^2 ROIs were selected to include areas with the highest tumour content and representative of the median CD3$^+$ content of the region.

In the discovery cohort, a total of 38 regions were stained and ablated using IMC panel I, 22 regions of the validation cohort using IMC panel II, and an additional 17 regions from DB patients belonging to both cohorts were stained and ablated with the third panel IMC panel III (Supplementary tables 2_I, 5_I).

For samples in the discovery cohort, ROIs were selected in the regions with the highest tumour content once for high CD3$^+$ infiltration regions and once for low CD3$^+$ infiltration regions. In samples UH6, UH9 and UH12, all four regions were analysed, while for UH4 and ROI, spanning two high CD3$^+$ regions and one spanning the two low CD3$^+$ regions were analysed together to be consistent with WES and RNA-seq analyses (Supplementary table 2_I).

The two regions with the highest difference in CD3 infiltration were selected in the validation cohort. In UH18, only one ROI was ablated because of the high levels of necrosis in the region, while only one region and thus one ROI was selected from the three biopsies UH24, UH28, UH29 because of the small size of the FFPE blocks (Supplementary table 2_I).

2.4 MIF of samples for testing SIMPLI and ICI response analysis

mIF was performed on slide D from 24 FFPE blocks: 13 samples from the discovery cohort and 11 from the validation cohort. The slide was stained with a panel of eight antibodies: CD74, Transcription Factor 7 (TCF7), PDL1, Ki67, PD1, GzB, CD68, CD8 and 4’,6-diamidino-2-phenylindole (DAPI) (Supplementary Table 5_I). A Ventana Discovery Ultra automated staining platform (Roche) was employed to run an automated Opal-based mIF staining protocol after optimisation of Opal-antibody pairing and dilution, incubation and denaturation times according to manufacturer’s instructions. Additionally, the expected tissue and cellular localisation of each marker and fluorophore brightness were assessed to minimise the effects of fluorescence spillage when optimising antibody-Opal pairing. Spectral libraries containing each Opal were scanned to allow autofluorescence isolation and spectral unmixing. The automating staining protocol was applied with an autostainer to each slide after a one-hour incubation at 60°C. The staining protocol consisted of deparaffinisation (EZ-Prep solution, Roche), heat induced epitope retrieval (DISC. CC1 solution, Roche) and seven sequential rounds of staining and denaturation. Each staining round required a one-hour incubation of the slide with the primary antibody, 12-minute incubation with a secondary antibody conjugated with horseradish peroxidase (HRP) (DISC. Omnimap anti-Ms HRP RUO or DISC. Omnimap anti-Rb HRP RUO, Roche) and a 16-minute incubation with the Opal reactive fluorophore (Akoya Biosciences). After each staining round, the slide underwent denaturation at °C for 8 minutes to remove the primary and secondary antibodies and the HRP from the previous cycle without affecting the fluorescent signal.

At the last staining round, the slide was incubated with Opal TSA-DIG reagent (Akoya Biosciences) for 12 minutes, followed by a one-hour incubation with Opal 780 reactive fluorophore (Akoya Biosciences). Finally, after counterstaining with DAPI (Akoya Biosciences), the slide was coverslipped using ProLong Gold antifade mounting media (Thermo Fisher Scientific).

All slides were scanned with the Vectra Polaris automated quantitative pathology imaging system (Akoya Biosciences). Six fields of view within the regions selected by the pathologist were scanned at 20x and 40x magnification using appropriate exposure times.

2.5 Image analysis to test SIMPLI

SIMPLI was tested on three types of high-multiplexed imaging data, described in the following paragraphs:

2.5.1 IMC analysis of normal colon mucosa to test SIMPLI.

2.5.2 Analysis of a normal appendix image to test SIMPLI.

2.5.3 Analysis of a multiplexed immunofluorescence image.

2.5.4 Analysis of CODEX images to test SIMPLI.

2.5.1 IMC analysis of normal colon mucosa to test SIMPLI

The normal colon mucosa images were analysed with two separate SIMPLI workflows: the first included all processes in the raw data processing step, the pixel-based analysis step, and the cell-based analysis step up to the cell masking process; the second workflow consisted of the raw data processing step without the data normalisation process and the pixel-based analysis step.

2.5.1.1 Raw image processing

The data extraction process was applied to the raw TXT files of the CLN1-CLN6 ROIs to extract 28 images per sample: 26 channels from the antibody panel and two from the DNA intercalators (Supplementary Table 2_S). Then, pixel intensities were rescaled from 0 to 1 using the 99th intensity percentile within each sample and channel as the maximum. After data normalization, the images underwent the thresholding and masking process, which applied Otsu thresholding[109] with a custom CellProfiler4 pipeline to generate background-free masks for each marker. The Immunoglobulin A (IgA) masks used for measuring IgA$^+$ areas in the pixel-based analysis step were generated at this stage using a three-class global Otsu thresholding[109] with two background classes after applying a Gaussian filter with a three-pixel diameter.

This process was also employed to generate the masks for the lamina propria and the epithelium. The first was produced from the Vimentin mask by filling holes <75-pixel in diameter. The second was produced from the sum of the Pan-keratin and E-cadherin masks by dilating the images with a three-pixel disk and filling holes <75-pixel in diameter. The lamina propria and epithelium masks were then added into a sum image, which was dilated with a three-pixel disk and underwent the filling of all holes <25-pixel in diameter. All positive features outside the lamina and epithelium compartments were removed with an opening operation with a 150-pixel radius disk. Then, the lamina propria mask was subtracted from the resulting image to generate the final mask for the epithelium.

The raw image processing step was performed again, skipping the data normalization process and performing the thresholding and masking process directly on the raw images. This process was performed with a custom CellProfiler4[90] pipeline applying manually selected sample-specific thresholds for the generation of the IgA, E-Cadherin, Pan-Keratin and Vimentin masks from the raw images. The resulting masks images were then used to generate lamina propria and epithelial masks for each sample as previously described.

2.5.1.2 Cell- and pixel-based analysis

The cell-based analysis step started with the deterministic segmentation with a custom CellProfiler4[90] pipeline. This pipeline consisted in the global Otsu thresholding[109] of the DNA1 image to identify the cell nuclei, and then the radial expansion of each nucleus for up to 10 pixels over a membrane mask derived from the sum IgA, CD3, CD68, CD11c and E-cadherin masks. After segmentation, only cells overlapping with the lamina propria mask by at least 30% were retained. This 30% overlap threshold was selected after inspection by an expert histologist. Then cells were assigned to one of four main immune cell populations with the cell masking process. Cells overlapping marker-specific masks by more than a threshold defined by an expert histologist were assigned to the corresponding population in this order: ≥15% of the IgA mask for IgA cells; ≥15% of the CD3 mask for T cells; ≥25% of the CD68 mask for macrophages; ≥15% of CD11c mask for dendritic cells.

The pixel-based analysis step was applied to the IgA masks derived from either the normalised or the raw images. The IgA$^+$ areas in the lamina propria, epithelium and both compartments combined were measured and normalised over the areas of the three compartments.

2.5.2 Analysis of a normal appendix image to test SIMPLI

The analysis of a normal appendix image was performed with SIMPLI using a workflow, including the raw image processing step and the cell-based analysis step. For this step, single-cell data extraction through deterministic segmentation and cell masking was followed by cell phenotyping with both unsupervised clustering and expression thresholding. Then a homotypic spatial analysis was performed on the phenotyped cells.

2.5.2.1 Raw image processing and single-cell data extraction

Images from the 26 channels from the panel antibodies and two DNA intercalators were extracted from the raw TXT file of the APP1 ROI with SIMPLI's data extraction process. The resulting images were then normalised to the 99th percentile through the data normalisation process and thresholded with CellProfiler4[90] in the thresholding and masking process. A background free mask was generated for each marker and employed as input for the cell-based analysis step.

For single-cell data extraction, both deep learning and deterministic segmentation were applied, and the single-cell data produced with the latter approach was employed as input for the downstream processes. The deep learning segmentation process was performed with the 2D_versatile_fluo model[73] applied to the DNA1 channel with a probability threshold of 0.0015 and an NMS threshold of 0.01. Deterministic segmentation was performed with a custom CellProfiler4[90] pipeline. First nuclei were segmented from the DNA1 channel, and then cells were isolated through watershed segmentation with the nuclei as seeds on a membrane mask consisting of the sum of the CD45, Pan-keratin and E-cadherin masks.

After segmentation, cell masking was performed to assign cells to the epithelium or immune cell populations. Cells were assigned to a given population if they overlapped the mask of the corresponding marker by at least 10%. This threshold was identified by a mucosal immunologist and applied in this order: CD3 mask for T cells; CD20 and CD27 masks for B cells; CD68 mask for macrophages; CD11c mask for dendritic cells; E-cadherin and Pan-keratin masks for epithelial cells.

2.5.2.2 Cell phenotyping and homotypic spatial analysis

The two cell phenotyping approaches provided by SIMPLI were then applied to T cells. The unsupervised clustering process was applied at resolutions between 0.1 and 1.0, with 0.05 intervals using the mean intensities of the following markers in each T cell: CD3, CD45RA, CD45RO, CD4, CD8, Ki67, and PD1. A manual inspection of the resulting clusters allowed the identification of the run, which produced the highest number of biologically meaningful clusters (resolution = 0.25), which was selected for downstream analyses. These clusters were then re-identified with the expression thresholding process using mean intensity thresholds defined by an expert histologist for the following markers: CD3 >0.06 for cluster 1; CD8a >0.125 for cluster 2; CD45RA >0.125 for cluster 3; CD4 >0.125 and CD45RO >0.15 for cluster 4; and CD4 > 0.1 and PD1 >0.15 for cluster 5.

The tissue's distribution of $CD4^+PD1^+$ T cells (cluster 5, resolution = 0.25) was analysed with SIMPLI's homotypic aggregations process. Aggregations were quantified with the following parameters: a minimum of five points per cluster and reachability equivalent to a density of at least 5 cells/mm^2.

2.5.3 Analysis of a multiplexed immunofluorescence image

After scanning, all fields of view were processed with inForm[110] to perform spectral unmixing and autofluorescence isolation using previously acquired spectral libraries. The resulting raw images from each field at 20X and 40X magnification were stitched to generate single raw images in TIFF format. These raw images were then used for visualisation or as input for image analysis with SIMPLI.

To test the performance of SIMPLI of mIF derived images, six 20x fields of view for a total of >5mm^2 patient UH1 were joined into a single ROI (CRC1, Supplementary Table 1_S). Single-channel images for the CD8, PD1, Ki67, PDL1, CD68, GzB and DAPI channels were extracted, and their intensities were independently rescaled from 0 to 1 with custom R scripts. The resulting single-channel TIFF images were then employed as input for a SIMPLI workflow, starting at the thresholding and masking process of the raw data processing. In this step, the background noise was removed from each image by applying Otsu thresholding[109] with a custom CellProfiler4[90] pipeline.

For the cell-based analysis, SIMPLI's deterministic cell segmentation process was used to identify cells by applying a global threshold to the thresholded DAPI image and selecting all objects with a diameter between four and 60 pixels. Then, the cell masking process was skipped, and all cells were employed as input for cell phenotyping with the expression thresholding process. In this process, PD1$^+$CD8$^+$ cells, CD68$^+$ cells and PDL1$^+$ cells identified using mean intensity thresholds identified by a mucosal immunologist: 0.01 for CD8, 0.005 for PD1, 0.01 for CD68 and 0.01 for PDL1.

SIMPLI's heterotypic spatial analysis process was applied to calculate the distribution of minimum distances between PDL1$^+$ cells and PD1$^+$CD8$^+$ cells and between CD68$^+$ cells and PDL1$^+$ and PDL1$^-$ cells. All PDL1$^+$ cells and PD1$^+$CD8$^+$ cells at a distance from each other lower than double the maximum cell radius (24 pixels = 12μm) were considered proximal, while all other cells were classified as distal. The expression levels of GzB and Ki67 in CD8$^+$PD1$^+$ cells CD68 and Ki67 in PDL1$^+$ cells were compared across proximal and distal cells with a two-sided Wilcoxon rank-sum test.

2.5.4 Analysis of CODEX images to test SIMPLI

The 35 selected images were then converted to single-channel TIFF files, and their pixel intensities were normalised by rescaling from 0 to 1 within each channel and sample with custom R scripts. The normalised images were used as input for cell-based analysis with SIMPLI. The deterministic cell segmentation process was applied to each of the 35 images using a custom CellProfiler4[90] pipeline. First, a global threshold was applied to the HOECHST channel to identify the nuclei as all positive objects with a diameter between 5 and 40 pixels. Then, each nucleus was then expanded by 5 pixels in all directions to define the cell area. The resulting cells were assigned to ten phenotypes with SIMPLI's expression thresholding process according to the following thresholds of mean marker expression within each cell: Caudal Type Homeobox 2 (CDX2) >0.15 or Mucin 1 (MUC1) >0.15 or Cytokeratin >0.15 for tumour cells; CD34 >0.15 or CD31 >0.15 for endothelial cells; Vimentin >0.1 for other stromal cells; CD11c >0.3 for DCs; CD38 >0.26 for B cells; CD4 >0.13 and CD3 >0.1 for CD4$^+$ T cells; CD4 >0.12 and Forkhead Box P3 (FOXP3) >0.5 and CD3 >0.1 for Tregs; CD8 >0.16 and CD3 >0.1 for CD8$^+$ T cells, and CD68 >0.11 for macrophages.

After cell phenotyping, the minimum distances between macrophages, CD8$^+$ T cells, CD4$^+$ T cells, Tregs, and B cells to tumour cells and endothelial cells were calculated using the heterotypic aggregations process. The resulting distance distributions were then compared between CLR and DII colorectal cancer subtypes with two-sided Wilcoxon rank-sum tests, and the False Discovery Rate (FDR) was calculated with the Benjamini-Hochberg correction. Differences between distance distributions with FDR < 0.1 were considered biologically relevant only if the difference between the median distances of the two subtypes was greater than 8μm, corresponding to the diameter of B and T lymphocytes[111]. Additionally, a permutation test was performed by randomly shuffling the identities of all cells in each sample 10000 times to derive an expected distribution of differences in distances between CLR and DII cells. This expected distribution was compared to the observed values to estimate if the observed difference was compatible with a random distribution.

2.6 IMC analysis of ICI-treated colorectal cancers

The colorectal cancer image analysis was performed with SIMPLI, using Ilastick[91] v. 1.3.0 for the filtering of background signals for specific markers. The cell-based analysis was performed with the same workflow for both the discovery and validation cohort for the single-cell data extraction processes.

2.6.1 Raw image processing and pixel-based analysis

The raw image processing step was performed on all the 77 ROIs from both cohorts together (Supplementary Table 2_I), with two consecutive SIMPLI workflows. In the first workflow, single-channel TIFF images for each antibody and the two DNA intercalators were produced from the raw IMC[71] MCD and TXT files with the raw data extraction process. Then these images were normalised by sample and by channel with the data normalisation process.

After visual inspection by a mucosal immunologist, the normalised images for PD1, PDL1, GzB, CD45RA, TIM3, VISTA, TCF7, CD134, CD206, and FOLR2 underwent processing with Ilastick[91] v. 1.3.0 to generate probability masks for each pixel to belong to the signal or background noise. For this purpose, a random forest classifier was trained for each marker using a closely related marker as a reference (CD3 for PD1; Vimentin for PDL1; CD8 and CD15 for GZB; CD45 and CD45RO for CD45RA; CD68 for Folate Receptor Beta (FOLR2) and CD206). These masks were then used as an additional input to the second SIMPLI workflow employed for image processing. For B2M, an additional mask to be used only for the pixel-based analysis was generated from the normalised images with an *ad hoc* threshold of 0.5 pixel intensity. For regions UH19_87, UH27_96, and UH27_97, the CD3 masking threshold was adjusted to 0.175, 0.15, and 0.15 after manual inspection.

The second SIMPLI workflow consisted of the thresholding and masking process. In this process, the normalised images generated in the first workflow were and the probability masks produced with Ilastick[91]. These images were thresholded to produce the masks for each marker. This process was also used to produce the tumour and stroma masks. The tumour masks were generated as the sum of the Pan-keratin and E-cadherin masks for all regions except UH18_103 and UH22_112, where only E-cadherin was used. The resulting images were smoothened with a Gaussian filter and filled all holes <30 pixels in diameter. The stroma masks were generated from the sum of the Vimentin, SMA and DNA masks in the discovery cohort, and the Vimentin, CD68, CD11c, CD3, CD27 and CD45 masks in the validation cohort, the resulting images had all holes <20 pixels in diameter filled up. The stroma and tumour masks were then summed to produce a tissue mask for each region.

The SIMPLI pixel-based analysis step was performed in two distinct workflows, one for the discovery cohort and one for the validation cohort. The measured areas were normalised over the total tissue area or the area of the five main immune populations (T cells, B cells, macrophages, dendritic cells and neutrophils) for the discovery cohort and T cell and macrophages only for the validation cohort.

2.6.2 Single-cell data extraction

Single-cell data extraction was performed as a single SIMPLI workflow for the discovery and the validation cohorts. Deterministic cell segmentation was performed with a custom CellProfiler4 pipeline. First, segmentation using local Otsu thresholding[109] on the product of the two DNA images were used to identify the nuclei. Second, a membrane mask was generated from the sum of all membrane markers for the membrane markers: CD3, CD20, CD27, CD16, CD11c, CD15, SMA, CD34, Vimentin and Pan-keratin for the discovery cohort and CD45, Pan-keratin and E-cadherin for the validation cohort. Then the membrane mask was used as a base for the radial expansion of the nuclei by up to 10 pixels to generate the cell masks. Only cells overlapping with the tissue mask by at least 10% of their area were retained. Finally, the mean intensity of all markers was measured in each of the retained cells.

After cell segmentation, cell identities were identified as belonging to one of five main populations or as tumour cells with the cell masking process. This assignment was performed with the following overlap thresholds applied in order: ≥25% of the CD3$^+$ mask for T cells; ≥10% of CD11c$^+$ CD68- mask for dendritic cells; >10% of the sum of CD68$^+$ CD11c$^+$ and CD68$^+$ CD11c$^-$ masks for macrophages; ≥5% of IgA$^+$, IgM$^+$, CD20$^+$, and CD27$^+$ mask for B cells; and ≥25% of CD15$^+$ mask for neutrophils. Cells not overlapping with any of these markers were defined as tumour cells if they overlapped ≥80% with the tumour mask or left unassigned otherwise. Subsequently, PD1$^+$ cells were identified as a subset of T cells overlapping the PD1 by at least ≥1% of their area, while PDL1$^+$ cells were identified as cells whose area overlapped the PDL1 mask by at least 10%. All these overlap thresholds were identified by the manual inspection of a mucosal immunologist.

2.6.3 Cell phenotyping in the discovery and validation cohorts

Single-cell phenotype clustering was performed for the discovery cohort for T cells, B cells, macrophages, dendritic cells, neutrophils, PD1$^+$ and PDL1$^+$ cells separately with SIMPLI's unsupervised clustering process. Independent clustering was used to compare the relative abundance of cell subpopulations between hypermutated and non-hypermutated CRCs or DB- and nDB-CRCs using samples from Pembrolizumab and Nivolumab treated patients alone or combined. The clustering was based on the mean expression of a set of markers typical of that population for each main population (Supplementary table 5_I).

Unsupervised clustering was performed for each population at ten different resolution values from 0.1 to 1.0 with 0.1 increments. After manual inspection of the clustering output at different resolutions, the one with the highest number of biologically meaningful clusters was chosen for each cell population.

CD74$^+$ macrophages and CD8$^+$GzB$^+$ CD8$^+$Ki67$^+$ T cells were identified with SIMPLI's expression thresholding process. Specific thresholds selected by a mucosal immunologist were applied to the mean marker intensities in each cell: 0.1 for CD74; 0.1 for CD8; 0.05 for GzB; 0.15 for Ki67. CD8$^+$ T cells positive for both the Ki67 and the GzB threshold were classified as either CD8$^+$GzB$^+$ or CD8$^+$Ki67$^+$ T cells according to which of the two markers had the highest mean intensity value.

Single-cell clustering was performed on all T cells in the validation cohort using all 17 T cell markers included in IMC panel II (Supplementary table 5_I). The distribution of cells within each cluster over the total cells in the phenotyped population was compared between DB- and nDB-CRCs or hypermutated and non-hypermutated CRCs using a two-sided Wilcoxon rank-sum test, applying the Benjamini-Hochberg correction to obtain the FDR.

In the ROIs from regions stained with IMC panel III, CD74$^+$ macrophages were identified using a threshold of 0.35 on the mean intensity of CD74 with SIMPLI's expression thresholding process. The unsupervised clustering of the identified CD74$^+$ macrophages was performed using 16 macrophage markers (Supplementary table 5_I).

2.6.4 Homotypic and heterotypic spatial analyses

The homotypic spatial analysis process of SIMPLI was applied to $CD68^+CD74^+$ macrophages identified by unsupervised clustering in the discovery cohort and by expression thresholding in the validation cohort. For this analysis, high-density clusters were identified as clusters with a minimum of five points and a reachability equivalent to a density of at least 5 cells/10,000μm^2.

SIMPLI's heterotypic spatial analysis process was applied to measure all the minimum distances between the $CD8^+GzB^+$ and $CD8^+Ki67^+$ T cell phenotypes and the $CD68^+CD74^+$ macrophage phenotype. This process was then repeated for $CD8^+GzB^+PD1^+$ or $CD8^+Ki67^+PD1^+$ T cells, $CD68^+CD74^+PDL1^+$ cells and the distance distributions were compared using a two-sided Wilcoxon rank-sum test.

.

Chapter 4. Analysis of response to immune checkpoint inhibitors in colorectal cancer

Chapter 3. SIMPLI: Single-cell Identification from MultiPLexed Images

3.1 Contributions

In this study[99], I developed the SIMPLI software with support from Mohamed Reda Keddar and Damjan Temelkovski. I also analysed the data together with Lucia Montorsi, Amelia Acha-Sagredo, Michael J. Pitcher, Jo Spencer., and Francesca D. Ciccarelli. Finally, I wrote the manuscript with Francesca D. Ciccarelli, and all authors reviewed and approved its final version.

Francesca D. Ciccarelli acquired the funding conceived and supervised the study with support from Jo Spencer. Manuel Rodriguez-Justo, Gianluca Basso, and Luigi Laghi selected and clinically assessed the samples used in the study, and Manuel Rodriguez-Justo performed their pathological assessment. Lucia Montorsi stained the sample slides for IMC[71], while Amelia Acha-Sagredo performed the mIF experiments.

3.2 A SIMPLI (Single-cell Identification from MultiPLexed Images) approach for spatially resolved tissue phenotyping at single-cell resolution

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-28470-x OPEN

A SIMPLI (Single-cell Identification from MultiPLexed Images) approach for spatially-resolved tissue phenotyping at single-cell resolution

Michele Bortolomeazzi[1,2], Lucia Montorsi[1,2], Damjan Temelkovski[1,2], Mohamed Reda Keddar[1,2], Amelia Acha-Sagredo[1,2], Michael J. Pitcher[3], Gianluca Basso[4,5], Luigi Laghi[4,6], Manuel Rodriguez-Justo[7], Jo Spencer[3] & Francesca D. Ciccarelli[1,2]✉

Multiplexed imaging technologies enable the study of biological tissues at single-cell resolution while preserving spatial information. Currently, high-dimension imaging data analysis is technology-specific and requires multiple tools, restricting analytical scalability and result reproducibility. Here we present SIMPLI (Single-cell Identification from MultiPLexed Images), a flexible and technology-agnostic software that unifies all steps of multiplexed imaging data analysis. After raw image processing, SIMPLI performs a spatially resolved, single-cell analysis of the tissue slide as well as cell-independent quantifications of marker expression to investigate features undetectable at the cell level. SIMPLI is highly customisable and can run on desktop computers as well as high-performance computing environments, enabling workflow parallelisation for large datasets. SIMPLI produces multiple tabular and graphical outputs at each step of the analysis. Its containerised implementation and minimum configuration requirements make SIMPLI a portable and reproducible solution for multiplexed imaging data analysis. Software is available at "SIMPLI [https://github.com/ciccalab/SIMPLI]".

[1] Cancer Systems Biology Laboratory, The Francis Crick Institute, London NW1 1AT, UK. [2] School of Cancer and Pharmaceutical Sciences, King's College London, London SE11UL, UK. [3] School of Immunology and Microbial Sciences, King's College London, London SE19RT, UK. [4] Laboratory of Molecular Gastroenterology, IRCCS Humanitas Research Hospital, Rozzano 20089 MI, Italy. [5] Genomic Unit, IRCCS Humanitas Research Hospital, Rozzano 20089 MI, Italy. [6] Department of Medicine and Surgery, University of Parma, Parma 43121 PR, Italy. [7] Department of Histopathology, University College London Cancer Institute, London WC1E 6JJ, UK. ✉email: francesca.ciccarelli@crick.ac.uk

A detailed investigation of tissue composition and function in health and disease requires spatially resolved, single-cell approaches that precisely quantify cell types and states as well as their interactions in situ. Recent technological advances have enabled to stain histological sections with multiple tagged antibodies that are subsequently detected using fluorescence microscopy or mass spectrometry[1]. High-dimensional imaging approaches such as imaging mass cytometry (IMC)[2], multiplexed ion beam imaging (MIBI)[3], co-detection by indexing (CODEX)[4], multiplexed immunofluorescence (mIF, including cycIF)[5] and multiplexed immunohistochemistry (mIHC)[6,7] enable quantification and localisation of cells in sections from formalin-fixed paraffin-embedded (FFPE) tissues, including clinical diagnostic samples. This is of particular value for mapping the tissue-level characteristics of disease conditions and predicting the outcome of therapies that depend on the tissue environment, such as cancer immunotherapy. For example, a recent IMC phenotypic screen of breast cancer subtypes revealed the association between the heterogeneity of somatic mutations and that of the tumour microenvironment[8]. Similarly, a CODEX-based profile of FFPE tissue microarrays from high-risk colorectal cancer patients correlated PD1$^+$CD4$^+$ T cells with patient survival[9].

The analysis of multiplexed images requires the conversion of pixel intensity data into single-cell data, which can then be characterised phenotypically, quantified comparatively and localised spatially in the tissue. Currently available tools are technology specific and cover only some steps of the whole analytical workflow (Table 1). For example, several computational approaches have been developed to process raw images and extract single-cell data either interactively (Ilastik[10], CellProfiler4[11], CODEX Toolkit[4]) or via command line (imcyto[12], ImcSegmentationPipeline[13]). Distinct sets of tools can then perform cell phenotyping (CellProfiler Analyst[14], Cytomapper[15], Immunocluster[16]) or analyse cell–cell spatial interactions (CytoMap[17], ImaCytE[18], SPIAT[19], neighbouRhood[20]). Similarly, a few tools enable direct pixel-based analysis through pixel classification[10] or quantification of pixel positive areas[11]. Despite such a variety of tools, none of them can perform all of the required analytical steps in a common pipeline. Two exceptions are histoCAT ++[21] and QuPath[22], which however have been developed specifically for interactive use and are not well suited for the analysis of large datasets. Moreover, all of these tools rely on ad hoc configuration files and input formats, making the analysis challenging for users with limited computational skills and restricting the scalability, portability and reproducibility in different computing environments.

Here we introduce SIMPLI (Single-cell Identification from MultiPLexed Images), a tool that combines processing of raw images, extraction of single-cell data, and spatially resolved quantification of cell types or functional states into a single pipeline (Table 1). This is achieved through the integration of well-established tools and newly developed scripts into the same workflow, enabling ad hoc configurations of the analysis while ensuring interoperability between its different parts. SIMPLI can be run on desktop computers as well as on high-performance-computing environments, where it can be easily applied to large datasets due to automatic workflow parallelisation. To demonstrate the flexibility of SIMPLI to work with different technologies and experimental conditions, we analyse the phenotypes and spatial distribution of cells in different tissues (human colon, appendix, colorectal cancer) using multiplexed images obtained with distinct technologies (IMC, mIF, CODEX).

Results

Overview of the SIMPLI analytical workflow. SIMPLI performs the analysis of multiplexed imaging data in three steps (Methods, Fig. 1) integrating well-established and newly developed

Table 1 Features of representative tools for the analysis of multiplexed imaging data.

Computational tool	Image processing	Cell segmentation	Cell phenotyping preselected	Cell phenotyping unsupervised	Spatial analysis homotypic	Spatial analysis heterotypic	Pixel analysis	Parallelisation	Imaging technologies
SIMPLI	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1-6
CODEX Toolkit[4]	Yes	Yes	Yes	Yes	Yes	Yes	No	No	3
CellProfiler4[11]	Yes	Yes	Yes	No	No	Partial	Yes	Yes	1-6
HistoCAT++[21]	Yes	Yes	Yes	Yes	No	Yes	No	No	1,2,4,5
QuPath[22]	Yes	Yes	Yes	No	No	No	Yes	Yes	1-6
Cytomapper[15]	Partial	Yes	Yes	No	No	No	Yes	Yes	1-5
Ilastik[10]	No	Yes	Yes	No	Partial	No	Partial	Yes	1-6
ImcSegmentationPipeline[13]	Yes	Yes	No	No	No	Partial	No	Yes	1
Imcyto[12]	Yes	Yes	No	No	No	No	No	Yes	1
SPIAT[19]	No	No	Yes	No	Yes	Yes	No	No	2,5,6
Giotto[47]	No	No	No	Yes	Yes	Yes	No	No	1-6
ImaCytE[18]	No	No	No	Yes	No	Yes	No	No	1
CellProfiler Analyst[14]	No	No	Yes	No	No	No	No	No	1-6
Immunocluster[16]	No	No	No	Yes	No	No	No	No	1
NeighbouRhood[20]	No	No	No	No	No	Yes	No	No	1-5
CytoMAP[17]	No	No	No	No	No	Yes	No	No	2

For each tool, reported are the steps of the analytical workflow that it can perform, whether it can be parallelised and the multiplexed imaging platform it can be applied to (1: IMC; 2: mIF; 3: CODEX; 4: MIBI; 5: mIHC; 6: spatial transcriptomic visualisation). A method was considered compatible with a given imaging technology if this was reported in the original publication or other studies.

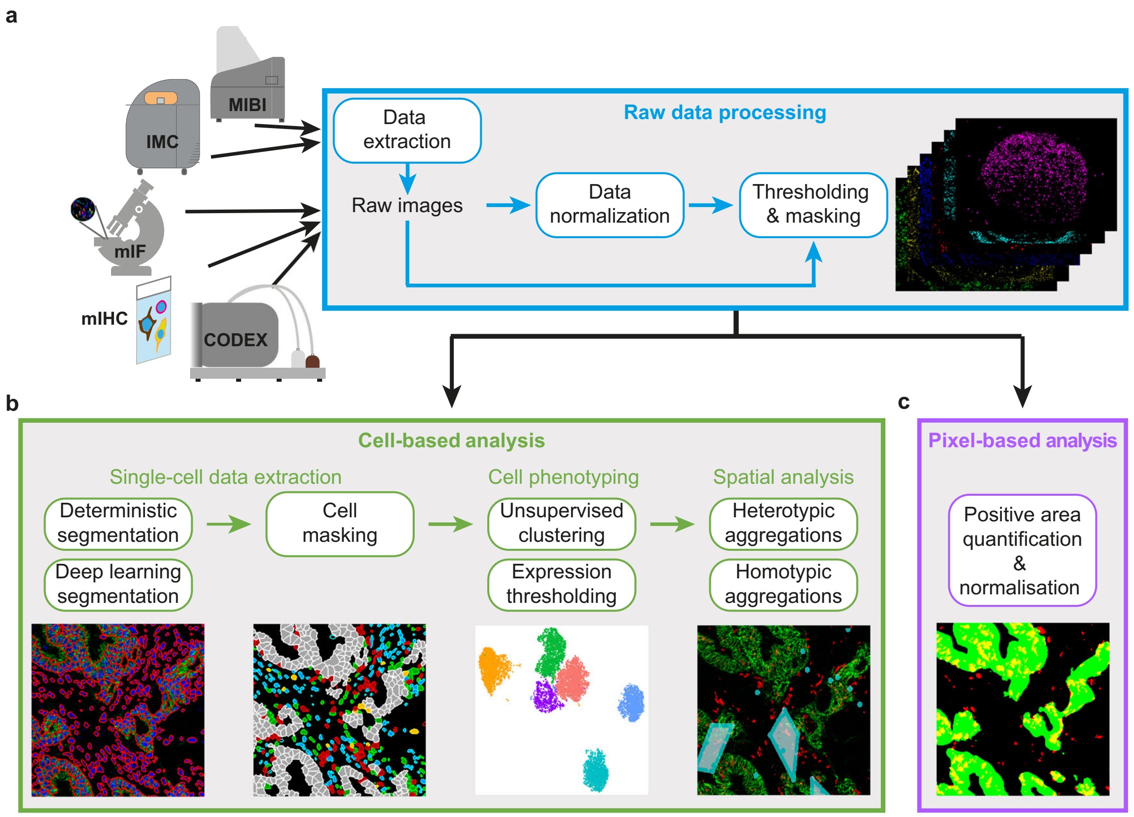


Fig. 1 Schematics of the SIMPLI workflow. a Raw images are extracted from IMC or MIBI data or directly imported from other imaging technologies. After their optional normalisation, these images are thresholded to remove the background noise and produce tissue compartments or marker masks. The resulting images can be analysed using a cell-based or a pixel-based approach. **b** In the cell-based analysis, single cells are segmented with deterministic or deep learning models and phenotyped using unsupervised or supervised approaches. The distribution of cells in the tissue can then be investigated through a spatial analysis of homotypic or heterotypic aggregations. **c** In the pixel-based approach, areas positive for a user-defined combination of markers are measured and normalised over the area of the whole image or of the masks defining compartments or areas positive for certain markers.

standalone processes (Supplementary Fig. 1). Each process can be run independently or even skipped with the possibility of using alternative input data at each point of the workflow.

The first step of SIMPLI consists of processing raw data from single or multi-channel images or text files from a variety of high-dimensional imaging technologies (Fig. 1a and Supplementary Fig. 1a). After data extraction, pixel values for each marker can be optionally normalised by rescaling them in each sample. This allows the user to apply the same thresholds for background noise reduction across samples. Alternatively, the normalisation step can be skipped and sample-specific thresholds can be applied directly to individual, non-normalised images to minimise the effect of non-uniform staining. This is recommended for example if markers have low signal-to-noise ratios and the resulting thresholds may be too restrictive or if platform-specific normalisation is required. In the last step of data processing, masks for specific tissue compartments or markers are derived using a fully customisable pipeline based on CellProfiler4[11], where the user can apply filters, thresholds and morphological operations to each image. The resulting processed images can then be analysed at the cell (Fig. 1b) and pixel (Fig. 1c) levels.

The cell-based analysis aims to investigate the qualitative and quantitative cell composition of the tissue and is composed of (1) single-cell data extraction, (2) cell phenotyping and (3) spatial analysis of cell–cell distances (Fig. 1b).

To extract cell data, SIMPLI implements single-cell segmentation using either a deterministic[11] or a deep learning[23] approach (Supplementary Fig. 1b). The former enables deterministic filtering based on cells size and shape, as well as marker intensities. The latter applies pre-trained models (either provided by SIMPLI or supplied by the user) to identify cells with high accuracy. After cell segmentation, SIMPLI produces the masks of the individual cells and calculates the expression values for each marker in each cell. Cells belonging to tissue compartments or positive for certain markers can then be identified based on their overlap with the previously derived tissue or marker masks.

To define the cell phenotypes, SIMPLI uses two alternative approaches (Supplementary Fig. 1b). The first applies unsupervised clustering to all cells or preselected subsets of cells (for example those mapping to specific tissue compartments or positive for certain markers) using marker expression levels. This leads to the unbiased classification of cells into clusters with similar expression profiles indicating similar phenotypes. The second approach identifies cells with designated phenotypes by applying combinations of user-defined thresholds to the expression values of the markers of interest. These thresholds can be identified through an expert-guided examination of the original images or using the visualisation plots produced by SIMPLI. The two approaches can be used independently or as cross-validation of the cell phenotypes.

To identify cell aggregations within the same (homotypic) or across different (heterotypic) cell types, SIMPLI implements a spatial analysis of the distance between cells within the imaged tissue (Supplementary Fig. 1b). In the case of homotypic aggregations, SIMPLI identifies groups of cells of the same type within a user-defined distance and visually localises them as clusters in the tissue image. In the case of heterotypic aggregations, SIMPLI computes the distance distribution between distinct cell types and compares them across cell types and experimental conditions. Observed distance distributions can also be compared to expected distributions obtained by randomly reshuffling the cell identities in each sample.

The pixel-based approach implemented in SIMPLI enables quantification of areas positive for a specific marker or combination of markers, independently of their association with cells (Fig. 1c and Supplementary Fig. 1c). The obtained marker-positive areas are then normalised over the area of the whole image, or those of specific tissue compartments or positive for certain markers using the predefined masks, to allow comparisons across samples. The pixel-based analysis is useful for the investigation of tissue features that are not detectable at the cell level. For instance, extracellular or secreted proteins cannot be quantified with approaches dependent on cell segmentation. In addition, being completely cell agnostic, the pixel-based analysis can provide independent validation of cell-based observations.

SIMPLI generates tables, plots and images as outputs of each process, thus enabling the visualisation of results at every step of the analysis.

IMC quantification of secreted and cell-associated IgA in human colon. To test its performance and versatility, we applied SIMPLI to four case studies of multiplexed images obtained with different technologies and with diverse origin, size and resolution of the tissue sections (Table 2).

As a first case study, we used SIMPLI to compare the levels of secreted and cell-associated immunoglobulin A (IgA), the major immunoglobulin isotype in intestinal mucosa[24], from IMC-derived multiplexed images of normal human colon. We stained six colon sections (CLN1-CLN6, Supplementary Data 1) with 26 antibodies marking T cells, macrophages, dendritic cells and B cells as well as stromal components (Supplementary Data 2) and ablated one region of interest (ROI) per sample.

Using SIMPLI, we extracted and normalised the 28 single-channel images (26 antibodies and two DNA intercalators) for each of the six ROIs and combined them into a single image per ROI (Fig. 2a). This normalisation enabled the selection of a single threshold for each marker to be used across all samples, thus reducing the complexity of the analysis configuration. By applying these thresholds to the E-cadherin and vimentin expression, we obtained the masks for the epithelium and the lamina propria, respectively (Fig. 2b). We used these masks to assign cells to the two compartments and normalise marker values or positive areas in the downstream analyses.

We then used the pixel-based approach to quantify both the IgA expressed by the plasma cells resident in the diffuse lymphoid tissue of the lamina propria as well as the secreted IgA undergoing transcytosis to traverse the epithelial compartment (Fig. 2b). As expected, most secreted IgA was localised in the epithelial crypts with only minimal presence of IgA$^+$ area in the surface epithelium (Supplementary Fig. 2a). Quantification of the normalised IgA$^+$ areas in the two compartments (Supplementary Fig. 2b) confirmed higher IgA$^+$ levels in the lamina propria than in the epithelium (Fig. 2c). To assess the impact of image normalisation performed in the data processing step, we repeated the same analysis starting from the raw images and applying sample-specific thresholds to remove the background noise. The resulting IgA levels correlated linearly with those obtained from normalised images (Supplementary Fig. 2c), showing that data normalisation does not impact the results.

Next, we quantified the IgA$^+$ plasma cells in the lamina propria using the cell-based approach. First, we performed single-cell segmentation with the deterministic approach and retained only cells overlapping for at least 30% or their area with the lamina propria mask (Fig. 2d and Supplementary Fig. 2d). We verified that varying the threshold of the overall had a minimal impact on the proportion of cells assigned to the lamina propria (Supplementary Fig. 2e). We then identified IgA$^+$ plasma cells, T cells, macrophages, and dendritic cells resident in the lamina propria according to the highest overlap between the cell area and the mask of each immune cell population (Fig. 2e). Again, we verified that the relative proportion of these cell populations changed only minimally varying the threshold of the overlap with the lamina propria mask (Supplementary Fig. 2f). Finally, we quantified the four immune cell populations across the six samples and observed that IgA$^+$ plasma cells constitute approximately 25% of all immune cells (Fig. 2f). This is consistent with previous quantifications of the fraction of plasma cells over the total mononucleated cells in the lamina propria of healthy individuals[25].

The relative proportion of IgA$^+$ plasma cells positively correlated with the normalised IgA$^+$ area in the lamina propria, demonstrating that the quantification from the single-cell analysis is supported by the cell agnostic measurements at the pixel level (Fig. 2g).

Localisation of T follicular helper cells in IMC images of a germinal centre. As a second case study, we used SIMPLI to spatially localise the immune cell populations within a FFPE section of the healthy human appendix (APP1, Supplementary Data 1). After staining the tissue section with 28 markers (26 antibodies and two DNA intercalators, Supplementary Data 2), we performed IMC and used SIMPLI to extract and normalise the single-channel images from the raw IMC data. The resulting combined image revealed a germinal centre in the B cell area and follicle-associated epithelium forming the boundary with the appendiceal lumen (Fig. 3a).

We performed single-cell segmentation with both approaches implemented in SIMPLI and observed high overlap in the identified cells (Supplementary Fig. 3a), indicating a good concordance between the two methods. We then classified 7573 cells obtained with the deterministic segmentation approach in immune and epithelial cells based on the highest overlap with the corresponding masks obtained in the data processing step (Fig. 3b, c). We obtained similar proportions of cells starting from the raw data and applying the z-score normalisation and k-means clustering as implemented in Histocat[26] (Supplementary Fig. 3b), again demonstrating that the normalisation implemented in SIMPLI does not impact the downstream analysis.

Next, we used both methods implemented in SIMPLI to further phenotype the T cells identified within the ROI. First, we applied unsupervised clustering using seven markers of T cell function (Supplementary Data 2). After inspection of the resulting clusters at different resolution levels, we selected 0.25 resolution that returned five distinct cell clusters (Fig. 3d). Based on the marker expression profiles, we assigned cluster 1 to CD4$^+$ T cells, cluster 2 to CD8$^+$CD45RO$^+$ T cells, cluster 3 to CD4$^+$CD45RA$^+$ T cells, cluster 4 to CD4$^+$CD45RO$^+$ T cells and cluster 5 to PD1$^+$CD4$^+$ T cells (Fig. 3e). The latter likely represented a set of PD1$^+$ T follicular helper cells known to be located in the germinal centre[27]. Interestingly, at higher resolution levels, cluster 5 was further divided into two smaller clusters showing PD1 high and low expression (Supplementary Fig. 3c). Similarly, clusters 1 and 2 were further divided into smaller subpopulation based on CD4 and CD45RO expression levels, respectively (Supplementary Fig. 3c). Therefore, although higher resolution levels increase the granularity of cell phenotyping, the unsupervised clustering approach implemented in SIMPLI is robust in identifying similar phenotyping clusters independently of the chosen resolution.

Table 2 Description of the case studies used to test SIMPLI.

Case study	Imaging technology	Analysed samples (*n*)	Channels (*n*)	ROI (mm^2)	Resolution (µm/pixel)	HPC platform	CPU time (h)	Elapsed real time (h)	RAM (GB)	Processes
1 (Fig. 2)	IMC	6	28	1.00	1.00	SGE	00:20:41	00:06:10	4.1	Raw data processing Cell masking Single-cell quantification Pixel intensity comparison
2 (Fig. 3)	IMC	1	28	1.00	1.00	SLURM	00:06:25	00:05:30	4.2	Raw data processing Cell masking Unsupervised clustering Expression thresholding Homotypic cell distances
3 (Fig. 4)	mIF	1	7	5.45	0.50	SLURM	00:11:45	00:08:23	16.7	Thresholding and masking Expression thresholding Heterotypic cell distances
4 (Fig. 5)	CODEX	35	58	1.13	0.38	SGE	02:32:35	00:26:01	22.5	Expression thresholding Heterotypic cell distances

For each case study, listed are the imaging technologies used to generate the tissue images, the number of samples and markers used, the size of the analysed region of interest (ROI), the resolution of the obtained images, the high-performance (HPC) platform and the computational resources employed to perform the analysis. These include the central processing unit (CPU) time and the elapsed real time, as well as the maximum random access memory (RAM) memory used. Finally, the specific analytical processes performed in each case study are also listed (single-cell segmentation was performed in all of them).
SGE Sun Grid Engine, *SLURM* Simple Linux Utility for Resource Management.

We re-identified the PD1$^+$ T follicular helper cells with the second phenotyping approach based on expression thresholding of CD4 and PD1. Starting from all T cells, we first extracted CD4$^+$ T cells (≥0.1 CD4 expression, Fig. 3f) and, within those, we further identified PD1$^+$ cells (≥0.15 PD1 expression, Fig. 3g). Both thresholds were chosen after manual inspection of the histological images. The expression profile of the resulting PD1$^+$CD4$^+$ T cells (Fig. 3h) closely recapitulated that of cluster 5 (Fig. 3e). We repeated the same analysis for clusters 1–4 confirming the high overlap between cells in unsupervised clusters and those re-identified using marker expression thresholds (Supplementary Fig. 3d). Moreover, these cells showed similar expression profiles (Supplementary Fig. 3e) and spatial localisation (Supplementary Fig. 3f), indicating that cell phenotypes identified with unsupervised clustering can be confirmed through user-guided thresholding of marker expression.

Finally, we investigated the spatial localisation of PD1$^+$ T follicular helper cells within the ROI by analysing their homotypic aggregations. This allowed us to localise a single high-density cluster containing 84% of PD1$^+$CD4$^+$ T cells within the germinal centre (Fig. 3i). This distribution of PD1$^+$CD4$^+$ T cells was in accordance with the localisation of T helper cells in the follicles of secondary lymphoid organs[27] and was confirmed by the histological inspection of the tissue image (Fig. 3j).

mIF analysis of spatially resolved cell–cell interactions in rectal cancer. As a third case study, we applied SIMPLI to the spatial analysis of mIF-derived images of a rectal cancer sample (CRC1, Supplementary Data 1) stained with anti CD8, PD1, Ki67, PDL1, CD68, GzB and 4',6-diamidino-2-phenylindole (DAPI) antibodies (Supplementary Data 2). We focused on a 5-mm^2 ROI, rich in T cells and located at the invasive margin of the tumour (Fig. 4a). This allowed us to characterise the cell–cell interactions between PDL1$^+$ cells and PD1$^+$CD8$^+$ T cells at the tumour boundary in a larger ROI, supporting the scalability of SIMPLI to the analysis of large regions (Table 2).

After image normalisation and single-cell segmentation, we identified PDL1$^+$ and PD1$^+$CD8$^+$ cells by applying expert-defined thresholds to PDL1 (≥0.01), CD8 (≥0.01), and PD1 (≥0.005) expression levels, respectively. We extracted 2026 PDL1$^+$ cells (Fig. 4b) and 3177 CD8$^+$ cells, 94 of which also expressed PD1 (Fig. 4c). The two sets of PDL1$^+$ and PD1$^+$CD8$^+$ cells constituted 3.7% and 0.2% of all cells in the analysed region, respectively (Fig. 4d). We confirmed similar proportions of PDL1$^+$ and PD1$^+$CD8$^+$ cells by performing signal unmixing, cell segmentation and cell phenotyping with the Inform tissue analysis software[28] (Akoya Biosciences, Fig. 4e).

We characterised the spatial relationship between these cells, focusing on the ones in close proximity to each other. Using the Euclidean distances between their centroids, we identified 35 PDL1$^+$ cells and 21 PD1$^+$CD8$^+$ T cells at a distance lower than 12 µm apart, which corresponded to twice the maximum cell radius length. We considered these cells proximal enough to be engaging in PD1-PDL1 mediated interactions. By comparing PD1$^+$CD8$^+$ T cells proximal to PDL1$^+$ cells and PD1$^+$CD8$^+$ T cells distal to PDL1$^+$ cells, we found no difference in the expression of cytotoxicity (GzB) or proliferation (ki67) markers (Fig. 4f). This is in line with the broad range of cytotoxic activity in this T cell subset observed in colorectal cancer[29]. On the contrary, PDL1$^+$ cells proximal to PD1$^+$CD8$^+$ T cells expressed higher levels of CD68 than PDL1$^+$ cells distal to PD1$^+$CD8$^+$ T cells (Fig. 4g), suggesting spatial proximity between PDL1$^+$ macrophages and PD1$^+$CD8$^+$ T cells. To validate this observation, we identified 1392 macrophages by applying an expert-defined threshold to CD68 expression value (≥0.01, Fig. 4h). We then classified these macrophages as PDL1$^-$ and

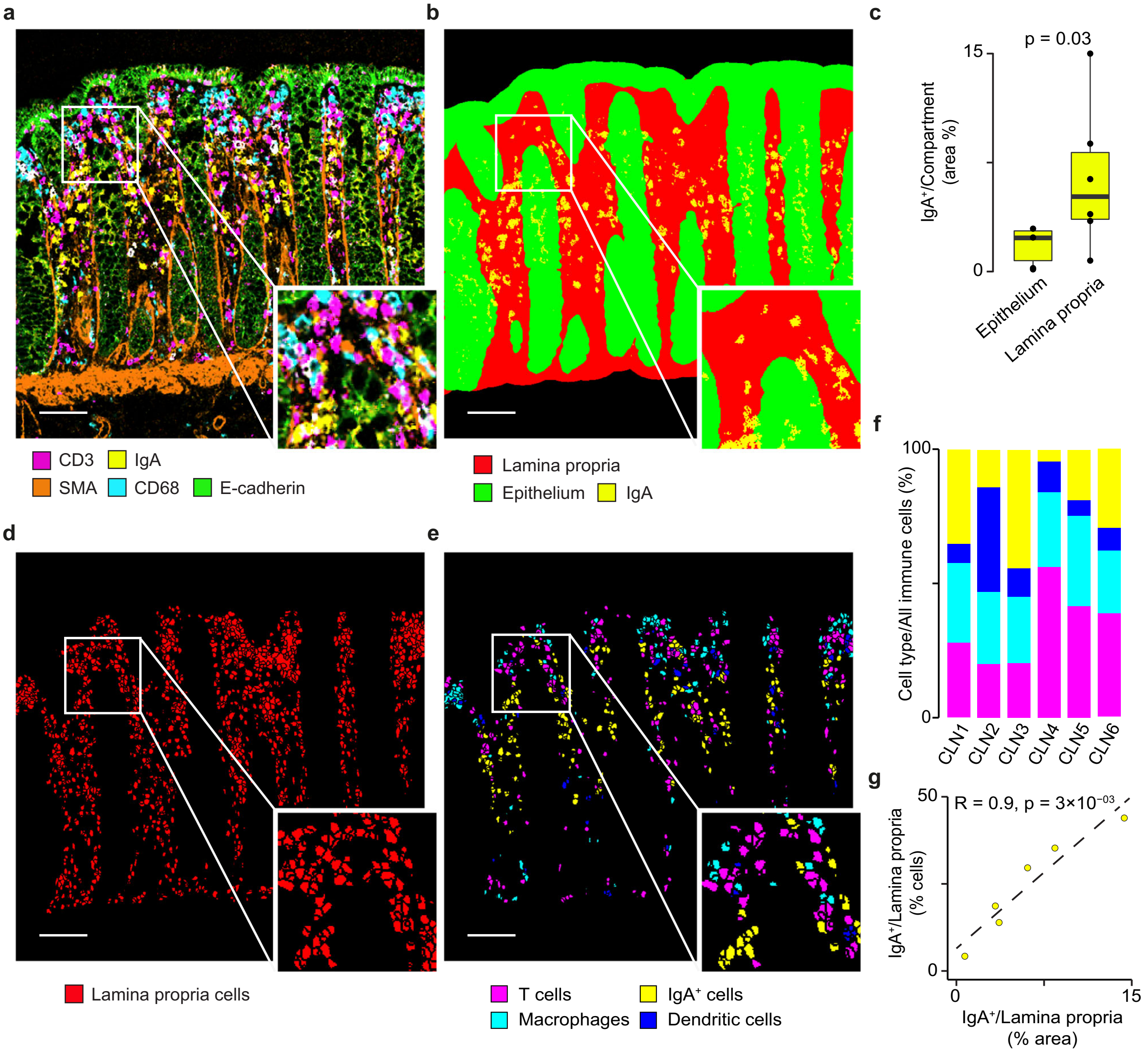
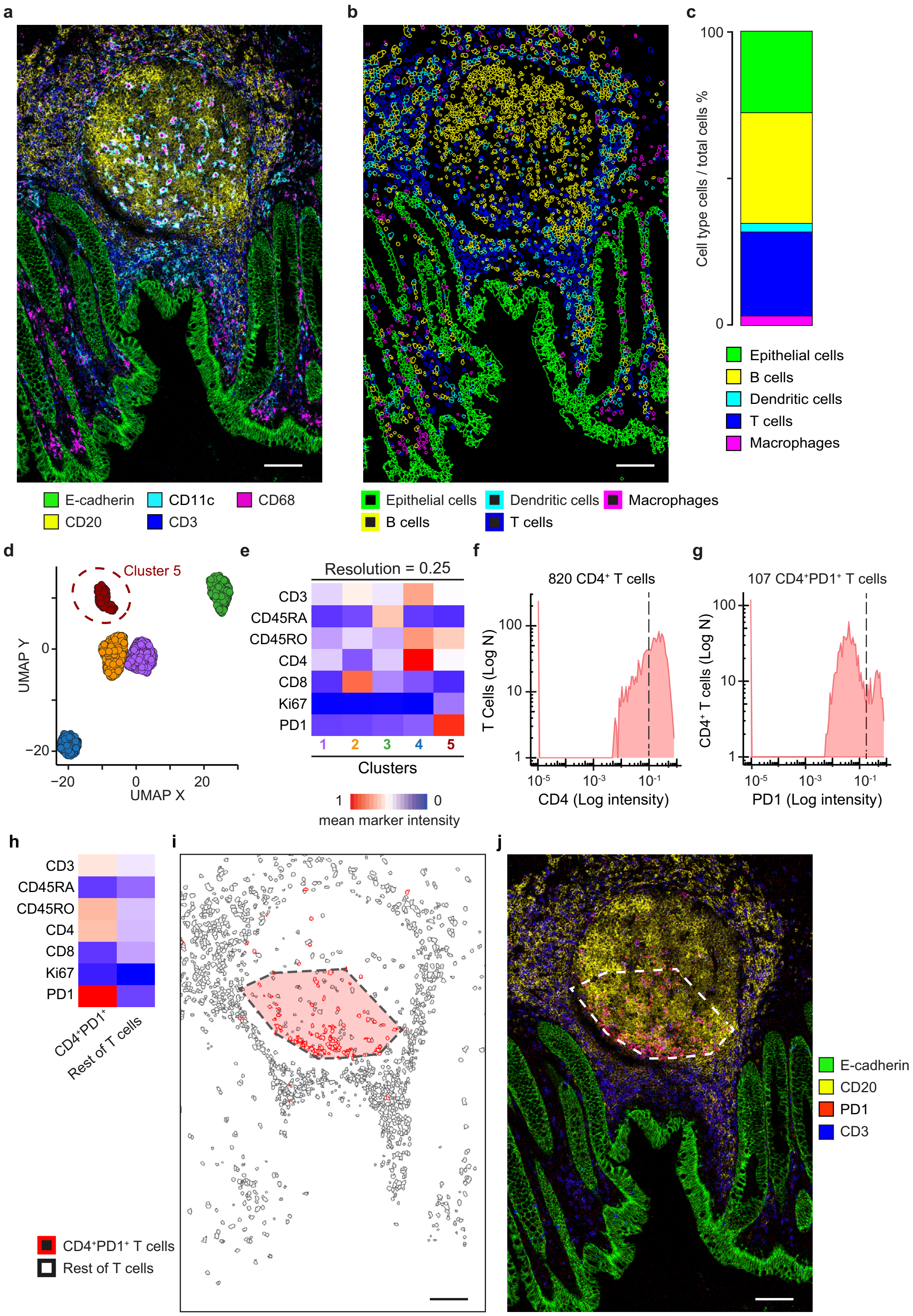


Fig. 2 IgA quantification in human colon mucosa. a IMC image of a representative sample (CLN6) of normal colon mucosa after extraction and normalisation of raw data. **b** Masks defining the lamina propria and the epithelial compartments overlaid with IgA$^+$ areas. Lamina propria and epithelial masks were obtained by thresholding the vimentin and E-cadherin channels, respectively. **c** Comparison of normalised IgA$^+$ areas in the lamina propria and epithelial compartments in six independent biological samples (CLN1-CLN6). Normalised areas were measured as the proportion of IgA$^+$ area over the lamina propria and epithelium masks, respectively. Data are presented as a box centred around the median and extending from the first to the third quartile. Whiskers represent the minimum and maximum values. An exact *p* value was calculated using a two-sided Wilcoxon test. **d** Outlines of the cells in the lamina propria. After single-cell segmentation, all cells overlapping with the lamina propria mask by at least 30% of their area were considered as cells resident in the lamina propria. **e** Outlines of immune cells resident in the lamina propria identified according to the highest overlap between their area and the masks for IgA$^+$ cells, T cells, macrophages and dendritic cells. **f** Relative proportions of T cells, IgA$^+$ cells, macrophages and dendritic cells over all immune cells in the lamina propria across CLN1-CLN6. **g** Correlation between normalised IgA$^+$ area and the proportion of IgA$^+$ cells over the total immune cells in the lamina propria in six independent biological samples (CLN1-CLN6). Pearson correlation coefficient R and associated *p* value based on Fisher's Z transform are shown. Images in panels (**a**), (**b**), (**d**), (**e**) were derived from a representative sample (CLN6, Supplementary Data 1). CD3 and T cells, magenta; IgA and IgA$^+$ cells, yellow; Smooth Muscle Actin (SMA), orange; CD68 and macrophages, cyan; E-cadherin and epithelial cells, green; Lamina propria and lamina propria cells, red; Dendritic cells, blue. Scale bar in all images = 100 µm. Source data are provided as a Source Data file.

PDL1$^+$ cells, respectively, using 0.1 PDL1 expression threshold. Comparing the distance of the resulting two populations from the nearest PD1$^+$CD8$^+$ T cells, we confirmed that PDL1$^+$CD68$^+$ macrophages were significantly closer to PD1$^+$CD8$^+$ T cells than PDL1$^-$CD68$^+$ macrophages (Fig. 4i). By inspecting the imaged tissue at ×40 magnification, we confirmed the localisation of PDL1$^+$CD68$^+$ macrophages in close proximity to PD1$^+$CD8$^+$ cells, as well as the presence of both PD1$^+$CD8$^+$GzB$^-$ T cells and PD1$^+$CD8$^+$GzB$^+$ T cells proximal to PDL1$^+$ cells (Fig. 4j).

Comparison of cell distances in CODEX images of colorectal cancer subtypes. As a fourth case study, we used SIMPLI to compare the distances between immune cells and tumour or



endothelial cells in Crohn's-like reaction (CLR) and diffuse inflammatory infiltration (DII) colorectal cancer subtypes[9]. The high-dimensional imaging data were derived from 35 colorectal cancer samples (17 CLRs and 18 DIIs, Supplementary Data 1) and were obtained using CODEX with a 56 marker panel[9] (Supplementary Data 2). Such a large number of antibodies enabled the identification and spatial localisation of T cells, B cells, plasma cells, macrophages, NK cells, granulocytes, dendritic cells, tumour cells, neuroendocrine cells, smooth muscle, nerves, lymphatic and blood vessels (Fig. 5a).

Fig. 3 Single-cell characterisation of T cells in a human germinal centre. a IMC image of a normal appendix (APP1) showing a central germinal centre with the columnar epithelium delimiting the appendiceal lumen. **b** Outlines of T cells, B cells, macrophages, dendritic and epithelial cells identified through the highest overlap with the respective masks. **c** Proportions of T cells, B cells, macrophages, dendritic and epithelial cells over all cells. **d** UMAP plot of 1466 T cells grouped in five clusters resulting from unsupervised clustering according to the expression of seven markers of T cell function (Supplementary Data 2). Cluster 5 (circled) corresponds to PD1$^+$CD4$^+$ T cells. **e** Expression profiles of the five clusters identified in (**d**). The mean intensity value of each marker across all cells is reported. The colour scale was normalised across all markers and cells. **f** Density plots of CD4 expression in T Cells. Cells with ≥0.1 CD4 expression were considered as CD4$^+$ T cells. **g** Density plot of PD1 expression in CD4$^+$ T cells. Cells with ≥0.15 PD1$^+$ expression were considered as PD1$^+$CD4$^+$ T cells. Thresholds for CD4 and PDL1 were identified through histological inspection of the PD1 channel images. **h** Expression profiles of the PD1$^+$CD4$^+$ T cells and the rest of T cells. For both populations, the mean intensity value of each marker across all cells is shown. The colour scale was normalised across all markers and cells. **i** Position map of T cells within the ROI. The area of a high-density cluster of ≥5 PD1$^+$CD4$^+$ T cells per 10,000 μm^2 is highlighted in red. **j** IMC image showing the localisation of the PD1 signal within the ROI. Images in (**a**), (**b**), (**i**), and (**j**) were derived from a single experiment (APP1, Supplementary Data 1). Panels (**a**), (**b**), (**c**), (**i**), and (**j**): E-cadherin and epithelial cells, green; CD11c and Dendritic cells, cyan; CD68 and macrophages, magenta; CD20 and B cells, yellow; CD3 and T cells, blue; PD1 and PD1$^+$CD4$^+$ cells, red. Panels (**d**) and (**e**): cluster 1, violet; cluster 2, orange; cluster 3, green; cluster 4, blue; cluster 5, red. Scale bar for all images = 100 μm. Source data are provided as a Source Data file.

We processed the raw data from the original study, including normalisation. We then performed single-cell segmentation and quantified the main cell types identified in the original study[9] by applying expert-defined thresholds to the expression of markers representative of each population (CDX2, MUC1 or cytokeratin for tumour cells; CD34 or CD31 for endothelial cells; vimentin for stromal cells; CD11c for dendritic cells; CD38 for B cells; CD3 and CD4 for CD4$^+$ T cells; CD3, CD4 and FOXP3 for Tregs; CD3 and CD8 for CD8$^+$ T cells, CD68 for macrophages). The obtained relative proportions of immune cells across all samples were highly concordant with those reported in the original study[9] (Fig. 5b).

We then measured the distances of the main immune cell types from tumour cells and blood vessels by performing a heterotypic spatial analysis as implemented in SIMPLI. First, we calculated the distances of each macrophage, CD8$^+$ T cell, CD4$^+$ T cell, Treg and B cell to the nearest tumour cell or endothelial cell using the coordinates of the cell centroids. From these, we derived the corresponding distance distributions from the nearest tumour cell or endothelial cell in each sample. Finally, we compared the resulting distributions between 17 CLR and 18 DII colorectal cancer subtypes. After correcting for multiple testing, we considered biologically relevant only differences between the median distances of the two sample subtypes bigger than 8 μm, corresponding to the diameter of B and T lymphocytes[30]. With this approach, we found that Tregs were significantly closer to tumour cells in DII (median distance = 22.4 μm) compared to CLR (35.6 μm, Fig. 5c). On the contrary, B cells were more proximal to blood vessels in CLR (33.5 μm) than in DII (43.3 μm, Fig. 5d). We further supported these results with a permutation test, where we re-labelled randomly the identities to all cells in each sample for 10,000 times to derive an expected distribution of differences in distances between CLR and DII cells. The comparisons of observed values to the expected distributions, confirmed that Tregs were significantly closer to tumour cells in DII (Fig.5e) while B cells were significantly closer to blood vessels in CLR (Fig. 5f). Since the spatial randomness used as a baseline for the permutation test is an approximation of the highly organised structure of biological tissues, we sought further support this result through independent inspection of the spatial distributions of B cells in CLRs (Fig. 5g) and DII (Fig. 5h) in the histological images.

This result, not reported in the original study, showcases the discovery potential of the quantitative analysis of spatial relationships between cell populations implemented in SIMPLI. In addition, the SIMPLI graphical representations of the tissue composition as an overlay of cell boundaries colour-coded by cell populations greatly facilitate the visual inspection of their spatial interactions in their original tissue context.

Discussion

SIMPLI is an open-source, customisable and technology-independent tool for the analysis of multiplexed imaging data. It enables the processing of raw images, the extraction of cell data and the spatially resolved quantification of cell types or functional states as well as a cell-independent analysis of tissues at the pixel level, all within a single platform (Table 1). Moreover, it gives high flexibility to the user who can decide whether to skip processes implemented in SIMPLI and replace them with external tools to then re-start the pipeline at any point.

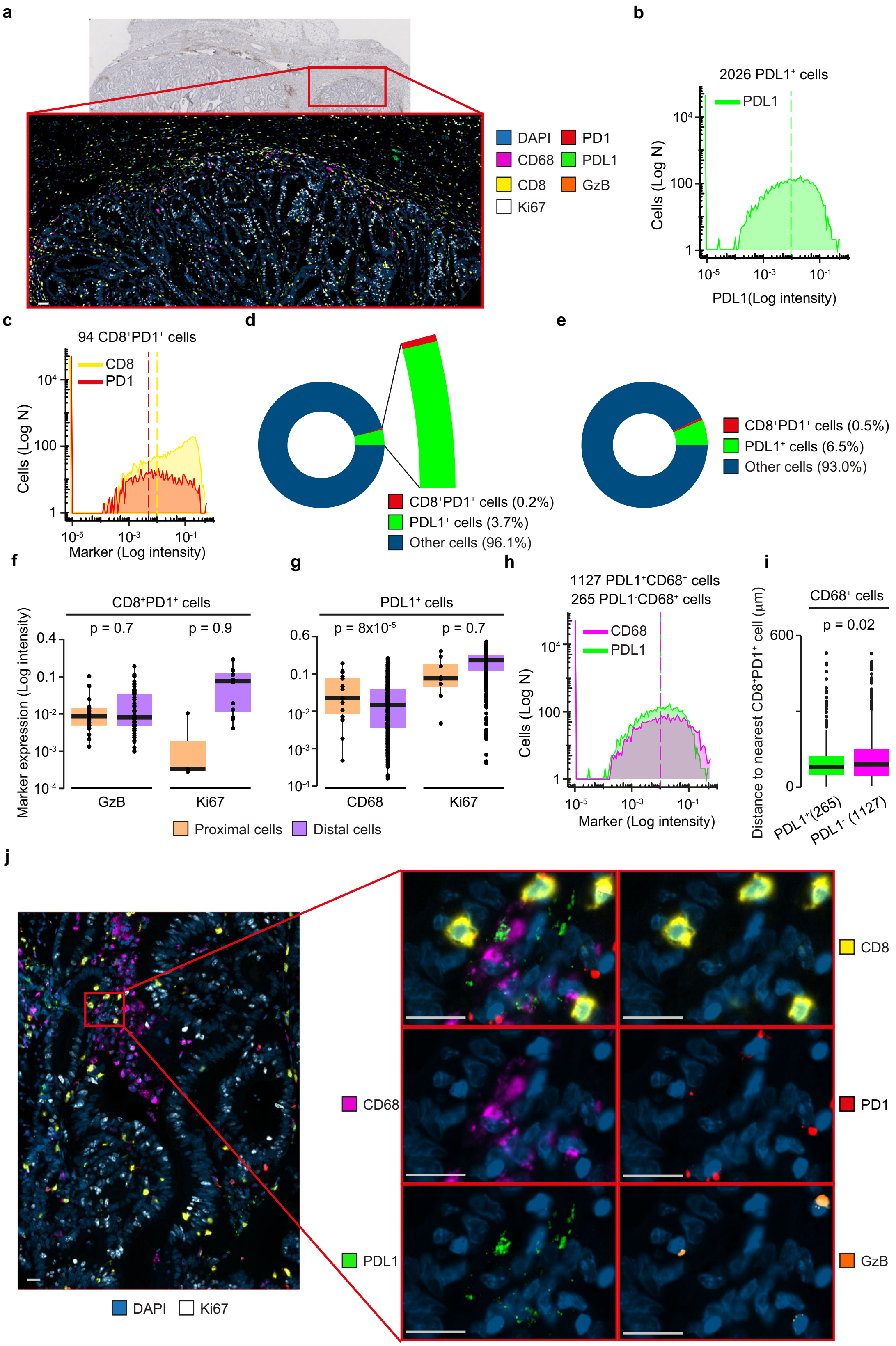
In comparison to currently available software, SIMPLI increases the portability, scalability and reproducibility of the analysis (Table 2). Moreover, it can easily accommodate specific analytical requirements across a wide range of tissues and imaging technologies at different levels of resolution and multiplexing through user-friendly configuration files. SIMPLI interoperates with multiple software and programming languages by leveraging workflow management and containerisation. This makes the inclusion of additional algorithms, features and imaging data formats easy to implement. For example, as possible future developments, SIMPLI may include alternative methods of cell segmentation, pixel and cell classification or a Graphical User Interface for interactive data visualisation. For this reason, we will maintain SIMPLI and its documentation up-to-date and will further expand it to leverage new tools as they become adopted by the community. Similarly, feedback from users will be collected through the dedicated GitHub repository.

Multiplexed imaging methods have proven to be a powerful approach for the study of tissues through the in-depth characterisation of cell phenotypes and interactions. SIMPLI, which was recently able to reveal differences in the composition of the microenvironment between colorectal cancers responsive and resistant to anti-PD1 immunotherapy[31], represents an effort to make these analyses more accessible to a wider community. This will enable the exploitation of highly multiplexed imaging technologies for multiple applications, ranging from basic life science and pharmaceutical research to precision medical use in the clinics.

Methods

All patients enrolled in this study provided written informed consent in accordance with approved institutional guidelines (University College London Hospital, REC Reference: 20/YH/0088; Istituto Clinico Humanitas, REC Reference: ICH-25-09).

SIMPLI description and implementation. SIMPLI's workflow is divided into three steps (raw image processing; cell-based analysis; pixel-based analysis), which are constituted of multiple standalone processes (Fig. 1 and Supplementary Fig. 1). Processes can be executed sequentially or independently from the command line or through a configuration file that can be edited with any text editor. This allows the user to skip some of them and use alternative input data for downstream analyses. In addition, parameters and options can be specified through the same



configuration files without the need to set up tool-specific input files in any specific directory structure.

Raw data from IMC or MIBI experiments (.mcd or.txt files) are converted into single or multi-channel.tiff images with imctools[32]. Data from other multiplexed imaging platforms may be supplied directly as raw single or multi-channel tiff images (Supplementary Fig. 1a). Raw images can be thresholded individually to minimise the effect of non-uniform staining and then used directly for the cell- and pixel-based analyses. Alternatively, they can be first normalised across samples by rescaling pixel values of each channel up to the 99th percentile of the distribution using the EBImage[33] package and custom R scripts. Normalised images can then be processed with CellProfiler4[11] to generate thresholded images and masks of tissue compartments or markers to be used in the following steps. In this step, the user can apply a range of filters, thresholds and morphological operations to each image, according to the experimental plan.

Fig. 4 Characterisation of PDL1$^+$ and PD1$^+$ cells at the tumour invasive margin. a CD3 immunohistochemistry (main image) and sequential mIF image (zoom-in, ×20 magnification) of a rectal cancer sample (CRC1). The mIF image corresponded to a 5 mm^2 tissue area at the invasive margin of the tumour and was obtained by combining the pre-processed images of seven markers. Scale bar = 50 μm. **b** Density plot of PDL1 expression in CD8$^-$ cells. Cells with ≥0.01 PDL1 expression were considered as PDL1$^+$ cells. **c** Density plots of CD8 and PD1 expression in T cells. Cells with ≥0.01 CD8 expression and ≥0.005 PD1 expression were considered as PD1$^+$CD8$^+$ T cells. Expression thresholds were identified through histological inspection of PDL1, CD8 and PD1 channel images and are indicated as dotted lines in the corresponding plots. **d** Proportions of PD1$^+$CD8$^+$ cells and PDL1$^+$ cells over total cells, as measured using SIMPLI data processing, including normalisation. **e** Proportions of PD1$^+$CD8$^+$ cells and PDL1$^+$ cells over total cells, as measured using the Inform tissue analysis software package[28]. **f** Comparison of the mean intensity of GzB and Ki67 between PD1$^+$CD8$^+$ T cells proximal ($n = 21$) and distal ($n = 73$) to PDL1$^+$cells. Proximal PD1$^+$CD8$^+$ T cells were defined as those at less than 12 μm from a PDL1$^+$ cell. **g** Comparison of the mean intensity of CD68 and Ki67 between PDL1$^+$cells proximal ($n = 35$) and distal ($n = 1991$) to PD1$^+$CD8$^+$ T cells. Proximal PDL1$^+$ cells were defined as those at less than 12 μm from a PD1$^+$CD8$^+$ T cell. **h** Density plots of CD68 and PD1 expression in all cells. Cells with ≥0.01 CD68 and PDL1 expression were considered as PDL1$^+$CD68$^+$ cells. **i** Comparison of distance of PDL1$^+$ ($n = 265$) and PDL1$^-$ ($n = 1127$) CD68$^+$ macrophages to the nearest PD1$^+$CD8$^+$ T cell. Data in (**f**), (**g**) and (**i**) are presented as a box centred around the median and extending from the first to the third quartile. Whiskers represent the minimum and maximum values. An exact *p* value was calculated using a two-sided Wilcoxon test. **j** High-resolution (×40 magnification) mIF image of PD1$^+$CD8$^+$ T cells proximal to PDL1$^+$CD68$^+$ cells. Zoom in images show each marker separately and merged. Scale bar = 20 μm. Images in (**a**) and (**j**) were derived from a single experiment (CRC1, Supplementary Data 1). DAPI and other cells, blue; PD1 and PD1$^+$CD8$^+$ T cells, red; CD68 and PDL1$^-$ CD68$^+$ cells, magenta; PDL1 and PDL1$^+$CD68$^+$ cells, green; CD8, yellow; Granzyme B (GzB), orange; Ki67, white; proximal cells, pink; distal cells, violet. Source data are provided as a Source Data file.

Pixel-based and cell-based analyses can be run as single workflows or in parallel within the same run. Both of them provide multiple outputs of the various processes, including tabular text files, visualisation plots and comparisons across samples (Supplementary Fig. 1).

The cell-based analysis is composed of cell data extraction, cell phenotyping and spatial analysis (Supplementary Fig. 1b). The extraction of cell data starts with single-cell segmentation using CellProfiler4[11] or StarDist[23] with scikit-image[34] used for feature extraction. In the latter case, default models or user-provided trained models can be used. Cell segmentation returns (1) single-cell data consisting of the marker expression values and the coordinates of each cell in the ROI and (2) the ROI segmentation mask marking all the pixels belonging to each cell with its unique identifier. Cells mapping to tissue compartments or positive for certain markers can then be identified based on their overlap with the tissue compartments or marker masks derived in the previous step. These cells are visualised in the ROI as outlines, while their proportions are quantified in barplots and boxplots.

All cells or only those in specific tissue compartments or positive for certain markers can be further phenotyped using two approaches. The first consists of unsupervised clustering based on the marker expression values using Seurat[35]. Cells are represented as nodes in a k-nearest neighbour graph based on their Euclidean distances in a principal component analysis space. This graph is then partitioned into clusters using the Louvain algorithm[36] at user-defined levels of resolution leading to the unsupervised identification of cell phenotypes. Clusters of cell phenotypes are plotted as scatterplots in Uniform Manifold Approximation and Projection (UMAP)[37] space. The second phenotyping approach is based on user-defined thresholds of marker expression values that can be combined using logical operators for the identification of designated cell phenotypes. The distributions of cells are represented as density plots based on the marker expression levels. In both phenotyping approaches, the expression profiles of the cell types are plotted as heatmaps, their proportions quantified in barplots and boxplots and their locations in the ROI visualised as cell outlines.

Once cell populations and phenotypes have been identified, the spatial analysis investigates the distance between cells of the same (homotypic aggregations) or different (heterotypic aggregations) types. The homotypic and heterotypic spatial analyses can be run in parallel or singularly on one or more sets of cells. In the homotypic analysis, clusters of cells of the same type within a user-defined distance are identified with DBSCAN[38] as implemented in the fpc[39] R package. These homotypic cell aggregations are visualised as position maps, reporting cell location and high-density clusters in the ROI. In the heterotypic analysis, the cell distances, defined as the Euclidean distances between cell centroids, are computed using custom R scripts and visualised as density plots. The resulting distribution of cell distances can be compared between group of samples using a two-sided Wilcoxon test with Benjamini–Hochberg FDR correction. Observed distances can also be compared to the distribution of expected distances obtained by reshuffling cell identities in each sample randomly for a user-defined number of times (default value = 10,000 reshufflings, Supplementary Fig. 1b). The statistical significance of this comparison is evaluated with a two-tailed permutation test adjusted for multiple hypothesis testing with the Benjamini–Hochberg correction.

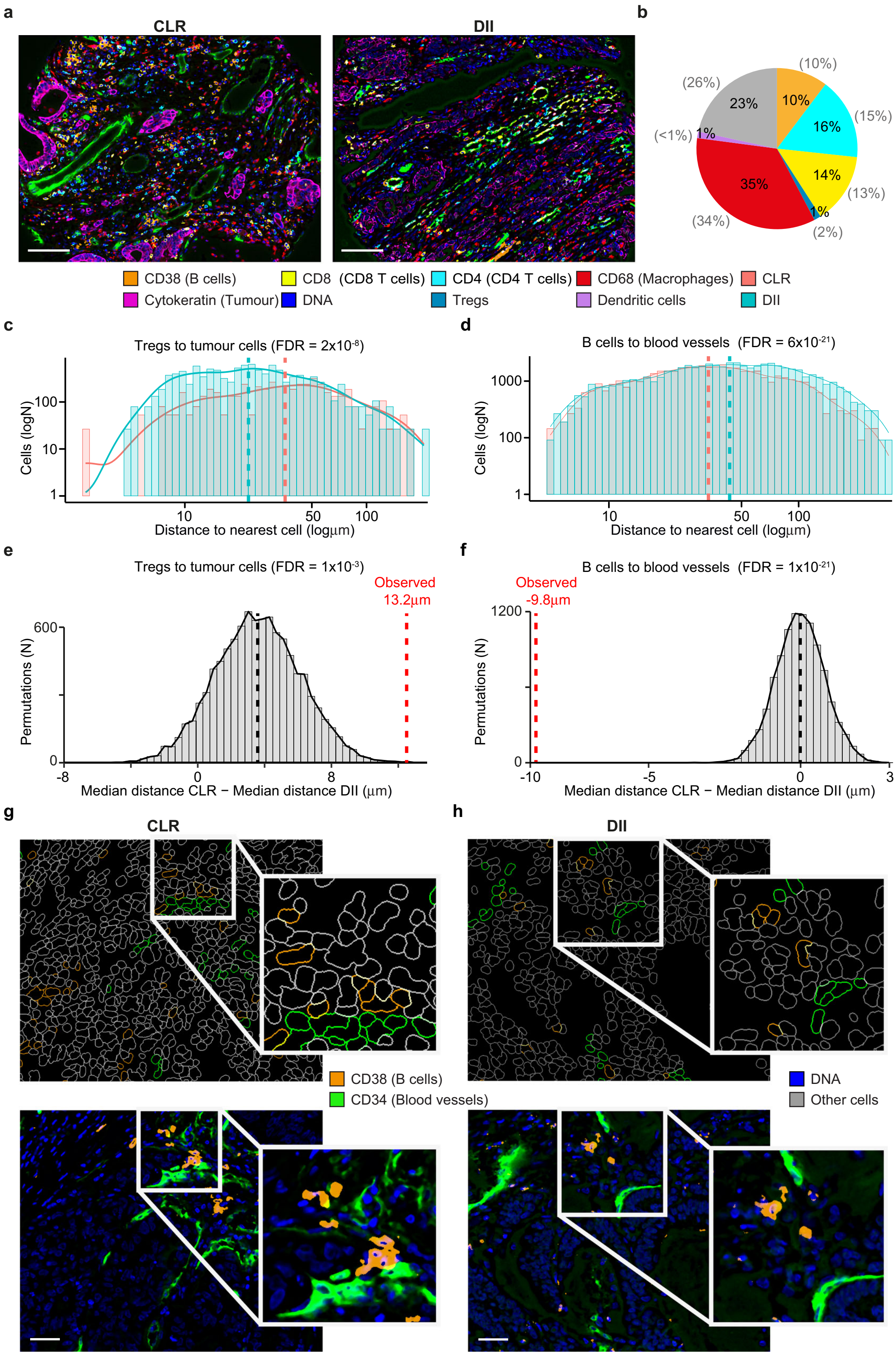
The pixel-based analysis quantifies areas positive for a user-defined combination of markers using the EBImage[33] package with custom R scripts (Supplementary Fig. 1c). These measurements are performed starting from the thresholded images produced in the raw image processing step (Supplementary Fig. 1a). The marker-positive areas obtained in this way are then normalised over the area of the whole image or specific tissue or marker compartments. The resulting normalised positive areas can then be quantified in barplots and boxplots.

SIMPLI is implemented as a Nextflow[40] pipeline employing Singularity containers[41] hosted on Singularity Hub[42] to manage all the libraries and software tools. This allows SIMPLI to automatically manage all dependencies, irrespective of the running platform. Nextflow also manages automatic parallelisation of all processes while still allowing the selection of parts of the analysis to execute.

Sample description. Six FFPE blocks of normal (non-cancerous) colon mucosa (CLN1-CLN6), one of normal appendix (APP1) and one of rectal cancer (CRC1) were obtained from eight individuals who underwent surgery for the removal of colorectal cancers (Supplementary Data 1). All blocks were reviewed by an expert pathologist (M.R.-J.).

Staining and IMC ablation of human colon mucosa and appendix. Four-μm-thick sections were cut from each block of samples CLN1-CLN6 and APP1 with a microtome and used for staining with a panel of 26 antibodies targeting the main immune, stromal and epithelial cell populations of the gastrointestinal tract (Supplementary Data 2). The optimal dilution of each antibody in the panel was identified by staining and ablating FFPE appendix sections. The resulting images were reviewed by a mucosal immunologist (J.S.) and the dilution giving the best signal to background ratio was selected for each antibody (Supplementary Data 2). To perform the staining for IMC, slides were dewaxed after a 1-h incubation at 60 °C, rehydrated and heat-induced antigen retrieval was performed with a pressure cooker in Antigen Retrieval Reagent-Basic (R&D Systems). Slides were incubated in a 10% BSA (Sigma), 0.1% Tween (Sigma), and 2% Kiovig (Shire Pharmaceuticals) Superblock Blocking Buffer (Thermo Fisher) blocking solution at room temperature for 2 h. Each antibody was added to a primary antibody mix at the selected concentration in blocking solution and incubated overnight at 4 °C. After two washes in PBS and PBS-0.1% Tween, the slides were treated with the DNA intercalator Cell-ID™ Intercalator-Ir (Fluidigm) (containing the two iridium isotopes 191Ir and 193Ir) 1.25 mM in a PBS solution. After a 30-min incubation, the slides were washed once in PBS and once in MilliQ water and air-dried. The stained slides were then loaded in the Hyperion Imaging System (Fluidigm) imaging module to obtain light-contrast high-resolution images of approximately 4 mm^2. These images were used to select the ROI in each slide. For CLN1-CLN6, 1 mm^2 ROIs were selected to contain the full thickness of the colon mucosa, with epithelial crypts in a longitudinal orientation. For APP1, a 1-mm^2 ROI containing a lymphoid follicle in its whole depth alongside a portion of lamina propria and of the epithelium was selected. ROIs were ablated at a 1 μm/pixel resolution and 200 Hz frequency.

IMC data analysis of human colon mucosa. Twenty-eight images from 26 antibodies (Supplementary Data 2) and two DNA intercalators were obtained from the raw.txt files of the ablated regions in CLN1-CLN6 using the data extraction process. Pixel intensities for each channel were normalised to the 99th percentile in all samples and Otsu thresholding was performed on the normalised images with a custom CellProfiler4 pipeline, which was employed also to generate the masks for the lamina propria (using the Vimentin channel including all <75-pixel large negative areas) and the epithelium (starting from the Pan-keratin and E-cadherin channels, dilatating the images with a three-pixel disk and the filling up of all <75-pixel large negative areas). These masks were then added into a sum image, which underwent dilatation with a three-pixel disk and filling up of all <25-pixel large negative areas. Positive features outside of the lamina and epithelium were removed with an opening operation using a 150-pixel radius and the lamina propria mask



was subtracted from the sum image to generate the final mask for the epithelial compartment. These masks and the thresholded images were used as input for the pixel-based and cell-based analysis processes. The IgA masks employed for the pixel analysis were generated using a three-class global Otsu thresholding with two background classes after applying a Gaussian filter with a 1.5-pixel large radius to remove high-intensity artefacts of that size, which we noticed after manual inspection of the images.

To evaluate the effect of normalisation on the downstream analysis, sample-specific thresholds were manually selected for IgA, E-Cadherin, Pan-Keratin and Vimentin and applied to the raw images. The resulting thresholded images were used to generate lamina propria and epithelial masks for each sample individually.

Pixel-level analysis was performed on the IgA masks derived from either the normalised or the raw images and IgA$^+$ areas in the tissue, lamina propria and epithelium were measured and normalised over the areas of the three compartments.

Fig. 5 Spatial localisation of immune cells in two colorectal cancer subtypes. a CODEX images of two representative CLR (CRC_12_24) and DII (CRC_31_16) colorectal cancer samples. **b** Proportions of $CD8^+$ T cells, $CD4^+$ T cells, Tregs, macrophages, dendritic cells, B cells and other mixed immune cell populations across the 35 analysed samples. Cell types were identified by applying expert-defined thresholds to the expression intensity of representative markers and normalised over the total non-cancer cells. These thresholds were derived through histological inspection of the channel images. The cell proportion corresponding to each population from the original study[9] is reported in brackets. Distance distribution of Tregs to the nearest tumour cell (**c**) and of B cells to the nearest endothelial cell (**d**) of CLR and DII samples. Distances between cell pairs were calculated using the cell centroids coordinates and the resulting distributions were compared between CRC subtypes using a two-sided Wilcoxon test. Benjamini–Hochberg FDR correction was applied for testing over ten cell type comparisons. Only differences of at least 8 µm and with FDR < 0.1 were considered significant. Dashed lines represent the medians of the distributions. Distribution of the expected differences between the median distances of Tregs to the nearest tumour cell (**e**) and of B cells to the nearest endothelial cell (**f**) in CLR and DII samples. Expected values were calculated with a permutation test, where cell identities were randomly reassigned for 10,000 times within each sample. The resulting median values were compared to the observed differences with a two-tailed permutation test adjusted for multiple hypothesis testing with the Benjamini–Hochberg correction. Single-cell outlines of B cells and blood vessels (upper panel) and associated images (lower panel) form a representative CLR (CRC_17_34) (**g**) and DII (CRC_15_29) (**h**) sample out of 35 colorectal cancer samples (Supplementary Data 1). CD38 and B cells, orange; CD8 and $CD8^+$ T cells, yellow; CD4 and $CD4^+$ T cells, cyan; CD68 and macrophages, red; cytokeratin and tumour cells, magenta; DNA, blue; Tregs, teal; dendritic cells, violet. Crohn's-like reaction (CLR) orange; diffuse inflammatory infiltration (DII), teal. Scale bar = 100 µm. Source data are provided as a Source Data file.

Cell-level analysis started with CellProfiler4 segmentation first on DNA1 with global Otsu thresholding to identify the cell nuclei. Then, cells were identified by radially expanding each nucleus for up to 10 pixels over a membrane mask derived from the IgA, CD3, CD68, CD11c and E-cadherin channels. After inspection by an expert histologist (J.S.), only cells overlapping with the lamina propria mask by at least 30% were retained.

Cell identities were assigned according to the highest overlap of the cell area with marker-specific thresholds defined by an expert histologist (J.S.): ≥15% of the IgA mask for IgA cells; ≥15% of the CD3 mask for T cells; ≥25% of the CD68 mask for macrophages; ≥15% of CD11c mask for dendritic cells.

IMC data analysis of human appendix. Images from the same 26 antibodies and two DNA intercalators used in the colon mucosa (Supplementary Data 2) were obtained from the raw.txt files of the ablated region in APP1, normalised to the 99th percentile and thresholded with CellProfiler4 as described above. For the cell-based analysis, nuclei were identified using the DNA1 channel and cells were isolated through watershed segmentation with the nuclei as seeds on a membrane mask summing up CD45, Pan-keratin and E-cadherin thresholded images.

Cells were assigned to the epithelium or to immune cell populations if they overlapped for ≥10% with the following masks: CD3 mask for T cells; CD20 and CD27 masks for B cells; CD68 mask for macrophages; CD11c mask for dendritic cells; E-cadherin$^+$ and Pan-keratin$^+$ masks for epithelial cells.

T cells were further phenotyped using unsupervised clustering at resolutions between 0.1 and 1.0, with 0.05 intervals and based on the cell marker intensity for CD3, CD45RA, CD45RO, CD4, CD8, Ki67 and PD1. The resulting clusters were manually inspected and the clustering with the highest number of biologically meaningful clusters (resolution = 0.25) was chosen. Clusters were re-identified using mean intensity thresholds defined by an expert histologist (J.S.) for the following markers: CD3 >0.06 for cluster 1; CD8a >0.125 for cluster 2; CD45RA >0.125 for cluster 3; CD4 >0.125 and CD45RO >0.15 for cluster 4; and CD4 >0.1 and PD1 >0.15 for cluster 5.

Homotypic aggregations of $PD1^+CD4^+$ T cells (cluster 5, resolution = 0.25) were computed using a minimum of five points per cluster and a reachability parameter corresponding to a density of at least 5 cells/mm^2.

CD3 staining and mIF of human rectal cancer. Two 4-µm-thick serial sections were cut from CRC1 FFPE block using a microtome. The first slide was dewaxed and rehydrated before carrying out HIER with Antigen Retrieval Reagent-Basic (R&D Systems). The tissue was then blocked and incubated with the anti-CD3 antibody (Dako, Supplementary Data 2) followed by horseradish peroxidase (HRP) conjugated anti-rabbit antibody (Dako) and stained with 3,3' diaminobenzidine substrate (Abcam) and haematoxylin. Areas with $CD3^+$ infiltration in the proximity of the tumour invasive margin were identified by a clinical pathologist (M.R.-J.)

The second slide was stained with a panel of six antibodies (CD8, PD1, Ki67, PDL1, CD68, GzB, Supplementary Data 2), Opal fluorophores and DAPI on a Ventana Discovery Ultra automated staining platform (Roche). Expected expression and cellular localisation of each marker as well as fluorophore brightness were used to minimise fluorescence spillage upon antibody-Opal pairing. Following a 1-h incubation at 60 °C, the slide was subjected to an automated staining protocol on an autostainer. The protocol involved deparaffinisation (EZ-Prep solution, Roche), HIER (DISC. CC1 solution, Roche) and seven sequential rounds of 1-h incubation with the primary antibody, 12 min incubation with the HRP-conjugated secondary antibody (DISC. Omnimap anti-Ms HRP RUO or DISC. Omnimap anti-Rb HRP RUO, Roche) and 16-min incubation with the Opal reactive fluorophore (Akoya Biosciences). For the last round of staining, the slide was incubated with Opal TSA-DIG reagent (Akoya Biosciences) for 12 min followed by Opal 780 reactive fluorophore for 1 h (Akoya Biosciences). A denaturation step (100 °C for 8 min) was introduced between each staining round in order to remove the primary and secondary antibodies from the previous cycle without disrupting the fluorescent signal. The slide was counterstained with DAPI (Akoya Biosciences) and coverslipped using ProLong Gold antifade mounting media (Thermo Fisher Scientific). The Vectra Polaris automated quantitative pathology imaging system (Akoya Biosciences) was used to scan the labelled slide. Six fields of view, within the area selected by the pathologist, were scanned at ×20 and ×40 magnification using appropriate exposure times and loaded into inForm[28] for spectral unmixing and autofluorescence isolation using the spectral libraries.

mIF data analysis. After spectral unmixing and merging of six ×20 fields of view for a total of >5 mm^2 ROI (Table 2), one single-tiff image was extracted for each marker and its intensity was rescaled from 0 to 1 with custom R scripts. The resulting single-tiff images were pre-processed to remove the background noise with Otsu thresholding in CellProfiler4 and used for cell segmentation by applying a global threshold to the DAPI channel and selecting all objects with a diameter between four and 60 pixels. $PD1^+CD8^+$ cells, $CD68^+$ cells and $PDL1^+$ cells were then identified using mean intensity thresholds of 0.01 for CD8, 0.005 for PD1, 0.01 for CD68 and 0.01 for PDL1. All thresholds were inspected by an expert histologist (J.S.). To crosscheck these results, images were analysed with the Inform[28] package. After spectral unmixing, images were segmented with the Adaptive Cell Segmentation option applied to the DAPI channel for nuclei identification ("relative intensity" = 0.1, "splitting sensitivity" = 0.1, "minimum size" = 5). Then $PD1^+CD8^+$ cells and $PDL1^+$ cells were identified.

The distributions of minimum distances between $PDL1^+$ cells and $PD1^+CD8^+$ cells were calculated from the coordinates of the centroids of each cell in the image. All $PDL1^+$ cells and $PD1^+CD8^+$ cells at a distance from each other lower than double the maximum cell radius (24 pixels = 12 µm) were considered as proximal. All other cells were classified as distal.

CODEX data analysis. A published dataset of colorectal CODEX images[9] was downloaded from The Cancer Imaging Archive (https://doi.org/10.7937/tcia.2020.fqn0-0326). It consisted of processed CODEX data from 35 colorectal cancer samples divided in two groups (CLR and DII) according to the peritumoral inflammatory levels and the presence of tertiary lymphoid structures[9]. For each sample, four.tiff images were available representing four 0.6-mm spots from two 70-core tissue microarrays. These images were hyperstacks of 58 channels including 56 antibodies (Supplementary Data 2) and two DNA markers with a resolution of 377 nm/pixel. After the manual review of all 140 spots, one representative image per sample was selected, having the best focus and containing both tumour and peritumoural immune infiltrates.

The single-channel tiff files for each selected image were extracted and the pixel intensities were rescaled from 0 to 1 with a custom R script. Using SIMPLI, single-cell segmentation was performed in each of the 35 images by applying a global threshold to the HOECHST channel to identify the nuclei and retain all objects with a diameter between 5 and 40 pixels. Each nucleus was then expanded by 5 pixels in all directions to define the cell area.

Resulting single cells were assigned to ten phenotypes according to the mean cell expression of CDX2 >0.15 or MUC1 >0.15 or cytokeratin >0.15 for tumour cells; CD34 >0.15 or CD31 >0.15 for endothelial cells; vimentin >0.1 for other stromal cells; CD11c >0.3 for dendritic cells; CD38 >0.26 for B cells; CD4 >0.13 and

CD3 >0.1 for CD4$^+$ T cells; CD4 >0.12 and FOXP3 >0.5 and CD3 >0.1 for Tregs; CD8 >0.16 and CD3 >0.1 for CD8$^+$ T cells, and CD68 >0.11 for macrophages. The heterotypic spatial analysis was performed by calculating the minimum distances of macrophages, CD8$^+$ T cells, CD4$^+$ T cells, Treg cells, and B cells to tumour cells and endothelial cells using the coordinates of the cell centroids. Only comparisons where the difference of the median cell–cell distances between the two histological subtypes was greater than 8 µm, corresponding to the diameter of B and T cells[30], were retained, no samples or cells were excluded from the analysis. As further support, a permutation test for each of the retained comparisons was run by re-assigning cell identities randomly in each sample 10,000 times. The resulting expected random distributions were compared to the observed values using a two-tailed permutation test and corrected for multiple testing.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The imaging mass cytometry data of human colon mucosa generated in this study have been deposited in the Zenodo database under accession code "5545882"[43]. The imaging mass cytometry data of the human appendix generated in this study have been deposited in the Zenodo database under accession code "5545760"[44]. The multiplex immunofluorescence data of human colorectal cancer generated in this study have been deposited in the Zenodo database under accession code "5545864"[45]. All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding author upon reasonable request. Source Data are provided with this paper.

Code availability

SIMPLI's code, documentation and an example dataset are available at "SIMPLI [https://github.com/ciccalab/SIMPLI]"[46]. The software code is protected by copyright. No permission is required from the rights-holder for non-commercial research uses. Commercial use will require a license from the rights-holder. For further information contact translation@crick.ac.uk who will reply within 5 business days.

Received: 7 April 2021; Accepted: 25 January 2022;
Published online: 09 February 2022

References

1. Parra, E. R., Francisco-Cruz, A. & Wistuba II State-of-the-art of profiling immune contexture in the era of multiplexed staining and digital analysis to study paraffin tumor tissues. *Cancers* **11**, 247 (2019).
2. Giesen, C. et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11**, 417–422 (2014).
3. Angelo, M. et al. Multiplexed ion beam imaging of human breast tumors. *Nat. Med.* **20**, 436–442 (2014).
4. Goltsev, Y. et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* **174**, 968–981.e915 (2018).
5. Lin, J. R., Fallahi-Sichani, M., Chen, J. Y. & Sorger, P. K. Cyclic immunofluorescence (CycIF), a highly multiplexed method for single-cell imaging. *Curr. Protoc. Chem. Biol.* **8**, 251–264 (2016).
6. Bauman, T. M. et al Quantitation of protein expression and co-localization using multiplexed immuno-histochemical staining and multispectral imaging. *J. Vis. Exp.* e53837 (2016).
7. Morrison, L. E. et al. Brightfield multiplex immunohistochemistry with multispectral imaging. *Lab. Investig.* **100**, 1124–1136 (2020).
8. Jackson, H. W. et al. The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).
9. Schürch, C. M. et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell* **182**, 1341–1359.e1319 (2020).
10. Berg, S. et al. ilastik: interactive machine learning for (bio)image analysis. *Nat. Methods* **16**, 1226–1232 (2019).
11. McQuin, C. et al. CellProfiler 3.0: next-generation image processing for biology. *PLoS Biol.* **16**, e2005970 (2018).
12. van Maldegem, F. et al. Characterisation of tumour microenvironment remodelling following oncogene inhibition in preclinical studies with imaging mass cytometry. *Nat. Commun.* **12**, 5906 (2021).
13. Zanotelli, V. R. T. & Bodenmiller, B. ImcSegmentationPipeline: A pixelclassification based multiplexed image segmentation pipeline. https://doi.org/10.5281/zenodo.3841961 (2017).
14. Jones, T. R. et al. CellProfiler Analyst: data exploration and analysis software for complex image-based screens. *BMC Bioinforma.* **9**, 482 (2008).
15. Eling, N., Damond, N., Hoch, T. & Bodenmiller, B. cytomapper: an R/Bioconductor package for visualization of highly multiplexed imaging data. *Bioinformatics* **36**, 5706–5708 (2020).
16. Opzoomer, J. W. et al. ImmunoCluster provides a computational framework for the nonspecialist to profile high-dimensional cytometry data. *eLife* **10**, e62915 (2021).
17. Stoltzfus, C. R. et al. CytoMAP: a spatial analysis toolbox reveals features of myeloid cell organization in lymphoid tissues. *Cell Rep.* **31**, 107523 (2020).
18. Somarakis, A., Unen, V. V., Koning, F., Lelieveldt, B. & Höllt, T. ImaCytE: visual exploration of cellular micro-environments for imaging mass cytometry data. *IEEE Trans. Vis. Computer Graph.* **27**, 98–110 (2021).
19. Yang, T. et al. SPIAT: an R package for the spatial image analysis of cells in tissues. Preprint at *bioRxiv* 2020.2005.2028.122614 (2020).
20. NeighbouRhood. https://github.com/BodenmillerGroup/neighbouRhood (2019).
21. Catena, R., Montuenga, L. M. & Bodenmiller, B. Ruthenium counterstaining for imaging mass cytometry. *J. Pathol.* **244**, 479–484 (2018).
22. Bankhead, P. et al. QuPath: open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
23. Schmidt, U., Weigert, M., Broaddus, C., Myers, G. *Cell Detection with Star-Convex Polygons*. (Springer International Publishing, 2018).
24. Pabst, O. & Slack, E. IgA and the intestinal microbiota: the importance of being specific. *Mucosal Immunol.* **13**, 12–21 (2020).
25. Dorn, I., Schlenke, P., Mascher, B., Stange, E. F. & Seyfarth, M. Lamina propria plasma cells in inflammatory bowel disease: intracellular detection of immunoglobulins using flow cytometry. *Immunobiology* **206**, 546–557 (2002).
26. Schapiro, D. et al. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nat. Methods* **14**, 873–876 (2017).
27. Song, W. & Craft, J. T follicular helper cell heterogeneity: time, space, and function. *Immunological Rev.* **288**, 85–96 (2019).
28. Kramer, A. S. et al. InForm software: a semi-automated research tool to identify presumptive human hepatic progenitor cells, and other histological features of pathological significance. *Sci. Rep.* **8**, 3418 (2018).
29. Zhang, L. et al. Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature* **564**, 268–272 (2018).
30. Strokotov, D. et al. Is there a difference between T- and B-lymphocyte morphology? *J. Biomed. Opt.* **14**, 064036 (2009).
31. Bortolomeazzi, M. et al. Immunogenomics of colorectal cancer response to checkpoint blockade: analysis of the KEYNOTE 177 trial and validation cohorts. *Gastroenterology* **161**, 1179–1193 (2021).
32. imctools. https://github.com/BodenmillerGroup/imctools (2017).
33. Pau, G., Fuchs, F., Sklyar, O., Boutros, M. & Huber, W. EBImage—an R package for image processing with applications to cellular phenotypes. *Bioinformatics* **26**, 979–981 (2010).
34. Van der Walt, S. et al. scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).
35. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
36. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.* **2008**, P10008 (2008).
37. Leland, M., John, H., Nathaniel, S. & Lukas, G. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
38. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *KDD-96 Proceedings*. (1996).
39. Henning, C. fpc. https://cran.r-project.org/package=fpc (2020).
40. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
41. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: scientific containers for mobility of compute. *PLoS One* **12**, e0177459 (2017).
42. Sochat, V. V., Prybol, C. J. & Kurtzer, G. M. Enhancing reproducibility in scientific computing: Metrics and registry for Singularity containers. *PLoS One* **12**, e0188511 (2017).
43. Bortolomeazzi, M. et al. Imaging Mass Cytometry of human normal colon mucosa (CLN1-6) from: A SIMPLI (Single-cell Identification from MultiPLexed Images) approach for spatially resolved tissue phenotyping at single-cell resolution. https://doi.org/10.5281/zenodo.5545882 (2021).
44. Bortolomeazzi, M. et al. Imaging Mass Cytometry Images (APP1) from: A SIMPLI (Single-cell Identification from MultiPLexed Images) approach for spatially resolved tissue phenotyping at single-cell resolution. https://doi.org/10.5281/zenodo.5545760 (2021).
45. Bortolomeazzi, M. et al. Vectra Polatis image of human colorectal cancer (CRC1) from: A SIMPLI (Single-cell Identification from MultiPLexed Images) approach for spatially resolved tissue phenotyping at single-cell resolution. https://doi.org/10.5281/zenodo.5545864 (2021).

46. Bortolomeazzi, M. et al. A SIMPLI (Single-cell Identification from MultiPLexed Images) approach for spatially-resolved tissue phenotyping at single-cell resolution. *ciccalab/SIMPLI* https://doi.org/10.5281/zenodo.5807230 (2021).
47. Dries, R. et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22**, 78 (2021).

Acknowledgements

We thank Sharavan Vishaan Venkateswaran for testing SIMPLI. This work was supported by Cancer Research UK (C43634/A25487, F.D.C.), the Cancer Research UK King's Health Partners Centre at King's College London (C604/A25135, F.D.C.), the Cancer Research UK City of London Centre (C7893/A26233, F.D.C.), innovation programme under the Marie Skłodowska-Curie (grant agreement No CONTRA-766030, F.D.C.) and the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001002, F.D.C.), the UK Medical Research Council (FC001002, F.D.C.), and the Wellcome Trust (FC001002, F.D.C.) and Crohn's and Colitis UK (M2019/3, J.S.). For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Author contributions

F.D.C. conceived and directed the study with support of J.S. M.B. developed the software with the help of D.T. and M.R.K. L.M. and A.A.-S performed the experiments. M.B, L.M, A.A.-S., M.J.P., J.S. and F.D.C. analysed the data. M.R.-J, G.B. and L.L. identified the samples and provided clinical assessments. M.R.-J. performed pathological assessments. M.B. and F.D.C. wrote the manuscript with contribution from A.A.-S., L.M. and M.J.P. All authors approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-28470-x.

Correspondence and requests for materials should be addressed to Francesca D. Ciccarelli.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

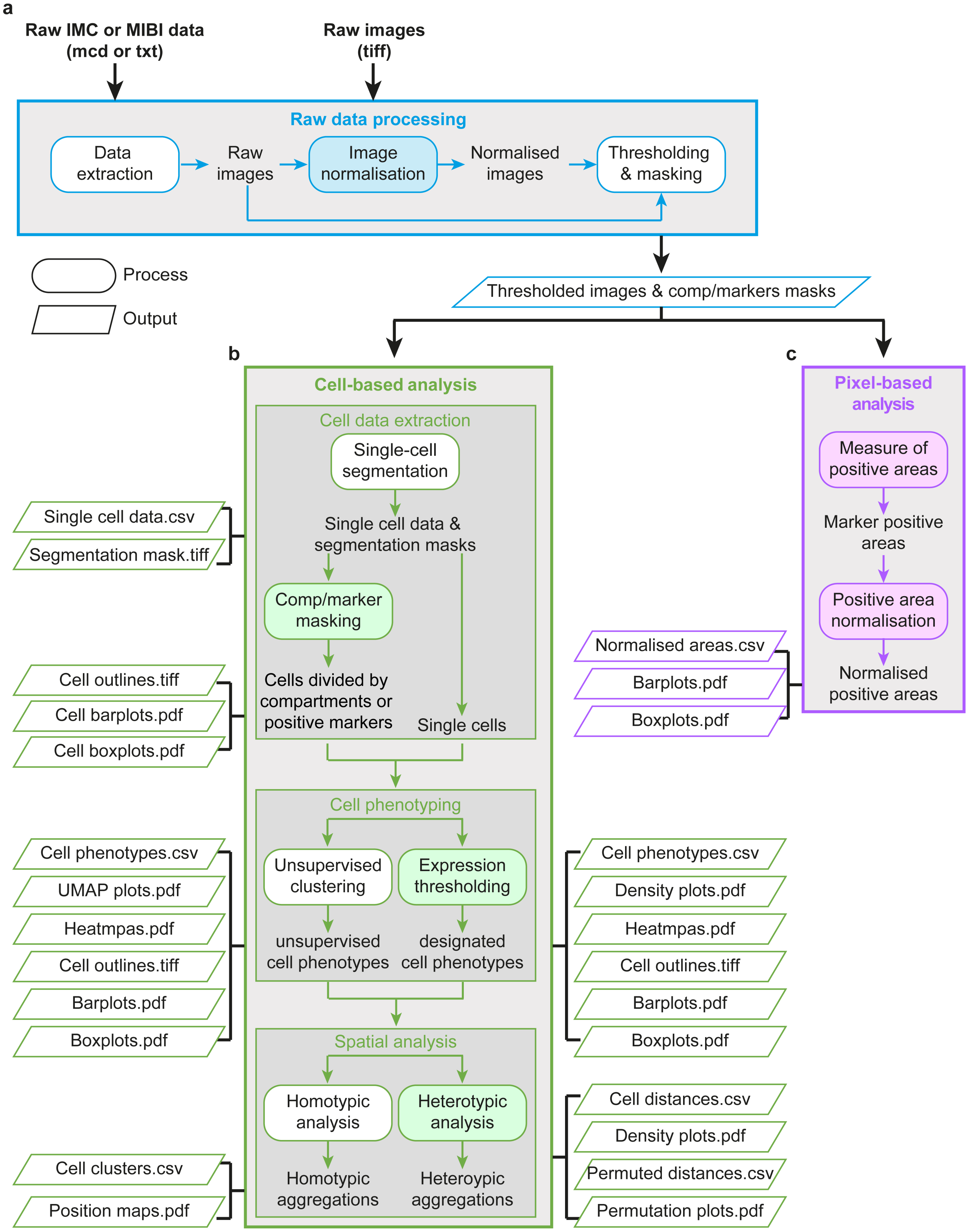
Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Figures

A SIMPLI (Single-cell Identification from MultiPLexed Images) approach for spatially resolved tissue phenotyping at single-cell resolution.

Supplementary Figure 1. SIMPLI workflow diagram



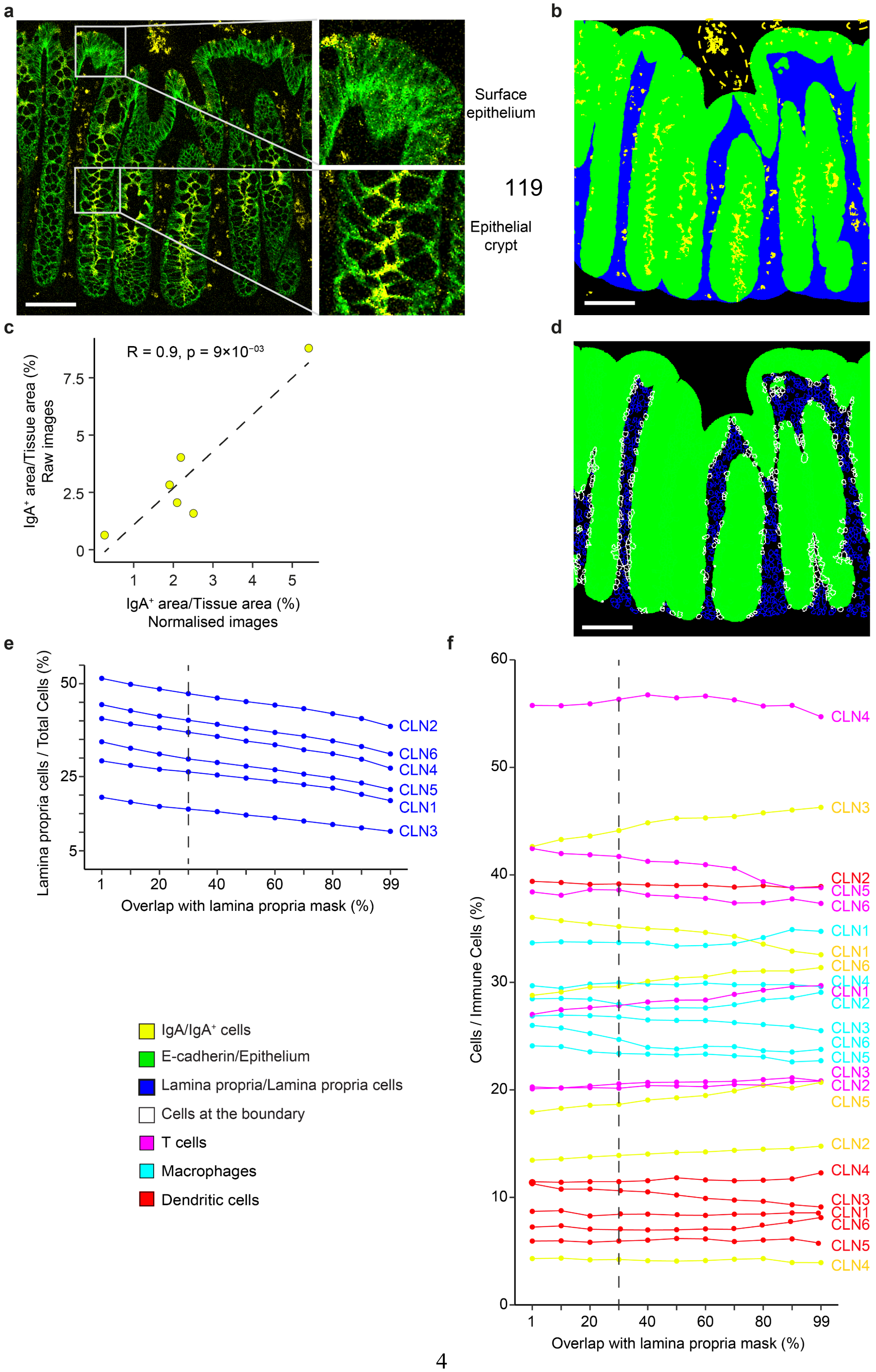
SIMPLI's workflow is divided into three main steps: raw data processing (**a**), cell-based analysis (**b**) and pixel-based analysis (**c**). Each step is divided in multiple stand-alone processes (rectangles), which rely on established tools and libraries (white) or newly developed codes (blue, green and pink), and produces multiple outputs (parallelograms).

a. Raw data processing. Raw data from IMC or MIBI (.mcd or .txt) are extracted using imctools[1]. Resulting images or original .tiff images from other imaging platforms are normalised using custom scripts and thresholded with a containerised headless instance of CellProfiler[2] to produce tissue compartments or marker masks as well as images for the following steps.

b. Cell-based analysis. This step is divided into cell data extraction, cell phenotyping and spatial analysis. Single cells are identified through single-cell segmentation using CellProfiler[2] or StarDist[3] with default or user-provided trained models. Cells belonging to tissue compartments or positive for certain markers can be identified based on their overlap with the tissue compartments or marker masks derived in the previous step. Subsequently, cell phenotypes are refined using unsupervised clustering with Seurat[4] or applying expression thresholds to one or more markers using *ad hoc* scripts. Finally, the spatial distribution of homotypic cell aggregations is performed with DBSCAN[5], while heterotypic cell aggregations are investigated using custom scripts. Additionally, a permutation test can be performed to assess whether the observed distance distributions differ from random distributions.

c. Pixel-based analysis. Areas positive for a specific marker or combination of markers are measured from the thresholded images and normalised over the area of the whole image or tissue compartments. These normalised values can then be compared across datasets. All processes in this step are performed using *ad hoc* scripts integrated in the pipeline.

Supplementary Figure 2. Pixel analysis and cell masking of human colon mucosa



a. IMC image of normal colon mucosa in CLN6 after data extraction and normalisation. Zoom-ins illustrate examples of surface epithelium and epithelial crypts. IgA$^+$ pixels are concentrated in the epithelial crypts where most of IgA transcytosis takes place.

b. Distribution of IgA$^+$ pixels in CLN6. Epithelium and lamina propria masks were generated as described in the Methods and superimposed to the mask of the IgA channel. Only IgA$^+$ pixels within the two compartments were retained for the pixel analysis, thus excluding likely artefacts (dotted circles). Scale bar in (a) and (b) = 100μm.

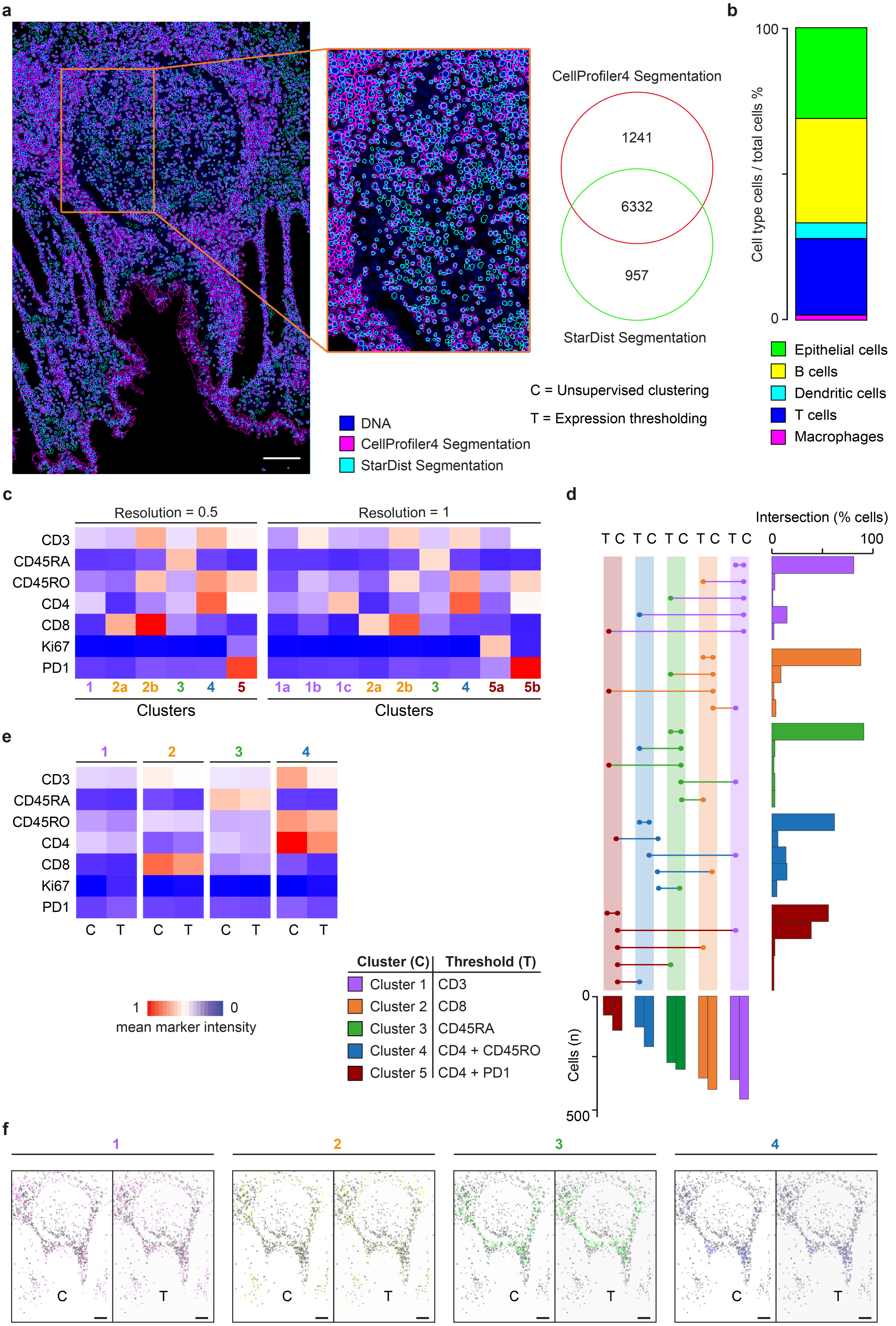
c. Correlation between IgA$^+$ areas measured from raw and normalised images across n = 6 biologically independent samples. Pearson correlation coefficient R and associated p-value based on Fisher's Z transform are shown.

d. Cells at the boundary between epithelium and lamina propria in CLN6. These were defined as cells with a partial overlap with both masks and their assignment to either compartment depends on the overlap threshold.

Parallel plots of all cells (**e**) and only immune cells (**f**) resident in the lamina propria at various thresholds of overlap (1% to 99% of the total cell area) across n = 6 biologically independent samples. Dotted lines represent the value chosen for the downstream analysis in Figure 2e,f (30%).

Images in panels (**a**), (**b**), (**d**) were derived from a representative sample (CLN6, Supplementary Data 1). IgA and IgA$^+$ cells, yellow; E-cadherin and epithelial cells, green; Lamina propria and lamina propria cells, blue; cells at the boundary, white; T cells, magenta; macrophages, cyan; Dendritic cells, red.

Supplementary Figure 3. Comparison of T cell phenotypes in human appendix



Cluster (C)	Threshold (T)
Cluster 1	CD3
Cluster 2	CD8
Cluster 3	CD45RA
Cluster 4	CD4 + CD45RO
Cluster 5	CD4 + PD1

a. Comparison of single-cell segmentations of APP1 (Supplementary Data 1) obtained with CellProfiler4[2] (magenta) and StarDist[3] (cyan) superimposed over the normalised DNA masks (blue). The two segmentations were performed as described in the Methods leading to the majority of cells identified by both approaches.

b. Proportions of T cells (blue), B cells (yellow), macrophages (magenta), dendritic cells (cyan) and epithelial cells (green) over all cells from non-normalised images. The expression values of each marker were normalised as a z-score value. Cell types were identified by K-means clustering (k = 6 on CD3, CD68, CD11c, Pan-Keratin and E-Cadherin).

c. Expression profiles of T cell subpopulations identified using unsupervised clustering with resolution of 0.5 and 1.0. Clusters are numbered as in Figure 3e showing how increasing the resolution splits bigger clusters obtained at lower resolution.

d. Percentage of cells shared between clustering-derived (C) and thresholding-derived (T) phenotypes. Number of cells identified by the two classification methods in each population are also reported in the lateral bars.

e. Comparison of the expression profiles of T cell subpopulations identified using unsupervised clustering (C) at 0.25 resolution and expression thresholding (T) of representative markers. For each population in (**c**) and (**e**), the mean intensity value of the markers across all cells is shown. The colour scale was normalised across all markers and cells, independently for each analysis. A total of n = 1,466 T cells from n = 1 biological sample were analysed.

f. Position map of T cells in APP1 colour-coded according to the phenotype obtained through unsupervised clustering or expression thresholding (cluster 1 = violet, cluster 2 = orange, cluster 3 = green, cluster 4 = blue, cluster 5 = red). Scale bar = 100μm.

Images in panels (**a**) and (**f**) were derived from a single sample (APP1, Supplementary Data 1).

References

1. imctools. *https://githubcom/BodenmillerGroup/imctools*, (2017).

2. McQuin C*, et al.* CellProfiler 3.0: Next-generation image processing for biology. *PLOS Biology* **16**, e2005970 (2018).

3. Schmidt U, Weigert M, Broaddus C, Myers G. Cell Detection with Star-Convex Polygons. (ed^(eds). Springer International Publishing (2018).

4. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411-420 (2018).

5. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Kdd* (ed^(eds) (1996).

Chapter 4. Analysis of response to immune checkpoint inhibitors in colorectal cancer

4.1 Contributions

In this study[107], I analysed the image analysis method and analysed the IMC[71] data, with support from Mohamed Reda Keddar for the density analysis of CD74$^+$ macrophages. I also supported Lucia Montorsi in the quantification of T cell infiltration from immunohistochemistry derived images. Francesca D. Ciccarelli, Mohamed Reda Keddar, Lucia Montorsi, Damjan Temelkowski, Subin Choi, Amelia Acha-Sagredo, Lorena Benedetti, Manuel Rodriguez-Justo, Kai-Keen Shiu and Jo Spencer and I wrote the manuscript, and all authors read and approved its final version.

Francesca D. Ciccarelli acquired the funding and conceived and directed the study with support from Jo Spencer. Victoria Kunene, Elisa Fontana., Hendrick-Tobias Arkenau and Kai-Keen Shiu selected the patients and provided clinical assessments, while Manuel Rodriguez-Justo performed the pathological assessments. Lucia Montorsi performed all immunostaining and analysed the resulting immunohistochemistry images. Robert Goldstone, Sophia Ward, Gareth A. Wilson, Maise Al Bakir and Charles Swanton provided the protocol for FFPE WES. Lorena Benedetti and Amelia Acha-Sagredo macro-dissected the regions from the FFPE blocks and prepared the collected tissue for the WES, RNA-seq and TCR-seq. Subin Choi and Damjan Temelkowski analysed the WES data.

Chapter 4. Analysis of response to immune checkpoint inhibitors in colorectal cancer

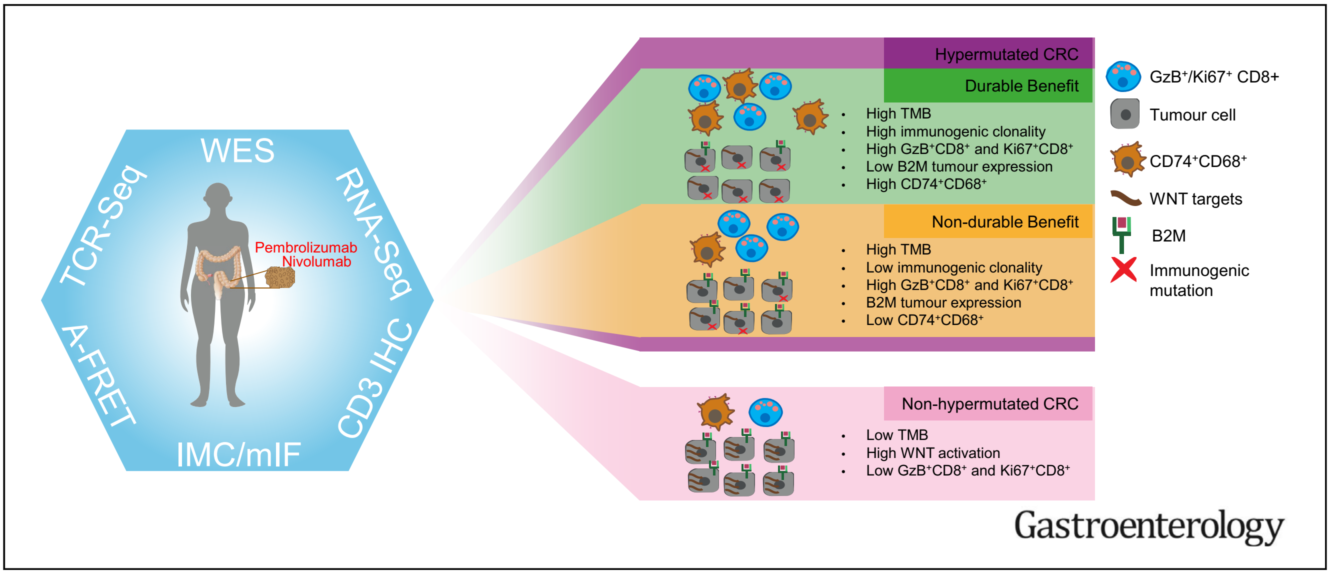
Mohamed Reda Keddar analysed the RNA-seq and TCR-seq data with support from Susan John. Lucia Montorsi performed the staining for IMC with support from Nedyalko Petrov and Katrina Todd, who also performed the acquisition. James Miles, Banafshe Larijani and Peter Parker performed the A-FRET experiments and contributed their interpretation. Amelia Acha-Sagredo performed the mIF experiments with support from Patty Way, Jonny Kohl, Tamara Denner and Emma Nye.

4.2 Immunogenomics of Colorectal Cancer Response to Checkpoint Blockade: Analysis of the KEYNOTE 177 Trial and Validation Cohorts

Immunogenomics of Colorectal Cancer Response to Checkpoint Blockade: Analysis of the KEYNOTE 177 Trial and Validation Cohorts

Michele Bortolomeazzi,[1,2,*] **Mohamed Reda Keddar,**[1,2,*] **Lucia Montorsi,**[1,2,*]
Amelia Acha-Sagredo,[1,2] Lorena Benedetti,[1,2] Damjan Temelkovski,[1,2] Subin Choi,[1,2]
Nedyalko Petrov,[3] Katrina Todd,[3] Patty Wai,[4] Johannes Kohl,[4] Tamara Denner,[5] Emma Nye,[5]
Robert Goldstone,[6] Sophia Ward,[6] Gareth A. Wilson,[7,8] Maise Al Bakir,[7,8] Charles Swanton,[7,8]
Susan John,[9] James Miles,[10] Banafshe Larijani,[10,11,12] Victoria Kunene,[13] Elisa Fontana,[14]
Hendrik-Tobias Arkenau,[14,15] Peter J. Parker,[2,16] Manuel Rodriguez-Justo,[17] Kai-Keen Shiu,[18]
Jo Spencer,[9] and Francesca D. Ciccarelli[1,2]

[1]Cancer Systems Biology Laboratory, The Francis Crick Institute, London, United Kingdom; [2]School of Cancer and Pharmaceutical Sciences, King's College London, London, United Kingdom; [3]Biomedical Research Centre, Guy's and St. Thomas' National Health Service Trust, London, United Kingdom; [4]State-Dependent Neural Processing Laboratory, The Francis Crick Institute, London, United Kingdom; [5]Experimental Histopathology, The Francis Crick Institute, London, United Kingdom; [6]Advanced Sequencing Facility, The Francis Crick Institute, London, United Kingdom; [7]Cancer Evolution and Genome Instability Laboratory, The Francis Crick Institute, London, United Kingdom; [8]Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, United Kingdom; [9]School of Immunology and Microbial Sciences, King's College London, London, United Kingdom; [10]FASTBASE Solutions S.L, Derio, Spain; [11]Cell Biophysics Laboratory, Ikerbasque, Basque Foundation for Science, Research Centre for Experimental Marine Biology and Biotechnology & Biophysics Institute, University of the Basque Country, Leioa, Bizkaia, Spain; [12]Centre for Therapeutic Innovation, Cell Biophysics Laboratory, Department of Pharmacy and Pharmacology & Department of Physics, University of Bath, Bath, United Kingdom; [13]Medical Oncology, University Hospitals Birmingham National Health Service Foundation Trust, Birmingham, United Kingdom; [14]Drug Development Unit, Sarah Cannon Research Institute UK, London, United Kingdom; [15]Department of Oncology, University College Hospital, London, United Kingdom; [16]Protein Phosphorylation Laboratory, The Francis Crick Institute, London, United Kingdom; [17]Department of Histopathology, University College London Cancer Institute, London, United Kingdom; [18]Department of Gastrointestinal Oncology, University College London Hospital National Health Service Foundation Trust, London, United Kingdom



BACKGROUND & AIMS: Colorectal cancer (CRC) shows variable response to immune checkpoint blockade, which can only partially be explained by high tumor mutational burden (TMB). We conducted an integrated study of the cancer tissue and associated tumor microenvironment (TME) from patients treated with pembrolizumab (KEYNOTE 177 clinical trial) or nivolumab to dissect the cellular and molecular determinants of response to anti- programmed cell death 1 (PD1) immunotherapy. **METHODS:** We selected multiple regions per tumor showing variable T-cell infiltration for a total of 738 regions from 29 patients, divided into discovery and validation cohorts. We performed multiregional whole-exome and RNA sequencing of the tumor cells and integrated these with T-cell receptor sequencing, high-dimensional imaging mass cytometry, detection of programmed death-ligand 1 (PDL1) interaction in situ, multiplexed immunofluorescence, and computational spatial analysis of the TME. **RESULTS:** In hypermutated CRCs, response to anti-PD1 immunotherapy was not associated with TMB but with high clonality of immunogenic mutations, clonally expanded T cells, low activation of Wnt signaling, deregulation of the interferon gamma pathway, and active immune escape mechanisms. Responsive hypermutated CRCs were also rich in cytotoxic and proliferating $PD1^{+}CD8$ T cells interacting with $PDL1^{+}$ antigen-presenting macrophages. **CONCLUSIONS:** Our study clarified the limits of TMB as a predictor of response of CRC to anti-PD1 immunotherapy. It identified a population of antigen-presenting macrophages interacting with CD8 T cells that consistently segregate with response. We therefore concluded that anti-PD1 agents release the PD1-PDL1 interaction between CD8 T cells and macrophages to promote cytotoxic antitumor activity.

Keywords: Anti-PD1 Immunotherapy; Tumor Mutational Burden; Wnt Signaling; Interferon Gamma; CD8 T cells.

CLINICAL AT

WHAT YOU NEED TO KNOW

BACKGROUND AND CONTEXT

Response of colorectal cancer to immune checkpoint blockade is highly variable, and molecular and cellular determinants of response remain poorly understood.

NEW FINDINGS

Tumor mutational burden is insufficient to predict response in colorectal cancer. Additional predictors are clonal immunogenic mutations, clonally expanded T cells, low Wnt activation, active immune escape, and high CD8 T cells and antigen-presenting macrophage infiltration.

LIMITATIONS

Due to the restricted use of anti-programmed cell death 1 immunotherapy in hypermutated colorectal cancers, our study has a limited patient cohort size. Additional data from prospective studies are needed.

IMPACT

Colorectal cancer stratification based on tumor mutational burden is limited and may be improved by accounting for other predictors, including the abundance of antigen-presenting macrophages in proximity to CD8 T cells.

Anticancer therapy based on immune checkpoint blockade has driven a paradigm shift in the treatment of several cancer types.[1] Pembrolizumab and nivolumab, 2 antibodies targeting programmed cell death 1 (PD1) expressed on T cells, have shown efficacy in advanced hypermutated colorectal cancers (CRCs).[2] Response is thought to depend on rich immune infiltration and high tumor mutation burden (TMB) leading to increased production of peptide neoantigens.[3] However, despite pervasive tumor immunogenicity, response is highly variable, and approximately half of patients with hypermutated CRCs show no benefit from treatment.[4]

We have dissected the extent to which TMB, cancer dysfunctional genes and pathways, as well as the qualitative and quantitative immune composition of the tumor microenvironment (TME) influence response to immune checkpoint blockade. To reproduce the most common clinical scenario where metastatic biopsies are not routinely taken, we performed a high-dimensional and multiregional profile of primary CRCs or local relapses from 29 patients, divided into a discovery and a validation cohort. The discovery cohort was composed of patients with metastatic disease treated with pembrolizumab as first-line therapy within the KEYNOTE 177 phase III clinical trial[5] or nivolumab. Most patients did not receive previous treatment, which offered the ideal opportunity to identify critical factors for response to treatment in cancer genetic and transcriptional dysregulation and immune microenvironment composition. We then extended the study to a more heterogenous validation cohort of patients who received anti-PD1 agents alone or in combination and as first-line therapy or in a chemorefractory setting to assess the general validity of our findings.

Methods

Patient Populations

Formalin-fixed paraffin-embedded blocks were obtained from surgical resections of the primary tumor or local relapse of 16 patients (UH1–UH16, discovery cohort) and 13 patients (UH17–UH29, validation cohort). UH1 through UH10 were treated with pembrolizumab as part of the KEYNOTE 177 clinical trial (NCT02563002)[5] and UH11 through UH16 were treated with nivolumab as first-line therapy. UH17 through UH19 were part of the KEYNOTE 177 trial, UH26 received pembrolizumab, UH20 through UH25 and UH29 were treated with nivolumab, UH27 received ipilimumab in combination with nivolumab and then nivolumab alone, and UH28 received nivolumab and then ipilimumab in combination with nivolumab.

* Authors share co-first authorship.

Abbreviations used in this paper: A-FRET, amplified Förster resonance energy transfer; B2M, beta-2-microglobulin; CD, cluster of differentiation; CRC, colorectal cancer; DB-CRC, durable benefit colorectal cancer; GzB, granzyme B; IMC, imaging mass cytometry; mIF, multiplexed immunofluorescence; nDB-CRC, no durable benefit colorectal cancer; PD1, programmed cell death 1; PDL1, programmed death-ligand 1; RNA-seq, RNA sequencing; SNV, single-nucleotide variants; T-cell, receptor β-chain sequencing (TCR-seq); TCGA, The Cancer Genome Atlas; TMB, tumor mutational burden; TME, tumor microenvironment; WES, whole-exome sequencing.

Most current article

0016-5085
https://doi.org/10.1053/j.gastro.2021.06.064

Response to therapy was assessed using Response Evaluation Criteria In Solid Tumors 1.1.[6] Patients were considered to achieve durable benefit if the disease did not progress for at least 12 months after receiving immunotherapy, and no durable benefit if the disease progressed within 12 months. Further details on treatment and other clinical parameters, including tumor staging and prior lines of treatment, are reported as Supplementary Methods and Supplementary Table 1.

CD3 and H&E Staining

Cluster of differentiation (CD) 3 immunostaining was performed on several slides across the depth of each analyzed tumor block, for a total of 418 regions. Slides were digitally acquired at 20× resolution and loaded into QuPath[7] to quantify the number of $CD3^+$ cells/mm^2. H&E staining was performed on 13 additional slides of the validation cohort.

Imaging Mass Cytometry

Imaging mass cytometry (IMC) was performed in 77 regions of the discovery and validation cohorts using 3 panels of 42 antibodies in total. IMC data analysis was done with SIMPLI.[8] Positive areas for combinations of markers were quantified and normalized over the tissue area or the area of selected immune populations. After segmentation, cell identities were assigned according to the highest overlap with marker-specific masks. Unsupervised cell clustering was performed with Seurat[9] and used to compare the relative abundances of cell subpopulations between tumor groups. High-density clusters of $CD68^+CD74^+$ cells were identified using DBSCAN.[10]

Multiplexed Immunofluorescence

Multiplexed Immunofluorescence (mIF) was performed in 24 whole slides of the discovery and validation cohorts using an automated Opal-based mIF staining protocol with 8 antibodies. Fluorescently labeled slides were scanned, and images were loaded into inForm (Akoya Biosciences) for spectral unmixing and autofluorescence isolation.

Whole-Exome Sequencing

Whole-exome sequencing (WES) was conducted on 32 macrodissected tumor regions and matched normal tissue of the discovery cohort. Sequencing data were aligned using BWA MEM.[11] Somatic single-nucleotide variants (SNVs) and indels were called using Strelka.[12] ANNOVAR[13] was used to annotate exonic or splicing SNVs and indels, and damaging SNVs and indels were identified as previously described.[14] Copy number analysis was done using ASCAT[15] and integrated with gene expression data. Amplified genes, deleted genes, heterozygously deleted genes with a damaging mutation in the other allele, and copy number neutral genes with at least 1 damaging mutation were considered as damaged genes. Immunogenic mutations were predicted using Polysolver[16] and NeoPredPipe,[17] and their clonality was assessed using PyClone.[18]

RNA Sequencing

We conducted 3′-RNA sequencing (RNA-seq) on 88 macrodissected regions of the discovery and validation cohorts. Raw reads were processed using the Lexogen QuantSeq 3′ messenger RNA-seq pipeline.[19] Differential gene expression was assessed using DESeq2.[20] Pathway enrichment analysis of differential expressed genes was done using MetaCore 20.3 build 70200 (Clarivate Analytics).

T-Cell Receptor β-Chain Sequencing

T-cell receptor β-chain sequencing (TCR-seq) was performed on 28 macrodissected regions of the discovery cohort. Genomic DNA was submitted to Adaptive Biotechnologies (Seattle, WA) for nonlymphoid tissue (survey level) TCR-seq.[21] Data were analyzed using the immunoSEQ Analyzer toolset.

PD1-PDL1 Amplified Förster Resonance Energy Transfer

In situ interaction between PD1 and programmed death-ligand 1 (PDL1) was measured in 58 regions of the discovery cohort via amplified Förster resonance energy transfer (A-FRET)[22] at FASTBASE Solutions (Derio, Spain). FRET efficiency was calculated from 793 optical fields of view to cover the whole surface of the regions analyzed. The results were expressed as the median fields of view values per region. Detailed protocols and methods are provided in the Supplementary Methods.

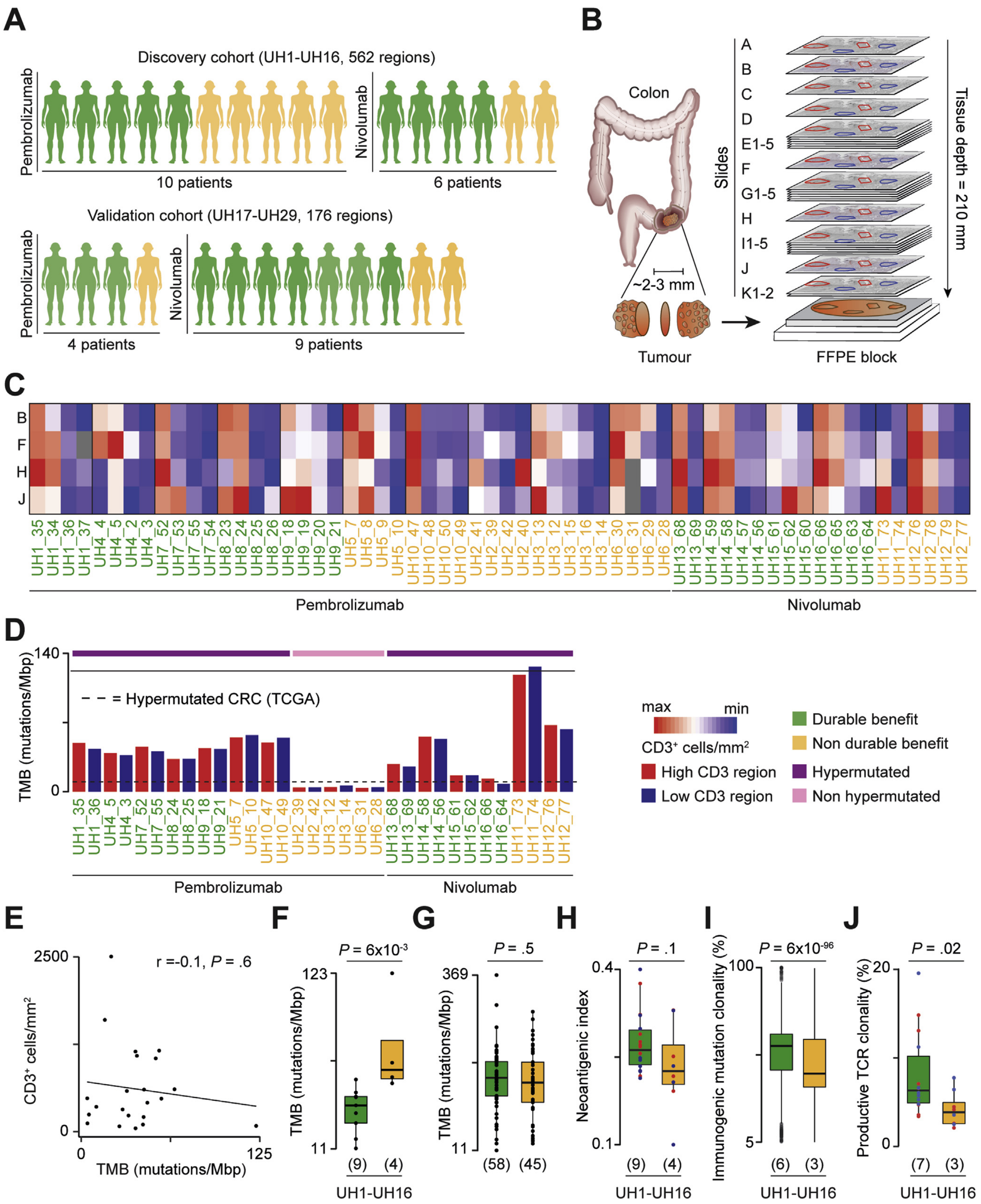
Results

Response of Hypermutated Colorectal Cancers Is Associated With Clonal Immunogenic Mutations and Clonally Expanded T Cells

To assess how immune infiltration correlates with tumor genetic and transcriptional alterations in CRC, we performed a multiomic and multiregional profile of 24 sequential slides (A–K) from formalin-fixed paraffin-embedded tumor blocks, for a total of 562 regions from 16 patients of the discovery cohort (Figure 1*A*). Ten of these patients received pembrolizumab (UH1–UH10) and 6 nivolumab (UH11–UH16) as a first-line treatment in advanced metastatic setting. According to Response Evaluation Criteria In Solid Tumors 1.1, 9 patients achieved durable benefit, and 7 had no durable benefit from the treatment (Supplementary Table 1). We validated the main findings of the study in 176 additional regions from 13 patients with CRC (UH17–UH29, Supplementary Figure 1*A*) treated with anti-PD1 agents alone or in combinations with other immune checkpoint inhibitors as first-line therapy or in a chemorefractory setting (Supplementary Table 1). Ten of them reached a durable benefit, and 3 had no durable benefit (Figure 1*A*).

Because T cells are the effector cells that mediate the response to anti-PD1 immunotherapy, we selected multiple regions per block with variable T-cell content in proximity to the tumor infiltrating margins (Supplementary Figure 1*B*). These regions were then projected in all sequential slides to perform additional CD3 immunohistochemistry for quantification of T-cell variability in the 3-dimensions of the tumor as well as IMC, mIF, WES, RNA-seq, TCR-seq, and A-FRET detection of the PD1-PDL1 interaction in situ (Figure 1*B*).

CLINICAL AT


A
Discovery cohort (UH1-UH16, 562 regions)
Pembrolizumab
10 patients
Nivolumab
6 patients
Validation cohort (UH17-UH29, 176 regions)
Pembrolizumab
4 patients
Nivolumab
9 patients
B
Colon
~2-3 mm
Tumour
Slides
A
B
C
D
E1-5
F
G1-5
H
I1-5
J
K1-2
Tissue depth = 210 mm
FFPE block
C
B
F
H
J
UH1_35 UH1_34 UH1_36 UH1_37 UH4_4 UH4_5 UH4_2 UH4_3 UH7_52 UH7_53 UH7_55 UH7_54 UH8_23 UH8_24 UH8_25 UH8_26 UH9_18 UH9_19 UH9_20 UH9_21 UH5_7 UH5_8 UH5_9 UH5_10 UH10_47 UH10_48 UH10_50 UH10_49 UH2_41 UH2_39 UH2_42 UH2_40 UH3_13 UH3_12 UH3_15 UH3_16 UH3_14 UH6_30 UH6_31 UH6_29 UH6_28 UH13_68 UH13_69 UH14_59 UH14_58 UH14_57 UH14_56 UH15_61 UH15_62 UH15_60 UH16_66 UH16_65 UH16_63 UH16_64 UH11_73 UH11_74 UH12_76 UH12_78 UH12_79 UH12_77
Pembrolizumab
Nivolumab
D
TMB (mutations/Mbp)
140
0
- - = Hypermutated CRC (TCGA)
UH1_35 UH1_36 UH4_5 UH4_3 UH7_52 UH7_55 UH8_24 UH8_25 UH9_18 UH9_21 UH5_7 UH5_10 UH10_47 UH10_49 UH2_39 UH2_42 UH3_12 UH3_14 UH6_31 UH6_28 UH13_68 UH13_69 UH14_58 UH14_56 UH15_61 UH15_62 UH16_66 UH16_64 UH11_73 UH11_74 UH12_76 UH12_77
Pembrolizumab
Nivolumab
max
min
CD3⁺ cells/mm²
🟥 High CD3 region
🟦 Low CD3 region
🟩 Durable benefit
🟨 Non durable benefit
🟪 Hypermutated
🟪 Non hypermutated
E
2500
0
CD3⁺ cells/mm²
r =-0.1, P = .6
0
125
TMB (mutations/Mbp)
F
P = 6x10⁻³
123
11
TMB (mutations/Mbp)
(9) (4)
UH1-UH16
G
P = .5
369
11
TMB (mutations/Mbp)
(58) (45)
H
P = .1
0.4
0.1
Neoantigenic index
(9) (4)
UH1-UH16
I
P = 6x10⁻⁹⁶
100
5
Immunogenic mutation clonality (%)
(6) (3)
UH1-UH16
J
P = .02
20
0
Productive TCR clonality (%)
(7) (3)
UH1-UH16

As a first analysis, we compared T-cell infiltration between and within tumors (Supplementary Figure 1*C*). In both the discovery (Figure 1*C*) and validation (Supplementary Figure 1*D*) cohorts, we observed widespread intertumor and intratumor heterogeneity of T-cell infiltration, with up to a 38-fold difference in CD3$^+$ cell densities between patients and up to a 20-fold difference between regions of the same patient (Supplementary Table 2). To investigate how heterogeneity in T-cell infiltration correlated with TMB, we performed multiregional WES in the discovery cohort by selecting 2 regions per patient, 1 with high and 1 with low T-cell infiltration (Supplementary Table 2). TMB was comparable between regions of the same patient (Figure 1*D*) and did not correlate with T-cell density across samples (Figure 1*E*). Similar lack of correlation was observed in hypermutated CRCs from The Cancer Genome Atlas (TCGA) (Supplementary Figure 1*E*), indicating TMB independence of T-cell heterogeneity.

WES also showed that the TMB in 3 patients (UH2, UH3, and UH6) was lower than 12 mutations/megabase pair (TCGA lower bound of CRC hypermutated phenotype[23]) despite negative MLH1 and PMS2 immunostaining and consistent with resistance to treatment (Supplementary Table 1). All patients of the validation cohort had hypermutated CRCs (Supplementary Table 1).

Given that approximately 50% of patients with hypermutated CRC do not respond to immunotherapy, we compared TMB between hypermutated CRCs with durable benefit (DB-CRCs) and those with no durable benefit (nDB-CRCs) to assess the role of TMB as a marker of response within hypermutated CRC. Surprisingly, in the discovery cohort, DB-CRCs had a significantly lower TMB than nDB-CRCs (Figure 1*F*). When adding hypermutated CRCs from the validation cohort and published studies,[24–28] we observed no significant difference between DB- and nDB-CRCs (Figure 1*G*). Together with the lack of response in non-hypermutated CRCs (Supplementary Table 1), these results indicate that a TMB below 12 mutations/megabase pair is a predictor of resistance to anti-PD1 immunotherapy in CRC. Above this threshold, TMB is not a predictor of response.

To understand whether the proportion of cancer-associated neoantigens differed between responders and nonresponders, we predicted how many cancer mutations were potentially immunogenic in each patient. In the discovery cohort, the ratio between immunogenic mutations and all mutations (neoantigenic index) was similar between DB- and nDB-CRCs (Figure 1*H*). However, we observed a high number of clonal immunogenic mutations in DB-CRCs (Figure 1*I*), indicating expansion of tumor cells with the same potential immune targets. Similar results were observed using an external data set of hypermutated CRCs treated with immune checkpoint inhibitors[25] (Supplementary Figure 1*F*). Consistent with dominant antigenic targets, the productive TCR repertoire was also more clonal in DB-CRCs (Figure 1*J*).

Therefore, although hypermutated CRCs responding to anti-PD1 agents do not have more mutations than those failing to respond, they have significantly more clonally expanded immunogenic mutations and T-cell clones.

Durable-Benefit Colorectal Cancers Show Widespread Immune Dysregulation and Silencing of the Beta-2-Microglobulin Gene

To dissect CRC molecular determinants of response to anti-PD1 agents in CRC, we compared genetic and transcriptional dysregulations between hypermutated and non-hypermutated CRCs as well as between DB- and nDB-CRCs.

Genes of the Wnt pathway were frequently damaged (Supplementary Figure 2*A*, Supplementary Table 3) and transcriptionally deregulated (Figure 2*A*, Supplementary

◀

Figure 1. Study design and quantification of tumor heterogeneity. (*A*) Description of the study cohorts. Clinical benefit from the treatment was assessed with Response Evaluation Criteria In Solid Tumors 1.1. (*B*) Experimental design: 24 sequential slides from formalin-fixed paraffin-embedded (FFPE) CRC blocks before treatment were used for multiregional CD3 immunohistochemistry (*slides A*, *B*, *F*, *H*, and *J*), IMC (*slide C*), mIF (*slide D*), WES (*slides E1–E5*), RNA-seq (*slides G1–G5*), TCR-seq (*slides I1–I5*), and A-FRET detection of PD1-PDL1 interaction in situ (*slides K1–K2*). Multiple regions with variable CD3 infiltration were identified in *slide A* and projected to all other slides. (*C*) Quantification of CD3$^+$ cells/mm^2 from immunohistochemistry staining in 60 regions of the discovery cohort using Qupath.[7] Values were normalized within each patient. The *gray boxes* indicate missing measures. (*D*) TMB of 32 sequenced regions in the discovery cohort. The *dotted line* corresponds to the TMB threshold of hypermutated CRC (12 mutations/megabase pairs).[23] (*E*) Correlation between CD3$^+$ cells/mm^2 from immunohistochemistry staining of *slide F* (discovery) and *slide E* (validation) and TMB across samples. Average CD3$^+$ cell density across multiple regions per slide is reported. For the discovery cohort, TMB was calculated as the average between the 2 sequenced regions. For the validation cohort, TMB was obtained from the FM1 test.[40] Pearson correlation coefficient *R* and associated *P* value are shown. (*F*) Comparison of TMB between DB- and nDB-CRCs of the discovery cohort and (*G*) in hypermutated CRCs from the validation cohort (Supplementary Table 1) and published studies.[24–28] For[26,28] response was unavailable and the overall survival from the start of immunotherapy was used to define DB (≥12 months) and nDB (<12 months). (*H*) Comparison of neoantigenic index (ratio of predicted immunogenic mutations over all nonsilent mutations) and (*I*) clonality of immunogenic mutations in 17 regions with >30% tumor purity (Supplementary Table 2). Regions with lower purity were excluded because of unreliable mutation clonality assessment.[41] Results hold true even when using all regions (data not shown). (*J*) Comparison of productive clonality of TCR beta rearrangements between DB- and nDB-CRCs with available data (Supplementary Table 2). The number of patients in each tumor group is reported in *brackets*. Distributions were compared using the 2-sided Wilcoxon's rank sum test. The *horizontal line* in the middle of each *box* indicates the median; the *top* and *bottom borders* of the box mark the 75th and 25th percentiles, respectively, and the *vertical lines* mark points within 1.5 the inter-quartile range.

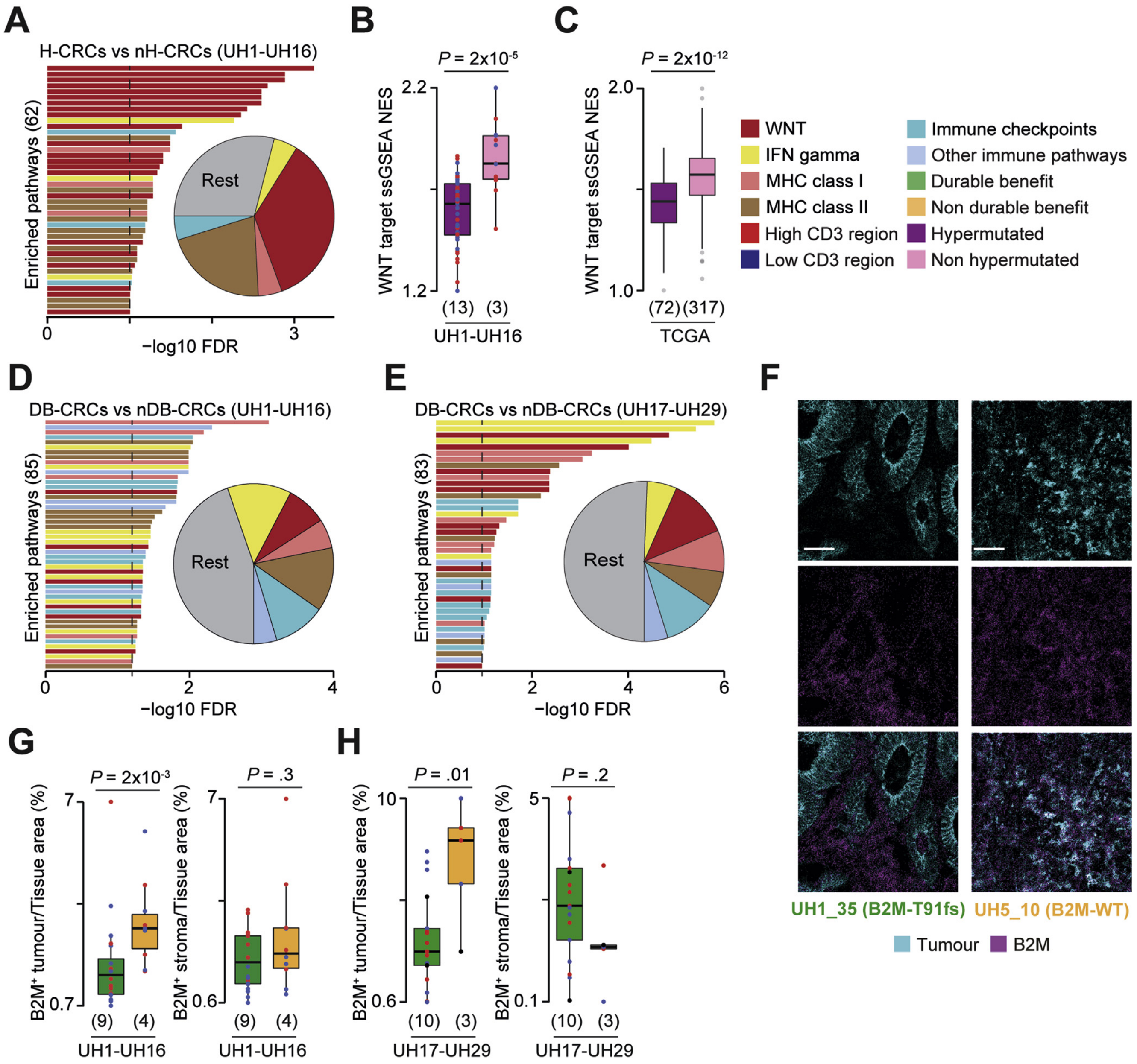


Figure 2. Cancer and immune aberrations across CRC groups. (*A*) Representative enriched pathways in differentially expressed genes between hypermutated and non-hypermutated CRCs of the discovery cohort. The false discovery rate (FDR) was calculated using Benjamini-Hochberg correction. Proportions of immune-related pathways over all enriched pathways are reported as pie chart. Normalized enrichment scores (NES) from single sample Gene Set Enrichment Analysis (ssGSEA)[42] of 68 transcriptional targets of the Wnt pathway[29,43] between hypermutated and non-hypermutated CRCs from (*B*) the discovery cohort and (*C*) TCGA. Representative pathways enriched in differentially expressed genes between DB- and nDB-CRCs from the (*D*) discovery and (*E*) validation cohorts. (*F*) Representative IMC images of CRCs with mutated and wild-type (WT) B2M protein. Scale bar = 50 μm. Comparison of normalized tumor and stroma B2M$^+$ areas between DB- and nDB-CRCs of the (G) discovery and (*H*) validation cohorts. Number of patients in each tumor group is reported in brackets. Distributions were compared using the 2-sided Wilcoxon's rank sum test. IFN, interferon; MHC, major histocompatibility complex. The *horizontal line* in the middle of each *box* indicates the median; the *top* and *bottom borders* of the box mark the 75th and 25th percentiles, respectively, and the *vertical lines* mark minimum and maximum of all the data.

Table 4) in hypermutated compared with non-hypermutated CRCs. Moreover, Wnt downstream targets were significantly downregulated in hypermutated compared with non-hypermutated CRCs (Figure 2*B*) as confirmed in TCGA (Figure 2*C*). Transcriptional Wnt activation is known to reduce T-cell infiltration,[29,30] suggesting a potential impact on the TME composition of these tumors.

Genes encoding members of the interferon gamma pathway, antigen presentation machinery, and other immune-related processes were damaged (Supplementary Figure 2*A*)

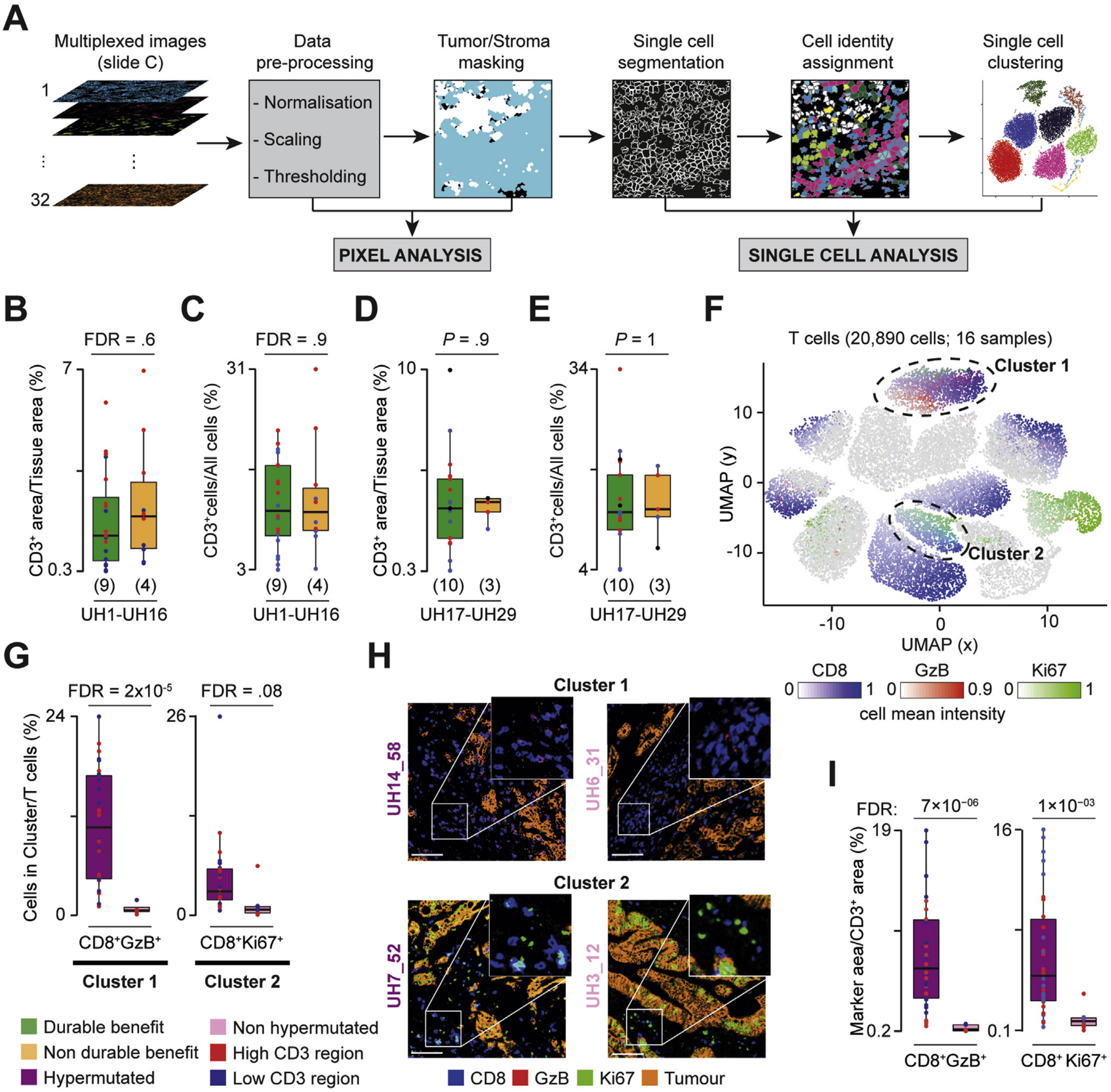


Figure 3. Comparison of T cells infiltrates between CRC groups. (*A*) IMC analysis workflow using SIMPLI.[8] For each region, images of the markers used (Supplementary Table 5) were preprocessed to extract pixel intensities. Masks for tumor and stroma were derived and used for the pixel analysis. Each region was segmented into single cells that were assigned to tumor or stroma, phenotypically identified through expression of representative markers, and used for single cell clustering. (*B*) Comparison of normalized CD3$^+$ areas and (*C*) CD3$^+$ cells between DB and nDB-CRCs in the discovery cohort. Benjamini-Hochberg false discovery rate (FDR) correction was applied for testing over 5 immune populations. (*D*) Comparison of normalized CD3$^+$ areas and (*E*) CD3$^+$ cells between DB- and nDB-CRCs in the validation cohort. (*F*) Uniform Manifold Approximation and Projection (UMAP) map of 20,890 T cells in 38 regions from 16 CRCs of the discovery cohort. Cells were grouped in 13 clusters based on the expression of 12 phenotypic markers using Seurat[9] (Supplementary Table 8) and colored according to the mean intensities of representative markers. The *circles* indicate the 2 clusters enriched in hypermutated CRCs. (*G*) Proportions of cluster 1 (CD8$^+$GzB$^+$ cells) and cluster 2 (CD8$^+$Ki67$^+$ cells) over the total T cells in hypermutated and non-hypermutated CRCs. Distributions were compared using the 2-sided Wilcoxon's rank sum test. Benjamini-Hochberg FDR correction was applied for testing over 13 clusters. (*H*) IMC-derived images of tumor-associated markers (E-cadherin and pan-keratin) and CD8 and GzB or CD8 and Ki67 in 2 representative samples. *Scale bar* = 100 μm. (*I*) Comparisons of normalized CD8$^+$GzB$^+$ and CD8$^+$Ki67$^+$ areas between hypermutated and non-hypermutated CRCs. Distributions were compared using the 2-sided Wilcoxon's rank sum test. Benjamini-Hochberg FDR correction was applied for testing over 25 combinations of T-cell markers (Supplementary Table 6). The *horizontal line* in the middle of each *box* indicates the median; the *top* and *bottom borders* of the box mark the 75th and 25th percentiles, respectively, and the *vertical lines* mark minimum and maximum of all the data.

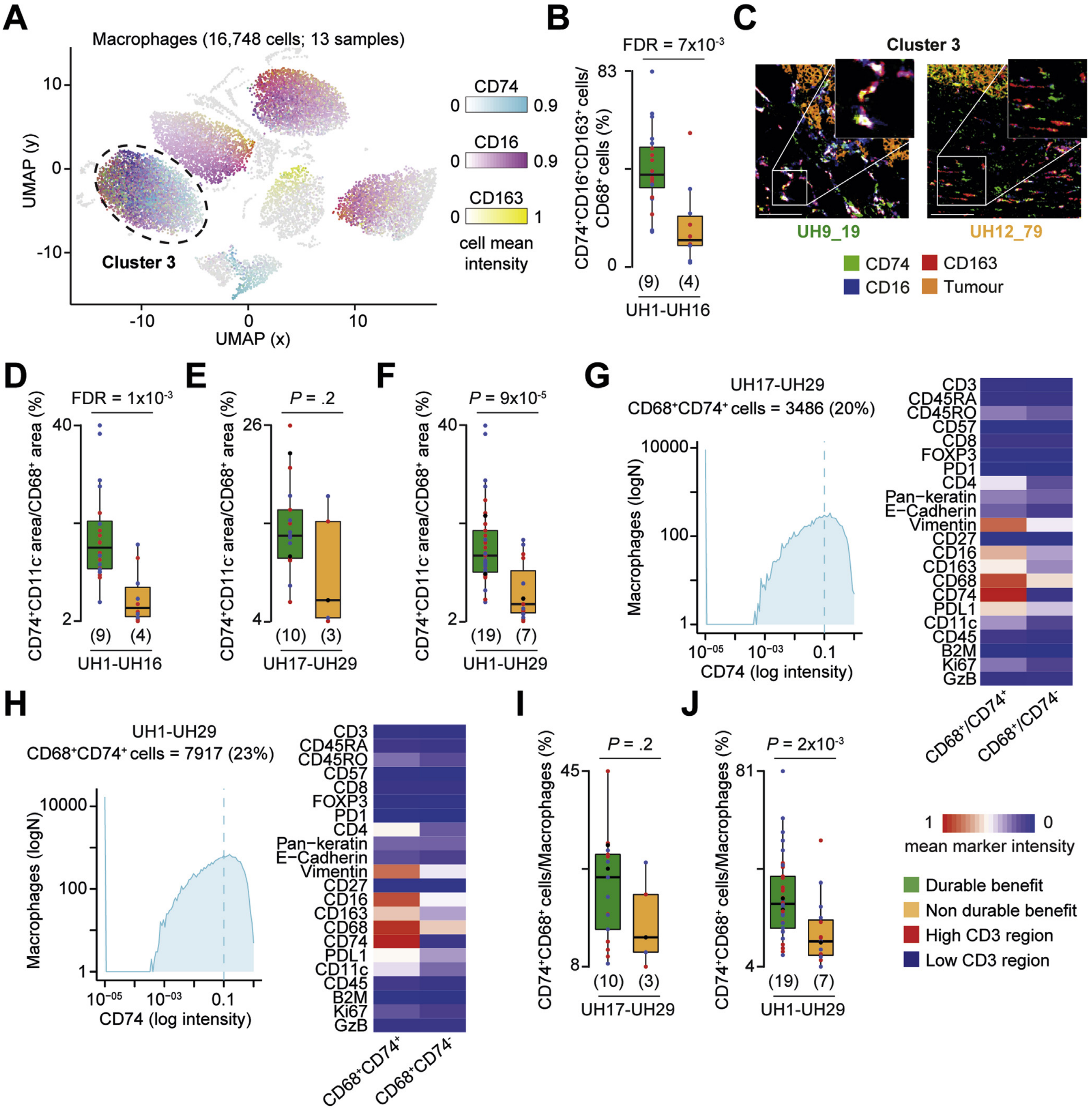


Figure 4. Difference in CD74$^+$ macrophages between DB- and nDB-CRCs. (*A*) Uniform Manifold Approximation and Projection (UMAP) map of 16,748 macrophages in 30 regions from 13 hypermutated CRCs in the discovery cohort. Cells were grouped in 9 clusters based on the expression of 11 phenotypic markers using Seurat[9] (Supplementary Table 8) and colored according to the mean intensities of representative markers. The *circle* indicates the cluster enriched in DB-CRCs. (*B*) Proportions of cluster 3 (CD68$^+$CD74$^+$ cells) over the total macrophages in DB- and nDB-CRCs. Distributions were compared using the 2-sided Wilcoxon's rank sum test. Benjamini-Hochberg false discovery rate (FDR) correction was applied for testing over 9 clusters. (*C*) IMC-derived images of CD74, CD16, and CD163 and tumor-associated markers (E-cadherin and pan-keratin) in 2 representative samples. *Scale bar* = 100 μm. (*D*) Comparisons of normalized CD74$^+$ area between DB- and nDB-CRCs in the discovery, (*E*) validation, and (*F*) both cohorts using the 2-sided Wilcoxon's rank sum test. For the discovery cohort, Benjamini-Hochberg FDR correction was applied for testing over 9 combinations of macrophage markers (Supplementary Table 6). (*G*) CD74$^+$ macrophages in the validation and (*H*) combined cohorts were identified by applying a threshold of 0.1 CD74 expression to all macrophages after IMC image histologic inspection. Mean marker intensities in CD74$^+$ and CD74$^-$ macrophages are reported and normalized across all markers and cells. (*I*) Comparison of normalized of CD74$^+$ macrophages between DB- and nDB-CRCs in the validation and (*J*) combined cohorts. Distributions were compared using the 2-sided Wilcoxon's rank sum test. The number of patients in each tumor group is reported in brackets. The *horizontal line* in the middle of each *box* indicates the median; the *top* and *bottom borders* of the box mark the 75th and 25th percentiles, respectively, and the *vertical lines* mark minimum and maximum of all the data.

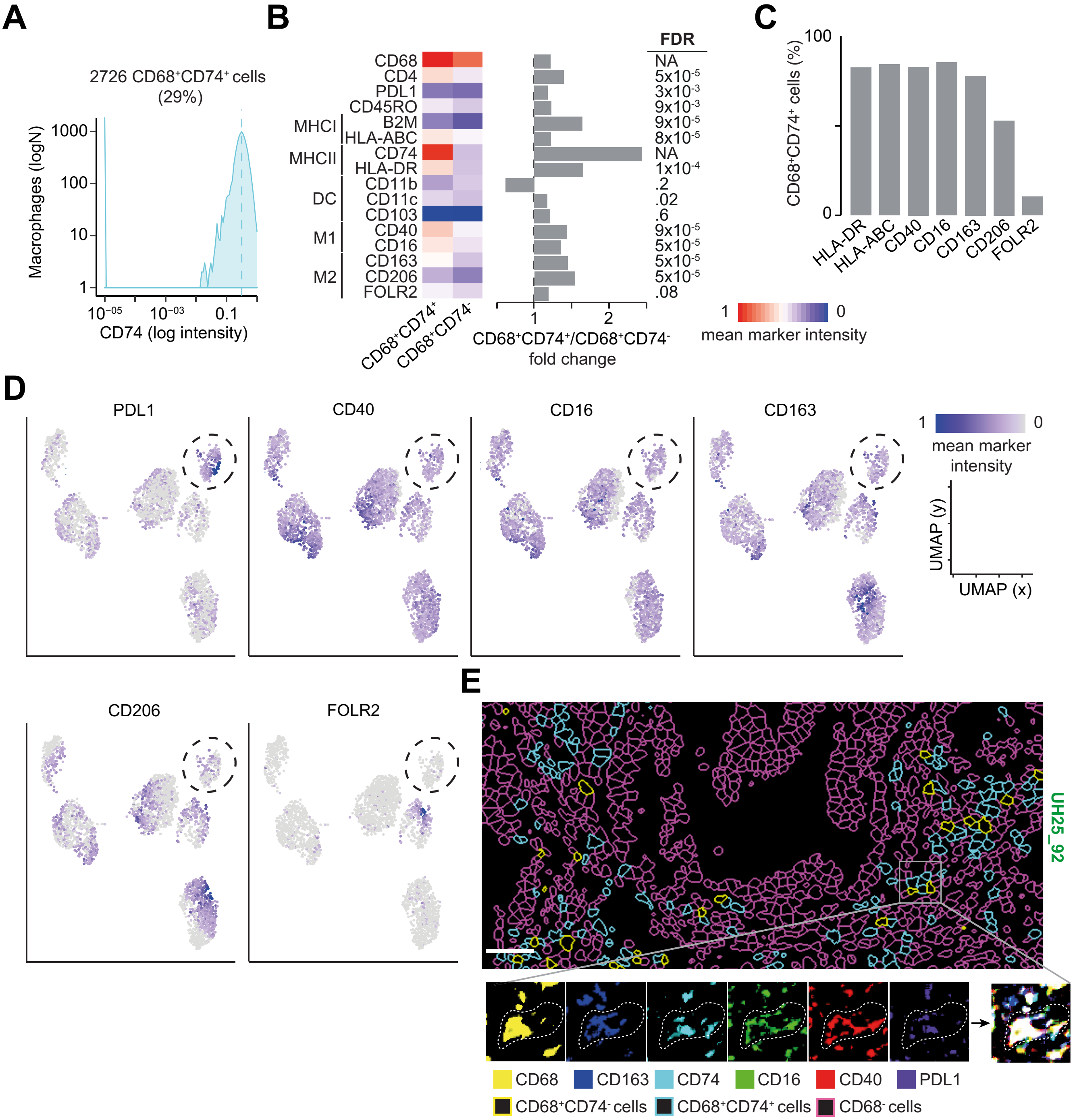
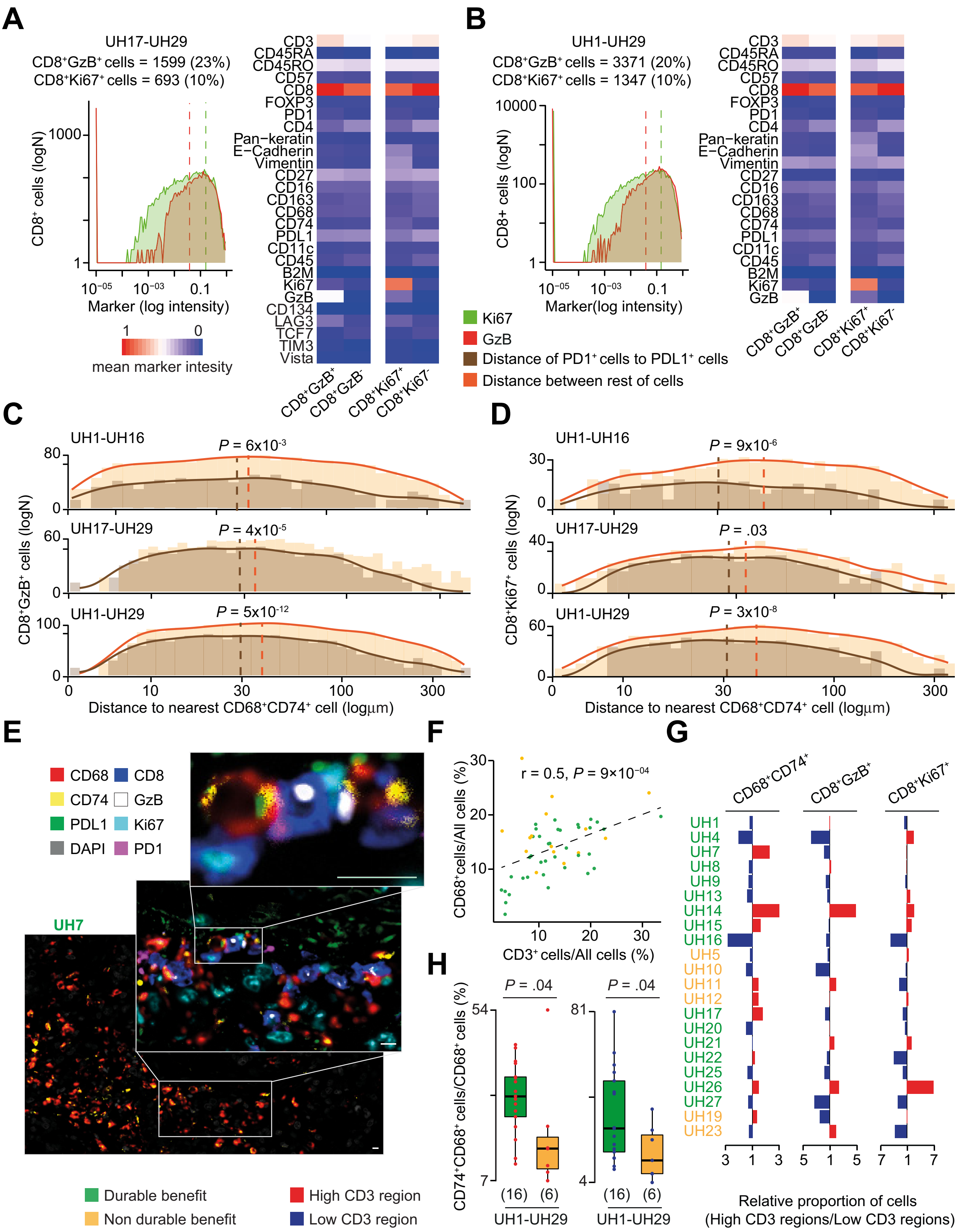


Figure 5. Functional characterization of CD68^{+}CD74^{+} cells. (*A*) CD68^{+}CD74^{+} cells in 10 DB-CRCs from both cohorts were identified by applying a threshold of 0.35 CD74 expression to all CD68^{+} cells after IMC image histologic inspection. (*B*) Mean marker intensities in CD68^{+}CD74^{+} and CD68^{+}CD74 cells. Values were normalized across all markers and cells. Marker distributions were compared with the 2-sided Wilcoxon's rank sum test, and Benjamini-Hochberg false discovery rate (FDR) correction was applied to account for testing over 14 markers. Fold change between the mean expression in CD68^{+}CD74^{+} and CD68^{+}CD74 cells is reported. DC, dendritic cells. (*C*) Percentage of CD68^{+}CD74^{+} cells expressing selected markers associated with antigen presentation and M1 and M2 phenotypes. (*D*) Uniform Manifold Approximation and Projection (UMAP) maps of 2726 CD68^{+}CD74^{+} cells in 17 regions from 10 DB-CRCs. Cells were grouped in 6 clusters based on the expression of 16 phenotypic markers using Seurat[9] and colored according to the mean intensities of representative markers. The *circle* indicates a CPDL1-expressing cluster. (*E*) Single-cell segmentation (*upper panel*) and IMC images (*lower panels*) of selected CD68^{+}CD74^{+} cell-associated markers from a representative DB-CRC. The *right bottom panel* reports the combination of all the selected markers. *Scale bar* = 100 μm.


A
UH17-UH29
CD8+GzB+ cells = 1599 (23%)
CD8+Ki67+ cells = 693 (10%)
CD8+ cells (logN)
1000
1
10−05 10−03 0.1
Marker (log intensity)
1 0
mean marker intesity
CD3 CD45RA CD45RO CD57 CD8 FOXP3 PD1 CD4 Pan−keratin E−Cadherin Vimentin CD27 CD16 CD163 CD68 CD74 PDL1 CD11c CD45 B2M Ki67 GzB CD134 LAG3 TCF7 TIM3 Vista
CD8+GzB+ CD8+GzB- CD8+Ki67+ CD8+Ki67-
B
UH1-UH29
CD8+GzB+ cells = 3371 (20%)
CD8+Ki67+ cells = 1347 (10%)
CD8+ cells (logN)
10000
100
1
10−05 10−03 0.1
Marker(log intensity)
CD3 CD45RA CD45RO CD57 CD8 FOXP3 PD1 CD4 Pan−keratin E−Cadherin Vimentin CD27 CD16 CD163 CD68 CD74 PDL1 CD11c CD45 B2M Ki67 GzB
CD8+GzB+ CD8+GzB- CD8+Ki67+ CD8+Ki67-
Ki67
GzB
Distance of PD1+ cells to PDL1+ cells
Distance between rest of cells
C
UH1-UH16 P = 6x10-3
UH17-UH29 P = 4x10-5
UH1-UH29 P = 5x10-12
CD8+GzB+ cells (logN)
80 0 60 0 100 0
0 10 30 100 300
Distance to nearest CD68+CD74+ cell (logμm)
D
UH1-UH16 P = 9x10-6
UH17-UH29 P = .03
UH1-UH29 P = 3x10-8
CD8+Ki67+ cells (logN)
30 0 40 0 60 0
0 10 30 100 300
Distance to nearest CD68+CD74+ cell (logμm)
E
CD68 CD8
CD74 GzB
PDL1 Ki67
DAPI PD1
UH7
F
r = 0.5, P = 9×10−04
CD68+cells/All cells (%)
30 20 10
10 20 30
CD3+ cells/All cells (%)
G
CD68+CD74+ CD8+GzB+ CD8+Ki67+
UH1 UH4 UH7 UH8 UH9 UH13 UH14 UH15 UH16 UH5 UH10 UH11 UH12 UH17 UH20 UH21 UH22 UH25 UH26 UH27 UH19 UH23
3 1 3 5 1 5 7 1 7
Relative proportion of cells
(High CD3 regions/Low CD3 regions)
H
P = .04 P = .04
CD74+CD68+cells/CD68+cells (%)
54 7 81 4
(16) (6) (16) (6)
UH1-UH29 UH1-UH29
Durable benefit
Non durable benefit
High CD3 region
Low CD3 region

or transcriptionally dysregulated in hypermutated DB-CRCs compared with nDB-CRCs in the discovery (Figure 2*D*, Supplementary Table 4) and validation (Figure 2*E*) cohorts. Interestingly, 2 DB-CRCs showed a clonal truncating mutation (T91fs) in the beta-2-microglobulin (B2M) gene encoding the invariable subunit of the major histocompatibility complex class I complex (Supplementary Figure 2*B*). Because a B2M antibody was part of the IMC panel (Supplementary Table 5), we could assess that a *B2M* truncating mutation led to no protein expression in the tumor compared with a widespread B2M expression in CRCs with wild-type B2M (Figure 2*F*). In general, B2M protein expression was significantly reduced in the tumor but not in the stroma of DB-CRCs in the discovery (Figure 2*G*) and validation (Figure 2*H*) cohorts as well as in both cohorts combined (Supplementary Figure 2*C*). In melanoma, B2M loss has been associated with resistance to immune checkpoint inhibitors.[31] Our data indicate an opposite association in CRC, supporting similar recent observations.[32]

Hypermutated Colorectal Cancers Are Enriched in Cytotoxic and Proliferating CD8 T Cells

To understand the role of TME in the response to anti-PD1 agents, we analyzed multiple tumor regions of the discovery cohort with IMC (Supplementary Figure 3*A*) using markers for T cells, macrophages, neutrophils, dendritic cells, and B cells as well as the tissue structure (Supplementary Table 5). After regional ablation and image processing, we verified that the relative proportion of stroma and tumor cells was similar across samples (Supplementary Figure 3*B*–*F*). We then applied 2 independent and complementary analytical approaches. In one, we compared the normalized pixel area of individual or combined markers (pixel analysis, Figure 3*A*). In the other, we applied single-cell segmentation, assigned cell identities, and compared the relative abundance of immune cell populations identified through unsupervised single-cell clustering (single-cell analysis, Figure 3*A*). Outcomes of all analyses were validated by independent histologic assessment of unprocessed images.

Hypermutated DB- and nDB-CRCs showed no difference in normalized CD3$^+$ area (Figure 3*B*, Supplementary Table 6) or proportion of CD3$^+$ cells (Figure 3*C*, Supplementary Table 7), confirming that overall T-cell infiltration does not correlate with TMB (Figure 1*E*) or response to therapy. To further investigate whether DB- and nDB-CRCs differed in specific T-cell subpopulations, we performed single-cell clustering using 12 T-cell markers (Supplementary Table 5). We found no qualitative or quantitative differences in T-cell subpopulations between DB- and nDB-CRCs (Supplementary Table 8, Supplementary Figure 4).

Given their relevance to immune checkpoint inhibitors, we further profiled T cells in the validation cohort by adding 5 markers of T-cell function to the 12 used previously (Supplementary Table 5). We confirmed no significant difference in the normalized CD3$^+$ area or proportion of CD3$^+$ cells between DB- and nDB-CRCs of the validation cohort (Figure 3*D* and *E*). Moreover, single-cell clustering with all 17 phenotypic markers of T cells confirmed no difference in T-cell infiltrates between DB -and nDB-CRCs (Supplementary Table 8).

We repeated the same comparison between hypermutated and non-hypermutated CRCs of the discovery cohort. In this case, we found 2 clusters of CD8 T cells (cluster 1, expressing granzyme B [GzB], and cluster 2, expressing Ki67) significantly higher in hypermutated CRCs (Figure 3*F*–*H*). Pixel analysis confirmed these results (Figure 3*I*, Supplementary Figure 3*G*).

Our analysis identified the cytotoxic and proliferating CD8 T-cell subpopulations that are specifically enriched in hypermutated CRCs, confirming recent reports[33] and likely due to Wnt low activation observed in these samples (Figure 2*A* and *B*). No qualitative or quantitative differences in any subpopulation of T cells were detected between hypermutated DB- and nDB-CRCs, which were both rich in CD8 T cells.

Hypermutated Durable Benefit Colorectal Cancers Are Enriched in CD74$^+$ Macrophages

To further investigate the association of TME with response, we compared the relative abundance of all other main immune populations between hypermutated and non-

Figure 6. Interaction between CD74$^+$ macrophages and GzB$^+$Ki67$^+$ CD8 T cells. (*A*) CD8$^+$GzB$^+$ and CD8$^+$Ki67$^+$ T cells in the validation and (*B*) combined cohorts were identified by applying a threshold of 0.05 GzB and 0.15 Ki67 expression to CD8 T cells, after IMC image histologic inspection. Markers of mean intensities in CD8$^+$GzB$^+$ or CD8$^+$Ki67$^+$ and CD8$^+$GzB$^-$ or CD8$^+$Ki67$^-$T cells were normalized across all markers and cells. (*C*) Distance distributions of CD8$^+$GzB$^+$ or (*D*) CD8$^+$Ki67$^+$ to the nearest CD74$^+$ macrophage in the discovery, validation, and combined cohorts. Distances between cells were divided into 1.1-μm bins, and the density curves fitting the histograms were measured. Distributions of PD1$^+$ or PDL1$^+$ and the rest of the cells were compared using the 2-sided Wilcoxon's rank sum test. The *dashed lines* represent medians of the distributions. (*E*) High-resolution mIF image of a representative CRC with a highlighted cluster of CD74$^+$ macrophages (*main image*) and their interactions with CD8$^+$GzB$^+$ and CD8$^+$Ki67$^+$ T cells (*zoom-ins*). The image was scanned at original magnification ×40. *Scale bars* = 10 μm. (*F*) Correlation between normalized T cells and macrophages in 26 DB- and nDB-CRCs of the discovery and validation cohorts. Pearson correlation coefficient *R* and associated *P* value are shown. (*G*) Ratios of normalized CD74$^+$ macrophages, CD8$^+$GzB$^+$, and CD8$^+$Ki67$^+$ T cells between regions with high and low T-cell infiltration. For samples with more 2 regions, the total cells in the high or low regions were normalized and used to compute the ratio. (*H*) Comparisons of normalized of CD74$^+$ macrophages between DB- and nDB-CRCs of the combined cohorts considering only high (*left*) and low (*right*) T-cell infiltration regions. Distributions were compared using the 2-sided Wilcoxon's rank sum test. The *horizontal line* in the middle of each *box* indicates the median; the *top* and *bottom borders* of the box mark the 75th and 25th percentiles, respectively, and the *vertical lines* mark points within 1.5 the inter-quartile range.

hypermutated CRCs or DB- and nDB-CRCs of the discovery cohort.

We found no difference in dendritic cells, neutrophils, and B cells between hypermutated and non-hypermutated CRCs (Supplementary Table 8). However, we observed proportionally higher CD68$^+$CD74$^+$ cells in DB-CRCs than in nDB-CRCs (cluster 3, Figure 4*A*–*C*), which was confirmed by pixel analysis (Figure 4*D*). To validate these results, we profiled the macrophages also in the validation cohort. Pixel analysis confirmed a higher normalized CD68$^+$CD74$^+$ area in the validation samples alone (Figure 4*E*) and together with the discovery cohort (Figure 4*F*). To identify CD68$^+$CD74$^+$ cells, we applied a threshold of 0.1 CD74 expression to all macrophages in the validation cohort (Figure 4*G*) and in all hypermutated CRCs (Figure 4*H*). We verified that CD68$^+$CD74$^+$ cells identified in this way matched phenotypically to cells in cluster 3 of the discovery cohort (Supplementary Figure 5*A*–*C*). Comparing the proportion of CD68$^+$CD74$^+$ cells between DB-CRCs and nDB-CRCs we found that it was higher in DB-CRCs of the validation cohort alone (Figure 4*I*) and when all hypermutated CRCs were analyzed together (Figure 4*J*). Therefore, we found that CD68$^+$CD74$^+$ cells are associated with response to anti-PD1 immunotherapy in CRC.

To further characterize these cells, we profiled selected DB-CRCs from both cohorts (Supplementary Table 2) with 16 additional markers (Supplementary Table 5) and identified CD68$^+$CD74$^+$ cells applying a threshold on CD74 expression (Figure 5*A*). All stained markers, except those associated with dendritic cell functions, were more expressed in CD68$^+$CD74$^+$ cells than in CD68$^+$CD74$^-$ cells (Figure 5*B*). HLA-ABC, HLA-DR, CD40, CD16, and CD163 were expressed in >80% of CD68$^+$CD74$^+$ cells, whereas M2-associated markers, such as CD206 and FOLR2, were specific to smaller subsets (Figure 5*C*). This expression profile suggested that CD68$^+$CD74$^+$ macrophages may have a T-cell–activating phenotype. Interestingly, approximately 40% of them expressed both M1 and M2 markers, consistent with phenotypic plasticity (Supplementary Figure 5*D*). Single-cell clustering identified 6 distinct groups of CD68$^+$CD74$^+$ cells, one of which expressed high levels of PDL1, together with CD40, CD16, and CD163, but not CD206 and FOLR2 (Figure 5*D*). Independent histologic inspection confirmed coexpression of these markers in CD68$^+$CD74$^+$ cells (Figure 5*E*).

Finally, we compared the normalized PD1 or PDL1 protein expression between DB- and nDB-CRCs and found no significant differences in the discovery (Supplementary Figure 6*A*), validation (Supplementary Figure 6*B*), and combined (Supplementary Figure 6*C*) cohorts. This was supported by gene expression analysis (Supplementary Figure 6*D* and *E*) and single-cell clustering, which detected no qualitative or quantitative differences in PD1$^+$ or PDL1$^+$ cells (Supplementary Table 8). In general, the expression of both PD1 and PDL1 genes was low (Supplementary Figure 6*F*), as confirmed also in TCGA CRCs, where their expression was significantly lower than in melanoma and lung cancer (Supplementary Figure 6*G*). Consistent with their low expression, we could detect PD1-PDL1 protein complex formation in only a minority of regions (Supplementary Table 2), and A-FRET intensity was lower than in melanoma and renal cancer.[22] The proportion of regions with detectable PD1-PDL1 complex was significantly less in hypermutated than in non-hypermutated CRCs, whereas there was no difference between DB- and nDB-CRCs (Supplementary Figure 6*H*).

Our analyses suggest that a subset of antigen-presenting macrophages with a T-cell– activating phenotype may play a key role in CRC response to anti-PD1 immune therapy. The overall expression of PD1 and PDL1 is low at the gene and protein levels and they show no association with response, indicating that unlike other cancer types,[2] they are not biomarkers of response in CRC.

CD74$^+$PDL1$^+$ Macrophages Interact with PD1$^+$ Cytotoxic and Proliferating CD8 T Cells

Our deep investigation of immune infiltrates showed that hypermutated DB-CRCs are immune hot tumors, with high levels of CD74$^+$ macrophages compared with nDB-CRCs as well as of cytotoxic and proliferating T cells associated with the hypermutated phenotype. Because CD74$^+$ macrophages also expressed PDL1 while the 2 CD8 T-cell populations expressed PD1 (Supplementary Table 8, Supplementary Figure 4), we asked whether these cells were proximal in the TME and interacted through PD1-PDL1 contact.

To interrogate this, we identified CD8$^+$GzB$^+$ and CD8$^+$Ki67$^+$ cells in the validation cohort (Figure 5*A*) and in all hypermutated CRCs (Figure 6*B*) by applying a threshold of 0.05 (GzB) and 0.15 (Ki67) expression to all CD8 T cells. We verified that these cells were phenotypically similar to clusters 1 (CD8$^+$GzB$^+$ cells) and 2 (CD8$^+$Ki67$^+$ cells) of the discovery cohort (Supplementary Figure 7). These 2 populations did not selectively express any additional T-cell markers used in the validation cohort, except the immune checkpoint protein LAG3 (Figure 6*A*). The absence of TCF7 expression in proliferating CD8 T cells suggested that they do not have stem-like characteristics and are not analogous to recently described intratumoral T-cell developmental niches.[34,35]

After identifying the CD8$^+$GzB$^+$ and CD8$^+$Ki67$^+$ T-cell subpopulations, we measured the centroid distance between them and CD68$^+$CD74$^+$ cells. We then measured the distance between CD8$^+$GzB$^+$PD1$^+$ or CD8$^+$Ki67$^+$PD1$^+$ and CD68$^+$CD74$^+$PDL1$^+$ cells and found that they were closer than to other cells in the discovery, validation, and combined cohorts (Figure 6*C* and *D*). Moreover, a substantial fraction of CD74$^+$ macrophages (52% in DB-CRCs and 32% in nDB-CRCs) aggregated in high-density clusters composed of $\geq$5 cells/10,000 μm^2[34]. These computationally identified clusters of CD74$^+$ macrophages also contained CD8$^+$GzB$^+$ and CD8$^+$Ki67$^+$ cells (Supplementary Figure 8). The existence of these clusters was confirmed through independent histologic inspection (Supplementary Figure 9), which also detected direct interactions between CD8$^+$GzB$^+$PD1$^+$ or CD8$^+$Ki67$^+$PD1$^+$ and CD68$^+$CD74$^+$PDL1$^+$ cells. To confirm these interactions at higher resolution, we performed mIF with 8 key markers defining CD74$^+$CD68$^+$, GzB$^+$CD8$^+$, and Ki67$^+$CD8$^+$ cells (Supplementary Table 5). We confirmed the presence of clusters of CD74$^+$ macrophages in close proximity to CD8$^+$GzB$^+$ and CD8$^+$Ki67$^+$ T cells and detected their interaction via PD1-PDL1 contact (Figure 6*E*, Supplementary Figure 10).

Taking advantage of the multiregional profiles, we asked how the observed intratumor T-cell heterogeneity (Figure 1*C*, Supplementary Figure 1*D*) affected the distinctive infiltration pattern of DB-CRCs. First, we observed that tumor regions rich in T cells were also rich in macrophages (Figure 6*F*), indicating that intratumor heterogeneity involves a more general pattern of coinfiltration. Next, we investigated how the 3 key populations of DB-CRCs (CD8$^+$GzB$^+$, CD8$^+$Ki67$^+$, and CD68$^+$CD74$^+$ cells) were distributed across regions of the same tumor. We observed that their relative proportions were highly variable between high and low infiltrate regions and that no clear pattern could be seen discriminating DB- and nDB-CRCs (Figure 6*G*). Despite such a heterogeneous composition of the immune infiltrates, we observed consistently higher proportion of CD74$^+$ macrophages in DB-CRCs than in nDB-CRCs independently of T-cell infiltration levels (Figure 6*H*).

Our data consistently indicate that CD74$^+$ macrophages differ between DB- and nDB-CRCs across cohorts and regions. We therefore propose that their interaction with CD8$^+$GzB$^+$PD1$^+$ and CD8$^+$Ki67$^+$PD1$^+$ cells through PDL1 is key to confer durable benefit from treatment.

Discussion

In this study, we integrated multiregional genomic, transcriptomic, histopathologic, and immune-phenotypic data to characterize the tumor-immune interactions determining response of CRC to immune checkpoint blockade.

After extensive unsupervised investigation of variability in leukocyte subpopulations between hypermutated DB- and nDB-CRCs using multiple approaches, we found CD74$^+$ macrophages were the only immune cell population that consistently segregated with response in DB-CRCs. This is remarkable, given the observed genetic and immune inter- and intratumor heterogeneity and the diversified treatment history and suggests that CD74$^+$ macrophages could be further developed as a robust predictor of response in a broad range of patients. These macrophages express PDL1 and are in close proximity to PD1$^+$ CD8 T cells, indicating that the PD1/PDL1 interaction between these cells may restrain CD8 T-cell function and may be the one that anti-PD1 antibodies break to release cytotoxic antitumor activity.

The high cytotoxic CD8 infiltration in hypermutated CRCs is likely enabled by the low activation of the Wnt pathway, resulting in an immune hot environment. To evade immune elimination, hypermutated DB-CRCs develop immune escape mechanisms via genetic inactivation or transcriptional repression of antigen-presenting genes. Interestingly, unresponsive hypermutated CRCs do not show such a pervasive disruption of the antigen presentation machinery, despite comparably high levels of CD8 infiltration. The molecular mechanisms by which these tumors survive the attack of cytotoxic CD8 T cells need further investigation, although a possible explanation could reside in their significantly reduced proportion of CD74$^+$ macrophages.

Similarly, further investigations are required to explain how tumors lacking B2M can respond to immunotherapy. In B2M-null CRC mice, response to anti-PD1 agents relies on CD4 T cells rather than CD8 T cells.[36] Although we did not observe any difference in CD4 T cells between DB- and nDB-CRCs, this suggests that anti-PD1 agents may act through several mechanisms, including antigen-independent T-cell activation or reinduction of B2M expression.

Our study also highlights cancer-specific traits of response to anti-PD1 immunotherapy. We show that in CRC, high TMB is necessary but not sufficient to achieve durable benefit and that above the critical threshold of the hypermutated phenotype, even CRCs with very high TMB may not respond to treatment. This is different from lung cancer and melanoma, where response always positively correlates with TMB.[37,38] In CRC, a low TMB is a marker of resistance, not because of a low neoantigenic load but because it is associated with a higher activation of the Wnt pathway leading to immune cold tumors.

Moreover, while the impairment of antigen presentation in immune hot tumors is shared across cancer types,[39] the association of B2M loss with response and the overall low PD1 and PDL1 expression are specific traits of CRC. This suggests that universal predictors of response to immunotherapy may not exist and that the specific genetics of the tumor as well as the features of the TME should be considered. In the case of CRC, these may include clonal immunogenic mutations and expanded T cells, low activation of the Wnt pathway, and high infiltration of CD8 T cells coupled with CD74 macrophages.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at https://doi.org/10.1053/j.gastro.2021.06.064.

References

1. Ribas A, Wolchok JD. Cancer immunotherapy using checkpoint blockade. Science 2018;359:1350–1355.
2. Le DT, Uram JN, Wang H, et al. PD-1 blockade in tumors with mismatch-repair deficiency. N Engl J Med 2015;372:2509–2520.
3. **Van den Eynde M, Mlecnik B, Bindea G**, et al. The link between the multiverse of immune microenvironments in metastases and the survival of colorectal cancer patients. Cancer Cell 2018;34:1012–1026.e3.
4. Ganesh K, Stadler ZK, Cercek A, et al. Immunotherapy in colorectal cancer: rationale, challenges and potential. Nat Rev Gastroenterol Hepatol 2019;16:361–375.

5. André T, Shiu K-K, Kim TW, et al. Pembrolizumab in microsatellite-instability–high advanced colorectal cancer. N Engl J Med 2020;383:2207–2218.
6. Hoos A, Eggermont AM, Janetzki S, et al. Improved endpoints for cancer immunotherapy trials. J Natl Cancer Inst 2010;102:1388–1397.
7. Bankhead P, Loughrey MB, Fernandez JA, et al. QuPath: open source software for digital pathology image analysis. Sci Rep 2017;7:16878.
8. Bortolomeazzi M, Montorsi L, Temelkovski D, et al. A SIMPLI (Single-cell Identification from MultiPLexed Images) approach for spatially resolved tissue phenotyping at single-cell resolution. bioRxiv 2021;2021. 04. 01.437886.
9. Butler A, Hoffman P, Smibert P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36:411–420.
10. Ester M, Kriegel H-P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining 1996:226–231.
11. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at: https://arxiv.org/abs/1303.3997 2013.
12. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods 2018;15:591–594.
13. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164.
14. **Mourikis TP, Benedetti L, Foxall E**, et al. Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma. Nat Commun 2019; 10:3101.
15. **Van Loo P, Nordgard SH**, Lingjaerde OC, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A 2010;107:16910–16915.
16. Shukla SA, Rooney MS, Rajasagi M, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat Biotechnol 2015;33:1152–1158.
17. Schenck RO, Lakatos E, Gatenbee C, et al. Neo-PredPipe: high-throughput neoantigen prediction and recognition potential pipeline. BMC Bioinform 2019; 20:264.
18. Roth A, Khattra J, Yap D, et al. PyClone: statistical inference of clonal population structure in cancer. Nat Methods 2014;11:396–398.
19. Moll P, Ante M, Seitz A, et al. QuantSeq 3′ mRNA sequencing for RNA quantification. Nat Methods 2014; 11:i–i.
20. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.
21. Carlson CS, Emerson RO, Sherwood AM, et al. Using synthetic templates to design an unbiased multiplex PCR assay. Nat Commun 2013;4:2680.
22. **Sánchez-Magraner L, Miles J**, Baker CL, et al. High PD-1/PD-L1 checkpoint interaction infers tumor selection and therapeutic sensitivity to anti-PD-1/PD-L1 treatment. Cancer Res 2020;80:4244.
23. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330–337.
24. Goodman AM, Kato S, Bazhenova L, et al. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. Mol Cancer Ther 2017;16:2598.
25. Le DT, Durham JN, Smith KN, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science 2017;357:409–413.
26. **Samstein RM, Lee C-H, Shoushtari AN, Hellmann MD**, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nat Genet 2019;51:202–206.
27. **Schrock AB, Ouyang C**, Sandhu J, et al. Tumor mutational burden is predictive of response to immune checkpoint inhibitors in MSI-high metastatic colorectal cancer. Ann Oncol 2019;30:1096–1103.
28. Valero C, Lee M, Hoen D, et al. The association between tumor mutational burden and prognosis is dependent on treatment context. Nat Genet 2021;53:11–15.
29. **Grasso CS, Giannakis M, Wells DK**, et al. Genetic mechanisms of immune evasion in colorectal cancer. Cancer Discov 2018;8:730–749.
30. Spranger S, Bao R, Gajewski TF. Melanoma-intrinsic β-catenin signalling prevents anti-tumour immunity. Nature 2015;523:231–235.
31. **Sade-Feldman M, Jiao YJ**, Chen JH, et al. Resistance to checkpoint blockade therapy through inactivation of antigen presentation. Nat Commun 2017;8:1136.
32. Middha S, Yaeger R, Shia J, et al. Majority of B2M-mutant and-deficient colorectal carcinomas achieve clinical benefit from immune checkpoint inhibitor therapy and are microsatellite instability-high. JCO Precis Oncol 2019;3. PO.18.00321.
33. de Vries NL, van Unen V, Ijsselsteijn ME, et al. High-dimensional cytometric analysis of colorectal cancer reveals novel mediators of antitumour immunity. Gut 2020; 69:691–703.
34. Jansen CS, Prokhnevska N, Master VA, et al. An intratumoral niche maintains and differentiates stem-like CD8 T cells. Nature 2019;576:465–470.
35. Siddiqui I, Schaeuble K, Chennupati V, et al. Intratumoral Tcf1+PD-1+CD8+ T cells with stem-like properties promote tumor control in response to vaccination and checkpoint blockade immunotherapy. Immunity 2019; 50:195–211.e10.
36. Germano G, Lu S, Rospo G, et al. CD4 T cell dependent rejection of beta 2 microglobulin null mismatch repair deficient tumors. Cancer Discov 2021;11:1844–1859.
37. Gurjao C, Tsukrov D, Imakaev M, et al. Limited evidence of tumour mutational burden as a biomarker of response to immunotherapy. bioRxiv 2020;2020. 09.03.260265.
38. Wood MA, Weeder BR, David JK, et al. Burden of tumor mutations, neoepitopes, and other variants are weak

predictors of cancer immunotherapy response and overall survival. Genome Med 2020;12:33.

39. **McGranahan N, Rosenthal R**, Hiley CT, et al. Allele-specific HLA loss and immune escape in lung cancer evolution. Cell 2017;171:1259–1271.e11.

40. **Frampton GM, Fichtenholtz A**, Otto GA, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol 2013;31:1023–1031.

41. **Tarabichi M, Salcedo A, Deshwar AG**, et al. A practical guide to cancer subclonal reconstruction from DNA sequencing. Nat Methods 2021;18:144–155.

42. Barbie DA, Tamayo P, Boehm JS, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 2009;462:108.

43. Herbst A, Jurinovic V, Krebs S, et al. Comprehensive analysis of β-catenin target genes in colorectal carcinoma cell lines with deregulated Wnt/β-catenin signaling. BMC Genomics 2014;15:74.

Author names in bold designate shared co-first authorship.

Received February 5, 2021. Accepted June 22, 2021.

Correspondence
Address correspondence to: Francesca D. Ciccarelli, PhD, Cancer Systems Biology Laboratory, The Francis Crick Institute, London NW1 1AT, United Kingdom. e-mail: francesca.ciccarelli@crick.ac.uk; or Jo Spencer, PhD, School of Immunology and Microbial Sciences, King's College London, London SE1 9RT, United Kingdom. e-mail: jo.spencer@kcl.ac.uk.

Acknowledgments
The authors thank Professor Rebecca Fitzgerald (University of Cambridge) and Professor Toby Lawrence (King's College London) for critical comments to the manuscript. The authors acknowledge technical support from the Advanced Sequencing, the Flow Cytometry and the Experimental Histopathology Platforms of The Francis Crick Institute, the National Institute for Health Research Biomedical Research Centre Immune Monitoring Core Facility Centre based at Guy's and St Thomas' National Health Service Foundation Trust and King's College London, Christopher Applebee and Sánchez-Magraner for the FRET analysis, and Joe Brock for graphical illustrations. The views expressed are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research or the United Kingdom Department of Health. The results published here are in part based upon data generated by The Cancer Genome Atlas managed by the National Cancer Institute and National Human Genome Research Institute. WES (EGAD00001006165) and RNA sequencing (EGAD00001006164) data have been deposited in the European Genome-phenome Archive. TCR-seq (https://doi.org/10.5281/zenodo.3744371) and IMC (https://doi.org/10.5281/zenodo.3743253) data have been deposited in Zenodo. The code used in this study and all supporting data are available upon request.

CRediT Authorship Contributions
Michele Bertolomeazzi, MRes (Formal analysis: Equal; Software: Lead; Writing – original draft: Equal).
Reda Keddar, MRes (Formal analysis: Equal; Methodology: Lead; Writing – original draft: Supporting).
Lucia Montorsi, PhD (Data curation: Equal; Formal analysis: Equal; Writing – original draft: Supporting).
Amelia Acha, PhD (Data curation: Equal; Formal analysis: Equal; Writing – original draft: Supporting).
Lorena Benedetti, PhD (Data curation: Supporting; Formal analysis: Supporting).
Damjan Temelkowski, PhD (Methodology: Supporting; Software: Supporting).
Subin Choi, MRes (Data curation: Supporting).
Nedyalko Petrov, MRes (Methodology: Supporting).
Katrina Todd, PhD (Methodology: Supporting).
Patty Way, PhD (Methodology: Supporting).
Jonny Kohl, PhD (Methodology: Supporting).
Tamara Denner, PhD (Methodology: Supporting).
Emma Nye, PhD (Methodology: Supporting).
Robert Goldstone, PhD (Methodology: Supporting).
Sophia Ward, PhD (Methodology: Supporting).
Gareth Wilson, PhD (Methodology: Supporting).
Maise Al Bakir, PhD (Methodology: Supporting).
Charles Swanton, PhD (Methodology: Supporting).
Susan John, PhD (Formal analysis: Supporting).
James Miles, MRes (Methodology: Supporting).
Banafshe Larijani, PhD (Methodology: Supporting).
Victoria Kunene, PhD (Resources: Supporting).
Elisa Fontana, PhD (Resources: Supporting).
Toby Arkenau, PhD (Resources: Supporting).
Peter Parker, PhD (Methodology: Supporting).
Manuel Rodriguez-Justo, PhD (Formal analysis: Supporting; Investigation: Supporting; Resources: Lead; Writing – original draft: Supporting).
Kai-Keen Shiu, MD PhD (Investigation: Supporting; Resources: Lead; Writing – original draft: Supporting).
Jo Spencer, PhD (Conceptualization: Equal; Formal analysis: Equal; Writing – original draft: Equal).
Francesca Ciccarelli, PhD (Conceptualization: Equal; Funding acquisition: Lead; Investigation: Lead; Supervision: Lead; Writing – original draft: Equal).

Conflicts of interest
The authors disclose no conflicts.

Funding
This work was supported by the Francis Crick Institute which receives its core funding from Cancer Research UK (FC001002, FC001169, FC001745, FC001130), the UK Medical Research Council (FC001002, FC001169, FC001745, FC001130), and the Wellcome Trust (FC001002, FC001169, FC001745, FC001130). Francesca D. Ciccarelli is supported by Cancer Research UK (C43634/A25487), Guys and St Thomas Charity (R170504), the European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska-Curie grant agreement No. CONTRA-766030, the Cancer Research UK King's Health Partners Centre at King's College London (C604/A25135), and the Cancer Research UK City of London Centre (C7893/A26233).

Supplementary Methods

Patient cohorts, treatment and experimental design

FFPE blocks were obtained from the resection of the primary tumour or local relapse of 16 patients (discovery cohort) and 13 patients (validation cohort) treated with immune checkpoint inhibitors in the setting of metastatic CRC until disease progression, unacceptable toxicity or completion of treatment (Supplementary Table 1). In the discovery cohort, patients UH1-UH10 were treated with Pembrolizumab (200 mg every three weeks) as part of the KEYNOTE 177 clinical trial (ClinicalTrials.gov, NCT02563002)[1], while patients UH11-UH16 were treated with Nivolumab (240mg every two weeks). In the validation cohort, patients UH17-UH19 were part of the KEYNOTE 177 trial, UH26 received Pembrolizumab (2mg/kg every three weeks) and patients UH20-UH25 and UH29 were treated with Nivolumab (240 mg every two weeks). Patient UH27 received Ipilimumab (1mg/kg) in combination with Nivolumab (3mg/kg) every three weeks for four cycles followed by Nivolumab alone (240 mg every two weeks). Patient UH28 received Nivolumab (3mg/kg) for two cycles, then Ipilimumab (1mg/kg) in combination with Nivolumab every three weeks for three cycles. Patients treated with Nivolumab were enrolled in the UK wide Bristol Myers Squibb Individual Patient Supply Request Programme as per Article 5/1 of Article Directive 2001/83/EC. All patients were consented at the UCL Cancer Institute Pathology Biobank - REC reference 15/YH/0311.

Response to therapy was assessed using the formal Response Evaluation Criteria in Solid Tumours (RECIST)[2] version 1.1. Patients were considered to achieve durable benefit (DB) if the disease did not progress for at least 12 months after commencing immunotherapy; no durable benefit (nDB) if the disease progressed within 12 months. Twelve-month cut-off was considered clinically better than the progression-free survival from chemotherapy as first line treatment of metastatic stable (8.3 months[3]) or hypermutated (8.2 months[1]) CRC.

Twenty-four sequential sections were cut from each FFPE tumour block of samples UH1-UH16 using a microtome. Sections were then used for CD3 staining (slides A, B, F, H and J); Imaging Mass Cytometry (IMC, slide C); multiplexed Immunofluorescence (mIF, slide D); Whole Exome Sequencing (WES, slides E1-5); RNA sequencing (RNA-seq, slides G1-5); T-Cell Receptor Beta sequencing (TCR-seq, Slides I1-5) and to detect PD1-PDL1 interaction *in situ* (slides K1-2). For samples

UH17-UH23 and UH25-UH27, 11 sequential sections were used for CD3 staining (slides A, E and G); Haematoxylin and Eosin (HE) staining (slide B); IMC (slide C); mIF (slide D) and RNAseq (slides F1-5). Since UH24, UH28 and UH29 were biopsies, RNAseq could not be performed and only four sequential sections were used for CD3 and HE staining (slides A and B); IMC (slide C) and mIF (Slide D). Sections used for CD3 and HE staining, IMC, mIF and A-FRET were 4μm thick, while those used for DNA and RNA extraction were 10μm thick. Tumour content was assessed by a board-certified surgical pathologist (M.R.J.).

CD3 staining and quantification

CD3 staining was performed upon slide dewaxing and heat-induced epitope retrieval (HIER) using Antigen Retrieval Reagent-Basic (R&D Systems). Tissues were blocked and incubated first with anti CD3 antibody (Dako, Supplementary Table 5) and subsequently with horseradish peroxidase conjugated anti rabbit antibody (Dako). They were then stained with 3,3' diaminobenzidine (DAB) substrate (Abcam) and haematoxylin. Slide A was reviewed by a certified pathologist (M.R-J) to identify two to four regions per slide with variable CD3$^+$ infiltration (for a total of 90 regions, Supplementary Table 2) in proximity to the invasive margins of the tumour (Supplementary Figure 1B).

Digital acquisition of CD3 stained slides was performed using Hamamatsu Nanozoomer (Hamamatsu Photonics) or Axioscan Z1 (Zeiss) at 20x resolution. The whole slide images were then loaded into QuPath[4] v0.2.0-m4 to quantify CD3$^+$ infiltration within each region. The “Estimate Stain Vector” function was run as pre-processing step to increase the contrast between DAB and haematoxylin. The outlines of the regions delimited by the pathologist in slide A and projected in all other slides (Supplementary Figure 1C). The regions were then divided into 0.09 mm^2 large tiles and CD3$^+$ cells were quantified within each tile using the “positive cell detection” function. The median value of CD3$^+$ cells per mm^2 across all tiles was considered as representative of CD3$^+$ infiltration for that region. For slides B and I, CD3$^+$ cells were also quantified for the whole tumour region.

Imaging mass Cytometry (IMC)

Three panels of 42 antibodies in total were assembled to represent the main immune and stromal populations of the gut TME (IMC panels I, II and III, Supplementary Table 5). Twenty-five of these antibodies were already metal-tagged (Fluidigm), while 17 were purchased in a carrier-free form, tested via immunohistochemistry and tagged using the Maxpar X8 metal conjugation kit (Fluidigm). To identify the optimal dilution for each antibody, concentrations ranging from 1/50 to 1/5,000 were tested in FFPE appendix sections. After staining and ablation, images were visualised using MCD Viewer (Fluidigm) and the concentration giving the best signal to background ratio was selected (Supplementary Table 5).

IMC was performed in 38 regions of the discovery cohort using IMC panel I, in 22 regions of the validation cohort using IMC panel II, and in additional 17 regions of selected samples from both cohorts with IMC panel III (Supplementary Table 2). In the discovery cohort, the two regions (one with low and one with high CD3 infiltration) with the highest tumour content were selected except for UH4 UH6, UH9 and UH12. For UH6, UH9 and UH12, all four regions were analysed, for UH4, the two high and two low CD3 regions were analysed together, to be consistent with WES and RNA-seq analyses (see below). In the validation cohort, the two regions with the highest difference in CD3 infiltration were selected except for the three biopsies (UH24, UH28 and UH29) and UH18, where only one region was analysed. Slides were incubated for one hour at 60°C, dewaxed, rehydrated and subjected to HIER using a pressure cooker and Antigen Retrieval Reagent-Basic (R&D Systems). Tissues were blocked in a solution containing 10% BSA (Sigma), 0,1% Tween (Sigma), 1:50 Kiovig (Shire Pharmaceuticals) Superblock Blocking Buffer (Thermo Fisher) for two hours at room temperature. The primary antibody mix was prepared in blocking solution at the selected concentration for each antibody and incubated overnight at 4°C. Slides were then washed twice in PBS and PBS-0.1% Tween and incubated with 2 isotopes (^{191}Ir and ^{193}Ir) of DNA intercalator Cell-ID™ Intercalator-Ir (Fluidigm) 1.25mM diluted in PBS for 30 minutes at room temperature. Slides were then washed once in PBS and once in MilliQ water and air-dried. Stained slides were loaded in the Hyperion Imaging System (Fluidigm) imaging module to obtain light-contrast high resolution images of approximately 4 mm^2. For each region, a 1 mm^2 area with high tumour content and representative of the median CD3$^+$ content of the region was selected for laser ablation (Supplementary Figure 3A) at 1 μm/pixel resolution and 200Hz frequency.

IMC pixel analysis

IMC data analysis was performed with SIMPLI[5] as summarised in Figure 3A. For each of the 77 ablated regions, TIFF images from each antibody and two DNA intercalators were obtained from the raw IMC .mcd and .txt files using imctools[6]. Pixel intensities for each channel were normalised to the 99th percentile of the intensity distribution and the obtained values scaled between 0 and 1. Background pixels were removed using global thresholding with CellProfiler[7] 3.1.8. After visual inspection, channels for PD1, PDL1, GzB, CD45RA, TIM3, Vista, TCF7, CD134, CD206 and FOLR2 were further filtered using probability masks produced with Ilastick[8] 1.3.0. For this purpose, random forest classifiers were trained using closely related markers (CD3 for PD1; Vimentin for PDL1; CD8 and CD15 for GZB; CD45 and CD45RO for CD45RA; CD68 for FOLR2 and CD206). The resulting background probability masks were converted into binary images with CellProfiler[7] 3.1.8 and applied to the original normalised images to remove the background. Custom R scripts were used to count the positive pixels in all processed images for each channel. The sum of all positive pixels for a channel constituted the positive area for that channel. Given B2M low expression, an ad-hoc threshold was applied on the normalized intensity and all pixels higher than 0.5 were considered as positive. For regions UH19_87, UH27_96, and UH27_97, the CD3 masking threshold was adjusted after manual inspection to 0.175, 0.15 and 0.15 respectively.

Tumour masks were generated with CellProfiler[7] 3.1.8 summing up the Pan-keratin and E-cadherin channels for all regions except UH18_103 and UH22_112, where only E-cadherin was used. The resulting images were smoothened with a Gaussian filter and filling up all <30 pixel negative areas. The stroma masks were obtained using the Vimentin, SMA and DNA channels in the discovery cohort and Vimentin, CD68, CD11c, CD3, CD27 and CD45 channels in the validation cohort. All <20 pixel negative areas were filled up. The tissue mask for each region corresponded to the sum of tumour and stroma masks. Pixel analysis was performed by normalising the positive areas for each marker or combination of markers over the total tissue area or the area of the five main immune populations (T cells, B cells, macrophages, dendritic cells and neutrophils) for the discovery cohort and for T cell and macrophages only for the validation cohort (Supplementary Table 6).

IMC single cell analysis

Single cell analysis was based on cell segmentation, assignment of cell identity and phenotype clustering (Figure 3A).

Cell segmentation was performed with CellProfiler[7] 3.1.8 identifying nucleus and membrane of each cell in each region. First, the two DNA channels were multiplied and used for nucleus segmentation using local Otsu thresholding. Second, all channels for the membrane markers (CD3, CD20, CD27, CD16, CD11c, CD15, SMA, CD34, Vimentin and Pan-keratin for UH1-UH16 and CD45, Pan-keratin and E-cadherin for UH17-UH29) were used to obtain membrane images. Cell masks were then generated by radially expanding each nucleus up to 10 pixels on the membrane mask and only cells overlapping with the tissue mask were retained. Finally, the mean intensity of all markers was measured in each cell in each region.

Cell identities were assigned according to the maximum overlap of the cell area with marker-specific thresholds identified by the histologist (J.S.) after image manual inspection. These thresholds were: ≥25% of CD3$^+$ mask for T cells; ≥10% of CD11c$^+$ CD68- mask for dendritic cells; >10% of the sum of CD68$^+$ CD11c$^+$ and CD68$^+$ CD11c- masks for macrophages; ≥5% of IgA$^+$, IgM$^+$, CD20$^+$, and CD27$^+$ mask for B cells; and ≥25% of CD15$^+$ mask for neutrophils. Cells that did not overlap with any of these markers were defined as tumour cells if they overlapped ≥80% with the tumour mask or were left unassigned otherwise. Within CD3$^+$ cells, PD1$^+$ cells were identified as those showing ≥1% overlap with the PD1 mask. PDL1$^+$ cells were identified as those overlapping ≥10% of the PDL1 mask.

For the discovery cohort, single cell phenotype clustering was performed for T cells, B cells, macrophages, dendritic cells, neutrophils, PD1$^+$ and PDL1$^+$ cells separately using Seurat[9] 2.4 with custom R scripts for IMC data analysis. Independent clustering was used to compare the relative abundance of cell subpopulations between hypermutated and non-hypermutated CRCs or DB- and nDB-CRCs using Pembrolizumab and Nivolumab samples alone or combined. The total number of cells used in each clustering is shown in Supplementary Table 7. For each main population, the clustering was based on the mean expression of a set of markers typical of that population (Supplementary Table 5). The mean marker intensities across all cells were integrated using multiple canonical correlation analysis (CCA) and aligning the CCA subspaces to reduce the inter-sample variability. The resulting CCA vectors were then used as input for unsupervised clustering with ten values of resolution, ranging from 0.1 to 1.0. The ten resulting sets of clusters were manually inspected and the one with the highest number of biologically meaningful clusters was chosen. For the validation cohort, T cells, macrophages, PD1$^+$ and PDL1$^+$ cells were identified as described above. CD74$^+$ macrophages and CD8$^+$GzB$^+$ CD8$^+$Ki67$^+$ T cells were identified using specific expression thresholds on the mean cell intensity (0.1 for CD74; 0.1 for CD8; 0.05 for GzB; 0.15 for Ki67). CD8$^+$ T cells positive for both the Ki67 and the GzB threshold were identified as CD8$^+$GzB$^+$ or CD8$^+$Ki67$^+$ T cells according the marker with the highest intensity. T cells in the validation cohort underwent single cell clustering, using all 17 T cell markers (Supplementary Table 5). The distribution of cells within each cluster over the total cells was compared between DB- and nDB-CRCs or hypermutated and non-hypermutated CRCs using two-sided Wilcoxon rank sum test, correcting for FDR. All comparisons are shown in Supplementary Table 8.

For the 17 regions stained with IMC panel III, T cells, macrophages, dendritic cells and tumour cells were identified as described above (Supplementary Table 7). CD74$^+$ macrophages were identified using a threshold of 0.35 on the mean cell intensity. Single-cell clustering of the identified CD74$^+$ macrophages was performed using using Seurat[9] 2.4 with 16 macrophage markers (Supplementary Table 5).

IMC neighbour and cluster density analysis

The pixel coordinates of the centroid of each cell were extracted from the cell masks with CellProfiler[7] 3.1.8 and used to measure the Euclidean distances between each pair of cells in each region. High-density clusters of CD68$^+$CD74$^+$ within each region were identified using DBSCAN (Density-Based Spatial Clustering of Applications with Noise[10]) as implemented in the fpc R package version 2.2.5. Starting form cell pixel coordinates, highly dense clusters were defined as portions of the ablated regions with ≥5 CD68$^+$CD74$^+$per 10,000μm^2, corresponding to a minimum number of five points (MinPts) within a radius (eps) of 56.42μm.

Multiplexed Immunofluorescence (mIF)

mIF was performed on slide D of 24 DB- and nDB samples (Supplementary Table 2). An automated Opal-based mIF staining protocol was developed using a Ventana Discovery Ultra automated staining platform (Roche) with eight markers specific for CD8$^+$GzB$^+$PD1$^+$, CD8$^+$Ki67$^+$PD1$^+$ and CD68$^+$CD74$^+$PDL1$^+$ cells, DAPI and Opal fluorophores (Supplementary Table 5). Antibody dilution, incubation time and effect of denaturation steps as well as Opal dilution were assessed for each marker following manufacturer's instructions. The optimal antibody-Opal pairing was achieved considering the expected expression and cellular localisation of each marker and the fluorophore brightness to minimize fluorescence spillage. The final staining order was CD74, TCF7, PDL1, Ki67, PD1, GzB, CD68 and CD8.

Slides were baked for 1 hr at 60°C, loaded onto the autostainer and subjected to a fully automated staining protocol involving deparaffinisation (EZ-Prep solution, Roche), HIER (DISC. CC1 solution, Roche) and seven sequential rounds of 1 hr incubation with the primary antibody, 12 minutes incubation with the HRP-conjugated secondary antibody (DISC. Omnimap anti-Ms HRP RUO or DISC. Omnimap anti-Rb HRP RUO, Roche) and 16 minute incubation with the Opal reactive fluorophore (Akoya Biosciences). For the last round of staining, tissues were incubated with Opal TSA-DIG reagent (Akoya Biosciences) for 12 minutes and with Opal 780 reactive fluorophore for 1 hour (Akoya Biosciences). Before each round of staining, a denaturation step (100°C for 8 minutes) was introduced to remove the primary and secondary antibodies from the previous cycle without disrupting the fluorescent signal. Once the staining was completed, the slides were counterstained with 4',6-diamidino-2-phenylindole (DAPI, Akoya Biosciences) and coverslipped using ProLong Gold antifade mounting media (Thermo Fisher Scientific). Fluorescently labelled slides were scanned using a Vectra Polaris automated quantitative pathology imaging system (Akoya Biosciences). Spectral libraries were constructed with the inForm 2.4 image analysis software (Akoya Biosciences) following the manufacturer's instructions. Whole-slide scans were obtained at 20x and 40x magnification using appropriate exposure times, and several fields of views were selected per slide and loaded into inForm (Akoya Biosciences) for spectral unmixing and autofluorescence isolation using the spectral libraries.

DNA sequencing

All regions were macro-dissected with a needle under a stereo microscope using slide A as a guide (Supplementary Figure 1G). Genomic DNA was extracted from 32 tumour regions and 16 matched normal tissue of slides E1-5 of samples UH1-UH16 (Supplementary Table 2) and the two regions corresponding to those used for IMC were selected. For UH4, the two high and low CD3$^+$ regions were merged to obtain enough DNA for library preparation. DNA was extracted using GeneRead DNA FFPE kit (Qiagen) and DNA libraries were prepared using 50-200ng of genomic DNA with the KAPA HyperPrep kit (Roche). Protein-coding genes were captured using SureSelectXT Human All Exon V5 probes (Agilent) and sequenced on Illumina HiSeq 4000 using 100bp paired end read protocol, according to manufacturer's instructions. Approximately 100 million reads were generated per sample.

Raw reads were aligned to GRCh38 reference human genome using BWA MEM[11] v0.7.15 after pre-alignment quality control. Regions harbouring small insertions and deletions (indels) were re-aligned locally using GATK[12] v3.6. The resulting BAM files were sorted, merged, marked for duplicates and subjected to post-alignment quality control using Picard v2.10.1. The final mean depth of coverage was >70x for tumour and >30x for normal samples, considering only targeted exons as defined in the SureSelectXT BED file (50.5Mbp in total).

Somatic SNVs and indels were called using Strelka[13] v2.9.0 on the targeted exome extended 100bp in both directions. Mutations were retained if they had an Empirical Variant Scoring (EVS) >7 for SNVs and >6 for indels in at least one region of the same patient. Mean sensitivity in variant calling was >91% in all patients expect UH16 (33%) as assessed using 241 somatic mutations from FM1[14] or the patient pathological reports for comparison. Nineteen mutations in FM1 or pathological reports but missed by Strelka were added to the pool of somatic alterations after manual check.

For samples UH1-UH3, UH7-UH10 and UH12, copy number analysis was done using ASCAT[15] v2.5.2. To process WES data, AlleleCount[16] v4.0.0 was run on germline SNPs from 1000 Genomes Phase 3[17] after correction for GC bias. For each SNP, a custom script was used to calculate the LogR and B-Allele Frequency (BAF). SNPs with <6 reads were filtered out in all samples except UH1 and UH10 where 5 or 7 reads were used. Because of degraded starting DNA of samples UH4-UH6, UH11 and UH13-UH16, the DepthOfCoverage option of GATK[12] v3.6 was used to calculate SNP LogR and copy numbers for genomic segments were obtained using Copynumber R package. The gene copy number was derived from that of the genomic segment covering at least 25% of the gene length.

Prediction of damaged genes and immunogenic mutations

ANNOVAR[18] (release 16/04/2018) was used to annotate exonic or splicing SNVs and indels. All truncating mutations (stop-gains, stop-losses and frameshift indels) were considered as damaging. Non-truncating mutations (non-frameshift indels and missense SNVs) were considered damaging if predicted by at least five function-based methods or two conservation-based methods[19]. Mutations within two bps of a splicing junction were considered as damaging if predicted by at least one ensemble algorithm in dbNSFP. Gain of function mutations were predicted using OncodriveClust[20] for the discovery patients together, with default parameters and with false discovery rate (FDR) <10%.

A gene was considered amplified if its copy number was >1.4 times the sample ploidy or >2 if the ploidy was not available in both regions of the same patient, or if its CPM was >1.5 than in the other region of the same patient. A gene was considered as deleted if it had copy number = 0, CPM = 0 and had no mutations. A gene was considered deleted in heterozygosity if it had copy number = 1.

Amplified genes, deleted genes, heterozygously deleted genes with a damaging mutation in the other allele and copy number neutral genes with at least one damaging mutation were considered as damaged genes.

To predict putative immunogenic mutations from all somatic SNVs and indels, HLA typing of each patient was predicted from the normal BAM files using Polysolver[21] v4 (Supplementary Table 9). NeoPredPipe[22] was then used to predict neoantigens in expressed genes (CPM >0), with a strong HLA binding (rank <0.5%) and cross-referenced with known epitopes (UniProt reference proteome). SNVs or indels generating at least one neoantigen were considered as potentially immunogenic. The neoantigenic index was calculated for each region as:

$$Neoantigenic\ index = \frac{number\ of\ immunogenic\ mutations}{number\ of\ nonsilent\ mutations}$$

PyClone[23] v.0.13.1 was run to assess the clonality of predicted immunogenic mutations, defined as the proportion of tumour cells harbouring the mutation. PyClone was run independently for each region using tumour purity from the pathological assessment of slide A (Supplementary Table 2) and the gene copy number from ASCAT.

RNA sequencing

Total RNA was extracted from 58 macro-dissected regions of slides G1-5 in samples UH1-UH16 and 30 regions of slides F1-5 in samples UH17-UH23 and UH25-

UH27 (Supplementary Table 2) using the High Pure FFPE RNA isolation kit (Roche). For UH4, the two high and low CD3^{+} regions were merged to obtain enough RNA for library preparation. RNA libraries were prepared starting from 5-50 ng of RNA using the QuantSeq 3'mRNA-seq Library Prep kit FWD for Illumina (Lexogen) and sequenced on Illumina HiSeq 4000 using 75 or 100 bp single end reads, according to manufacturer's instructions. Approximately 5-40 million reads were generated per sample.

Raw reads were processed using the Lexogen QuantSeq 3' mRNA-seq pipeline with default parameters[24]. Reads were first trimmed to remove Illumina adapters and polyA tails using bbduk from BBMap[25] v36.20. Trimmed reads were then aligned to GRCh38 reference human genome using STAR[26] v2.5.2a. Between 50-98% of the initial reads were retained after alignment and quality check. Gene expression was quantified using HTSeq[27] v0.6.1p1 and the GDC h38 GENCODE v22 GTF annotation file. To account for differences in sequencing depth across regions raw counts were normalised to the counts-per-million (CPM) gene expression unit calculated as:

$$CPM_{ij} = \frac{RC_{ij}}{\sum_k RC_{ik}} \times 10^6$$

where RC_{ij} is the raw read count of gene (j) in region (i). Since for samples UH1-UH16 RNA-seq was performed in six batches, potential batch effect was corrected using removeBatchEffect function from the Limma package[28] v3.36.5 on the log2-transformed CPM matrix with default parameters.

Differential gene expression between tumour groups was assessed using DESeq2[29] v1.20.0 from the raw read counts with default parameters with alpha set to 5%. To account for the experimental and clinical variability across samples (Supplementary Table 1), uncorrelated batch and clinical co-variates were included in the analysis (discovery cohort: batch effect, prior lines of treatment and Lynch syndrome in the comparison of DB- and nDB-CRCs of the; batch effect in the comparison of hypermutated and non-hypermutated CRCs. Validation cohort: prior lines of treatment, Lynch syndrome, treatment type and TNM staging).

A gene was considered as differentially expressed if DESEq2 Wald test FDR was <5% and had a fold change |>2|. Differentially expressed genes were used for pathway enrichment analysis in the three comparisons using MetaCore v20.3 build 70200 (Clarivate Analytics). Pathway enrichment was assessed through overrepresentation analysis based on a hypergeometric test. A pathway was considered enriched if the FDR was <10%.

TCR sequencing

TCR-seq was performed in 28 macro-dissected regions in slides I1-5 of the discovery cohort, after excluding UH1, UH4 and UH5 because DNA was not sufficient (Supplementary Table 2). For UH12 all four regions were sequenced while for UH16 the two high and low CD3$^+$ regions were merged. For all remaining samples, the two regions corresponding to those used for IMC and WES were used.

DNA was extracted from macro-dissected regions (Supplementary Figure 1G) using GeneRead DNA FFPE kit (Qiagen) and submitted to Adaptive Biotechnologies (Seattle, USA) for non-lymphoid tissue (survey level) TCR-seq using a two-step, amplification bias-controlled multiplex PCR approach[30]. In the first step, V and J gene segments encoding the TCR beta CDR3 locus were amplified using reference gene primers to quantify total nucleated cells and measure the fraction of T cells in each sample. In the second step, proprietary barcodes and Illumina adapters were added. Finally, CDR3 and reference gene libraries were sequenced according to the manufacturer's instructions.

Raw reads were de-multiplexed and processed to remove adapter and primer sequences, identify and remove primer dimer, germline and other contaminant sequences. Resulting reads were clustered using both the relative frequency ratio between similar clones and a modified nearest-neighbour algorithm, to merge closely related sequences to correct for technical errors introduced through PCR and sequencing. The resulting reads were sufficient for annotating the V(N)D(N)J genes of each unique CDR3 and the translation of the encoded CDR3 amino acid sequence. V, D and J gene definitions were based on annotation in accordance with the IMGT database (www.imgt.org). The set of observed biological TCR Beta CDR3 sequences were normalised to correct for residual multiplex PCR amplification bias and quantified against a set of synthetic TCR Beta CDR3 sequence analogues[30]. Data was analysed using the immunoSEQ Analyzer toolset.

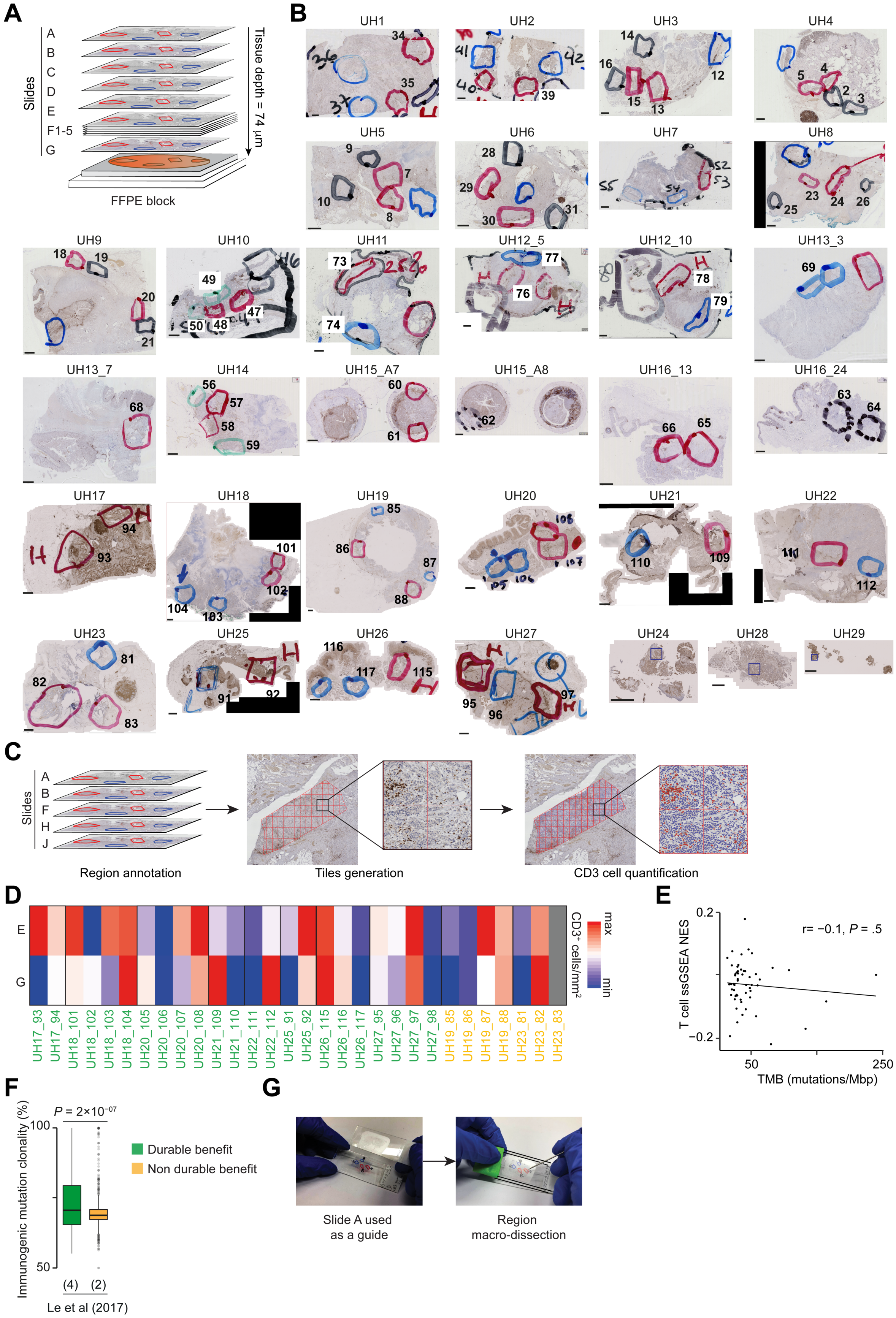
Detection of PD1-PDL1 interaction *in situ*

A total of 58 regions in slides K1-2 of the discovery cohort (Supplementary Table 2) were submitted to FASTBASE Solutions (Derio, Spain) to measure the interaction between PD1 and PDL1 *in situ* via amplified Förster Resonance Energy Transfer (A-FRET)[31]. Slides K1 were incubated overnight at 4 ºC with anti PD1 primary antibody (Supplementary Table 5) for donor only analyses. Slides K2 were stained with both anti PD1 and anti PDL1 primary antibodies for donor and acceptor analyses. Slides K1 were subsequently incubated with anti-mouse Fab-ATTO488 and slides K2 with both anti-mouse Fab-ATTO488 and anti-rabbit Fab-HRP. Alexa594 conjugated tyramide was added to slides K1 and K2 at 1/100 dilution in presence of 0,15% H_2O_2 and incubated at room temperature in the dark for 20 minutes. After washing in PBS and PBST twice, slides were mounted using Prolong Diamond Antifade Mount (Thermo Fisher), sealed and incubated at room temperature overnight before being transferred to a 4 ºC refrigerator for storage. FASTBASE Solutions SL frequency domain FLIM automated software programme was used to measure the excited-state lifetime of donor fluorophore (ATTO488) in both K1 and K2 slides. FRET efficiency E.was calculated as:

$$E = [1 - (tDA/tD)\ x100]$$

where *tDA* is donor lifetime in Slide K1 and *tD* is donor lifetime in Slide K2. *tDA* and *tD* values were collected for 793 optical fields of view (FOVs, with a median of 12 FOVs per region) in total to cover the whole surface of the regions analysed. Data were then collected in .csv files and imported into a macro spreadsheet programmed to calculate the A-FRET efficiency in each FOV. The results were finally expressed as the median FOV values per region.

Supplementary Figure 1. Experimental workflow, region selection and CD3 quantification



A. Experimental workflow for the analysis of the validation cohort. Eleven sequential sections from FFPE blocks of UH17-UH23 and UH25-UH27 were used for multiregional CD3 IHC (slides A, E and G), IMC with panel II (Table S5, slide C), mIF (slide D) and RNA-seq (slides F1-5). For the three biopsies (UH24, UH28, UH29) only IHC, IMC and mIF were performed.

B. Selected regions on slide A**.** Starting from slide A, CD3 staining was used by a board-certified pathologist (M.J.S.) to select multiple regions per sample with variable CD3 infiltrates and at the invasive margins of the tumour. For UH12, UH13, UH15, UH16 two different blocks were used. For biopsies (UH24, UH28 and UH29) shown are the IMC-ablated regions. Scale bar = 2mm.

C. Schematic of CD3 quantification. All slides immune-stained with anti-CD3 antibody were imported into QuPath[4]. Regions selected by the pathologist in slide A were projected into all other immune-stained slides and divided into 0.09 mm^2 tiles. CD3$^+$ cells were quantified within each tile and the CD3 content of a region was defined as the median number of CD3$^+$ cells per mm^2 across all tiles.

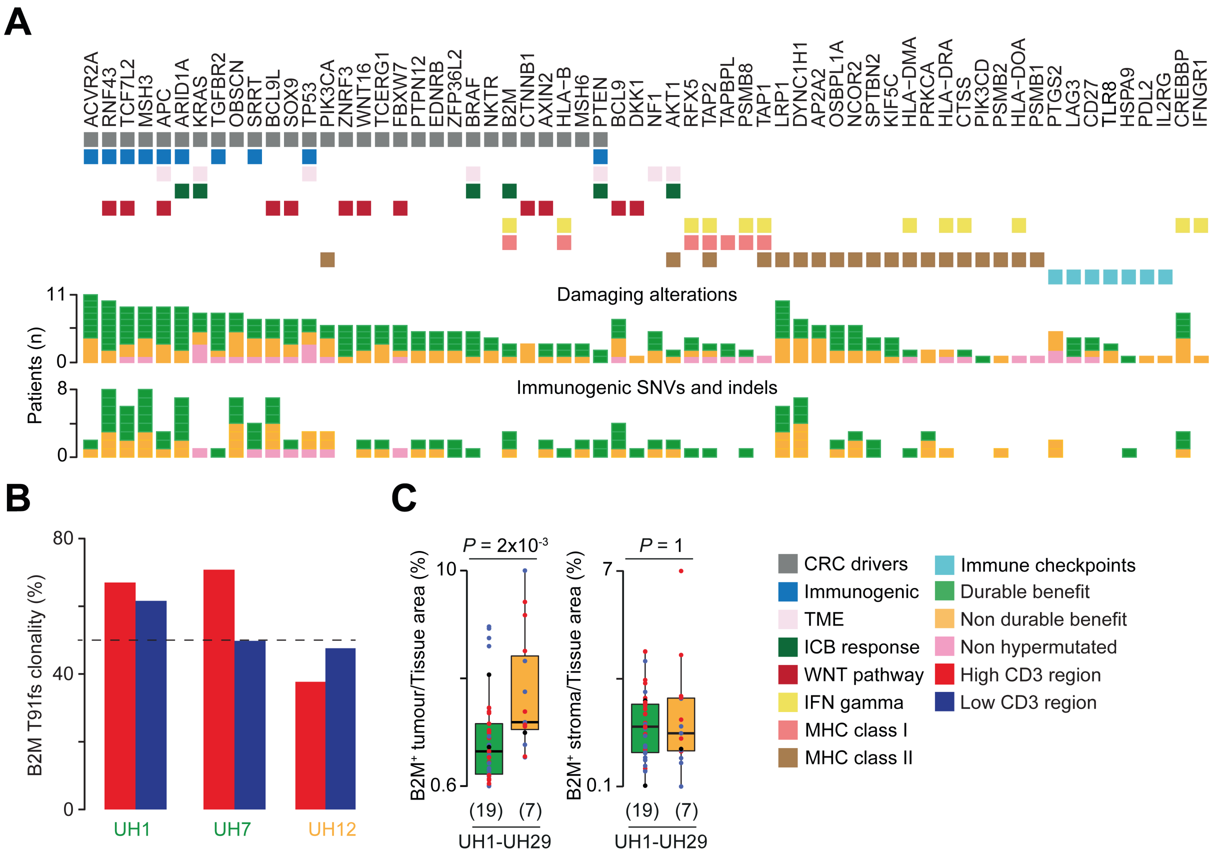
D. Quantification of CD3$^+$ cells/mm^2 from IHC staining in slides E and G in 30 regions of patients from the validation cohort using Qupath[4]. Values were normalised within each patient. Grey boxes indicate missing values.

E. Correlation between the T cell signature normalised enrichment scores (NES) from[32] and TMB in 56 hypermutated CRCs from TCGA. Pearson correlation coefficient and associated p-value are shown. ssGSEA, single sample gene set enrichment analysis.

F. Comparison of clonality of immunogenic mutations between four DB and two nDB-CRCs with purity >30% from[33]. For two samples (Subjects 33 and 36) the reference counts were randomly sampled from the rest of samples and the variant counts were subsequently calculated. Given the lack of copy number data, PyClone[23] was ran with minor_cn=0, major_cn=2, prior=total_copy_number. Due to lack of expression data, immunogenic mutations were considered as expressed if contained in 11,056 genes expressed in >30% of TCGA CRCs.

G. Macro-dissection for DNA and RNA extraction. Regions selected in slide A were used as a reference for the macro-dissection of all slides used for DNA and RNA extraction. Each slide was aligned to slide A using a stereo microscope and regions were manually dissected with a needle. The collected tissue was subsequently used for DNA or RNA extraction

Supplementary Figure 2. Genetic and TME features in DB and nDB-CRCs



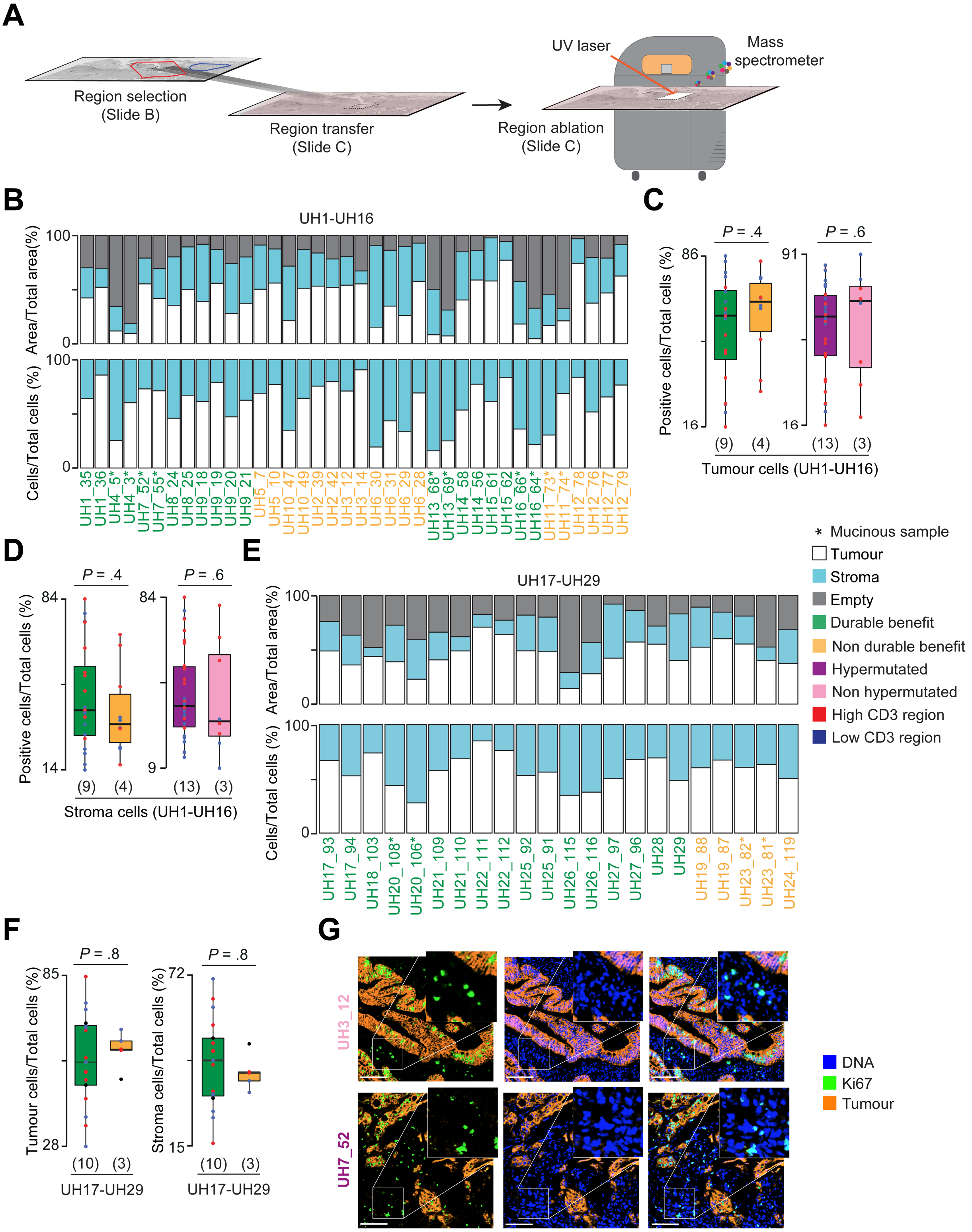
A. Predicted damaging alterations (truncating and missense damaging alterations, double hits, gene amplifications leading to increased expression, gene homozygous deletions) and immunogenic mutations of representative genes from a manually curated list of 647 genes including common CRC drivers[34]; genes whose alterations are immunogenic[35, 36]; genes that modify the TME[37-41], modulate the response to immune checkpoint inhibitors[38-40, 42-45], or encode members of WNT[41, 46] and IFN-gamma pathways (MetaCore Clarivate Analytics), components of the antigen presentation machinery via the major histocompatibility complex (MHC) class I[47, 48] or class II (MetaCore, Clarivate Analytics), and immune checkpoints[49, 50].

B. Clonality of *B2M* truncating mutation (T91fs) in all sequenced regions from two DB and one nDB-CRCs. Clonality was measured using Pyclone[23] after correction for purity and copy number alterations (Methods).

C. Comparison of tumour and stroma B2M$^+$ areas between DB- and nDB-CRCs in all analysed samples. Distributions were compared using two-sided Wilcoxon rank sum test and the number of patients in each group is reported in brackets.

TME, tumour microenvironment; ICB, Immune Checkpoint Blockade; IFN, Interferon; MHC, Major Histocompatibility Complex; IC, Immune Checkpoints.

Supplementary Figure 3. Proportion of tumour and stroma from IMC across samples



A. IMC experimental workflow. Representative 1mm^2 areas in slide B were projected into slide C using the macroscopic tissue structure as a reference. Slide D stained with the IMC antibody panel was loaded in the Hyperion Imaging System (Fluidigm) for regional ablation.

B. Proportions of tumour areas and cells in ablated regions of the discovery cohort. Areas not covered by stroma or tumour are depicted in grey.

Comparison of the proportion of tumour **(C)** and stroma cells **(D)** over total cells between DB and nDB-CRCs or hypermutated and non-hypermutated CRCs in the discovery cohort.

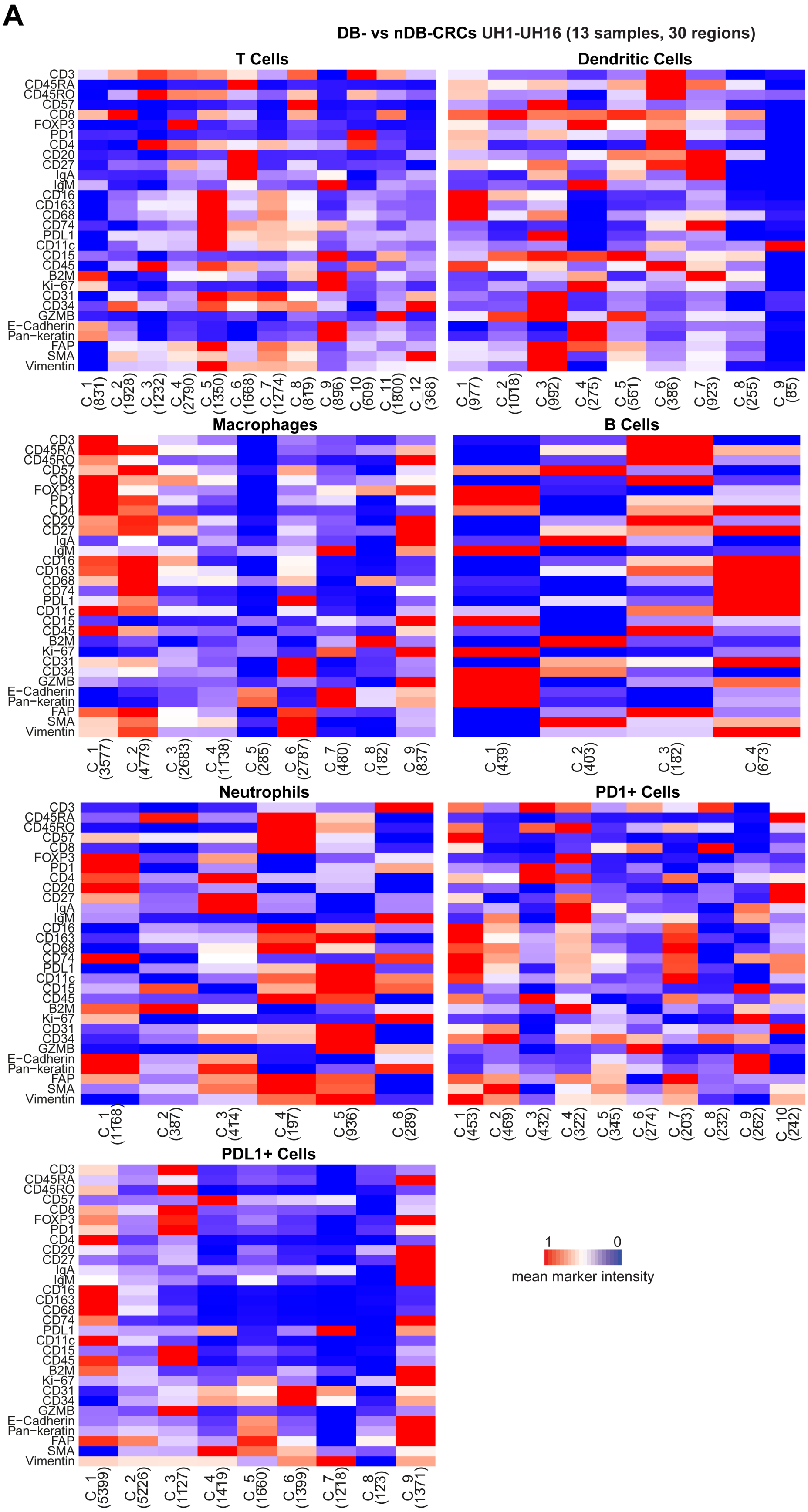
E. Proportions of tumour areas and cells in ablated regions of the validation cohort. Areas not covered by stroma or tumour are depicted in grey.

F. Comparison of the proportion of tumour and stroma cells over total cells between DB and nDB-CRCs in the validation cohort.

All distributions were compared using two-sided Wilcoxon rank sum test and the number of patients in each group is reported in brackets.

G. IMC-derived images of tumour-associated markers (E-cadherin and Pan-Keratin), Ki67 and DNA staining in two representative hypermutated and non-hypermutated CRCs. Scale bar = 100μm.

Supplementary Figure 4. Protein expression heatmaps from single cell clustering



T cells

C_1	CD3, CD8, Ki−67
C_2	CD3, CD8
C_3	CD3, CD45RO, CD4
C_4	CD3, CD45RO, FOXP3, CD4
C_5	CD3, CD45RO, CD8, CD4, CD16
C_6	CD3, CD45RA, CD8, CD4
C_7	CD4
C_8	CD3, CD45RO, CD57, CD8
C_9	E−Cadherin, Ki−67
C_10	CD3, CD45RO, PD1, CD4
C_11	CD3, CD8, GZMB
C_12	CD3

Dendritic Cells

C_1	CD16
C_2	CD16, CD11c
C_3	Vimentin
C_4	Pan−keratin, E−Cadherin, Ki−67
C_5	Pan−keratin, E−Cadherin, CD11c
C_6	CD4, CD11c
C_7	CD74
C_8	CD16, CD163, CD11c
C_9	CD11c

Macrophages

C_1	CD4, CD16, CD163, CD68
C_2	Vimentin, CD16, CD163, CD68 CD74
C_3	CD16, CD68
C_4	CD68
C_5	Pan−keratin, E−Cadherin, CD68
C_6	Vimentin, CD68
C_7	Pan−keratin, E−Cadherin
C_8	Pan−keratin, CD68
C_9	Pan−keratin

B Cells

C_1	Pan−keratin, E−Cadherin, Ki−67
C_2	CD45RA
C_3	CD45RA, CD20, CD27
C_4	CD45RA, CD27

Neutrophils

C_1	Pan−keratin, E−Cadherin, CD15, Ki−67
C_2	CD15
C_3	Pan−keratin, E−Cadherin, CD15
C_4	CD45RO, CD15
C_5	CD15, GZMB
C_6	Pan−keratin, CD15, Ki−67, GZMB

PD1+ Cells

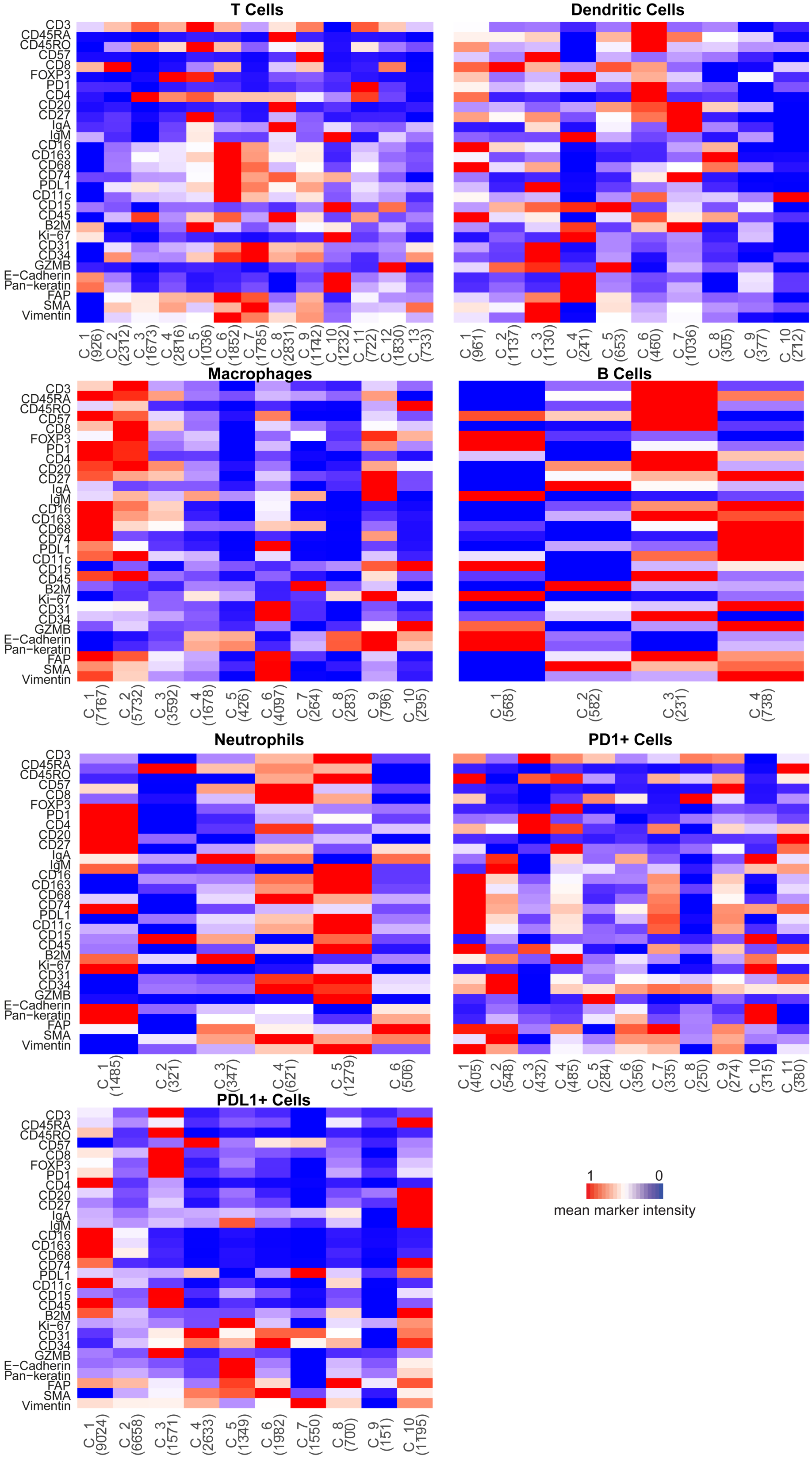
C_1	CD3, CD45RO, CD57, CD8, CD4, CD16
C_2	CD3, CD4
C_3	CD3, CD45RO, PD1, CD4
C_4	CD3, CD45RO, FOXP3, CD4
C_5	CD3, CD8
C_6	CD3, CD45RO, CD8, GZMB
C_7	CD3, CD45RO, CD4, CD16
C_8	CD3, CD45RO, CD8
C_9	CD4, E−Cadherin, Ki−67
C_10	CD3, CD45RA, CD45RO, CD4

PDL1+ Cells

C_1	Vimentin, CD16, CD163, CD68, CD74, PDL1
C_2	FAP, SMA, Vimentin, CD16, PDL1
C_3	CD45RO, SMA, Vimentin, PDL1
C_4	SMA, Vimentin, PDL1
C_5	SMA
C_6	SMA, Vimentin, CD31
C_7	Vimentin, PDL1, CD31
C_8	FAP
C_9	Vimentin, CD74, PDL1

B

Hypermutated vs non-Hypermutated-CRCs UH1-UH16 (16 samples, 38 regions)



T cells

C_1 CD3, CD8, Ki−67
C_2 CD3, CD8
C_3 CD3, CD45RO, CD4
C_4 CD3, CD45RO, FOXP3, CD4
C_5 CD3, CD45RO, FOXP3, CD4, CD27
C_6 CD3, CD45RO, CD8, CD4, CD16
C_7 CD4
C_8 CD3, CD45RO, CD45RA, CD8, CD4
C_9 CD3, CD45RO, CD57, CD8, CD4
C_10 E−Cadherin, Ki−67
C_11 CD3, CD45RO, PD1, CD4
C_12 CD3, CD8, GZMB
C_13 CD3

Dendritic Cells

C_1 CD4, CD16, CD11c
C_2 CD16, CD11c
C_3 Vimentin
C_4 Pan−keratin, E−Cadherin, Ki−67
C_5 FAP, CD11c
C_6 CD4, CD11c
C_7 CD74
C_8 CD16
C_9 Pan−keratin, E−Cadherin, CD11c
C_10 CD11c

Macrophages

C_1 Vimentin, CD16, CD163, CD68, CD74
C_2 CD16, CD68, CD11c
C_3 CD16, CD68
C_4 Pan−keratin, E−Cadherin, CD68, Ki−67
C_5 Pan−keratin, E−Cadherin
C_6 Vimentin, CD68
C_7 CD68
C_8 Pan−keratin, E−Cadherin, Ki−67
C_9 Pan−keratin, E−Cadherin, CD74
C_10 E−Cadherin

B Cells

C_1 CD45RA, CD27
C_2 Pan−keratin, E−Cadherin, Ki−67
C_3 CD45RA, Vimentin, CD20, CD27
C_4 CD45RA, CD27, CD74

Neutrophils

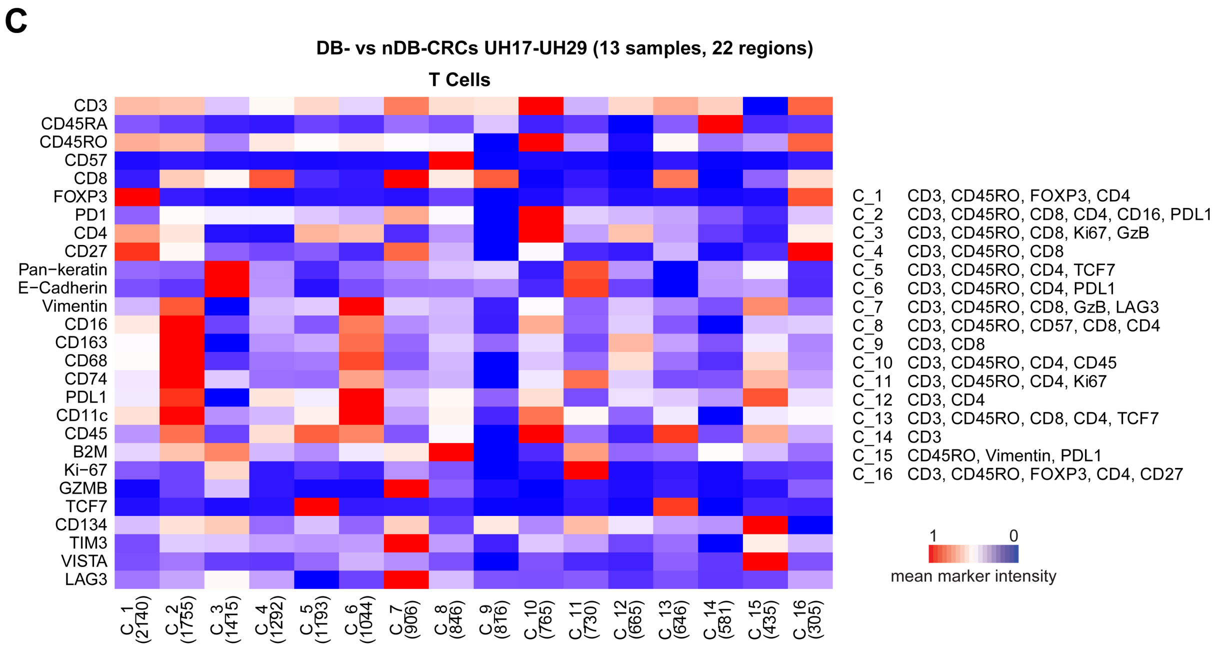
C_1 Pan−keratin, E−Cadherin, CD15, Ki−67
C_2 CD15
C_3 Pan−keratin, CD15
C_4 Pan−keratin, E−Cadherin, Vimentin, CD15
C_5 CD15, GZMB
C_6 Pan−keratin, E−Cadherin, CD15

PD1+ Cells

C_1 CD3, CD45RO, CD8, CD4, CD16, CD45
C_2 CD3, CD4
C_3 CD3, CD45RO, PD1, CD4, CD45
C_4 CD3, CD45RO, FOXP3, CD4
C_5 CD3, CD45RO, CD8, GZMB
C_6 CD3, CD8
C_7 CD3, CD45RO, CD4, CD16
C_8 CD3, CD45RO, CD8
C_9 CD3, CD45RO, CD57, CD8, CD4
C_10 CD4, E−Cadherin, Ki−67
C_11 CD3, CD45RA, CD45RO, CD4, CD45

PDL1+ Cells

C_1 Vimentin, CD16, CD68, PDL1
C_2 FAP, SMA, Vimentin, CD16, CD68
C_3 SMA, Vimentin
C_4 SMA, Vimentin, PDL1, CD31
C_5 SMA
C_6 SMA, Vimentin, CD31
C_7 Vimentin, PDL1, CD31
C_8 Vimentin, PDL1
C_9 FAP
C_10 Vimentin, CD74, PDL1, CD31

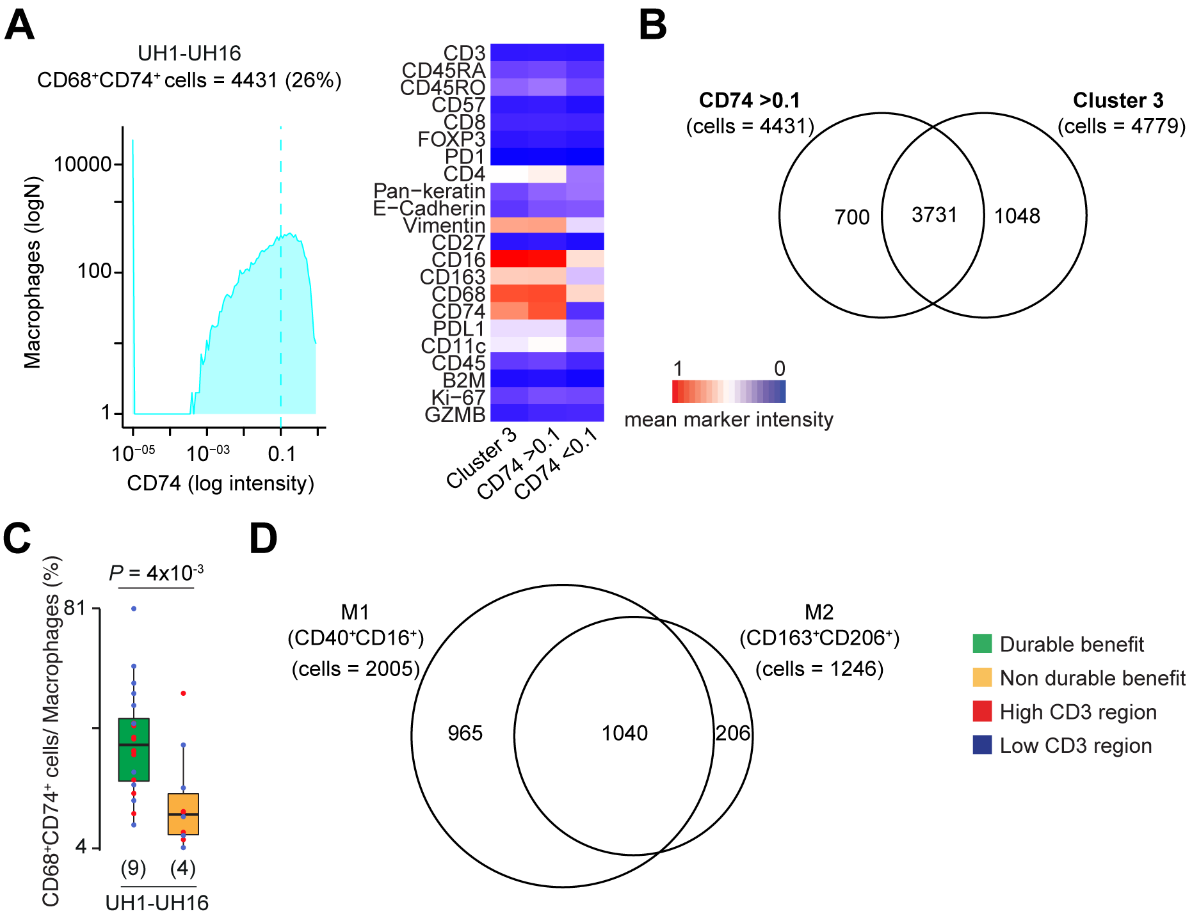


Subpopulations (clusters) of T cells, dendritic cells, macrophages, B cells, neutrophils, PD1^{+} cells, and PDL1^{+} cells were identified based on the expression of phenotypic markers using Seurat[9] in the discovery cohort. For each subpopulation, the mean value of 30 markers of IMC Panel I (Table S5) across the cells in that cluster is reported. For each cluster, the number of cells is reported in brackets. Single cell clustering was performed separately for DB vs nDB-CRCs **(A)** and hypermutated vs non-hypermutated CRCs **(B)**.

C. T cell subpopulations in the validation cohort. In this case, T cells were clustered using 17 markers and the mean value of 27 markers of IMC Panel II (Table S5) across the cells in that cluster is reported.

For each cell population the colour scale was normalized separately for each marker across all analysed clusters.

Supplementary Figure 5. CD74^{+} macrophage identification and phenotyping



A. CD74^{+} macrophages in the discovery cohort identified by applying 0.1 threshold on CD74 expression. The mean intensities of IMC markers in CD74^{+} and CD74^{-} macrophages are reported. Colour gradient was normalised across all markers and cells.

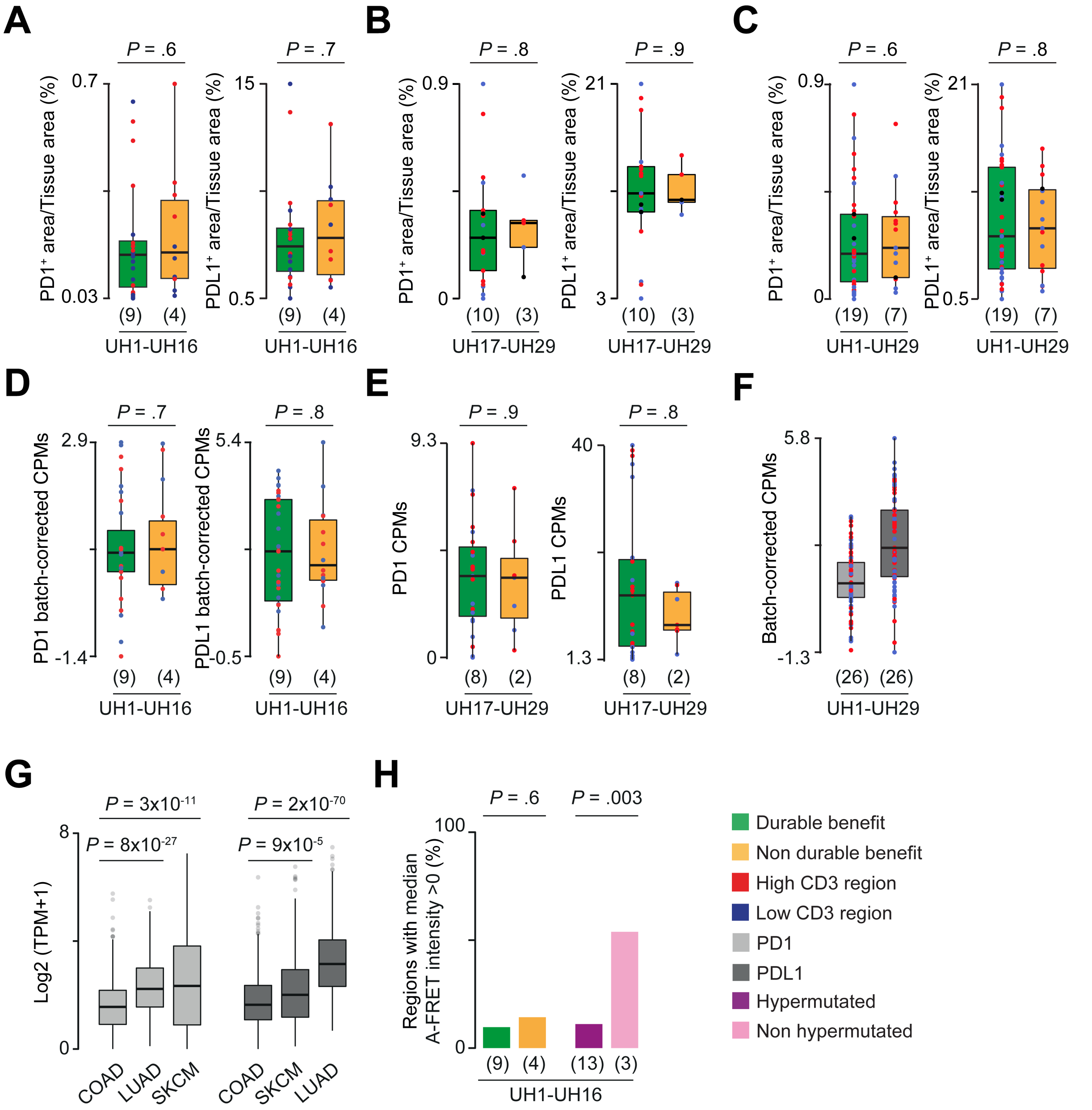
B. Overlap between CD74^{+} macrophages identified with 0.1 CD74 expression and cluster 3 in the discovery cohort.

C. Comparison of CD74^{+} macrophages between DB- and nDB-CRCs in the discovery cohort. CD74^{+} macrophages were identified by applying a threshold of 0.1 CD74 expression to all macrophages.

D. Overlap between CD74^{+} macrophages expressing M1 and M2-associated markers. Used thresholds after histological inspection of IMC images were: 0.2 for CD40; 0.1 for CD16; 0.15 for CD163 and 0.1 for CD206.

All distributions were compared using two-sided Wilcoxon rank sum test and the number of patients in each group is reported in brackets

Supplementary Figure 6. Comparison of PD1 and PDL1 gene and protein expression



Comparison of normalised PD1⁺ and PDL1⁺ areas between DB and nDB-CRCs from the discovery **(A)** and validation **(B)** cohorts.

C. Comparison of normalised PD1⁺ and PDL1⁺ positive areas in all samples analysed. Comparison of *PD1* and *PDL1* Counts Per Million (CPMs) expression levels in the discovery **(D)** and validation **(E)** cohorts. For the discovery samples batch correction was applied (Methods).

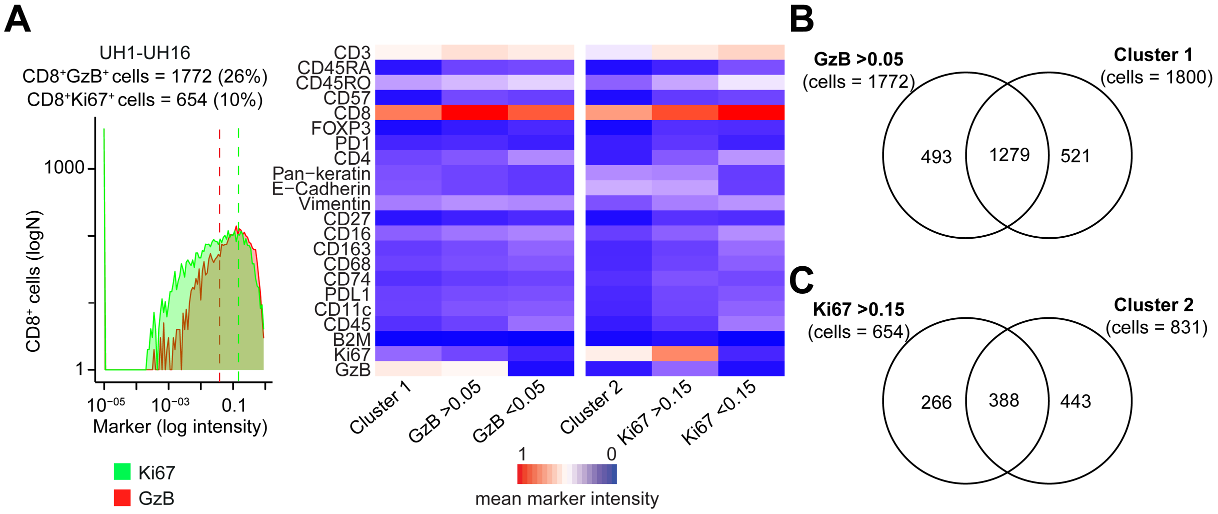
F. *PD1* and *PDL1* gene expression in all regions from all samples analysed. CPMs were obtained from RNAseq raw counts and corrected for batch effects.

G. Comparison of *PD1* and *PDL1* gene expression in colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD), and skin cutaneous melanoma (SKCM) from TCGA. Transcripts Per Million (TPM) were computed from raw read counts.

H. Percentage of regions with median A-FRET intensity higher than zero in DB and nDB-CRCs and hypermutated and non-hypermutated CRCs. Proportions were compared using Fisher's exact test.

All distributions were compared using two-sided Wilcoxon rank sum test and the number of patients in each group is reported in brackets.

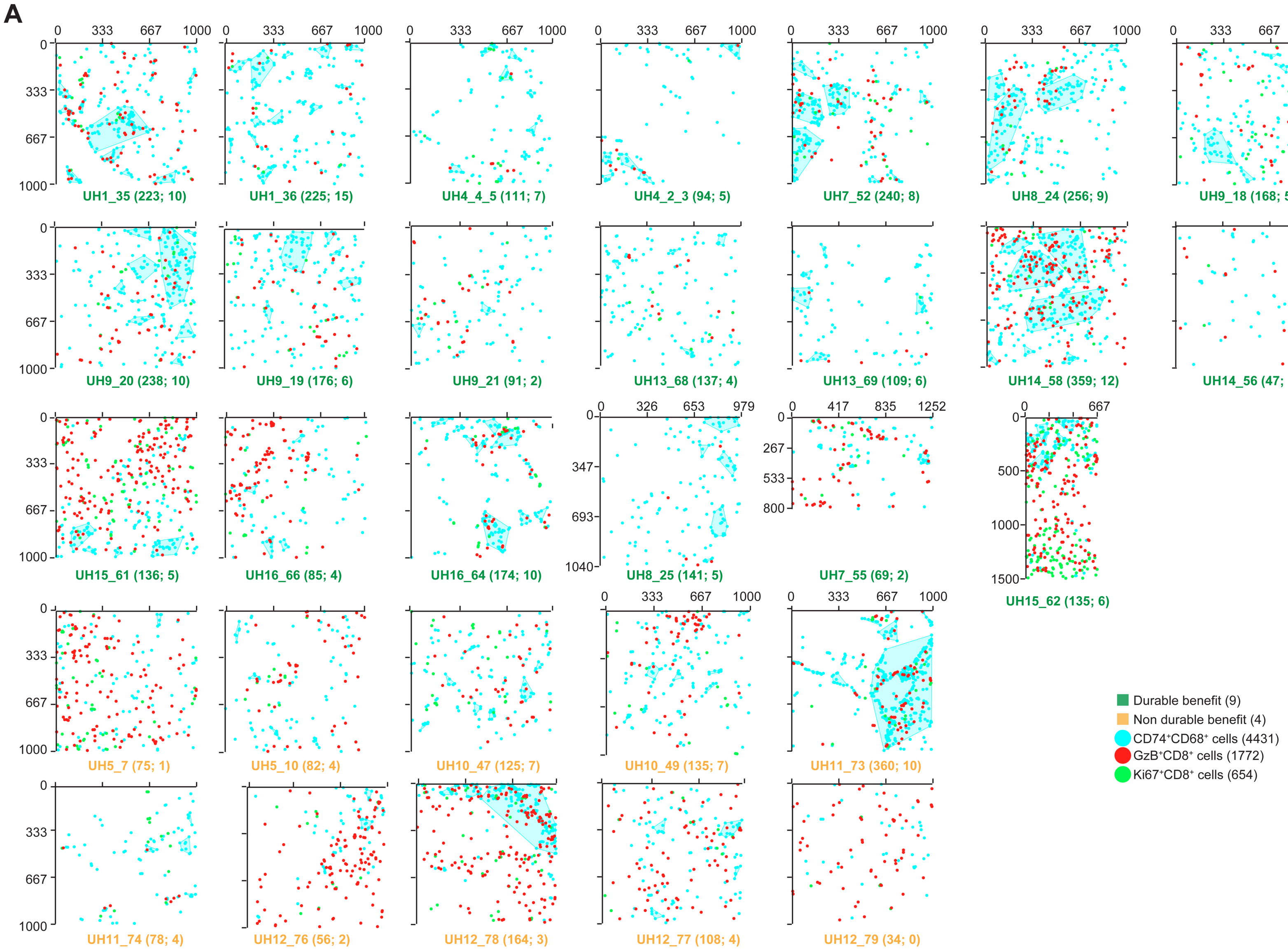
Supplementary Figure 7. Proliferating and cytotoxic CD8 T cell identification

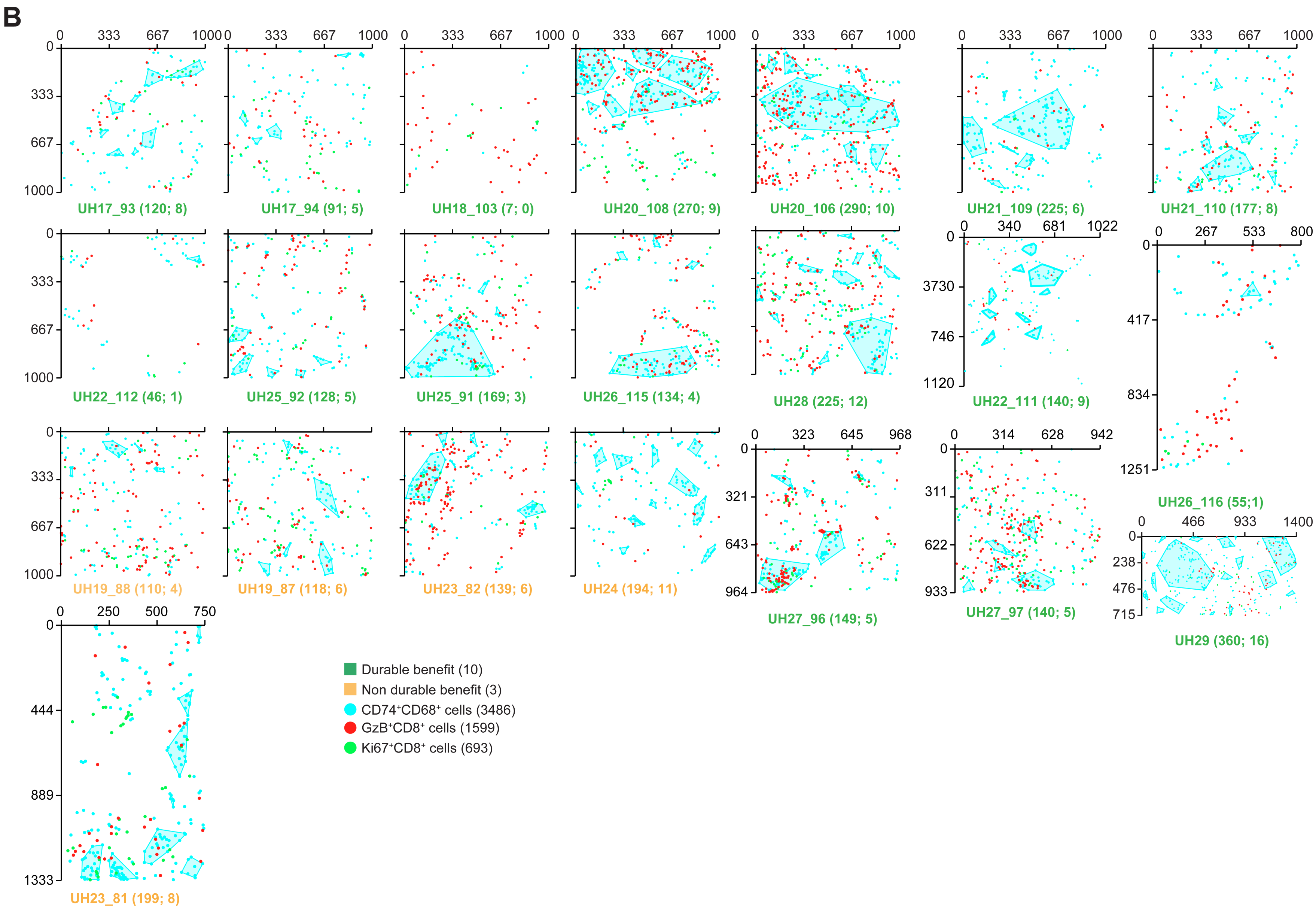


A. CD8$^+$GzB$^+$ and CD8$^+$Ki67$^+$ cells in the discovery cohort identified by applying a threshold of 0.05 GzB and 0.15 Ki67 expression to CD8 T cells, respectively. The mean intensities of IMC markers in CD8$^+$GzB$^+$ or CD8$^+$Ki67$^+$ and CD8$^+$GzB$^-$ or CD8$^+$Ki67$^-$ T cells are reported. Colour scale was normalised across all markers and cells.

Overlap between CD8$^+$GzB$^+$ cells and cluster 1 **(B)** or CD8$^+$Ki67$^+$ cells and cluster 2 **(C)** cells in the discovery cohort. Positive cells were identified using and expression threshold as described in **(A)**.

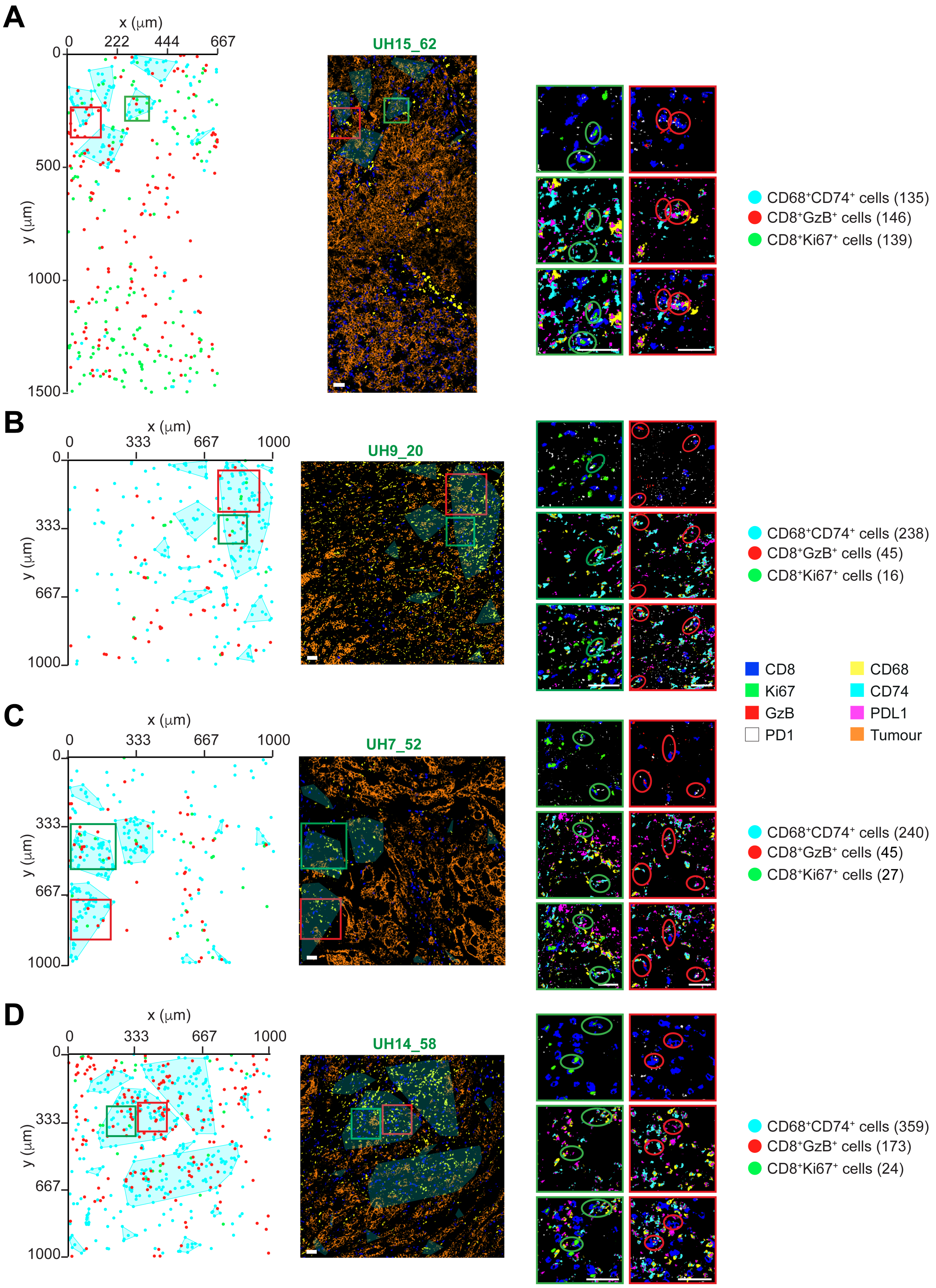
Supplementary Figure 8. High-density CD74$^+$ macrophage clusters

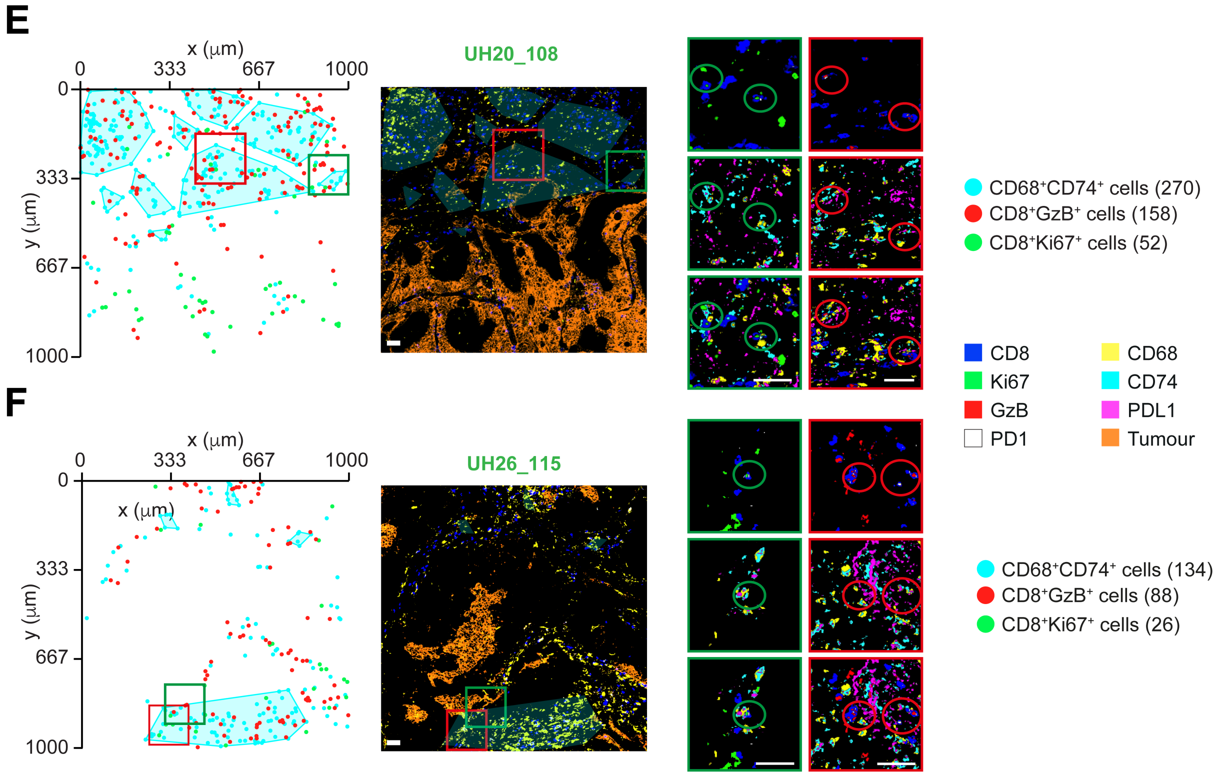



B
0 333 667 1000
UH17_93 (120; 8)
UH17_94 (91; 5)
UH18_103 (7; 0)
UH20_108 (270; 9)
UH20_106 (290; 10)
UH21_109 (225; 6)
UH21_110 (177; 8)
UH22_112 (46; 1)
UH25_92 (128; 5)
UH25_91 (169; 3)
UH26_115 (134; 4)
UH28 (225; 12)
0 340 681 1022
0 3730 746 1120
UH22_111 (140; 9)
0 267 533 800
0 417 834 1251
UH26_116 (55;1)
UH19_88 (110; 4)
UH19_87 (118; 6)
UH23_82 (139; 6)
UH24 (194; 11)
0 323 645 968
0 321 643 964
UH27_96 (149; 5)
0 314 628 942
0 311 622 933
UH27_97 (140; 5)
0 466 933 1400
0 238 476 715
UH29 (360; 16)
0 250 500 750
0 444 889 1333
UH23_81 (199; 8)
🟩 Durable benefit (10)
🟧 Non durable benefit (3)
● CD74⁺CD68⁺ cells (3486)
● GzB⁺CD8⁺ cells (1599)
● Ki67⁺CD8⁺ cells (693)

High-density cluster maps of CD68$^+$CD74$^+$ cells in 52 IMC regions from hypermutated CRCs of the discovery **(A)** and validation **(B)** cohorts. Clusters were identified from cell pixel coordinates as portions of the ablated region with ≥5 CD68$^+$/CD74$^+$ cells per 10,000μm^2 (Methods). CD8$^+$/GzB$^+$ and CD8$^+$/Ki67$^+$ cells were subsequently mapped. The number of CD68$^+$/CD74$^+$ cells and the number of high-density clusters are reported in brackets for each region.

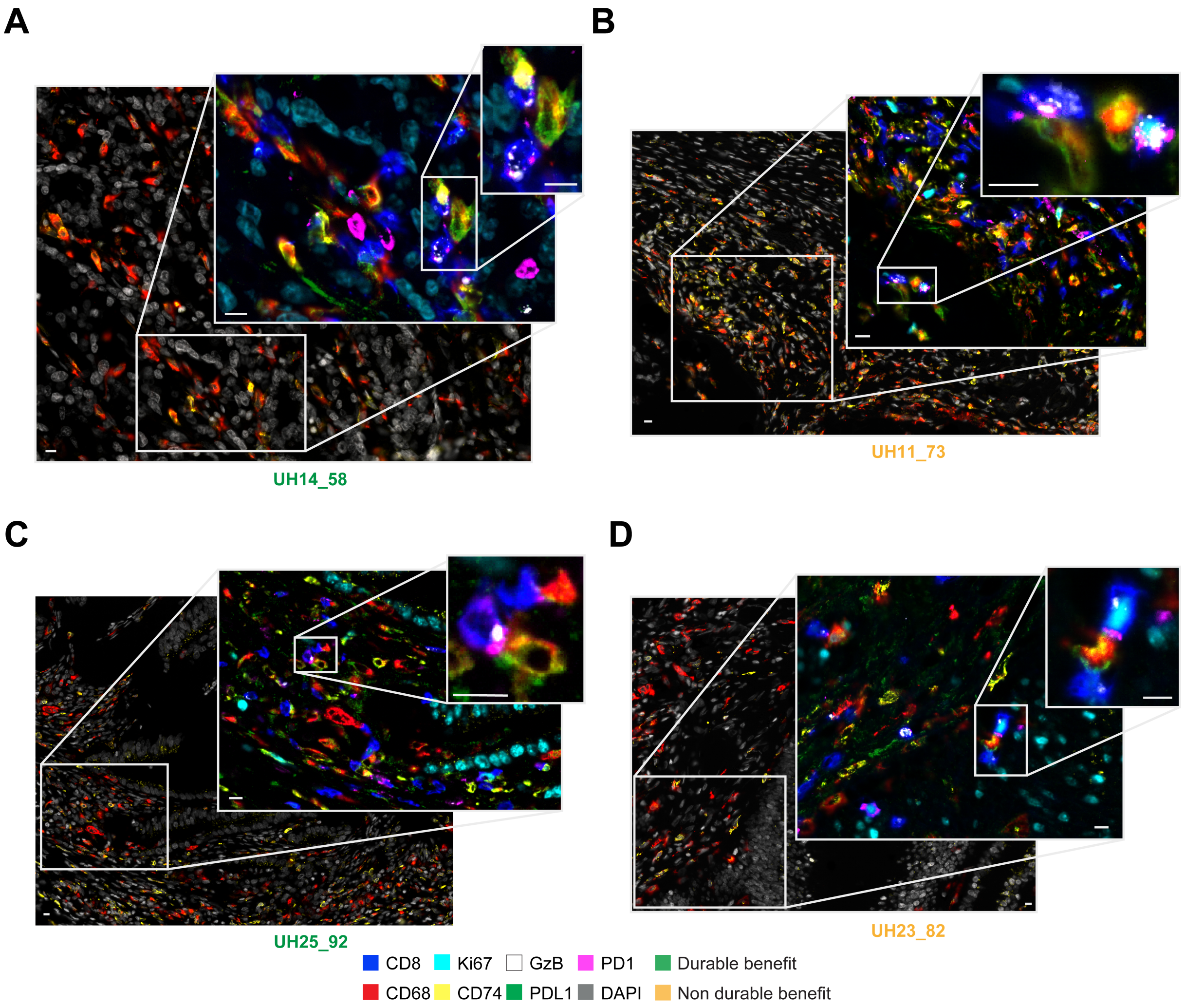
Supplementary Figure 9. Examples of interactions between CD74$^+$ macrophages and cytotoxic or proliferating CD8 T cells by IMC





High-density cluster maps of CD68$^+$/CD74$^+$ cells in representative IMC regions from the discovery **(A-D)** and validation **(E,F)** cohorts. Clusters identified computationally (left panels) as described in Supplementary Figure 8 and in Methods. Red and green squares indicate areas of interest that were identified independently via histological inspection (middle panels). In these area CD68$^+$/CD74$^+$/PDL1$^+$ cells interact with CD8$^+$/Ki67$^+$/PD1$^+$ cells (green) and CD8$^+$/GzB$^+$/PD1$^+$ cells (red). These areas are further detailed (right panels) to show the cellular interactions between CD68$^+$/CD74$^+$/PDL1$^+$ cells and CD8$^+$/Ki67$^+$/PD1$^+$ cells (green circles) and CD8$^+$/GzB$^+$/PD1$^+$ cells (red circles). Images were compiled overlaying single-marker images obtained applying a median filter. For each region, number of cells are reported in brackets. Scale bar = 50μm.

Supplementary Figure 10. Examples of interactions between CD74$^+$ macrophages and cytotoxic or proliferating CD8 T cells by mIF



High resolution (40x) images of cellular interactions between CD68$^+$CD74$^+$PDL1$^+$ cells and CD8$^+$PD1$^+$GzB$^+$ and CD8$^+$PD1$^+$Ki67$^+$GzB$^+$ cells within high-density clusters of CD68$^+$CD74$^+$ cells in representative DB- and nDB-CRCs from the discovery **(A,B)** and validation **(C,D)** cohorts. Scale bar = 10 µm.

References

1. André T, Shiu K-K, Kim TW, et al. Pembrolizumab in Microsatellite-Instability–High Advanced Colorectal Cancer. New England Journal of Medicine 2020;383:2207-2218.
2. Hoos A, Eggermont AM, Janetzki S, et al. Improved endpoints for cancer immunotherapy trials. J Natl Cancer Inst 2010;102:1388-97.
3. Shi Q, de Gramont A, Grothey A, et al. Individual patient data analysis of progression-free survival versus overall survival as a first-line end point for metastatic colorectal cancer in modern randomized trials: findings from the analysis and research in cancers of the digestive system database. J Clin Oncol 2015;33:22-8.
4. Bankhead P, Loughrey MB, Fernandez JA, et al. QuPath: Open source software for digital pathology image analysis. Sci Rep 2017;7:16878.
5. Bortolomeazzi M, Montorsi L, Temelkovski D, et al. A SIMPLI (Single-cell Identification from MultiPLexed Images) approach for spatially resolved tissue phenotyping at single-cell resolution. bioRxiv 2021:2021.04.01.437886.
6. imctools. https://github.com/BodenmillerGroup/imctools, 2018.
7. McQuin C, Goodman A, Chernyshev V, et al. CellProfiler 3.0: Next-generation image processing for biology. PLoS Biol 2018;16:e2005970.
8. Berg S, Kutra D, Kroeger T, et al. ilastik: interactive machine learning for (bio)image analysis. Nat Methods 2019;16:1226-1232.
9. Butler A, Hoffman P, Smibert P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36:411-420.
10. Ester M, Kriegel H-P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining 1996:226–231.
11. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at: https://arxiv.org/abs/1303.3997 2013.
12. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297-303.
13. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods 2018;15:591-594.
14. Frampton GM, Fichtenholtz A, Otto GA, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol 2013;31:1023-31.
15. Van Loo P, Nordgard SH, Lingjaerde OC, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A 2010;107:16910-5.
16. Castel SE, Levy-Moonshine A, Mohammadi P, et al. Tools and best practices for data processing in allelic expression analysis. Genome Biol 2015;16:195.
17. Consortium GP. A global reference for human genetic variation. Nature 2015;526:68-74.
18. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164.
19. Mourikis TP, Benedetti L, Foxall E, et al. Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma. Nat Commun 2019;10:3101.
20. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics 2013;29:2238-44.

21. Shukla SA, Rooney MS, Rajasagi M, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat Biotechnol 2015;33:1152-8.

22. Schenck RO, Lakatos E, Gatenbee C, et al. NeoPredPipe: high-throughput neoantigen prediction and recognition potential pipeline. BMC Bioinform 2019;20:264.

23. Roth A, Khattra J, Yap D, et al. PyClone: statistical inference of clonal population structure in cancer. Nat methods 2014;11:396-398.

24. Moll P, Ante M, Seitz A, et al. QuantSeq 3′ mRNA sequencing for RNA quantification. Nat Methods 2014;11:i-iii.

25. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. https://www.osti.gov/biblio/1241166 , 2014.

26. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15-21.

27. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 2015;31:166-169.

28. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.

29. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.

30. Carlson CS, Emerson RO, Sherwood AM, et al. Using synthetic templates to design an unbiased multiplex PCR assay. Nat Commun 2013;4:2680.

31. Sánchez-Magraner L, Miles J, Baker CL, et al. High PD-1/PD-L1 Checkpoint Interaction Infers Tumor Selection and Therapeutic Sensitivity to Anti-PD-1/PD-L1 Treatment. Cancer Research 2020;80:4244.

32. Thorsson V, Gibbs DL, Brown SD, et al. The Immune Landscape of Cancer. Immunity 2018;48:812-830 e14.

33. Le DT, Durham JN, Smith KN, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science 2017;357:409-413.

34. Repana D, Nulsen J, Dressler L, et al. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol 2019;20:1.

35. Turajlic S, Litchfield K, Xu H, et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol 2017;18:1009-1021.

36. Ballhausen A, Przybilla MJ, Jendrusch M, et al. The shared neoantigen landscape of MSI cancers reflects immunoediting during tumor evolution. Preprint at: https://www.biorxiv.org/content/10.1101/691469v1 2019.

37. Feng M, Jin JQ, Xia L, et al. Pharmacological inhibition of beta-catenin/BCL9 interaction overcomes resistance to immune checkpoint blockades by modulating Treg cells. Sci Adv 2019;5:eaau5240.

38. Coelho MA, de Carne Trecesson S, Rana S, et al. Oncogenic RAS Signaling Promotes Tumor Immunoresistance by Stabilizing PD-L1 mRNA. Immunity 2017;47:1083-1099 e6.

39. Charoentong P, Finotello F, Angelova M, et al. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. Cell Rep 2017;18:248-262.

40. Zhao J, Chen AX, Gartrell RD, et al. Immune and genomic correlates of response to anti-PD-1 immunotherapy in glioblastoma. Nat Med 2019;25:462-469.

41. Grasso CS, Giannakis M, Wells DK, et al. Genetic Mechanisms of Immune Evasion in Colorectal Cancer. Cancer Discov 2018;8:730-749.

42. Rizvi NA, Hellmann MD, Snyder A, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science 2015;348:124-8.

43. Sade-Feldman M, Jiao YJ, Chen JH, et al. Resistance to checkpoint blockade therapy through inactivation of antigen presentation. Nat Commun 2017;8:1136.

44. McGranahan N, Rosenthal R, Hiley CT, et al. Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. Cell 2017;171:1259-1271 e11.

45. Middha S, Yaeger R, Shia J, et al. Majority of B2M-Mutant and -Deficient Colorectal Carcinomas Achieve Clinical Benefit From Immune Checkpoint Inhibitor Therapy and Are Microsatellite Instability-High. JCO Precis Oncol 2019;3.

46. De Nicola F, Goeman F, Pallocca M, et al. Deep sequencing and pathway-focused analysis revealed multigene oncodriver signatures predicting survival outcomes in advanced colorectal cancer. Oncogenesis 2018;7.

47. Castro A, Ozturk K, Pyke RM, et al. Elevated neoantigen levels in tumors with somatic mutations in the HLA-A, HLA-B, HLA-C and B2M genes. BMC Med Genomics 2019;12:107.

48. Burr ML, Sparbier CE, Chan KL, et al. An Evolutionarily Conserved Function of Polycomb Silences the MHC Class I Antigen Presentation Pathway and Enables Immune Evasion in Cancer. Cancer Cell 2019;36:385-401 e8.

49. Demaria O, Cornen S, Daeron M, et al. Harnessing innate immunity in cancer therapy. Nature 2019;574:45-56.

50. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. Nat Rev Cancer 2012;12:252-64.

Chapter 5. The Network of Cancer Genes

5.1 Contributions

In this study[100], I collected and analysed the data on miRNA-target interactions, the evolutionary origins and orthologs of all human genes. Joel Nulsen, Lisa Dressler, Aikaterini Tourna, Francesca D. Ciccarelli, and I wrote the manuscript with contributions from Santhilata Kuppili Venkata and Dimitra Repana, and all authors reviewed and approved its final version.

Francesca D. Ciccarelli acquired the funding, conceived and supervised the study. Santhilata Kuppili Venkata collected and analysed gene duplicability. Lisa Dressler processed and analysed protein-protein interactions, protein complexes, and gene essentiality. Joel Nulsen processed and analysed RNA and protein expression and protein function. Dimitra Repana, Aikaterini Tourna, Anna Yakovleva, and Tommaso Palmieri curated the literature. Santhilata Kuppili Venkata and Joel Nulsen updated the database and website.

5.2 The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens

Genome Biology

DATABASE Open Access

The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens

Dimitra Repana[1,2†], Joel Nulsen[1,2†], Lisa Dressler[1,2†], Michele Bortolomeazzi[1,2†], Santhilata Kuppili Venkata[1,2†], Aikaterini Tourna[1,2], Anna Yakovleva[1,2], Tommaso Palmieri[1,2] and Francesca D. Ciccarelli[1,2*]

Abstract

The Network of Cancer Genes (NCG) is a manually curated repository of 2372 genes whose somatic modifications have known or predicted cancer driver roles. These genes were collected from 275 publications, including two sources of known cancer genes and 273 cancer sequencing screens of more than 100 cancer types from 34,905 cancer donors and multiple primary sites. This represents a more than 1.5-fold content increase compared to the previous version. NCG also annotates properties of cancer genes, such as duplicability, evolutionary origin, RNA and protein expression, miRNA and protein interactions, and protein function and essentiality. NCG is accessible at http://ncg.kcl.ac.uk/.

Keywords: Cancer genomics screens, Cancer genes, Cancer heterogeneity, Systems-level properties

Background

One of the main goals of cancer genomics is to find the genes that, upon acquiring somatic alterations, play a role in driving cancer (cancer genes). To this end, in the last 10 years, hundreds of cancer sequencing screens have generated mutational data from thousands of cancer samples. These include large sequencing efforts led by international consortia such as the International Cancer Genome Consortium (ICGC) [1] and The Cancer Genome Atlas (TCGA) [2]. Cancer genomes usually acquire thousands of somatic alterations and several methods have been developed to identify cancer genes from the pool of all altered genes [3, 4]. These methods have been applied to specific datasets from individual cancer types and to pooled datasets from several cancer types. This is the case for the Pan-Cancer Atlas project [5] and for the recent analysis of the whole set of TCGA samples [6], which accompanied the conclusion of the TCGA sequencing phase [7]. As more and more studies contribute to our knowledge of cancer genes, it becomes increasingly challenging for the research community to maintain an up-to-date overview of cancer genes and of the cancer types to which they contribute.

The Network of Cancer Genes (NCG) is a project launched in 2010 with the aim to gather a comprehensive and curated collection of cancer genes from cancer sequencing screens and to annotate their systems-level properties [8–11]. These define distinctive properties of cancer genes compared to other human genes [12] and include gene duplicability, evolutionary origin, RNA and protein expression, miRNA and protein interactions, and protein function and essentiality. NCG is based on the manual curation of experts who review studies describing cancer sequencing screens, extract the genes that were annotated as cancer genes in the original publications, and collect and analyze the supporting evidence.

Various other databases have been developed to analyze cancer data. Some of them focus on cancer alterations rather than cancer genes (COSMIC [13], DoCM

* Correspondence: francesca.ciccarelli@crick.ac.uk; francesca.ciccarelli@kcl.ac.uk
†Dimitra Repana, Joel Nulsen, Lisa Dressler, Michele Bortolomeazzi and Santhilata Kuppili Venkata contributed equally to this work.
[1]Cancer Systems Biology Laboratory, The Francis Crick Institute, London NW1 1AT, UK
[2]School of Cancer and Pharmaceutical Sciences, King's College London, London SE1 1UL, UK

[14], DriverDB [15], the Cancer Genome Interpreter [16], OncoKB [17], and cBIOPortal [18] among others). Other databases collect only cancer genes with a strong indication of involvement in cancer (the Cancer Gene Census, CGC [19]) and annotate specifically oncogenes or tumor suppressor genes (ONGene [20], TSGene [21]) or cancer genes in specific cancer types (CoReCG [22]). NCG differs from all the above resources because it does not focus on mutations, on particular groups of genes or cancer types. It instead compiles a comprehensive repository of mutated genes that have been proven or predicted to be the drivers of cancer. NCG is widely used by the community. Recent examples of its use include studies identifying and validating cancer genes [23–25] and miRNA cancer biomarkers [26]. NCG has also been used to investigate the effect of long noncoding RNAs on cancer genes [27] and to find cancer-related transcription factors [28].

Here, we describe the sixth release of NCG, which contains 2372 cancer genes extracted from 275 publications consisting of two sources of known cancer genes and 273 cancer sequencing screens. As well as mutational screens of individual cancer types, the collected publications now include four adult and two pediatric pan-cancer studies. In addition to an update of the systems-level properties of cancer genes already present in previous releases (gene duplicability, evolutionary origin, protein function, protein-protein and miRNA-target interactions, and mRNA expression in healthy tissues and cancer cell lines), NCG now also annotates the essentiality of cancer genes in human cell lines and their expression at the protein level in human tissues. Moreover, broader functional annotations of cancer genes in KEGG [29], Reactome [30], and BioCarta [31] are also provided.

The expert curation of a large number of cancer sequencing screens and the annotation of a wide variety of systems-level properties make NCG a comprehensive and unique resource for the study of genes that promote cancer.

Construction and content

The NCG database integrates information about genes with a known or predicted driver role in cancer. To facilitate the broad use of NCG, we have developed a user-friendly, interactive, and open-access web portal for querying and visualizing the annotation of cancer genes. User queries are processed interactively to produce results in a constant time. The front-end is connected to a database, developed using relational database management system principles [32] (Additional file 1: Figure S1). The web application for the NCG database was developed using MySQL v.5.6.38 and PHP v.7.0. Raw data for each of the systems-level properties were acquired from heterogeneous data sources and processed as described below. The entire content of NCG is freely available and can be downloaded from the database website.

Gene duplicability and evolutionary origin

Protein sequences from RefSeq v.85 [33] were aligned to the human genome assembly hg38 with BLAT [34]. From the resulting genomic alignments, 19,549 unique gene loci were identified and genes sharing at least 60% of the original protein sequence were considered to be duplicated [35] (Additional file 2: Table S1). Orthologous genes for 18,486 human genes (including 2348 cancer genes, Additional file 2: Table S1) in 2032 species were collected from EggNOG v.4.5.1 [36] and used to trace the gene evolutionary origin as previously described [37]. Genes were considered to have a pre-metazoan origin if their orthologs could be found in prokaryotes, unicellular eukaryotes, or opisthokonts [37].

Gene and protein expression

RNA-Seq data from healthy human tissues for 18,984 human genes (including all 2372 cancer genes, Additional file 2: Table S1) were derived from the non-redundant union of Protein Atlas v.18 [38] and GTEx v.7 [39]. Protein Atlas reported the average transcripts per million (TPM) values in 37 tissues, and genes were considered to be expressed in a tissue if their expression value was ≥ 1 TPM. GTEx reported the distribution of TPM values for individual genes in 11,688 samples across 30 tissue types. In this case, genes were considered to be expressed if they had a median expression value ≥ 1 TPM.

Gene expression data for all 2372 cancer genes in 1561 cancer cell lines were taken from the Cancer Cell Line Encyclopedia (CCLE, 02/2018) [40], the COSMIC Cancer Cell Line Project (CLP, v.84) [19], and a Genentech study (GNE, 06/2014) [41] (Additional file 2: Table S1). Gene expression levels were derived directly from the original sources, namely reads per kilobase million (RPKM) values for CCLE and GNE, and microarray *z*-scores for CLP. Genes were categorized as expressed if their expression value was ≥ 1 RPKM in CCLE or GNE and were annotated as over, under, or normally expressed in CLP, as determined by COSMIC.

The current release of NCG also includes protein expression from immunohistochemistry assays of healthy human tissues as derived from Protein Atlas v.18. Data were available for 13,001 human proteins including 1799 cancer proteins (Additional file 2: Table S1). Proteins were categorized as not detected or as having low, medium, or high expression in 44 tissues on the basis of staining intensity and fraction of stained cells [38]. In Protein Atlas, expression levels were reported in multiple cell types for each tissue. NCG retained the highest reported value as the expression level for that tissue.

Gene essentiality

Gene essentiality was derived from two databases, PICKLES (09/2017) [42] and OGEE v.2 [43], both of which collected data from CRISPR-Cas9 knockout and shRNA knockdown screens of human cell lines. In PICKLES, data from primary publications have been re-analyzed and genes were considered essential in a cell line if their associated Bayes factor was > 3 [44]. We therefore used this threshold to define essential genes. In OGEE, genes were labelled as "essential" or "not essential" according to their annotation in the original publications. Consistently, we retained the same annotation. From the non-redundant union of the two databases, essentiality information was available for a total of 18,833 genes (including all 2372 cancer genes) in 178 cell lines (Additional file 2: Table S1).

Protein-protein and miRNA-target interactions

Human protein-protein interactions were derived from four databases (BioGRID v.3.4.157 [45], MIntAct v.4.2.10 [46], DIP (02/2018) [47], and HPRD v.9 [48]). Only interactions between human proteins supported by at least one original publication were considered [8]. The union of all interactions from the four sources was used to derive a human protein-protein interaction network of 16,322 proteins (including 2203 cancer proteins, Additional file 2: Table S1) and 289,368 binary interactions. To control for a possibly higher number of studies on cancer proteins resulting in an artificially higher number of interactions, a network of 15,272 proteins and 224,258 interactions was derived from high-throughput screens reporting more than 100 interactions [11].

Data on human protein complexes for 8080 human proteins (including 1414 cancer proteins; Additional file 2: Table S1) were derived from the non-redundant union of three primary sources, namely CORUM (07/2017) [49], HPRD v.9 [48], and Reactome v.63 [30]. Only human complexes supported by at least one original publication were considered [11].

Experimentally validated interactions between human genes and miRNAs were downloaded from miRTarBase v.7.0 [50] and miRecords v.4.0 [51], resulting in a total of 14,649 genes (including 2034 cancer genes) and 1762 unique miRNAs (Additional file 2: Table S1). To control for the higher number of single-gene studies focussing on cancer genes, a dataset of high-throughput screens testing ≥ 250 different miRNAs was also derived (Additional file 2: Table S1).

Functional annotation

Data on functional categories (pathways) were collected from Reactome v.63 [30], KEGG v.85.1 [29], and BioCarta (02/2018) [31]. Data for BioCarta were extracted from the Cancer Genome Anatomy Project [52]. All levels of Reactome were included, and level 1 and 2 pathways from KEGG were added separately. Overall, functional annotations were available for 11,344 human proteins, including 1750 cancer proteins assigned to 2318 pathways in total.

Utility and discussion

Catalogue of known and candidate cancer genes

To include new cancer genes in NCG, we applied a modified version of our well-established curation pipeline [11] (Fig. 1a). We considered two main groups of cancer genes: known cancer genes whose involvement in cancer has additional experimental support and candidate cancer genes whose somatic alterations have a predicted cancer driver role but lack further experimental support.

As sources of known cancer genes, we used 708 genes from CGC v.84 [19] and 125 genes from a manually curated list [53]. Of the resulting 711 genes, we further annotated 239 as tumor suppressor genes (TSGs) and 239 as oncogenes (OGs). The remaining 233 genes could not be unambiguously classified because either they had conflicting annotations in the two original sources (CGC and [53]) or they were defined as both OGs and TSGs. Despite these two sources of known cancer genes have been extensively curated, 49 known cancer genes are in two lists of possible false positives [6, 54].

Next, we reviewed the literature to search for studies that (1) described sequencing screens of human cancers and (2) provided a list of genes considered to be the cancer drivers. This led to 273 original papers published between 2008 and March 2018, 98 of which were published since the previous release of NCG [11] and 42 of which came from ICGC or TGCA (Additional file 2: Table S2). Overall, these publications describe the sequencing screens of 119 cancer types from 31 primary anatomical sites as well as six pan-cancer studies (Additional file 2: Table S2). In total, this amounts to samples from 34,905 cancer donors. Each publication was reviewed independently by at least two experts and all studies whose annotation differed between the experts were further discussed. Additionally, 31 randomly selected studies (11% of the total) were re-annotated blindly by a third expert to assess consistency. The manual revision of the 273 studies led to 2088 cancer genes, of which 427 were known cancer genes and the remaining 1661 were candidate cancer genes (Fig. 1b). Compared to the previous release, this version of NCG constitutes a significant increase in the number of

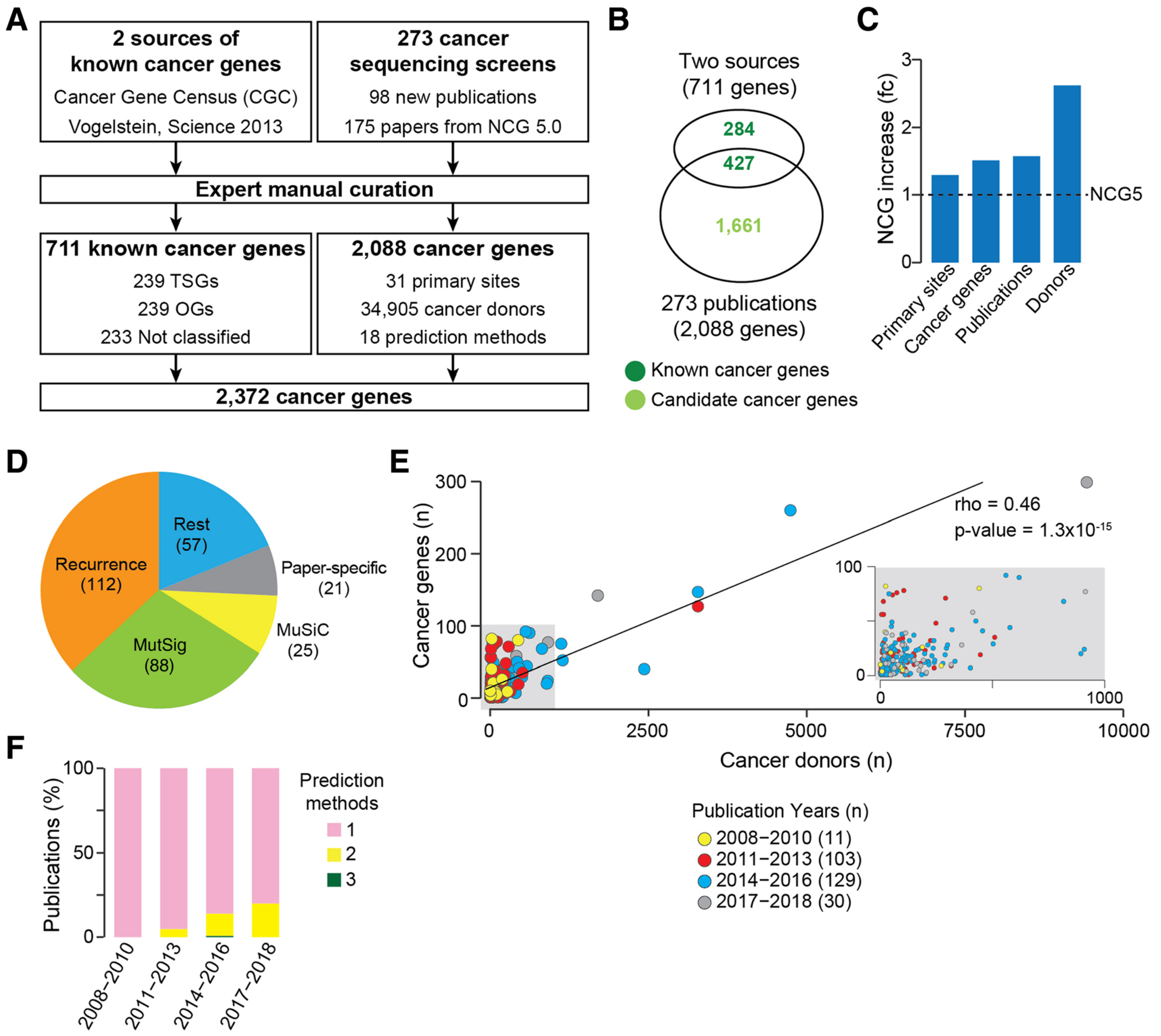


Fig. 1 Manual curation of cancer genes in NCG. **a** Pipeline used for adding cancer genes to NCG. Two sources of known cancer genes [19, 53] were integrated leading to 711 known cancer genes. In parallel, 273 publications describing cancer sequencing screens were reviewed to extract 2088 cancer genes. The non-redundant union of these two sets led to 2372 cancer genes currently annotated in NCG. **b** Intersection between known and candidate cancer genes in NCG. **c** Comparison of NCG content with the previous version [11]. **d** Pie chart of the methods used to identify cancer genes in the 273 publications. The total is greater than 273 because some studies used more than one method (Additional file 2: Table S2). **e** Cancer genes as a function of the number of cancer donors per study. The grey inset shows a magnification of the left bottom corner of the plot. **f** Number of methods used to identify cancer genes over time. PanSoftware used in one of the pan-cancer studies [6] was considered as a single method but is in fact a combination of 26 prediction tools

cancer primary sites (1.3-fold), cancer genes (1.5-fold), publications (1.6-fold), and analyzed donors (2.6-fold, Fig. 1c).

Based on literature evidence [6, 54], gene length, and function [10], 201 candidates were labelled as possible false-positive predictions. We further investigated the reasons why 284 known cancer genes were not identified as drivers in any of the 273 cancer sequencing screens. We found that these genes predispose to cancer rather than acquiring somatic alterations, are the chimeric product of gene fusions, are part of CGC Tier 2 (i.e., genes with lower support for their involvement in cancer), or were identified with different methods than sequencing. Eleven of these 284 genes are possible false positives [6, 54].

The annotation of a large number of studies allowed us to gain insights into how cancer genes have been identified in the last 10 years. Of the overall 18 prediction methods (Additional file 2: Table S2), the recurrence of a gene alteration within the cohort is the most widely used across screens (Fig. 1d). In this case, no further threshold of statistical significance or correction for the genome, gene, and cancer background mutation rate was applied, thus leading to possible false-positive predictions. Other frequently used prediction methods are MutSig [55], MuSiC [56], and ad hoc pipelines developed in the same publication (referred to as 'paper-specific'). Although they apply statistical methods to correct for the background mutation rate and reduce false positives, all of these approaches estimate the tendency of a gene to mutate more than expected within a cohort and therefore they all depend on sample size. Indeed, we observed an overall positive correlation between the number of cancer donors and the number of cancer genes (Fig. 1e). This confirms that the sensitivity of the approaches currently used to predict cancer genes is higher for large cohorts of samples. Finally, although the vast majority of analyzed studies tend to apply only one prediction method, more recent publications have started to use a combination of two or three methods (Fig. 1f).

Heterogeneity and specificity of cancer genes

The number of cancer genes and the relative proportion of known and candidate cancer genes vary greatly across cancer primary sites (Fig. 2a). More than 75% of cancer genes in cancers of the prostate, soft tissues, bone, ovary, cervix, thymus, and retina are known drivers. On the contrary, more than 75% of driver genes in cancers of the penis, testis, and vascular system are candidate cancer genes (Fig. 2a). This seems to be due to several factors including the sample size, the number of different methods that have been applied to identify cancer genes and the biology of each cancer type. For example, penis, vascular system, and testis cancers show a high proportion of candidate cancer genes. The corresponding cohorts have a small sample size and have been analyzed by one or two prediction methods. However, other cancer types showing equally high proportions of candidates (pancreas, skin, blood) have large sample sizes and were analyzed by several methods (Fig. 2b). Moreover, although the number of cancer genes is overall positively correlated with the number of sequenced samples (Figs. 1e and 2c), there are marked differences across primary sites. For example, ovary, bone, prostate, thyroid, and kidney cancers have substantially fewer cancer genes compared to cancers with similar numbers of cancer donors such as uterine, stomach, skin, and hepatobiliary cancers (Fig. 2c). This is likely due to variable levels of genomic instability and heterogeneity across cancer types of the same primary site. For example, in seven of the nine mutational screens of skin melanoma, a cancer type with high genomic instability [57], more than 50% of cancer genes are study-specific (Fig. 3a). Similarly, the 24 types of blood cancer are variable in terms of number of cancer genes, with diffuse large B-cell lymphoma having many more cancer genes than other blood cancers with higher numbers of cancer donors (Fig. 3b). In both cases, the use of the same method (i.e., MutSig in Fig. 3a and MuSiC in Fig. 3b) identified different cancer genes in different patient cohorts, highlighting the biological heterogeneity even across donors of the same cancer type.

Cancer genes, and in particular candidates, are highly cancer-specific (Fig. 3c). Hemicentin 1 (*HMCN1*) is the only candidate cancer gene to be significantly mutated in six primary sites (blood, brain, esophagus, large intestine, liver, and pancreas). A few known cancer genes are recurrently mutated across several primary sites, including *TP53* (25), *PIK3CA* (21), and *PTEN* (20; Fig. 3c). These are, however, exceptions, and the large majority of known and candidate cancer genes (64% of the total) are found only in one primary site, indicating high heterogeneity of cancer driver events. Similar specificity is also observed in terms of supporting publications. The majority of cancer genes are publication-specific, again with few exceptions including *TP53* (173), *PIK3CA* (87) and *KRAS* (86, Fig. 3d). Of note, the best-supported candidate gene is Titin (*TTN*, predicted in nine publications), which is a well-known false positive of recurrence-based approaches [55]. Interestingly, the scenario is different when analyzing the number of prediction methods that support cancer genes reported in at least two screens (Fig. 3e). In this case, few candidate and known cancer genes are identified by only one method, while the majority of them are supported by at least two (candidates) and three (known cancer genes) approaches. However, only six candidate cancer genes are supported by six methods, and *TP53* is the only cancer genes to be identified by 16 of the 18 methods (Fig. 3e).

Finally, the heterogeneity of the cancer driver landscape is reflected in the pan-cancer studies. Approximately 40% of the cancer genes from pan-cancer analyses were not previously predicted as drivers (Fig. 3f), despite the large majority of cancer samples having been already analyzed in the corresponding cancer-specific study. This is yet a further confirmation that current methods depend on the sample size and that a larger cohort leads to novel predictions. Only 35 cancer genes were shared across four pan-cancer re-analyses of adult tumors (Fig. 3g), suggesting that the prediction of cancer genes is highly method- and cohort-dependent. This is further confirmed by the poor overlap between cancer genes from adult and pediatric pan-cancer studies (Fig. 3h). In this case, however, it is also likely that different biological mechanisms are responsible for adult and childhood cancers.

Overall, our analysis of the cancer driver landscape suggests that the high heterogeneity of cancer genes observed across cancer types is due to a combination of

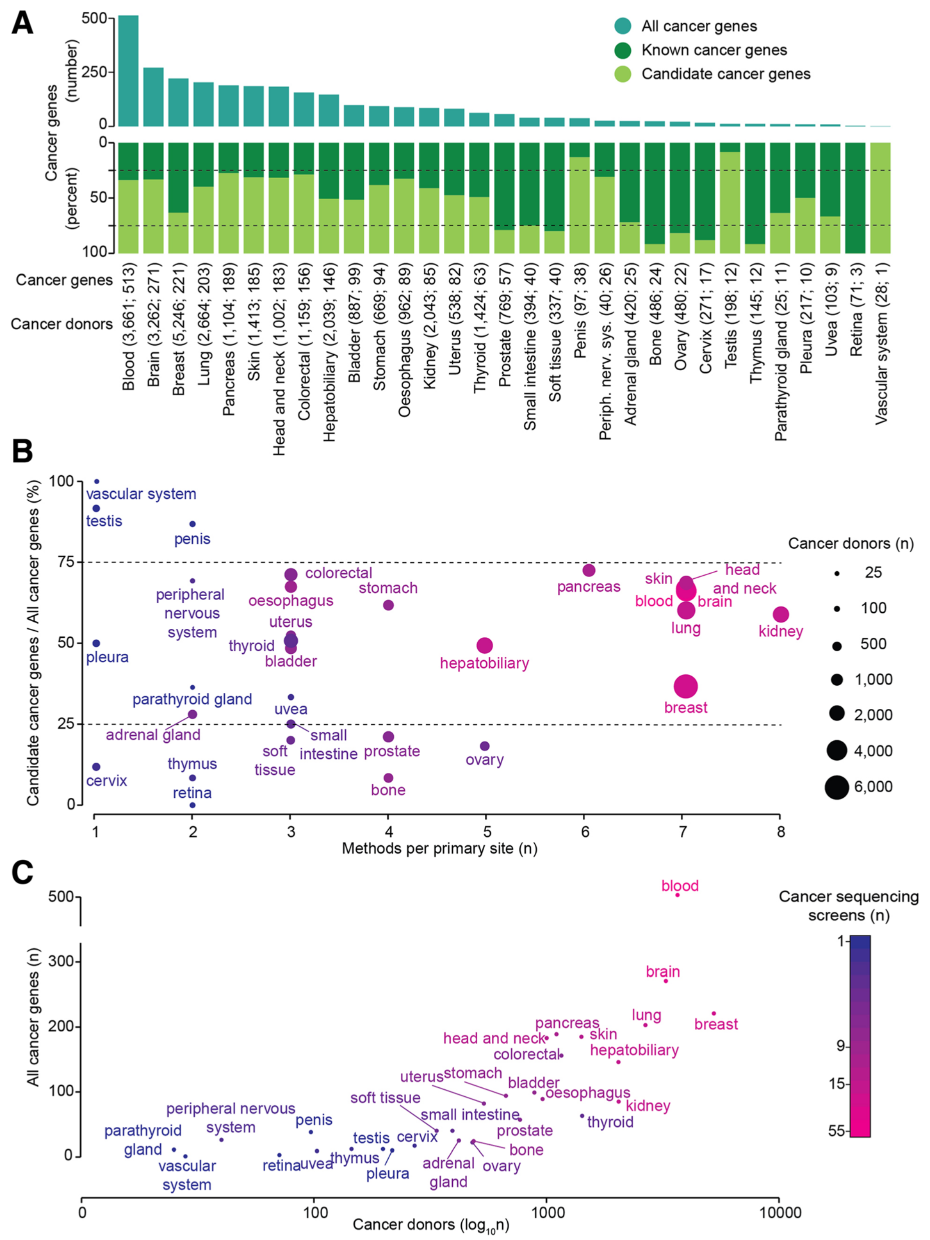


Fig. 2 Distribution of cancer genes across primary sites and cancer donors. **a** Number of total cancer genes and proportion of known and candidate cancer genes across the 31 tumor primary sites analyzed in the 267 cancer-specific studies. The number of cancer donors followed by the number of cancer genes is given in brackets for each primary site. **b** Proportion of candidate cancer genes over all cancer genes across the 31 tumor primary sites. The dot size is proportional to the donor cohort size. **c** Total number of cancer genes and cancer donors across the 31 tumor primary sites. The color scale in (**b**) and (**c**) indicates the number of screens for each primary site

sample size effect, prediction methods, and true biological differences across cancers.

Systems-level properties of cancer genes

In addition to collecting cancer genes from the literature, NCG also annotates the systems-level properties that distinguish cancer genes from other genes that are not implicated in cancer (Additional file 2: Table S1). We therefore compared each of these properties between cancer genes and the rest of human genes. We considered seven distinct groups of cancer genes. The first three were 711 known cancer genes, 1661 candidate cancer genes, and 2372 total cancer genes. After removing 201 possible false-positive predictions [6, 54] from the list of candidate cancer genes, we also identified two sets of candidate cancer genes with a stronger support. One was composed of 104 candidate cancer genes found in at least two independent screens of the same primary

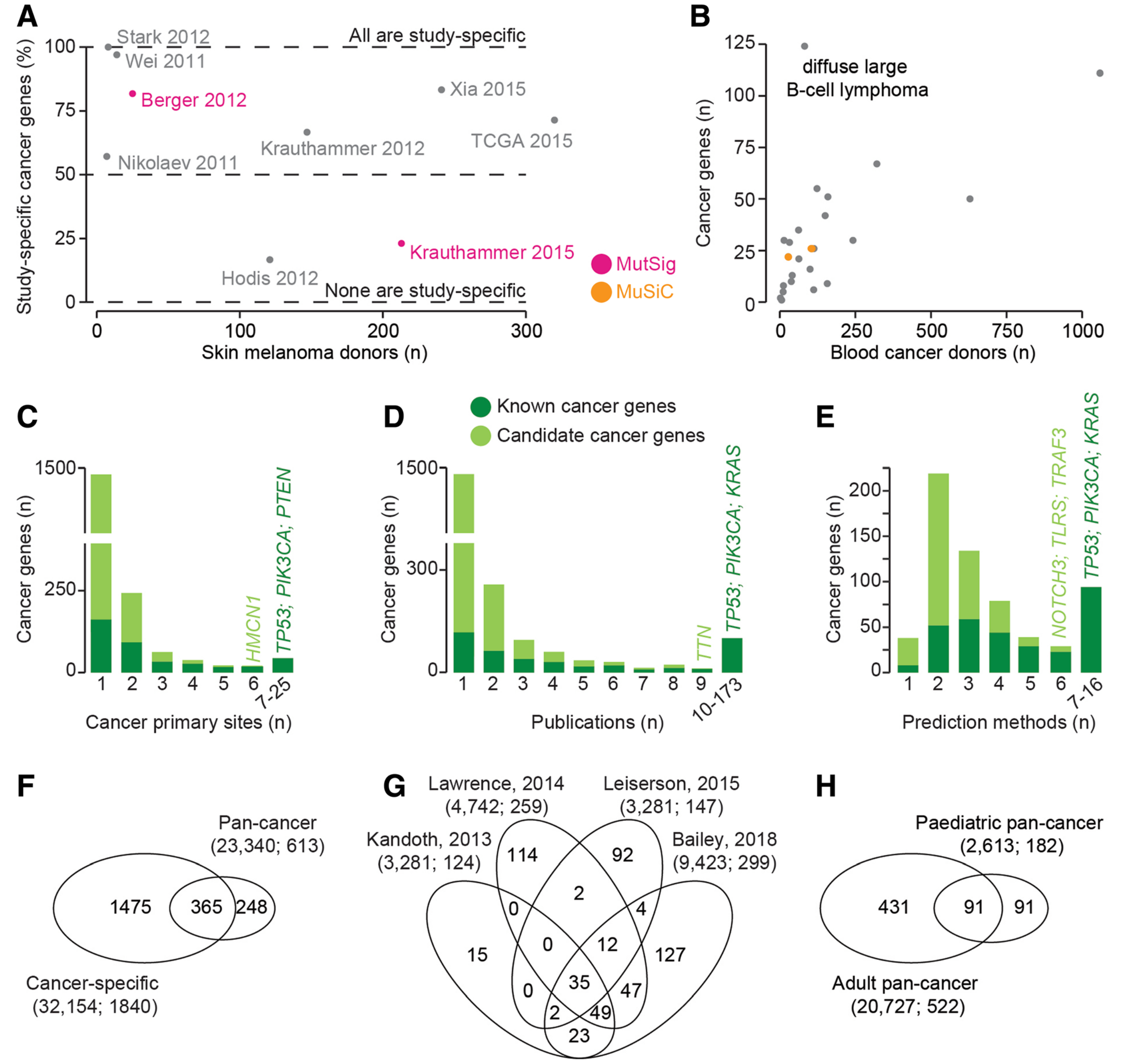


Fig. 3 Recurrence of cancer across primary sites and publications. **a** Proportion of study-specific cancer genes reported by each of the seven skin melanoma screens. **b** Total number of cancer genes and donors across 24 cancer types of the blood. The full list of blood cancer types is reported in Additional file 2: Table S2. **c** Number of primary sites in which each known or candidate cancer gene was reported to be a driver. **d** Number of publications in which each known or candidate cancer gene was reported to be a driver. **e** Number of methods used to predict cancer genes for drivers found in more than one publication. **f** Intersection of cancer genes in the cancer-specific and pan-cancer studies. **g** Venn diagram of cancer genes across the four pan-cancer studies of adult donors. **h** Intersection of cancer genes in pan-cancer screens of adult and pediatric donors. In **f**, **g**, and **h**, the number of donors followed by the total number of cancer genes are given in brackets

site. The other was formed of 711 candidate cancer genes identified in large cohorts composed of at least 140 donors (top 25% of the sample size distribution across screens). Finally, we compared the properties between 239 TSGs and 239 OGs.

As previously reported [35], we confirmed that a significantly lower fraction of cancer genes has duplicated copies in the human genome due to a high proportion of single-copied TSGs (Fig. 4a). The same trend was observed in both known and candidate cancer genes and is significant for the combination of the two gene sets. Interestingly, candidate cancer genes found in ≥ 2 screens show a high proportion of duplicated cancer genes (albeit not significant probably due to the small size of the group, Fig. 3d). This could suggest that several genes in this group may exert an oncogenic role.

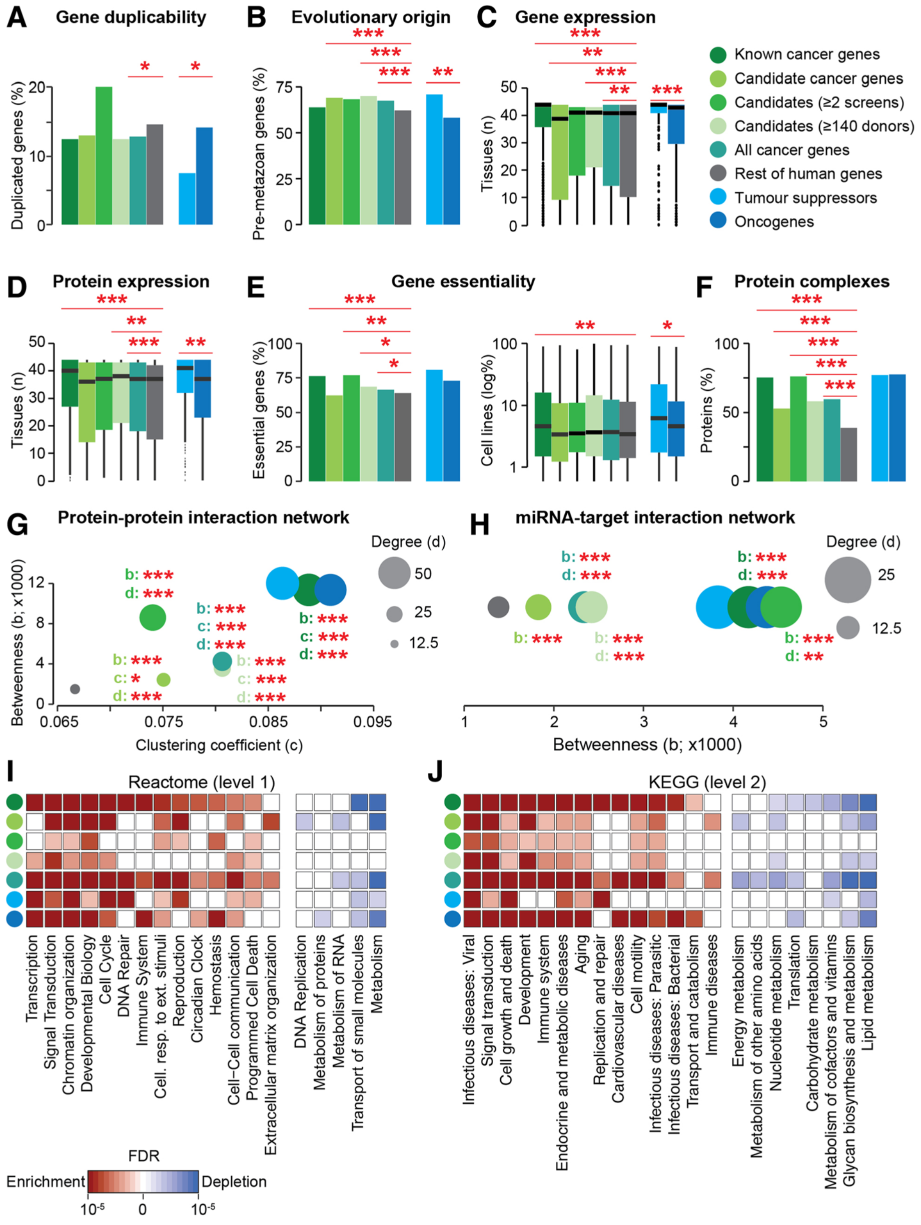


Fig. 4 Systems-level properties of cancer genes. **a** Percentage of genes with ≥ 1 gene duplicate covering ≥ 60% of the protein sequence. **b** Proportion of genes originating in pre-metazoan species. **c**, **d** Number of human tissues in which genes (**c**) and proteins (**d**) are expressed. In panel **c**, tissue types were matched between GTEx and Protein Atlas wherever possible, giving 43 unique tissues. In tissues represented in both datasets, genes were defined as expressed if they had ≥ 1 TPM in both datasets. Only genes present in both sources were compared (Additional file 2: Table S1). **e** Percentage of genes essential in ≥ 1 cell line and distribution of cell lines in which each gene is essential. Only genes with concordant annotation between OGEE and PICKLES were compared (Additional file 2: Table S1). **f** Percentage of proteins involved in ≥ 1 protein complex. **g** Median values of betweenness (centrality), clustering coefficient (clustering), and degree (connectivity) of human proteins in the protein-protein interaction network. **h** Median values of betweenness and degree of the target genes in the miRNA-target interaction network. The clustering coefficient is zero for all nodes, because interactions occur between miRNAs and target genes. Known, candidate, and all cancer genes were compared to the rest of human genes, while TSGs were compared to OGs. Significance was calculated using a two-sided Fisher test (**a**, **b**, **e**, **f**) or Wilcoxon test (**c**, **d**, **g**, **h**). *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. Enrichment and depletion of cancer genes in representative functional categories taken from level 1 of Reactome (**i**) and level 2 of KEGG (**j**). Significance was calculated comparing each group of cancer genes to the rest of human genes using a two-sided Fisher test. False discovery rates were calculated in each gene set separately. Only pathways showing enrichment or depletion are shown. The full list of pathways is provided in Additional file 2: Table S3

Cancer genes, and in particular candidate cancer genes, originated earlier in evolution (Fig. 4b) [37, 58, 59]. Known cancer genes alone do not differ from the rest due to the fact that OGs are significantly younger than TSGs (Fig. 4b).

Known cancer genes tend to be ubiquitously expressed at the mRNA (Fig. 4c) and protein (Fig. 4d) levels, and TSGs are more widely expressed than OGs. This trend is less clear when analyzing candidate cancer genes separately. Candidates with stronger support tend to resemble known cancer genes; however, the overall set of candidate cancer genes has a narrower tissue expression pattern at the gene and protein level (Fig. 4c, d).

A similar scenario is observed when analyzing gene essentiality. A higher fraction of cancer genes and, in particular of known cancer genes, is essential in at least one human cell line (Fig. 4e). Moreover, known cancer genes tend to be essential in a higher fraction of cell lines. Both measures of gene essentiality are higher in TSGs as compared to OGs (Fig. 4e). Candidate cancer genes with stronger support are again similar to known cancer genes but, when considered together, all candidate cancer genes are not significantly enriched in essential genes (Fig. 4e).

Proteins encoded by cancer genes are more often involved in protein complexes (Fig. 4f). They are also more connected (higher degree), central (higher betweenness), and clustered (higher clustering coefficient) in the protein-protein interaction network (Fig. 4g). We verified that this trend holds true also when using only data from high-throughput screens (Additional file 2: Table S2), thus excluding the possibility that the distinctive network properties of cancer proteins are due to their better annotation. These trends remain significant for all sets of cancer genes.

Cancer genes are regulated by a higher number of miRNAs (higher degree) and occupy more central positions (higher betweenness) in the miRNA-target interaction network (Fig. 4h). As above, these results remain valid also when only considering the miRNA-target network from high-throughput screens (Additional file 2: Table S2) and for any group of cancer genes considered.

Cancer genes are consistently enriched in functional categories such as signal transduction, chromatin reorganization, and cell cycle and depleted in others, such as metabolism and transport (Fig. 4i, Additional file 2: Table S3). Candidate cancer genes generally exhibit weaker enrichment than the other groups, most notably in DNA repair. Interestingly, however, extracellular matrix reorganization displays a specific enrichment for candidate cancer genes. Some functional categories are selectively enriched for OGs (e.g. development and immune system, Fig. 4j) or TSGs (e.g. DNA repair and programmed cell death, Fig. 4i). While annotations from Reactome and KEGG generally give concordant results, they differ significantly for gene transcription. In this case, Reactome shows a strong enrichment for cancer genes, while it is not significant in KEGG (Fig. 4i, j).

Overall our analyses confirm that cancer genes are a distinctive group of human genes. Despite their heterogeneity across cancer types and donors, they share common properties. Candidate cancer genes only share some of the properties of known cancer genes, such as an early evolutionary origin (Fig. 4b) and higher centrality and connectivity in the protein-protein and miRNA-target interaction networks (Fig. 4g, h). They do not differ from the rest of genes for all other properties. However, the two sets of candidate cancer genes with a stronger support overall maintain the vast majority of the distinctive properties of known cancer genes. This suggests that the current set of candidate cancer genes likely contains false positives and genes with weak support that do not resemble the properties of known cancer genes. This is further indicated when directly comparing the properties of known and candidate cancer genes (Additional file 2: Table S4). In this case, known cancer genes are significantly different for most properties when compared to the whole set of candidate cancer genes. However, these differences are reduced when the two sets of candidates with stronger support are used. Finally, TSGs and OGs constitute two distinct classes of cancer genes even based on their systems-level properties (Fig. 4).

Future directions

In the coming years, NCG will continue to collect new cancer genes and annotate their properties, including novel properties such as genetic interactions or epigenetic features for which large datasets are becoming available. So far, the cancer genomics community has focussed mostly on the identification of protein-coding genes with putative cancer driver activity. With the increasing availability of whole-genome sequencing data and a rising interest in non-coding alterations [27, 60], NCG will expand to also collect non-coding cancer drivers. Another direction for future development will be the analysis of clinical data, including therapeutic treatments, to link them to the altered drivers. This will contribute to the expansion of our knowledge of cancer driver genes in the context of their clinical relevance.

Conclusions

The present release of NCG describes a substantial advance in annotations of known and candidate cancer driver genes as well as an update and expansion of their systems-level properties. The extensive body of literature evidence collected in NCG enabled a systematic analysis of the methods used to identify cancer genes, highlighting their dependence on the number of cancer donors. We also confirmed the high heterogeneity of cancer genes within and across cancer types. The broad set of systems-level properties collected in NCG shows that cancer genes form a distinct group, different from the rest of human genes. For some of these properties, the differences observed for known cancer genes hold true also for candidate cancer genes, and TSGs show more pronounced cancer gene properties than OGs. Interestingly, these properties are shared by all cancer genes, independently of the cancer type or gene function. Therefore, focussing on genes with similar characteristics could be used for the identification and prioritization of new cancer driver genes [61]. In conclusion, the large-scale annotation of the systems-level properties of cancer genes in NCG is a valuable source of information not only for the study of individual genes, but also for the characterization of cancer genes as a group.

Additional files

Additional file 1: Figure S1. Schema of the NCG database. (PDF 338 kb)

Additional file 2: Table S1. Systems-level properties of cancer genes. Table S2 List of 273 publications describing cancer sequencing screens in NCG. Table S3 Enrichment and depletion of cancer genes in protein functional categories. Table S4 Comparison of systems-level properties between known and candidate cancer genes. (XLSX 122 kb)

Additional file 3: Lists of all cancer genes, known cancer genes, and candidate cancer genes. (XLSX 840 kb)

Abbreviations

CCLE: Cancer Cell Line Encyclopedia; CLP: Cell Line Project (COSMIC); GNE: Genentech; ICGC: International Cancer Genome Consortium; miRNA: MicroRNA; NCG: Network of Cancer Genes; OG: Oncogene; RPKM: Reads per kilobase million; TCGA: The Cancer Genome Atlas; TPM: Transcripts per million; TSG: Tumor suppressor gene

Acknowledgements

We thank all members of the Ciccarelli lab for useful feedback on the website and Elizabeth Foxall for proofreading of the manuscript.

Funding

This work was supported by Cancer Research UK [C43634/A25487] and by the Cancer Research UK King's Health Partners Centre at King's College London. Computational analyses were done using the High-Performance Computer at the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. JN is supported by the EPSRC Centre for Doctoral Training in Cross-Disciplinary Approaches to Non-Equilibrium Systems (CANES, EP/L015854/1). DR is supported by the Health Education England (HEE) Genomics Education Programme (GEP).

Availability of data and materials

The whole content of NCG can be freely downloaded from the website (http://ncg.kcl.ac.uk/). No license is required. The lists of the (candidate) cancer genes are also available in Additional file 3.
Original data were obtained from the following online sources:
BioCarta: https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways [62];
BioGRID: https://thebiogrid.org/ [63];
Cancer Cell Line Encyclopedia: https://portals.broadinstitute.org/ccle [64];
Cell Line Project: https://cancer.sanger.ac.uk/cell_lines [65];
CORUM: http://mips.helmholtz-muenchen.de/corum/ [66];
DIP: http://dip.doe-mbi.ucla.edu/dip/Main.cgi [67];
EggNOG: http://eggnogdb.embl.de/#/app/home [68];
Genentech: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2706/ [69];
GTEx: https://www.gtexportal.org/home/ [70];
HPRD: http://www.hprd.org/ [71];
KEGG: http://www.genome.jp/kegg/pathway.html [72];
MIntAct: https://www.ebi.ac.uk/intact/ [73];
miRecrods: http://c1.accurascience.com/miRecords/ [74];
miRTarBase: http://mirtarbase.mbc.nctu.edu.tw/php/index.php [75];
OGEE: http://ogee.medgenius.info/browse/ [76];
PICKLES: https://hartlab.shinyapps.io/pickles/ [77];
Protein Atlas: https://www.proteinatlas.org/ [78];
Reactome: https://reactome.org/ [79];
RefSeq: https://www.ncbi.nlm.nih.gov/refseq/ [80].

Authors' contributions

FDC conceived and supervised the study. SKV, DR, JN, LD, MB, AT, and FDC analyzed the data. SKV analyzed gene duplicability. MB processed evolutionary origins and miRNA-target interactions. LD processed protein-protein interactions, protein complexes, and gene essentiality. JN processed RNA and protein expression and protein function. DR, AT, AY, and TP curated the literature. SKV and JN updated the database and website. JN, LD, MB, AT, and FDC wrote the manuscript, with contributions from SKV and DR. All authors reviewed and approved the final version of the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

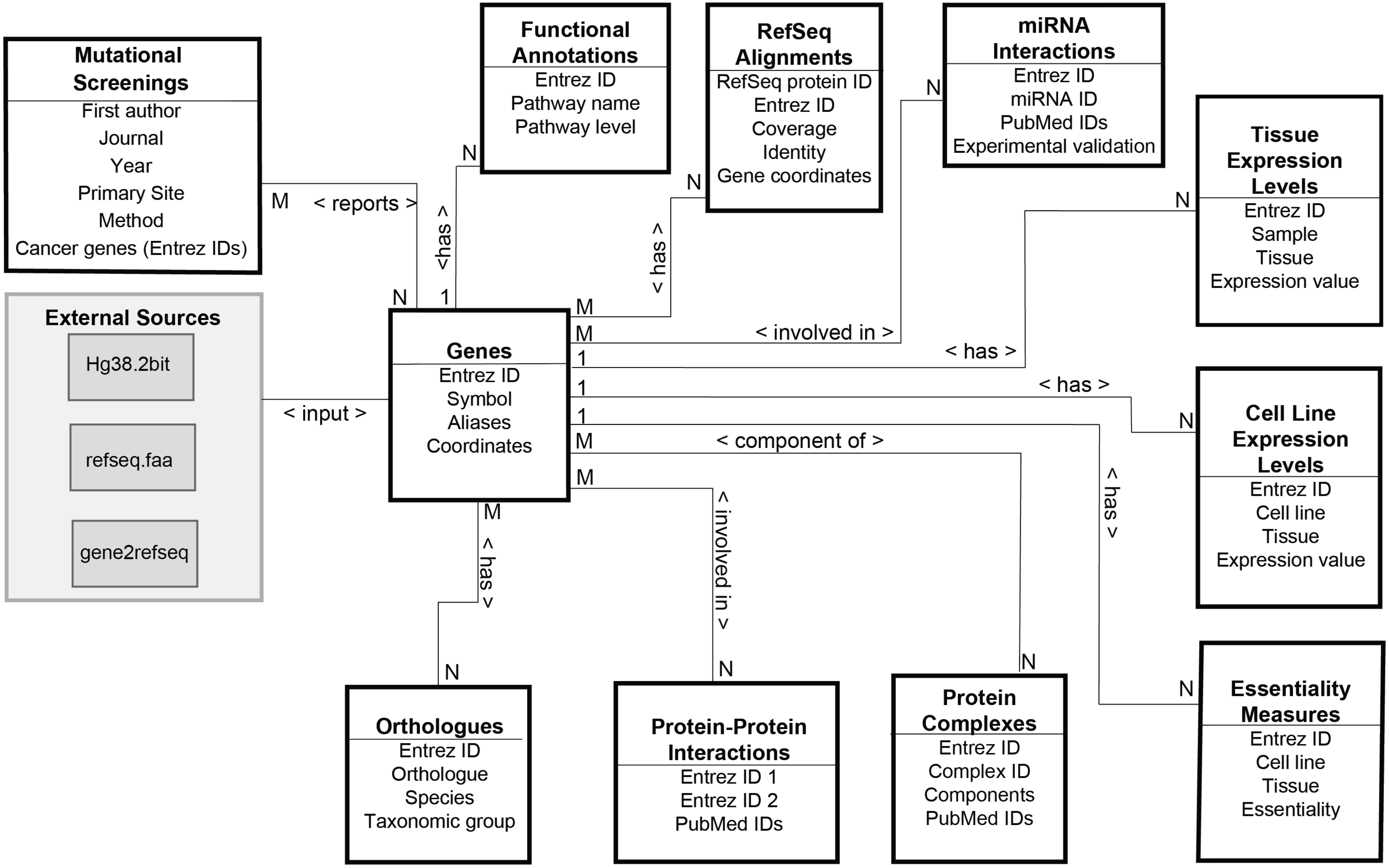
Received: 11 August 2018 Accepted: 12 December 2018
Published online: 03 January 2019

References

1. Consortium ICG. International network of cancer genome projects. Nature. 2010;464:993.
2. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol. 2015;19:A68.
3. Nakagawa H, Fujita M. Whole genome sequencing analysis for cancer genomics and precision medicine. Cancer Sci. 2018;109:513–22.
4. Poulos RC, Wong JW. Finding cancer driver mutations in the era of big data research. Biophys Rev. 2018;10:1–9.
5. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.
6. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173:371–85 e318.
7. Hutter C, Zenklusen JC. The Cancer genome atlas: creating lasting value beyond its data. Cell. 2018;173:283–5.
8. Syed AS, D'Antonio M, Ciccarelli FD. Network of Cancer Genes: a web resource to analyze duplicability, orthology and network properties of cancer genes. Nucleic Acids Res. 2010;38:D670–5.
9. D'Antonio M, Pendino V, Sinha S, Ciccarelli FD. Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes. Nucleic Acids Res. 2012;40:D978–83.
10. An O, Pendino V, D'Antonio M, Ratti E, Gentilini M, Ciccarelli FD. NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. Database. 2014;2014:bau015.

11. An O, Dall'Olio GM, Mourikis TP, Ciccarelli FD. NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. Nucleic Acids Res. 2016;44:D992–9.
12. Ciccarelli FD. The (r)evolution of cancer genetics. BMC Biol. 2010;8:74.
13. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45:D777–83.
14. Ainscough BJ, Griffith M, Coffman AC, Wagner AH, Kunisaki J, Choudhary MN, McMichael JF, Fulton RS, Wilson RK, Griffith OL, Mardis ER. DoCM: a database of curated mutations in cancer. Nat Methods. 2016;13:806–7.
15. Chung IF, Chen CY, Su SC, Li CY, Wu KJ, Wang HW, Cheng WC. DriverDBv2: a database for human cancer driver gene research. Nucleic Acids Res. 2016; 44:D975–9.
16. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, Tusquets I, Albanell J, Rodon J, Tabernero J, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 2018;10:25.
17. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, Rudolph JE, Yaeger R, Soumerai T, Nissan MH, et al. OncoKB: a precision oncology Knowledge Base. JCO Precis Oncol. 2017;1:1–16.
18. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2:401–4.
19. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer. 2004;4:177–83.
20. Liu Y, Sun J, Zhao M. ONGene: a literature-based database for human oncogenes. J Genet Genomics. 2017;44:119–21.
21. Zhao M, Kim P, Mitra R, Zhao J, Zhao Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic Acids Res. 2016; 44:D1023–31.
22. Agarwal R, Kumar B, Jayadev M, Raghav D, Singh A. CoReCG: a comprehensive database of genes associated with colon-rectal cancer. Database (Oxford). 2016;2016:1-9.
23. Andres-Leon E, Cases I, Alonso S, Rojas AM. Novel miRNA-mRNA interactions conserved in essential cancer pathways. Sci Rep. 2017;7:46101.
24. Jakel C, Bergmann F, Toth R, Assenov Y, van der Duin D, Strobel O, Hank T, Kloppel G, Dorrell C, Grompe M, et al. Genome-wide genetic and epigenetic analyses of pancreatic acinar cell carcinomas reveal aberrations in genome stability. Nat Commun. 2017;8:1323.
25. Woollard WJ, Pullabhatla V, Lorenc A, Patel VM, Butler RM, Bayega A, Begum N, Bakr F, Dedhia K, Fisher J, et al. Candidate driver genes involved in genome maintenance and DNA repair in Sezary syndrome. Blood. 2016;127:3387–97.
26. Xiong D, Pan J, Zhang Q, Szabo E, Miller MS, Lubet RA, You M, Wang Y. Bronchial airway gene expression signatures in mouse lung squamous cell carcinoma and their modulation by cancer chemopreventive agents. Oncotarget. 2017;8:18885–900.
27. Chiu HS, Somvanshi S, Patel E, Chen TW, Singh VP, Zorman B, Patil SL, Pan Y, Chatterjee SS, Cancer genome atlas research N, et al. Pan-Cancer analysis of lncRNA regulation supports their targeting of cancer genes in each tumor context. Cell Rep. 2018;23:297–312 e212.
28. Rosanova A, Colliva A, Osella M, Caselle M. Modelling the evolution of transcription factor binding preferences in complex eukaryotes. Sci Rep. 2017;7:7596.
29. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2016;45:D353–61.
30. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2018;46:D649–55.
31. Nishimura D. BioCarta. Biotech Software Intern Rep. 2001;2:117–20.
32. Elmasri R, Navathe S. Fundamentals of database systems. Boston: Addison-Wesley Publishing Company; 2010.
33. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–45.
34. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12:656–64.
35. Rambaldi D, Giorgi FM, Capuani F, Ciliberto A, Ciccarelli FD. Low duplicability and network fragility of cancer genes. Trends Genet. 2008;24: 427–30.
36. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2015;44:D286–93.
37. D'Antonio M, Ciccarelli FD. Modification of gene duplicability during the evolution of protein interaction network. PLoS Comput Biol. 2011;7: e1002029.
38. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347:1260419.
39. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N. The genotype-tissue expression (GTEx) project. Nat Genet. 2013;45:580.
40. Cancer Cell Line Encyclopedia Consortium, Genomics of Drug Sensitivity in Cancer Consortium: Pharmacogenomic agreement between two cancer cell line data sets**.** Nature 2015, 528**:**84.
41. Klijn C, Durinck S, Stawiski EW, Haverty PM, Jiang Z, Liu H, Degenhardt J, Mayba O, Gnad F, Liu J, et al. A comprehensive transcriptional portrait of human cancer cell lines. Nat Biotechnol. 2015;33:306–12.
42. Lenoir WF, Lim TL, Hart T. PICKLES: the database of pooled in-vitro CRISPR knockout library essentiality screens. Nucleic Acids Res. 2018;46:D776–80.
43. Chen WH, Lu G, Chen X, Zhao XM, Bork P. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. Nucleic Acids Res. 2017;45:D940–4.
44. Hart T, Tong AHY, Chan K, Van Leeuwen J, Seetharaman A, Aregger M, Chandrashekhar M, Hustedt N, Seth S, Noonan A, et al. Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. G3 (Bethesda). 2017;**7**:2719–27.
45. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, et al. The BioGRID interaction database: 2017 update. Nucleic Acids Res. 2017;45:D369–79.
46. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, et al. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014;42:D358–63.
47. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. Nucleic Acids Res. 2004;32: D449–51.
48. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al. Human protein reference database--2009 update. Nucleic Acids Res. 2009;37:D767–72.
49. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res. 2010;38:D497–501.
50. Chou C-H, Shrestha S, Yang C-D, Chang N-W, Lin Y-L, Liao K-W, Huang W-C, Sun T-H, Tu S-J, Lee W-H. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. 2017;46:D296–302.
51. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. 2009;37:D105–10.
52. Schaefer C, Grouse L, Buetow K, Strausberg RL. A new cancer genome anatomy project web resource for the community. Cancer J. 2001;7:52–60.
53. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013;339:1546–58.
54. Dulak AM, Stojanov P, Peng S, Lawrence MS, Fox C, Stewart C, Bandla S, Imamura Y, Schumacher SE, Shefler E, et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nat Genet. 2013;45:478–86.
55. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499:214–8.
56. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, et al. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 2012;22:1589–98.
57. Hao D, Wang L, Di LJ. Distinct mutation accumulation rates among tissues determine the variation in cancer risk. Sci Rep. 2016;6:19458.
58. Chu X-Y, Jiang L-H, Zhou X-H, Cui Z-J, Zhang H-Y. Evolutionary origins of cancer driver genes and implications for cancer prognosis. Genes. 2017;8:182.
59. Domazet-Loso T, Tautz D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. BMC Biol. 2010;8:66.

60. Li Y, Li L, Wang Z, Pan T, Sahni N, Jin X, Wang G, Li J, Zheng X, Zhang Y, et al. LncMAP: Pan-cancer atlas of long noncoding RNA-mediated transcriptional network perturbations. Nucleic Acids Res. 2018;46:1113–23.
61. D'Antonio M, Ciccarelli FD. Integrated analysis of recurrent properties of cancer genes to identify novel drivers. Genome Biol. 2013;14:R52.
62. Nishimura D: BioCarta**.** https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways. Accessed 16 Apr 2018.
63. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, et al: Biological General Repository for Interaction Datasets (BioGRID). re3data (https://doi.org/10.17616/r34c7g). Accessed 19 Feb 2018.
64. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, et al: The Cancer Cell Line Encyclopedia. https://portals.broadinstitute.org/ccle. Accessed 23 Mar 2018.
65. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L, et al: COSMIC Cancer Cell Line Project**.** https://cancer.sanger.ac.uk/cell_lines. Accessed 23 Mar 2018.
66. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW: The Comprehensive Resource of Mammalian protein complexes (CORUM). re3data (https://doi.org/10.17616/r3nk8k). Accessed 19 Feb 2018.
67. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: Database of Interacting Proteins (DIP). re3data (https://doi.org/10.17616/r3vg8d). Accessed 19 Feb 2018.
68. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M. eggNOG 4.5.1. FAIRsharing. https://doi.org/10.25504/fairsharing.j1wj7d. Accessed 16 Apr 2018.
69. Klijn C, Durinck S, Stawiski EW, Haverty PM, Jiang Z, Liu H, Degenhardt J, Mayba O, Gnad F, Liu J, et al: Genentech expression dataset.: EBI Array Express, accession number E-MTAB-2607 (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2607/). Accessed 23 Mar 2018.
70. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N. The genotype-tissue expression (GTEx) project. DataCite. https://doi.org/10.25491/j0aj-4w78. Accessed 23 Mar 2018.
71. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al: The Human Protein Reference Database (HPRD). re3data (https://doi.org/10.17616/r3mk9n). Accessed 19 Feb 2018.
72. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K: KEGG pathway database http://www.genome.jp/kegg/pathway.html. Accessed 16 Apr 2018.
73. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, et al: The MIntAct project. re3data (https://doi.org/10.17616/r3qs4r). Accessed 19 Feb 2018.
74. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T: miRecords v.4.0. http://c1.accurascience.com/miRecords/. Accessed 4 June 2018.
75. Chou C-H, Shrestha S, Yang C-D, Chang N-W, Lin Y-L, Liao K-W, Huang W-C, Sun T-H, Tu S-J, Lee W-H. miRTarBase v.7.0. FAIRsharing. https://doi.org/10.25504/fairsharing.f0bxfg. Accessed 4 June 2018.
76. Chen WH, Lu G, Chen X, Zhao XM, Bork P. Online Gene Essentiality database (OGEE). FAIRsharing. https://doi.org/10.25504/fairsharing.hsy066. Accessed 29 May 2018.
77. Lenoir WF, Lim TL, Hart T. Pooled In-vitro CRISPR Knockout Library Essentiality Screens (PICKLES) http://pickles.hart-lab.org/. Accessed 29 May 2018.
78. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al. The human protein atlas. FAIRsharing. https://doi.org/10.25504/fairsharing.tf6kj8. Accessed 23 Mar 2018.
79. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, et al: Reactome. re3data (https://doi.org/10.17616/r3ts52). Accessed 19 Feb 2018.
80. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al: Reference sequence (RefSeq) database release 85**.** FAIRsharing (https://doi.org/10.25504/FAIRsharing.4jg0qw). Accessed 11 Dec 2017.


Mutational Screenings
First author
Journal
Year
Primary Site
Method
Cancer genes (Entrez IDs)
Functional Annotations
Entrez ID
Pathway name
Pathway level
RefSeq Alignments
RefSeq protein ID
Entrez ID
Coverage
Identity
Gene coordinates
miRNA Interactions
Entrez ID
miRNA ID
PubMed IDs
Experimental validation
Tissue Expression Levels
Entrez ID
Sample
Tissue
Expression value
External Sources
Hg38.2bit
refseq.faa
gene2refseq
Genes
Entrez ID
Symbol
Aliases
Coordinates
Cell Line Expression Levels
Entrez ID
Cell line
Tissue
Expression value
Orthologues
Entrez ID
Orthologue
Species
Taxonomic group
Protein-Protein Interactions
Entrez ID 1
Entrez ID 2
PubMed IDs
Protein Complexes
Entrez ID
Complex ID
Components
PubMed IDs
Essentiality Measures
Entrez ID
Cell line
Tissue
Essentiality
M
< reports >
N
<has >
1
N
< has >
N
M
M
< involved in >
N
1
< has >
N
1
< has >
N
1
< has >
N
M
< component of >
N
M
< involved in >
N
M
< has >
N
< input >

Additional File 1: Figure S1. Schema of the NCG database

Entity-relationship diagram indicating one-to-many and many-to-many relationships between genes and other entities in the NCG database. The external source files used to generate the Genes entity are shown in grey.

Chapter 6. The Network of Cancer Genes and Healthy Drivers

6.1 Contributions

In this study[7], I collected and analysed the data on the evolutionary origins and orthologs of all human genes, and I analysed miRNA-target interactions together with Hrvoje Misetic. I performed data analysis on cancer gene annotations with Lisa Dressler, Mohamed Reda Keddar, Hrvoje Misetic, Giulia Sartini, Amelia Acha-Sagredo, and Francesca D. Ciccarelli. I also developed the database and website with Giulia Sartini, Jacki Goldman, Karen Ambrose, Mohamed Reda Keddar, Hrvoje Misetic, Lisa Dressler, Marc Pollit, Patrick Davis, and Amy Strange. Finally, I wrote the manuscript with Francesca D. Ciccarelli, Amelia Acha-Sagredo, Giulia Sartini, Lisa Dressler and Hrvoje Misetic, while all authors reviewed and approved its final version.

Lisa Dressler analysed data on protein-protein interactions, protein complex, gene essentiality, and cancer cell lines. Hrvoje Misetic analysed the TCGA data. Mohamed Reda Keddar analysed the gene duplicability. Giulia Sartini analysed the gene function, RNA and protein expression, and drug interactions. Amelia Acha-Sagredo, Lucia Montorsi, Neshika Wijewardhane, and Dimitra Repana curated the literature. Joel Nulsen analysed the germline variation. Francesca D. Ciccarelli conceived and supervised the study.

6.2 Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the Network of Cancer Genes (NCG) resource

Genome Biology

RESEARCH Open Access

Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the Network of Cancer Genes (NCG) resource

Lisa Dressler[1,2,3†], Michele Bortolomeazzi[1,2†], Mohamed Reda Keddar[1,2†], Hrvoje Misetic[1,2†], Giulia Sartini[1,2†], Amelia Acha-Sagredo[1,2†], Lucia Montorsi[1,2†], Neshika Wijewardhane[1,2], Dimitra Repana[1,2], Joel Nulsen[1,2], Jacki Goldman[4], Marc Pollitt[4], Patrick Davis[4], Amy Strange[4], Karen Ambrose[4] and Francesca D. Ciccarelli[1,2*]

* Correspondence: francesca.ciccarelli@crick.ac.uk
†Lisa Dressler, Michele Bortolomeazzi, Mohamed Reda Keddar, Hrvoje Misetic, Giulia Sartini, Amelia Acha-Sagredo and Lucia Montorsi contributed equally to this work.
[1]Cancer Systems Biology Laboratory, The Francis Crick Institute, London NW1 1AT, UK
[2]School of Cancer and Pharmaceutical Sciences, King's College London, London SE11UL, UK
Full list of author information is available at the end of the article

Abstract

Background: Genetic alterations of somatic cells can drive non-malignant clone formation and promote cancer initiation. However, the link between these processes remains unclear and hampers our understanding of tissue homeostasis and cancer development.

Results: Here, we collect a literature-based repertoire of 3355 well-known or predicted drivers of cancer and non-cancer somatic evolution in 122 cancer types and 12 non-cancer tissues. Mapping the alterations of these genes in 7953 pan-cancer samples reveals that, despite the large size, the known compendium of drivers is still incomplete and biased towards frequently occurring coding mutations. High overlap exists between drivers of cancer and non-cancer somatic evolution, although significant differences emerge in their recurrence. We confirm and expand the unique properties of drivers and identify a core of evolutionarily conserved and essential genes whose germline variation is strongly counter-selected. Somatic alteration in even one of these genes is sufficient to drive clonal expansion but not malignant transformation.

Conclusions: Our study offers a comprehensive overview of our current understanding of the genetic events initiating clone expansion and cancer revealing significant gaps and biases that still need to be addressed. The compendium of cancer and non-cancer somatic drivers, their literature support, and properties are accessible in the Network of Cancer Genes and Healthy Drivers resource at http://www.network-cancer-genes.org/.

Keywords: Driver genes, Somatic evolution, Cancer initiation, Systems-level properties

Background

Genetic alterations conferring selective advantages to cancer cells are the main drivers of cancer evolution and hunting for them has been at the core of international cancer genomic efforts [1–3]. Given the instability of the cancer genome, distinguishing driver alterations from the rest relies on analytical approaches that identify genes altered more frequently than expected or quantify the positive selection acting on them [4–6]. The results of these analyses have greatly expanded our understanding of the mechanisms driving cancer evolution, revealing high heterogeneity across and within cancers [7–9].

Recently, deep sequencing screens of non-cancer tissues have started to map positively selected genetic mutations in somatic cells that drive in situ formation of phenotypically normal clones [10, 11]. Many of these mutations hit cancer drivers, sometimes at a frequency higher than the corresponding cancer [12–16]. Yet, they do not drive malignant transformation. This conundrum poses fundamental questions on how genetic drivers of normal somatic evolution are related to and differ from those of cancer evolution. Addressing these questions will clarify the genetic relationship between tissue homeostasis and cancer initiation, with profound implications for cancer early detection.

To assess the extent of the current knowledge on cancer and non-cancer drivers, we undertook a systematic review of the literature and assembled a comprehensive repertoire of genes whose somatic alterations have been reported to drive cancer or non-cancer evolution. This allowed us to compare the current driver repertoire across and within cancer and non-cancer tissues and map their alterations in the large pancancer collection of samples from The Cancer Genome Atlas (TCGA). This revealed significant gaps and biases in our current knowledge of the driver landscape. We also computed an array of systems-level properties across driver groups, confirming the unique evolutionary path of driver genes and their central role in the cell.

We collected all cancer and non-cancer driver genes, together with a large set of their properties, in the Network of Cancer Genes and Healthy Drivers (NCG^{HD}) open-access resource.

Results

More than 3300 genes are canonical or candidate drivers of cancer and non-cancer somatic evolution

We conducted a census of currently known drivers through a comprehensive literature review of 331 scientific articles published between 2008 and 2020 describing somatically altered genes with a proven or predicted role in cancer or non-cancer somatic evolution (Fig. 1a). These publications included three sources of experimentally validated (canonical) cancer drivers, 311 sequencing screens of cancer (293) and non-cancer (18) tissues, and 17 pancancer studies (Additional file 1, Table S1). Each paper was assessed by at least two independent experts (Additional file 2, Fig. S1A-C) returning a total of 3355 drivers, 3347 in 122 cancer types and 95 in 12 non-cancer tissues, respectively (Fig. 1a). We further computed the systems-level properties of drivers and annotated their function, somatic variation, and drug interactions (Fig. 1a).

We reviewed the three sources of canonical cancer drivers [17–19] to exclude false positives (Additional file 3, Table S2) and fusion genes whose properties could not be

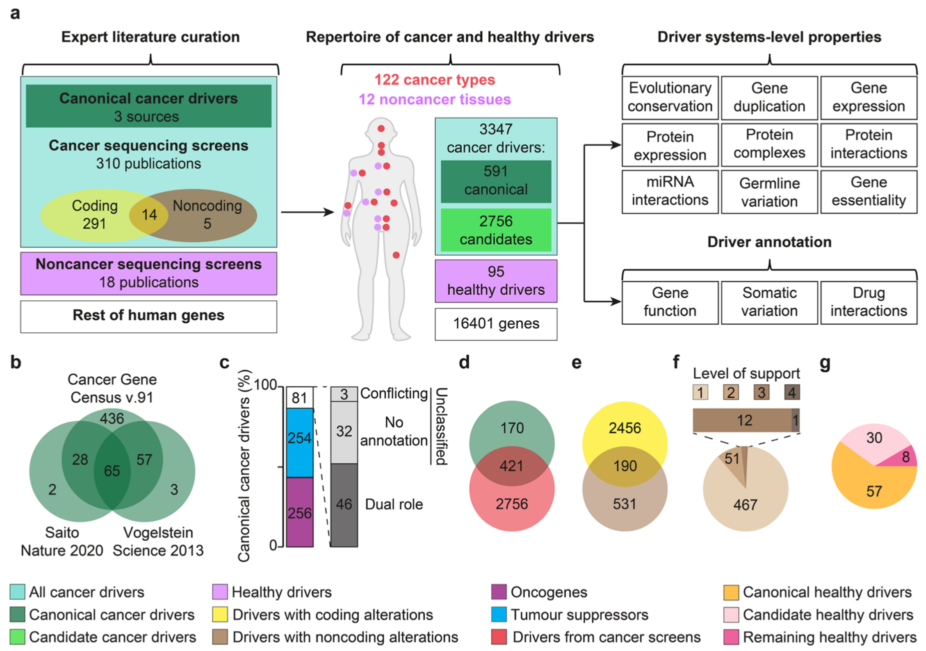


Fig. 1 Collection of a comprehensive repertoire of cancer and healthy drivers. **a** Literature review and driver annotation workflow. Expert literature curation of 331 publications led to a repertoire of cancer and healthy drivers in a variety of cancer and non-cancer tissues. Combining multiple data sources, a set of properties and annotations was computed for all these drivers. **b** Intersection of canonical drivers from three sources [17–19] that passed our manual curation. **c** Classification of canonical cancer drivers in tumor suppressors and oncogenes. Eighty-one cancer drivers had a dual role or could not be classified. **d** Intersection of canonical and candidate driver genes from 310 sequencing screens. Genes whose driver role had only statistical support were considered candidate cancer drivers. **e** Intersection between cancer drivers with coding and non-coding alterations. **f** Level of support for the driver role of 531 cancer genes with non-coding driver alterations only. Level 1 means that the gene was predicted as a driver only in one cancer sequencing screen; levels 2, 3, and 4 mean that it was predicted by two, three, or four screens or that it had experimental support. Experimental support was gathered from the 19 publications reporting non-coding cancer drivers (Additional file 1, Table S1) and from the CNCDatabase [20] and included in vitro and in vivo experiments, modification of gene expression, and survival association. **g** Proportion of healthy drivers that are also canonical or candidate cancer drivers, classified as canonical and candidate healthy drivers, respectively

mapped. Only 11% of the resulting 591 canonical drivers (Additional file 4, Table S3) were common to all three sources (Fig. 1b), indicating poor consensus even in well-known cancer genes. We further annotated the genetic mode of action for > 86% of canonical drivers, finding comparable proportions of oncogenes or tumor suppressors (Fig. 1c). The rest had a dual role or could not be univocally classified.

We extracted additional cancer drivers from the curation of 310 sequencing screens that applied a variety of statistical approaches (Additional file 2, Fig. S1 D) to identify cancer drivers among all altered genes. After removing possible false positives (Additional file 3, Table S2), the final list included 3177 cancer drivers, 2756 of which relied only on statistical support (candidate cancer drivers) and 421 were canonical drivers (Fig. 1d, Additional file 4, Table S3). Therefore, 170 canonical drivers have never been detected by any method, suggesting that they may elicit their role through non-mutational mechanisms or may fall below the detection limits of current approaches. Given the prevalence of cancer coding screens (Fig. 1a), only coding driver alterations have been reported for most genes (Fig. 1e) while 16% of them (531) were identified as drivers uniquely in non-coding screens. Since the prediction of drivers with non-coding alterations remains challenging, we further investigated the type of support that these genes had for their driver activity. The overwhelming majority of them (467 genes, 87%) have been predicted as drivers in only one screen. The remaining 64 genes are canonical drivers, have been predicted as drivers in multiple screens, or have additional experimental support for their driver activity (Fig. 1f).

Applying a similar approach (Additional file 2, Fig. S1 A-C), we reviewed 18 sequencing screens of healthy or diseased (non-cancer) tissues. They collectively reported 95 genes whose somatic alterations could drive non-malignant clone formation (healthy drivers). Interestingly, only eight of them were not cancer drivers (Fig. 1g, Additional file 4, Table S3), suggesting a high overlap between genetic drivers of cancer and non-cancer evolution. However, since many non-cancer screens only re-sequenced cancer genes or applied methods developed for cancer genomics (Additional file 2, Fig. S1E), this overlap may be overestimated.

The ability to capture cancer but not healthy driver heterogeneity increases with the donor sample size

To compare cancer and healthy drivers across and within tissues, we grouped the 122 cancer types and 12 non-cancer tissues into 12 and seven organ systems, respectively (the "Methods" section).

Despite the high numbers of sequenced samples (Additional file 5, Table S4) and detected drivers (Fig. 1), several lines of evidence indicated that our knowledge of cancer drivers is still incomplete. First, we detected a strong positive correlation between cancer drivers and donors overall (Fig. 2a) and in individual organ systems (Additional file 2, Fig. S2). This suggests that the current ability to identify new drivers depends on the number of samples included in the analysis. Second, candidates outnumbered canonical drivers in all organ systems except those with a small sample size or low mutation rate such as pediatric cancers, where only the most recurrent canonical drivers could be identified (Fig. 2b). Third, large donor cohorts enabled the detection of a broader representation of canonical drivers than small cohorts (Fig. 2c). For example, pooling thousands of samples together led to > 60% of canonical drivers being detected in adult pancancer re-analyses. Therefore, the size of the cohort influences the level of completeness and heterogeneity of the cancer driver repertoire. This is not surprising since all current approaches act at the cohort level, searching for positively selected genes altered more frequently than expected (Additional file2, Fig. S1D).

Our analysis also showed that the contribution of non-coding driver alterations remains largely unappreciated and non-coding drivers have not yet been reported in several tumors, including all pediatric cancers (Fig. 2d). Owing to the re-analysis of large whole-genome collections [21–26], almost 40% of adult pancancer drivers were instead modified by non-coding alterations (Fig. 2d). Hematologic and skin tumors also had a high proportion of non-coding driver variants thanks to screens focused on non-coding mutations [27, 28]. Therefore, the re-analysis of already available whole-genome data and further sequencing screens of non-coding variants are needed to fully appreciate their driver contribution.

Compared to cancer, sequencing screens of non-cancer tissues are still in their infancy, as reflected by the lower numbers of screened tissues and detected drivers

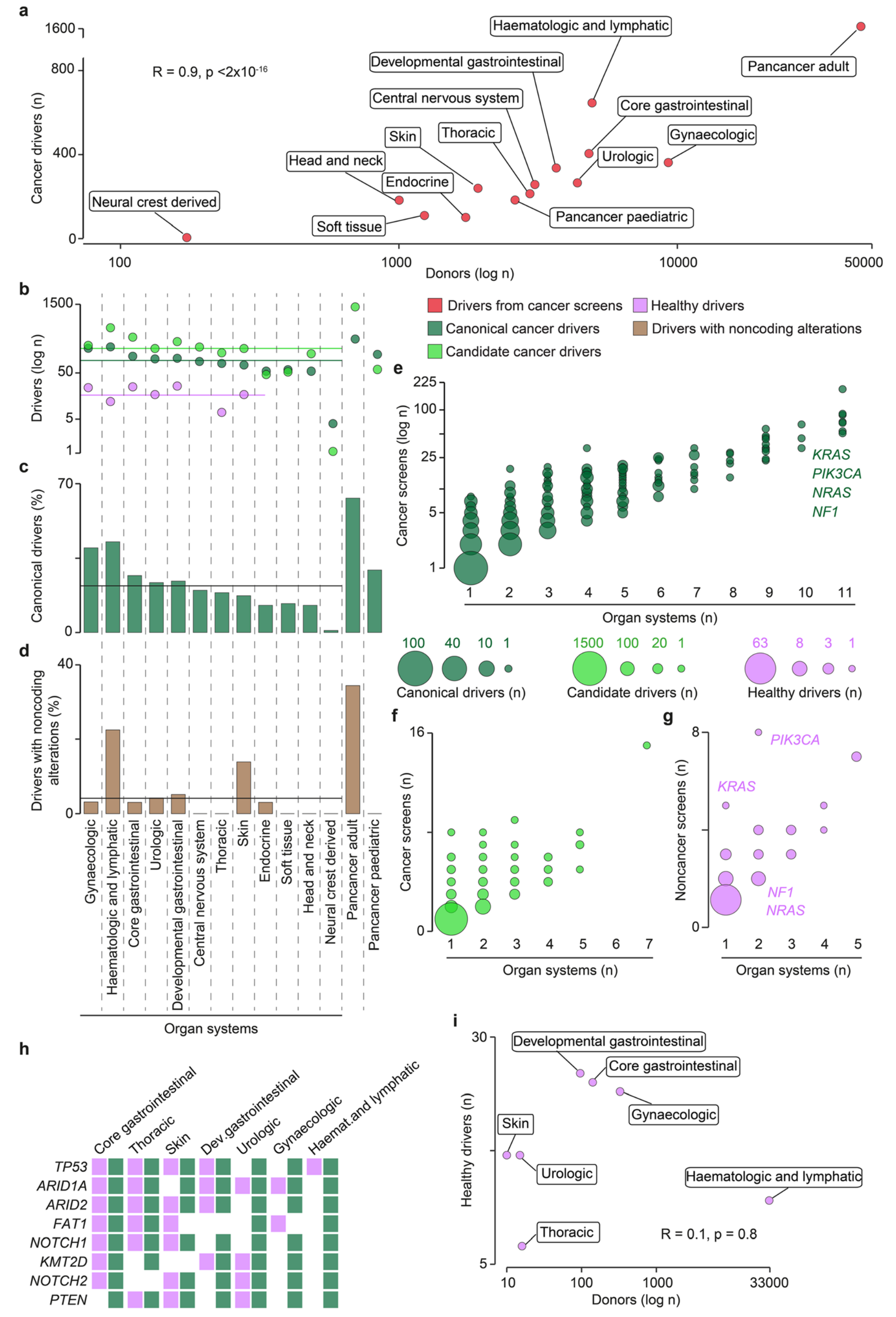


Fig. 2 Distribution of driver annotations by organ system. **a** Correlation between numbers of sequenced donors and identified cancer drivers across organ systems. Spearman correlation coefficient *R* and associated *p*-value are shown. **b** Number of canonical, candidate, and healthy drivers in each organ system. Horizontal lines indicate the median number of canonical (92), candidate (160), and healthy (17) drivers across organ systems. **c** Proportion of canonical drivers detected in each organ system over canonical drivers detected in all cancer screens (421). The horizontal line indicates the median across all organ systems (22%). **d** Proportion of genes with non-coding driver alterations over all cancer drivers in each organ system. The horizontal line indicates the median across all organ systems (4%). Number of canonical (**e**), candidate (**f**), and healthy (**g**) drivers across screens and organ systems. Representative genes with different recurrence between cancer and healthy tissues are indicated. **h** Organ system distribution of the top eight recurrent healthy drivers. The full list is provided as Additional file 6, Table S5. **i** Correlation between numbers of sequenced donors and identified healthy drivers across organ systems. Spearman correlation coefficient *R* and associated *p*-value are shown

(Fig. 2b). Despite this, some similarities and differences with cancer drivers could already be observed. Like cancer drivers (Fig. 2e, f, Additional file 6, Table S5), also healthy drivers were mostly organ-specific (Fig. 2g) and the most recurrent healthy drivers were also cancer drivers in the same organ system (Fig. 2h, Additional file 6, Table S5). However, some recurrent cancer drivers (*KRAS*, *PI3KCA*, *NRAS*, *NF1*) were reported to drive non-cancer clonal expansion only in one or two organ systems (Fig. 2g). Therefore, differences start to emerge at the tissue level between drivers of cancer and non-cancer evolution. Moreover, unlike cancer drivers, no correlation existed between the numbers of drivers and donors (Fig. 2i). This is likely affected by the lower number of non-cancer sequencing studies available so far. If additional studies will confirm the absence of correlation, this may indicate that the healthy driver repertoire is easier to saturate since fewer drivers are needed to initiate and sustain non-cancer clonal expansion [10, 11].

Alteration pattern hints at driver mode of action and confirms the incompleteness of the driver repertoire

To gain further insights into their mode of action, we mapped the type of alterations acquired by cancer and healthy drivers in 34 cancer types from TCGA. After predicting the damaging alterations in 7953 TCGA samples with matched mutation, copy number, and gene expression data (the “Methods” section), we identified the drivers with loss-of-function (LoF) and gain-of-function (GoF) alterations in these samples, respectively (Fig. 3a).

The comparison between canonical cancer drivers detected and undetected in sequencing screens (Fig. 1d) revealed that the latter were damaged in a significantly lower number of samples, due to fewer LoF alterations (Fig. 3b, Additional file 2, Fig. S3A). GoF alterations were instead comparable between the two groups, suggesting that current driver detection methods fail to identify drivers that undergo copy number gains but are rarely mutated.

We confirmed that the driver alteration patterns reflected their mode of action, with canonical tumor suppressors and oncogenes showing a prevalence of LoF and GoF alterations, respectively (Fig. 3c). Canonical drivers with a dual role resembled the alteration pattern of oncogenes while those still unclassified had a prevalence of LoF alterations, suggesting a putative tumor suppressor role (Fig. 3c). While all frequently altered (> 500 samples) oncogenes were overwhelmingly modified by GoF alterations (Additional file 7, Table S6), 16 of the 22 most frequently altered tumor suppressors had a prevalence of GoF alterations (Fig. 3d). In the majority of cases, this was due to different alteration patterns across organ systems (Additional file 2, Fig. S3B), and a possible oncogenic role has been documented for some others [29–38].

Since candidate drivers had no annotation of their mode of action, we reasoned that their alteration pattern could hint at their role as tumor suppressors or oncogenes. According to their prevalent pancancer alterations, 1318 candidates could be classified as putative tumor suppressors and 1405 as putative oncogenes (Additional file 7, Table S6). Interestingly, while candidates with predicted coding driver alterations showed similar distributions of LoF and GoF alterations (Fig. 3e), those with only non-coding driver alterations had a significantly lower occurrence of LoF alterations (Fig. 3f,

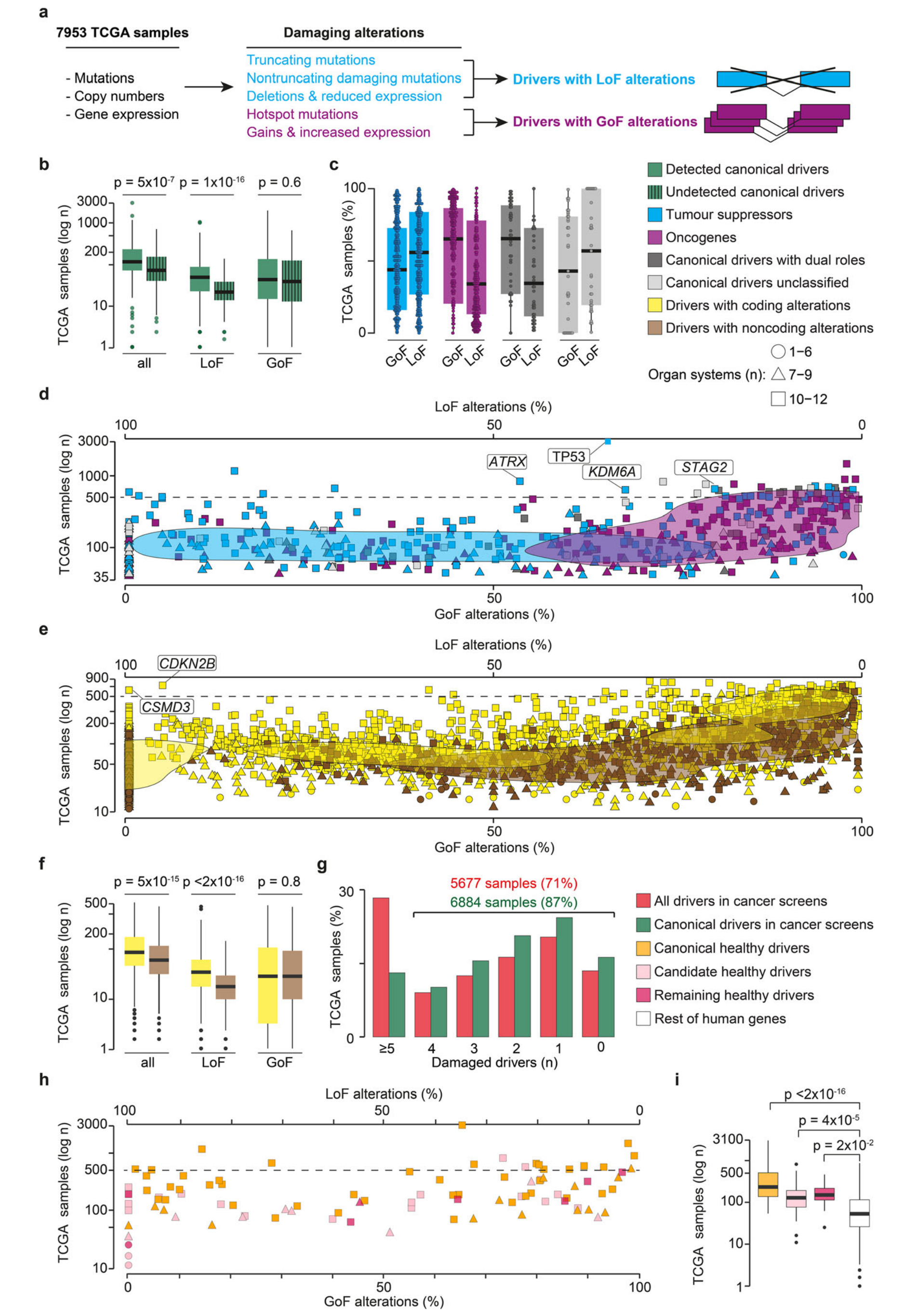


Fig. 3 (See legend on next page.)

(See figure on previous page.)
Fig. 3 Damaging alteration pattern of drivers in TCGA. **a** Identification of damaged drivers in 7953 TCGA samples. Mutations, gene deletions, and amplifications were annotated according to their predicted damaging effect. This allowed to distinguish drivers acquiring loss-of-function (LoF) or gain-of-function (GoF) alterations. **b** Number of TCGA samples with damaging alterations (all, LoF, GoF) in canonical drivers that were detected (421) or undetected (170) by cancer driver detection methods. **c** Proportion of TCGA samples with GoF and LoF alterations in tumor suppressors, oncogenes, and canonical drivers with a dual or unclassified role. Proportion of TCGA samples with GoF and LoF alterations in (**d**) canonical drivers and (**e**) candidate drivers. Genes mentioned in the text are highlighted. The two-dimensional Gaussian kernel density estimations were calculated for each driver group using the R density function. **f** Number of TCGA samples with damaging alterations (all, LoF, GoF) in drivers previously reported in coding and non-coding sequences. **g** Proportion of samples with variable numbers of all damaged drivers or only canonical drivers. **h** Proportion of TCGA samples with GoF and LoF alterations in healthy drivers. Canonical and candidate healthy drivers correspond to genes with a known or predicted cancer driver role. **i** Number of TCGA samples with damaged canonical, candidate, and remaining healthy drivers and the rest of human genes. All distributions were compared using a two-sided Wilcoxon rank-sum test

Additional file 2, Fig. S3C). This may suggest an activating role for their non-coding alterations too. Almost all candidates damaged in ≥ 500 samples (111/115) were putative oncogenes (Fig. 3e, Additional file 7, Table S6). Of the four putative tumor suppressors, *CSMD3* has a disputed cancer role [39–41] and a likely inflated mutation rate [42], while *CDKN2B* cooperates with its paralog *CDKN2A* to inhibit cell cycle [43], supporting its tumor suppressor role.

The number of damaged cancer drivers in individual TCGA samples confirmed that, despite all efforts, the current driver repertoire is still largely incomplete. The large majority of samples (71% and 87%, considering all drivers or only canonical drivers, respectively) had less than five damaged drivers, and ~ 15% of them had no damaged driver (Fig. 3g).

Given their high overlap with cancer drivers, most healthy drivers were recurrently damaged in cancer samples with no prevalence of GoF or LoF alterations (Fig. 3h, Additional file 7, Table S6). Interestingly, all healthy drivers, even the eight with no cancer involvement, were damaged in significantly more cancer samples than the rest of human genes (Fig. 3i). Moreover, 57% of TCGA samples had at least two altered drivers, one of which was a healthy driver, further supporting the hypothesis that more than one driver may be needed to promote the transformation of non-malignant clones into cancer [10, 11].

Properties of cancer and healthy drivers support their central role in the cell

A substantial body of work including our own [44–53] has shown that cancer drivers differ from the rest of the genes for an array of systems-level properties (Fig. 1a) that are a consequence of their unique evolutionary path and role in the cell. Using our granular annotation of drivers, we set out to check for similarities and differences across the driver groups.

We confirmed that cancer drivers, and in particular canonical drivers, were more conserved throughout evolution and less likely to retain gene duplicates than other human genes (Fig. 4a, Additional file 8, Table S7). They also showed broader tissue expression, engaged in a larger number of protein complexes, and occupied more central and highly connected positions in the protein-protein and miRNA-gene networks (Fig. 4a). We reported substantial differences between tumor suppressors and

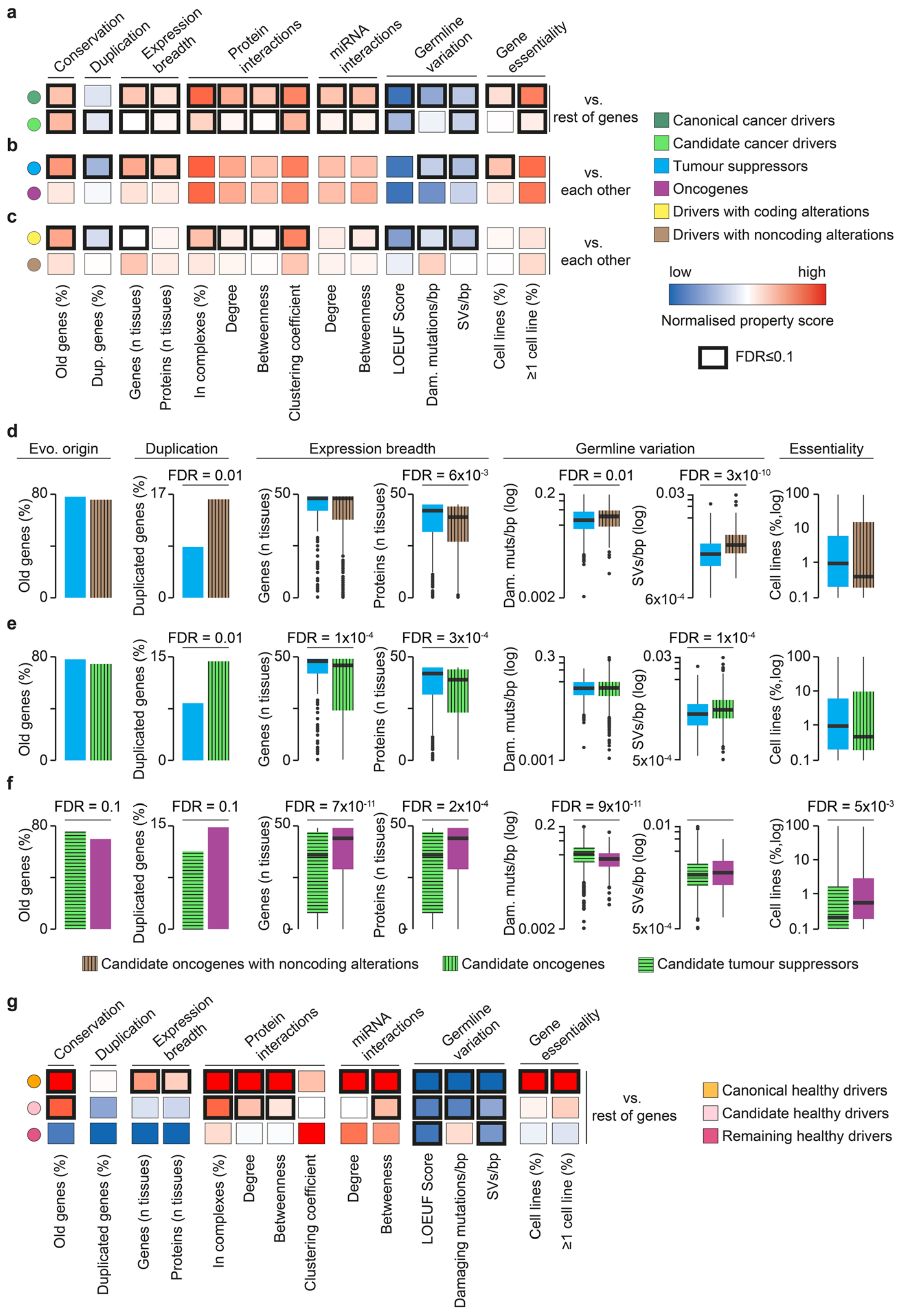


Fig. 4 (See legend on next page.)

(See figure on previous page.)
Fig. 4 Systems-level properties of cancer and healthy drivers. Comparisons of systems-level properties between (**a**) canonical or candidate cancer drivers and the rest of human genes, (**b**) tumor suppressors and oncogenes, and (**c**) cancer genes with coding driver alterations and cancer genes with non-coding driver alterations. The normalized property score was calculated as the normalized difference between the median (continuous properties) or proportion (categorical properties) values in each driver group and the rest of human genes (the "Methods" section). Comparisons of systems-level properties between (**d**) candidate oncogenes with non-coding driver alterations (324) and canonical tumor suppressors, (**e**) candidate oncogenes (1405) and canonical tumor suppressors, and (**f**) candidate tumor suppressors (1318) and canonical oncogenes. **g**. Comparisons of systems-level properties between canonical healthy, candidate healthy, and remaining healthy drivers and the rest of human genes. Proportions of old (pre-metazoan), duplicated, essential genes, and proteins involved in the complexes were compared using a two-sided Fisher's exact test. Distributions of gene and protein expression, protein-protein, miRNA-gene interactions, and germline variation were compared using a two-sided Wilcoxon rank-sum test. False discovery rate (FDR) was corrected for using Benjamini-Hochberg

oncogenes, with the former enriched in old and single-copy genes showing broader tissue expression (Fig. 4b, Additional file 8, Table S7).

We further expanded the systems-level properties of cancer drivers by exploring their tolerance towards germline variation, because this may indicate their essentiality. Using germline data from healthy individuals [54], we compared the loss-of-function observed/expected upper bound fraction (LOEUF) score, which quantifies selection towards LoF variation [54] as well as the number of damaging mutations and structural variants (SVs) per coding base pairs (bp) between drivers and the rest of genes (the "Methods" section). Cancer drivers, and in particular canonical drivers, had a significantly lower LOEUF score and retained fewer damaging germline mutations and SVs than the rest of the genes (Fig. 4a). This indicates that they are indispensable for cell survival in the germline. Selection against harmful variation was stronger in tumor suppressors than oncogenes (Fig. 4b). This was supported by a significantly higher proportion of cell lines where cancer drivers, and in particular tumor suppressors, were essential (Fig. 4a, b), as gathered from the integration of nine genome-wide essentiality screens [55–63] (the "Methods" section).

Genes with non-coding driver alterations had weaker systems-level properties than those with coding alterations (Fig. 4c, Additional file 8, Table S7) and the subset of them with > 50% GoF alterations resembled the property profile of oncogenes when compared to tumor suppressors (Fig. 4d, Additional file 8, Table S7). In general, all candidate drivers with a prevalence of GoF were similar to oncogenes, showing a higher proportion of duplicated genes, narrower tissue expression, and higher tolerance to germline variation than tumor suppressors (Fig. 4e, Additional file 8, Table S7). Conversely, candidate drivers with a prevalence of LoF were older, less duplicated, and less tolerant to germline variation than oncogenes (Fig. 4f, Additional file 8, Table S7).

Systems-level properties of healthy drivers varied according to the overlap with cancer drivers (Fig. 4g, Additional file 8, Table S7). Intriguingly, canonical healthy drivers showed stronger systems-level properties than any other group of drivers. In particular, they were enriched in evolutionarily conserved and broadly expressed genes encoding highly inter-connected proteins are regulated by many miRNAs. Moreover, these genes showed a strong selection against germline variation and high enrichment in essential genes (Fig. 4g). They therefore represent a core of genes with a very central role in the cell, whose modifications are not tolerated in the germline but are selected for in somatic cells because they confer selective growth advantages. Candidate healthy drivers and those not involved in cancer had a substantially different property profile (Fig. 4g). Although numbers are too low for any robust conclusion, it is tempting to speculate that genes able to initiate non-cancer clonal expansion but not tumorigenesis may follow a different evolutionary path.

The Network of Cancer Genes: an open-access repository of annotated drivers

We collected the whole repertoire of 3347 cancer and 95 healthy drivers, their literature support, and properties in the seventh release of the Network of Cancer Genes and Healthy Drivers (NCG^{HD}) database. NCG^{HD} is accessible through an open-access portal that enables interactive queries of drivers (Fig. 5a) as well as the bulk download of the database content.

In addition to the known or predicted mode of action and systems-level properties of cancer and healthy drivers, NCG^{HD} 7.0 also annotates their function, alteration pattern, and gene expression profile in TCGA and cancer cell lines, reported interactions with antineoplastic drugs, and potential role as treatment biomarkers (Fig. 5b). Altogether, this constitutes an extensive compendium of annotation of driver genes, including information relevant for planning experiments involving them.

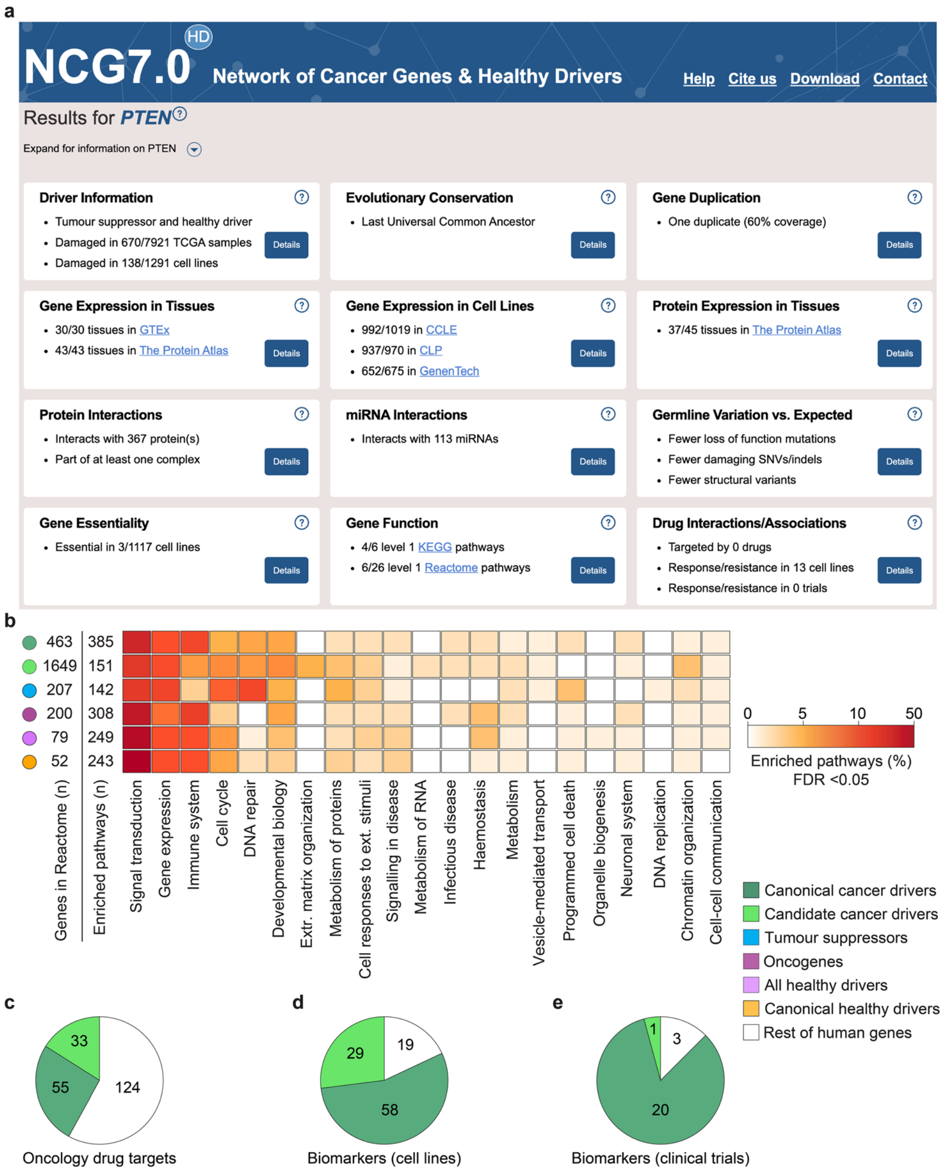
Functional gene set enrichment analysis showed that at least 60% of enriched pathways (FDR < 0.05) in any driver group converge to five broad functional processes (signal transduction, gene expression, immune system, cell cycle, and DNA repair, Fig. 5b, Additional file 9, Table S8). Within these, tumor suppressors showed a prevalence in cell cycle and DNA repair pathways, while oncogenes were enriched in the gene expression and immune system-related pathways (Additional file 9, Table S8). Healthy drivers closely resembled the functional profile of cancer drivers, given the high overlap (Fig. 5b). Because of the low number, it was not possible to assess the functional enrichment of healthy drivers not involved in cancer.

More than 9% of canonical cancer drivers are targets of anti-cancer drugs and cancer drivers constitute around 40% of their targets (Fig. 5c). Moreover, most of the genes used as biomarkers of resistance or response to treatment in cell lines (Fig. 5d) or clinical trials (Fig. 5e) are cancer drivers, with an overwhelming prevalence of canonical cancer drivers.

Discussion

The wealth of cancer genomic data and the availability of increasingly sophisticated analytical approaches for their interpretation have substantially improved the understanding of how cancer starts and develops. However, our in-depth analysis of the vast repertoire of drivers that have been collected so far shows clear limits in the current knowledge of the driver landscape.

The identification of drivers as genes under positive selection or with a higher than expected mutation frequency within a cohort of patients has biased the current cancer driver repertoire towards genes whose coding point mutations or small indels frequently recur across patients. This strongly impairs the ability to map the full extent of driver heterogeneity leading to an underappreciation of the driver contribution of rarely altered genes and those modified through non-coding or gene copy number alterations,



NCG7.0 HD Network of Cancer Genes & Healthy Drivers

Help Cite us Download Contact

Results for ***PTEN***

Expand for information on PTEN

Driver Information
- Tumour suppressor and healthy driver
- Damaged in 670/7921 TCGA samples
- Damaged in 138/1291 cell lines

Details

Evolutionary Conservation
- Last Universal Common Ancestor

Details

Gene Duplication
- One duplicate (60% coverage)

Details

Gene Expression in Tissues
- 30/30 tissues in GTEx
- 43/43 tissues in The Protein Atlas

Details

Gene Expression in Cell Lines
- 992/1019 in CCLE
- 937/970 in CLP
- 652/675 in GenenTech

Details

Protein Expression in Tissues
- 37/45 tissues in The Protein Atlas

Details

Protein Interactions
- Interacts with 367 protein(s)
- Part of at least one complex

Details

miRNA Interactions
- Interacts with 113 miRNAs

Details

Germline Variation vs. Expected
- Fewer loss of function mutations
- Fewer damaging SNVs/indels
- Fewer structural variants

Details

Gene Essentiality
- Essential in 3/1117 cell lines

Details

Gene Function
- 4/6 level 1 KEGG pathways
- 6/26 level 1 Reactome pathways

Details

Drug Interactions/Associations
- Targeted by 0 drugs
- Response/resistance in 13 cell lines
- Response/resistance in 0 trials

Details

Fig. 5 NCG^{HD} annotations of driver genes. **a** Example of the type of annotation provided in NCG^{HD} for cancer and healthy drivers (in this case *PTEN*). Annotation boxes can be expanded for further details, with the possibility of intersecting data interactively (for example, in the case of protein-protein or miRNA-gene interactions) and downloading data for local use. **b** Proportion of Reactome levels 2–8 enriched pathways mapping to the respective level 1 in each driver group. Enrichment was measured comparing the proportion of drivers in each pathway to that of the rest of human genes with a one-sided Fisher's exact test. FDR was calculated using Benjamini-Hochberg. The numbers of drivers and enriched Reactome pathways are reported for each group. Proportion of canonical and candidate cancer divers and rest of genes that are (**c**) targets of FDA-approved antineoplastic drugs or biomarkers of response or resistance to oncological drugs in (**d**) cancer cell lines and (**e**) clinical studies. The corresponding numbers for each group are also shown

particularly amplifications. It also results in a sizeable fraction of samples with very few or no cancer drivers. This gap can be solved by complementing cohort-level approaches with methods that account for all types of alterations and predict drivers in individual samples, for example identifying their network deregulations [64–66] or applying machine learning to identify driver alterations [67]. Alternatively, we have shown that systems-level properties capture the main features of cancer drivers, justifying their use for patient-level driver detection [68, 69].

Our comprehensive study has also shown that cancer sequencing screens have so far mostly focused on resequencing and analyzing the protein-coding portion of cancer genomes, leaving the contribution of non-coding drivers mostly uncovered. This bias may be addressed by performing additional cancer whole genome sequencing screens and improving analytical methods for the prediction of non-coding driver alterations.

Biases are starting to emerge also in the knowledge of healthy drivers. Many non-cancer sequencing screens only targeted cancer genes and healthy driver detection methods used so far were originally developed for cancer genomics. Both these factors may contribute at least in part to explain the high overlap between drivers of cancer and non-cancer evolution. An unbiased investigation of altered genes able to promote clonal expansion but not tumorigenesis could confirm whether their properties are indeed different from cancer drivers as suggested by our initial analysis on the few of them that have been identified so far. Additionally, the investigation of somatically mutated clones in non-cancer tissues has just started and new screens are continuously published. The integrated analysis of these new studies will broaden our understanding of non-cancer clonal expansion and further clarify its relationship with cancer transformation.

Our literature review did not cover driver genes deriving from chromosomal rearrangements or epigenetic changes because of their scattered annotations in the literature and difficulty in mapping their properties. Adding these genes to the repertoire when their knowledge will be mature will help close the gaps in the knowledge of the genetic drivers of tumorigenesis.

Conclusions

Our comprehensive analysis of cancer sequencing screens showed that the current repertoire of cancer driver genes is still incomplete and biased towards frequent mutations altering the gene coding sequence. This calls for the need for additional screens and methods to identify further coding and non-coding cancer drivers at single patient resolution. We confirmed the central role of cancer drivers within the cell, which is reflected in their evolutionary path and is shared by the majority of known healthy drivers. Further sequencing screens of healthy tissues are needed to clarify whether this is a feature of all genes whose mutations can driver non-cancer clonal expansion or there is a group of healthy drivers that underwent a different evolutionary path.

Methods

Literature curation

A literature search was carried out in PubMed, TCGA (https://www.cancer.gov/tcga) and ICGC (https://dcc.icgc.org/) to retrieve cancer screens published between 2018 and 2020 (Additional file 2, Fig. S1A). This resulted in 135 coding and 154 non-coding cancer screens. Of these, only 80 and 37 were retained after examining abstracts and full text, respectively. Criteria for removal were the absence of driver genes or driver detection methods and the impossibility to map non-coding driver alterations to genes. The 37 new cancer screens were added to 273 publications previously curated by our team

[70], totaling 310 publications (Additional file 1, Table S1). A similar literature search retrieved 24 sequencing screens of non-cancer tissues publications, 18 of which were retained after the abstract and full-text examination (Additional file 2, Fig. S1A; Additional file 1, Table S1). Each paper was reviewed independently by two experts and further discussed if annotations differed to extract the list of driver genes, the number of donors, the type of screen (whole-genome, whole-exome, target gene re-sequencing), the cancer or non-cancer tissues, and the driver detection method (Additional file 2, Fig. S1B).

Canonical cancer drivers were extracted from two publications [17, 18] and the Cancer Gene Census [71] v.91. In the latter case, all tiers 1 and 2 genes were retained, except those from genomic rearrangements leading to gene fusion (Additional file 2, Fig. S1B). Collected genes were further classified as tumor suppressor, oncogene, or having a dual role according to the annotation in the majority of sources. Genes with conflicting or unavailable annotation were left unclassified.

Drivers from cancer screens and canonical sources underwent further filtering (Additional file 2, Fig. S1C). First, they were intersected with a list of 148 possible false positives [18, 42]. After a manual check of the supporting evidence, two drivers were retained as canonical, five were considered as candidates, and 41 were removed (Additional file 3, Table S2). The three resulting lists (canonical drivers, drivers from cancer screens, and healthy drivers) were intersected to annotate canonical drivers in cancer screens, remaining drivers in cancer screens (candidate cancer drivers), canonical healthy drivers, candidate healthy drivers, and remaining healthy drivers (Additional file 2, Fig. S1C; Additional file 4, Table S3).

Cancer types and non-cancer tissues were mapped to organ systems using previous classification [72]. Cancer types not included in this classification were mapped based on their histopathology (retinoblastoma to central nervous system, vascular and peripheral nervous system cancers to soft tissue, penile tumors to urologic system).

Pancancer TCGA data

A dataset of 7953 TCGA samples with quality-controlled mutation (SNVs and indels), copy number, and gene expression data in 34 cancer types was assembled from the Genomic Data Commons portal I [73] (https://portal.gdc.cancer.gov/). Mutations were annotated with ANNOVAR [74] (April 2018) and dbNSFP [75] v3.0 and only those identified as exonic or splicing were retained. Damaging mutations included (1) truncating (stopgain, stoploss, frameshift) mutations, (2) missense mutations predicted by at least seven out of 10 predictors (SIFT [76], PolyPhen-2 HDIV [77], PolyPhen-2 HVAR, MutationTaster [78], MutationAssessor [79], LRT [80], FATHMM [81], PhyloP [82], GERP++RS [83], and SiPhy [84]), (3) splicing mutations predicted by at least one of two splicing-specific methods (ADA [75] and RF [75]), and (4) hotspot mutations identified with OncodriveCLUST [85] v1.0.0.

Copy number variant (CNV) segments, sample ploidy, and sample purity values were obtained from TCGA SNP arrays using ASCAT [86] v.2.5.2. Segments were intersected with the exonic coordinates of 19,756 human genes in hg19 and genes were considered to have CNV if at least 25% of their transcribed length was covered by a CNV segment. RNA-Seq data were used to filter out false-positive CNVs. Putative gene gains were defined as copy number (CN) > 2 times sample ploidy and the levels of expression were compared between samples with and without each gene gain using a two-sided Wilcoxon rank-sum test and corrected for multiple testing using Benjamini-Hochberg. Only gene gains with a false discovery rate (FDR) < 0.05 were retained. Homozygous gene losses had CN = 0 and fragments per kilobase per million (FPKM) values < 1 over sample purity. Heterozygous gene losses had CN = 1 or CN = 0 but FPKM values > 1 over sample purity. This resulted in 2,192,832 redundant genes damaged in 7921 TCGA samples.

In total, 518,115 genes were considered to acquire LoF alterations because they underwent homozygous deletion or had truncating, missense damaging, splicing mutations, or double hits (CN = 1 and LoF damaging mutation), while 1,674,717 genes were considered to acquire GoF alterations because they had a hotspot mutation or underwent gene gain with increased expression (Fig. 3a).

Systems-level properties

Protein sequences from RefSeq [87] v.99 were aligned to hg38 using BLAT [88]. Unique genomic loci were identified for 19,756 genes based on gene coverage, span, score, and identity [89]. Genes sharing at least 60% of their protein sequence were considered as duplicates [46].

Evolutionary conservation was assessed for 18,922 human genes using their orthologs in EggNOG [90] v.5.0. Genes were considered to have a pre-metazoan origin (and therefore conserved in evolution) if they had orthologs in prokaryotes, eukaryotes, or opisthokonts [53].

Gene expression for 19,231 genes in 49 healthy tissues was derived from the union of Protein Atlas [91] v.19.3 and GTEx [92] v.8. Genes were considered to be expressed in a tissue if their expression value was ≥ 1 transcript per million (TPM). Protein expression for 13,229 proteins in 45 healthy tissues was derived from Protein Atlas [91] v.19.3 retaining the highest value when multiple expression values were available.

A total of 542,397 non-redundant binary interactions between 17,883 proteins were gathered from the integration of five sources (BioGRID [93] v.3.5.185, IntAct [94] v.4.2.14, DIP [95] (February 2018), HPRD [96] v.9 and Bioplex [97] v.3.0). Data on 9476 protein complexes involving 8504 proteins were derived from CORUM [98] v.3.0, HPRD [96] v.9 and Reactome [99] v.72. Experimentally supported interactions between 14,747 genes and 1758 miRNAs were acquired from miRTarBase [100] v.8.0 and miRecords [101] v.4.0. Degree, betweenness, and clustering coefficient were calculated for protein and miRNA networks using the igraph R package [102] v.1.2.6.

The loss-of-function observed/expected upper bound fraction (LOEUF) score for 18,392 genes was obtained from gnomAD [54] v.2.1.1. Germline mutations (SNVs and indels) were obtained from the union of 2504 samples from the 1000 Genomes Project Phase 3 [103] v.5a and 125,748 samples from gnomAD [54] v.2.1.1. Mutations were annotated with ANNOVAR [74] (October 2019), and 18,812 genes were considered as damaged using the same definitions as for TCGA samples. A total of 32,558 germline SVs for 14,158 genes were derived using 15,708 samples from gnomAD [54] v.2.1.1. The numbers of damaging mutations and SVs per base pairs (bp) were calculated for each gene.

Essentiality data for 19,013 genes in 1122 cell lines were obtained integrating three RNAi knockdown and six CRISPR Cas9 knockout screens [55–63]. Genes with CERES [57] or DEMETER [63] scores < -1 or Bayes score [104] > 5 were considered as essential.

Proportions of duplicated, pre-metazoan, essential genes, and proteins engaging in complexes were compared between the gene groups using two-sided Fisher's exact test. Distributions of tissues where genes or proteins were expressed, protein and miRNA network properties, LOEUF scores, damaging mutations, and SVs per bp were compared between the gene groups using a two-sided Wilcoxon test. Multiple comparisons within each property were corrected using Benjamini-Hochberg. For each systems-level property in each driver group (*d*), a normalized property score was calculated as:

$$\text{Normalised property score} = \operatorname{sgn}(\Delta_d) \times \frac{|\Delta_d| - \min_t |\Delta_t|}{\max_t |\Delta_\mathrm{t}| - \min_t |\Delta_t|}$$

where *t* represents 11 gene groups (canonical drivers, candidate drivers, tumor suppressors, oncogenes, drivers with coding alterations, drivers with non-coding alterations, canonical healthy drivers, candidate healthy drivers, remaining healthy drivers, and the rest of human genes); $\operatorname{sgn}(\Delta_d)$ is the sign of the difference; and Δ_d indicates the difference of medians (continuous properties) or proportions (categorical properties) between each driver group and the rest of human genes. Minima and maxima were taken over all 11 gene groups for each property.

Pancancer cell line data

Mutation, CNV and gene expression data for 1291 cell lines were obtained from DepMap [56, 105] v. 20Q3. Mutations were functionally annotated using ANNOVAR [74] and LoF mutations were identified as described for TCGA samples. Hotspot mutations were detected using hotspot positions derived from TCGA. Homozygous gene deletions were defined as CN < 0.25 times cell line ploidy and expression < 1 TPM; heterozygous gene deletions were defined as $0.25 <$ CN < 0.75 times cell line ploidy; gene gains were defined as CN > 2 times cell line ploidy and significantly higher expression relative to cell lines with no gene gains. Genes with LoF or GoF alterations were defined as for TCGA samples. To map cell lines to organ systems, they were first associated with the TCGA cancer types and then the same classification as for TCGA was used [72].

Driver functional annotation

Gene functions were collected for 11,778 proteins from Reactome [99] v.72 and KEGG [106] v.94.1 (levels 1 and 2). Driver enrichment in Reactome pathways (levels 2–8) compared to the rest of human genes was assessed using a one-sided Fisher's exact test and corrected for multiple testing with Benjamini-Hochberg. Enriched pathways were then mapped to the corresponding Reactome level 1.

Drug interactions

A total of 247 FDA-approved, antineoplastic, and immunomodulating drugs targeting 212 human genes were downloaded from DrugBank [107] v.5.1.8. Genetic biomarkers of response and resistance to drugs in cancer cell lines were obtained from Genomics of Drug Sensitivity in Cancer (GDSC) [108] v.8.2. Of those, only 467 associations with FDR ≤ 0.25 involving 129 drugs and 106 genes were retained. Genetic biomarkers of response and resistance in clinical studies were obtained from the Variant Interpretation for Cancer Consortium Meta-Knowledgebase [109] v.1. A total of 868 associations between drugs and genomic features involving 64 anti-cancer drugs and drug combinations and 24 human genes were retained [109].

Database and website implementation

All annotations of driver genes were entered into a relational database based on MySQL [110] v.8.0.21 connected to a web interface enabling interactive retrieval of information through gene identifiers. The frontend was developed with PHP [111] v.7.4.15. The interactive displays of miRNA-gene and protein-protein interactions were implemented with the R packages Shiny [112] v.1.6.0 and igraph [102] v.1.2.6 and ran on Shiny Server v1.5.16.958.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-022-02607-z.

Additional file 1: Table S1. Publications describing driver genes.

Additional file 2: Figure S1. Literature search, review and annotation workflow; **Figure S2.** Correlation between numbers of donors and cancer drivers in individual organ systems; **Figure S3.** Patterns of driver damaging alterations in TCGA samples.

Additional file 3: Table S2. Putative false positive cancer drivers.

Additional file 4: Table S3. Canonical cancer drivers.

Additional file 5: Table S4. Donors in cancer and noncancer sequencing screens.

Additional file 6: Table S5. Drivers reported in cancer and non-cancer screens.

Additional file 7: Table S6. Cancer and non-cancer drivers damaged in TCGA.

Additional file 8: Table S7. Systems-level properties of driver genes.

Additional file 9: Table S8. Proportion of enriched pathways across driver groups.

Additional file 10. Review history.

Acknowledgements

We thank Steve Hindmarsh and Stefan Boeing for their contribution to the development of the NCG database and website.

Peer review information

Anahita Bishop and Barbara Cheifet were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 10.

Authors' contributions

LD analyzed the protein-protein interactions, protein complex, gene essentiality, and cancer cell line data. MB analyzed the gene conservation. MRK analyzed the gene duplicability. HM analyzed the TCGA data. MB and HM analyzed the miRNA-target interactions. GS analyzed the gene function, RNA and protein expression, and drug interactions. AAS, LM, NW, and DR curated the literature. JN analyzed the germline variation. GS, MB, JG, and KA developed the database. MRK, HM, LD, MB, MP, PD, and AS developed the website. LD, MB, MRK, HM, GS, AAS, and FDC analyzed the data. FDC conceived and supervised the study. MB, AAS, GS, and FDC wrote the manuscript with contributions from LD and HM. The authors reviewed and approved the final manuscript.

Funding

This work was supported by the Cancer Research UK [C43634/A25487], the Cancer Research UK King's Health Partners Centre at King's College London [C604/A25135], the Cancer Research UK City of London Centre [C7893/A26233], the innovation programme under the Marie Skłodowska-Curie grant agreement [CONTRA-766030], the EPSRC Centre for Doctoral Training in Cross-Disciplinary Approaches to Non-Equilibrium Systems (CANES, EP/L015854/1), the Health Education England Genomics Education Programme, and the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001002), the UK Medical Research Council (FC001002), and the Wellcome Trust (FC001002). For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Availability of data and materials

The whole content of NCG^{HD} can be freely downloaded from the website (http://network-cancer-genes.org/). No license is required.
Original data were obtained from the following online sources:
1000 Genomes Project Phase 3 [103] v.5a: https://www.internationalgenome.org/category/phase-3/
BioGRID [93] v.3.5.185: https://thebiogrid.org/
Bioplex [97] v.3.0: https://bioplex.hms.harvard.edu/interactions.php
CORUM [98] v.3.0: http://mips.helmholtz-muenchen.de/corum/
Depmap [59, 60] v20Q3: https://depmap.org/portal/
DIP [95] (February 2018): https://dip.doe-mbi.ucla.edu/dip/Main.cgi
DrugBank [107] v.5.1.8: https://go.drugbank.com/
EggNog [90] v.5: http://eggnog5.embl.de/#/app/home
GDSC [108] v.8.2: https://www.cancerrxgene.org/
GnomAD [54] v.2.1.1: https://gnomad.broadinstitute.org/
GTEx [92] v.8: https://gtexportal.org/home/
HPRD [96] v.9: https://www.hprd.org/
IntAct [94] v.4.2.14: https://www.ebi.ac.uk/intact/home
KEGG [106] v.94.1: https://www.genome.jp/kegg/
Meta-KB [109] v.1: https://cancervariants.org/
MiRecords [101] v.8.0: http://c1.accurascience.com/miRecords/
MiTarBase [100] v.4.0: https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2022/php/index.php
NCI Genomics Data Commons Portal [73]: https://gdc.cancer.gov/
PICKLES [61]: https://pickles.hart-lab.org/
Protein Atlas [91] v.19.3: https://www.proteinatlas.org/
Reactome [99] v.72: https://reactome.org/
RefSeq [87] v.99: https://www.ncbi.nlm.nih.gov/refseq/

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Cancer Systems Biology Laboratory, The Francis Crick Institute, London NW1 1AT, UK. [2]School of Cancer and Pharmaceutical Sciences, King's College London, London SE11UL, UK. [3]Department of Medical and Molecular Genetics, King's College London, London SE1 9RT, UK. [4]Scientific Computing, The Francis Crick Institute, London NW1 1AT, UK.

Received: 30 August 2021 Accepted: 10 January 2022
Published online: 26 January 2022

References

1. Network CGAR. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061–8. https://doi.org/10.1038/nature07385.
2. International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. Nature. 2010;464:993–8.
3. Hutter C, Zenklusen JC. The Cancer Genome Atlas: creating lasting value beyond its data. Cell. 2018;173(2):283–5. https://doi.org/10.1016/j.cell.2018.03.042.
4. Pon JR, Marra MA. Driver and passenger mutations in cancer. Annu Rev Pathol. 2015;10(1):25–50. https://doi.org/10.1146/annurev-pathol-012414-040312.
5. Porta-Pardo E, Kamburov A, Tamborero D, Pons T, Grases D, Valencia A, et al. Comparison of algorithms for the detection of cancer drivers at subgene resolution. Nat Methods. 2017;14(8):782–8. https://doi.org/10.1038/nmeth.4364.
6. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020;20(10):555–72. https://doi.org/10.1038/s41568-020-0290-x.
7. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173(2):371–85 e18. https://doi.org/10.1016/j.cell.2018.02.060.
8. Consortium ITP-CAoWG. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82–93. https://doi.org/10.1038/s41586-020-1969-6.
9. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502(7471):333–9. https://doi.org/10.1038/nature12634.
10. Wijewardhane N, Dressler L, Ciccarelli FD. Normal somatic mutations in cancer transformation. Cancer Cell. 2021;39(2):125–9. https://doi.org/10.1016/j.ccell.2020.11.002.

11. Kakiuchi N, Ogawa S. Clonal expansion in non-cancer tissues. Nat Rev Cancer. 2021;21(4):239–56. https://doi.org/10.1038/s41568-021-00335-3.
12. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science. 2015;348(6237):880–6. https://doi.org/10.1126/science.aaa6806.
13. Tang J, Fewings E, Chang D, Zeng H, Liu S, Jorapur A, et al. The genomic landscapes of individual melanocytes from human skin. Nature. 2020;586(7830):600–5. https://doi.org/10.1038/s41586-020-2785-8.
14. Yokoyama A, Kakiuchi N, Yoshizato T, Nannya Y, Suzuki H, Takeuchi Y, et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature. 2019;565(7739):312–7. https://doi.org/10.1038/s41586-018-0811-x.
15. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age. Science. 2018;362(6417):911–7. https://doi.org/10.1126/science.aau3879.
16. Suda K, Nakaoka H, Yoshihara K, Ishiguro T, Tamura R, Mori Y, et al. Clonal expansion and diversification of cancer-associated mutations in endometriosis and normal endometrium. Cell Rep. 2018;24(7):1777–89. https://doi.org/10.1016/j.celrep.2018.07.037.
17. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science. 2013; 339(6127):1546–58. https://doi.org/10.1126/science.1235122.
18. Saito Y, Koya J, Araki M, Kogure Y, Shingaki S, Tabata M, et al. Landscape and function of multiple mutations within individual oncogenes. Nature. 2020;582(7810):95–9. https://doi.org/10.1038/s41586-020-2175-2.
19. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18(11):696–705. https://doi.org/10.1038/s41568-018-0060-1.
20. Liu EM, Martinez-Fundichely A, Bollapragada R, Spiewack M, Khurana E. CNCDatabase: a database of non-coding cancer drivers. Nucleic Acids Res. 2021;49(D1):D1094–D101. https://doi.org/10.1093/nar/gkaa915.
21. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82–93. https://doi.org/10.1038/s41586-020-1969-6.
22. Hornshoj H, Nielsen MM, Sinnott-Armstrong NA, Switnicki MP, Juul M, Madsen T, et al. Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival. NPJ Genom Med. 2018;3(1):1. https://doi.org/10.1038/s41525-017-0040-5.
23. Juul M, Bertl J, Guo Q, Nielsen MM, Switnicki M, Hornshoj H, et al. Non-coding cancer driver candidates identified with a sample- and position-specific model of the somatic mutation rate. Elife. 2017;6. https://doi.org/10.7554/eLife.21778.
24. Zhu H, Uuskula-Reimand L, Isaev K, Wadi L, Alizada A, Shuai S, et al. Candidate cancer driver mutations in distal regulatory elements and long-range chromatin interaction networks. Mol Cell. 2020;77(6):1307–21 e10. https://doi.org/10.1016/j.molcel.2019.12.027.
25. Lanzos A, Carlevaro-Fita J, Mularoni L, Reverter F, Palumbo E, Guigo R, et al. Discovery of cancer driver long noncoding RNAs across 1112 tumour genomes: new candidates and distinguishing features. Sci Rep. 2017;7(1):41544. https://doi.org/10.1038/srep41544.
26. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, Lopez-Bigas N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol. 2016;17(1):128. https://doi.org/10.1186/s13059-016-0994-0.
27. Cornish AJ, Hoang PH, Dobbins SE, Law PJ, Chubb D, Orlando G, et al. Identification of recurrent noncoding mutations in B-cell lymphoma using capture Hi-C. Blood Adv. 2019;3(1):21–32. https://doi.org/10.1182/bloodadvances.2018026419.
28. Hayward NK, Wilmott JS, Waddell N, Johansson PA, Field MA, Nones K, et al. Whole-genome landscapes of major melanoma subtypes. Nature. 2017;545(7653):175–80. https://doi.org/10.1038/nature22071.
29. Botlagunta M, Vesuna F, Mironchik Y, Raman A, Lisok A, Winnard P Jr, et al. Oncogenic role of DDX3 in breast cancer biogenesis. Oncogene. 2008;27(28):3912–22. https://doi.org/10.1038/onc.2008.33.
30. Pu J, Wang J, Qin Z, Wang A, Zhang Y, Wu X, et al. IGF2BP2 promotes liver cancer growth through an m6A-FEN1-dependent mechanism. Front Oncol. 2020;10:578816. https://doi.org/10.3389/fonc.2020.578816.
31. Sun X, Jia M, Sun W, Feng L, Gu C, Wu T. Functional role of RBM10 in lung adenocarcinoma proliferation. Int J Oncol. 2019;54(2):467–78. https://doi.org/10.3892/ijo.2018.4643.
32. Soussi T, Wiman KG. TP53: an oncogene in disguise. Cell Death Differ. 2015;22(8):1239–49. https://doi.org/10.1038/cdd.2015.53.
33. Yang MH, Chang SY, Chiou SH, Liu CJ, Chi CW, Chen PM, et al. Overexpression of NBS1 induces epithelial-mesenchymal transition and co-expression of NBS1 and Snail predicts metastasis of head and neck cancer. Oncogene. 2007;26(10): 1459–67. https://doi.org/10.1038/sj.onc.1209929.
34. Manandhar S, Kim CG, Lee SH, Kang SH, Basnet N, Lee YM. Exostosin 1 regulates cancer cell stemness in doxorubicin-resistant breast cancer cells. Oncotarget. 2017;8(41):70521–37. https://doi.org/10.18632/oncotarget.19737.
35. Li A, Zhu X, Wang C, Yang S, Qiao Y, Qiao R, et al. Upregulation of NDRG1 predicts poor outcome and facilitates disease progression by influencing the EMT process in bladder cancer. Sci Rep. 2019;9(1):5166. https://doi.org/10.1038/s41598-019-41660-w.
36. Meacham CE, Lawton LN, Soto-Feliciano YM, Pritchard JR, Joughin BA, Ehrenberger T, et al. A genome-scale in vivo loss-of-function screen identifies Phf6 as a lineage-specific regulator of leukemia cell growth. Genes Dev. 2015;29(5):483–8. https://doi.org/10.1101/gad.254151.114.
37. Sesen J, Casaos J, Scotland SJ, Seva C, Eisinger-Mathason TS, Skuli N. The bad, the good and eIF3e/INT6. Front Biosci (Landmark Ed). 2017;22:1–20.
38. Shi J, Zhang L, Zhou D, Zhang J, Lin Q, Guan W, et al. Biological function of ribosomal protein L10 on cell behavior in human epithelial ovarian cancer. J Cancer. 2018;9(4):745–56. https://doi.org/10.7150/jca.21614.
39. Liu P, Morrison C, Wang L, Xiong D, Vedell P, Cui P, et al. Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. Carcinogenesis. 2012;33(7):1270–6. https://doi.org/10.1093/carcin/bgs148.
40. Lai MW, Liang KH, Lin WR, Huang YH, Huang SF, Chen TC, et al. Hepatocarcinogenesis in transgenic mice carrying hepatitis B virus pre-S/S gene with the sW172* mutation. Oncogenesis. 2016;5(12):e273. https://doi.org/10.1038/oncsis.2016.77.

41. Cai C, Cooper GF, Lu KN, Ma X, Xu S, Zhao Z, et al. Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference. PLoS Comput Biol. 2019;15(7):e1007088. https://doi.org/10.1371/journal.pcbi.1007088.
42. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214–8. https://doi.org/10.1038/nature12213.
43. Hannon GJ, Beach D. pI5INK4B is a potentia| effector of TGF-β-induced cell cycle arrest. Nature. 1994;371:257–61.
44. Syed AS, D'Antonio M, Ciccarelli FD. Network of Cancer Genes: a web resource to analyze duplicability, orthology and network properties of cancer genes. Nucleic Acids Res. 2010;38(suppl_1):D670–D75. https://doi.org/10.1093/nar/gkp957.
45. Trigos AS, Pearson RB, Papenfuss AT, Goode DL. Somatic mutations in early metazoan genes disrupt regulatory links between unicellular and multicellular genes in cancer. eLife. 2019;8:e40947. https://doi.org/10.7554/eLife.40947.
46. Rambaldi D, Giorgi FM, Capuani F, Ciliberto A, Ciccarelli FD. Low duplicability and network fragility of cancer genes. Trends Genet. 2008;24(9):427–30. https://doi.org/10.1016/j.tig.2008.06.003.
47. Domazet-Loso T, Tautz D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. BMC Biol. 2010;8(1):66. https://doi.org/10.1186/1741-7007-8-66.
48. D'Antonio M, Ciccarelli FD. Integrated analysis of recurrent properties of cancer genes to identify novel drivers. Genome Biol. 2013;14(5):R52. https://doi.org/10.1186/gb-2013-14-5-r52.
49. Ostrow SL, Barshir R, DeGregori J, Yeger-Lotem E, Hershberg R. Cancer evolution is associated with pervasive positive selection on globally expressed genes. PLoS Genet. 2014;10(3):e1004239. https://doi.org/10.1371/journal.pgen.1004239.
50. An O, Dall'Olio GM, Mourikis TP, Ciccarelli FD. NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. Nucleic Acids Res. 2016;44:D992–9.
51. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. Bioinformatics. 2006; 22(18):2291–7. https://doi.org/10.1093/bioinformatics/btl390.
52. Xia J, Sun J, Jia P, Zhao Z. Do cancer proteins really interact strongly in the human protein-protein interaction network? Comput Biol Chem. 2011;35(3):121–5. https://doi.org/10.1016/j.compbiolchem.2011.04.005.
53. D'Antonio M, Ciccarelli FD. Modification of gene duplicability during the evolution of protein interaction network. PLoS Comput Biol. 2011;7(4):e1002029. https://doi.org/10.1371/journal.pcbi.1002029.
54. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43. https://doi.org/10.1038/s41586-020-2308-7.
55. Dempster JM, Rossen J, Kazachkova M, Pan J, Kugener G, Root DE, et al. Extracting biological insights from the project Achilles genome-Scale CRISPR screens in cancer cell lines. BioRxiv. 2019;720243. https://doi.org/10.1101/720243.
56. Broad D. DepMap 20Q3 Public, figshare. Dataset. 2020. https://doi.org/10.6084/m9.figshare.12931238.v1.
57. Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat Genet. 2017;49(12):1779–84. https://doi.org/10.1038/ng.3984.
58. Behan FM, Iorio F, Picco G, Goncalves E, Beaver CM, Migliardi G, et al. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. Nature. 2019;568(7753):511–6. https://doi.org/10.1038/s41586-019-1103-9.
59. DepMap Broad. Project SCORE processed with CERES. figshare. Dataset. 2019. https://doi.org/10.6084/m9.figshare.9116732.
60. DepMap Broad. DepMap GeCKO 19Q1. figshare. Fileset. 2019. https://doi.org/10.6084/m9.figshare.7668407.
61. Lenoir WF, Lim TL, Hart T. PICKLES: the database of pooled in-vitro CRISPR knockout library essentiality screens. Nucleic Acids Res. 2018;46(D1):D776–D80. https://doi.org/10.1093/nar/gkx993.
62. McFarland JM, Ho ZV, Kugener G, Dempster JM, Montgomery PG, Bryan JG, et al. Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. Nat Commun. 2018;9(1):4610. https://doi.org/10.1038/s41467-018-06916-5.
63. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a cancer dependency map. Cell. 2017;170(3):564–76 e16. https://doi.org/10.1016/j.cell.2017.06.010.
64. Bertrand D, Chng KR, Sherbaf FG, Kiesel A, Chia BK, Sia YY, et al. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. Nucleic Acids Res. 2015;43(7):e44. https://doi.org/10.1093/nar/gku1393.
65. Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, et al. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol. 2012;13(12):R124. https://doi.org/10.1186/gb-2012-13-12-r124.
66. Hou JP, Ma J. DawnRank: discovering personalized driver genes in cancer. Genome Med. 2014;6(7):56. https://doi.org/10.1186/s13073-014-0056-8.
67. Dong C, Guo Y, Yang H, He Z, Liu X, Wang K. iCAGES: integrated CAncer GEnome Score for comprehensively prioritizing driver genes in personal cancer genomes. Genome Med. 2016;8(1):135. https://doi.org/10.1186/s13073-016-0390-0.
68. Nulsen J, Misetic H, Yau C, Ciccarelli FD. Pan-cancer detection of driver genes at the single-patient resolution. Genome Med. 2021;13(1):12. https://doi.org/10.1186/s13073-021-00830-0.
69. Mourikis TP, Benedetti L, Foxall E, Temelkovski D, Nulsen J, Perner J, et al. Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma. Nat Commun. 2019;10(1):3101. https://doi.org/10.1038/s41467-019-10898-3.
70. Repana D, Nulsen J, Dressler L, Bortolomeazzi M, Venkata SK, Tourna A, et al. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol. 2019; 20(1):1. https://doi.org/10.1186/s13059-018-1612-0.
71. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019;47(D1):D941–D47. https://doi.org/10.1093/nar/gky1015.
72. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell. 2018;173:291–304.e6.
73. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a shared vision for cancer genomic data. N Engl J Med. 2016;375(12):1109–12. https://doi.org/10.1056/NEJMp1607591.
74. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164. https://doi.org/10.1093/nar/gkq603.

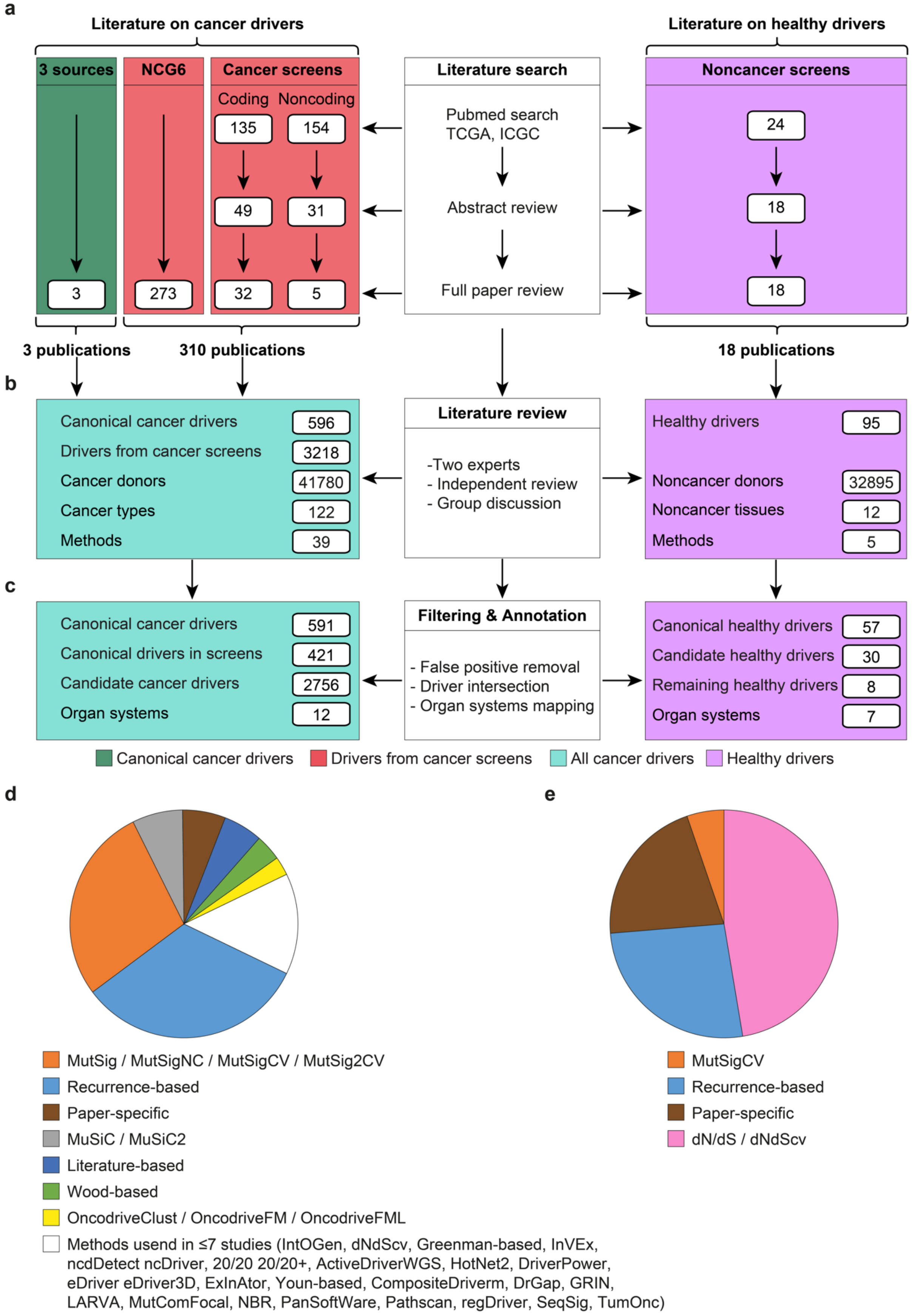
75. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Hum Mutat. 2016;37(3):235–41. https://doi.org/10.1002/humu.22932.
76. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4. https://doi.org/10.1093/nar/gkg509.
77. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7.20.
78. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7(8):575–6. https://doi.org/10.1038/nmeth0810-575.
79. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118. https://doi.org/10.1093/nar/gkr407.
80. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553–61. https://doi.org/10.1101/gr.092619.109.
81. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34(1):57–65. https://doi.org/10.1002/humu.22225.
82. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20(1):110–21. https://doi.org/10.1101/gr.097857.109.
83. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6(12):e1001025. https://doi.org/10.1371/journal.pcbi.1001025.
84. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;25(12):i54–62. https://doi.org/10.1093/bioinformatics/btp190.
85. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics. 2013;29(18):2238–44. https://doi.org/10.1093/bioinformatics/btt395.
86. Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A. 2010;107(39):16910–5. https://doi.org/10.1073/pnas.1009843107.
87. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–D45. https://doi.org/10.1093/nar/gkv1189.
88. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res. 2002;12(4):656–64.
89. Bhagwat M, Young L, Robison RR. Using BLAT to find sequence similarity in closely related genomes. Curr Protoc Bioinformatics. 2012;37(1):10.8.1–10.8.24. https://doi.org/10.1002/0471250953.bi1008s37.
90. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47(D1):D309–D14. https://doi.org/10.1093/nar/gky1085.
91. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419. https://doi.org/10.1126/science.1260419.
92. Consortium G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369(6509):1318–30. https://doi.org/10.1126/science.aaz1776.
93. Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res. 2019;47(D1):D529–D41. https://doi.org/10.1093/nar/gky1079.
94. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014;42(D1):D358–63. https://doi.org/10.1093/nar/gkt1115.
95. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. Nucleic Acids Res. 2004;32(90001):D449–51. https://doi.org/10.1093/nar/gkh086.
96. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database--2009 update. Nucleic Acids Res. 2009;37(Database):D767–72. https://doi.org/10.1093/nar/gkn892.
97. Huttlin EL, Bruckner RJ, Navarrete-Perea J, Cannon JR, Baltier K, Gebreab F, et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell. 2021;184(11):3022–40 e28. https://doi.org/10.1016/j.cell.2021.04.011.
98. Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, et al. CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res. 2019;47(D1):D559–D63. https://doi.org/10.1093/nar/gky973.
99. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020;48(D1):D498–503. https://doi.org/10.1093/nar/gkz1031.
100. Huang H-Y, Lin Y-C-D, Li J, Huang K-Y, Shrestha S, Hong H-C, et al. miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. Nucleic Acids Res. 2020;48:D148–D54.
101. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA–target interactions. Nucleic Acids Res. 2009;37(Database):D105–D10. https://doi.org/10.1093/nar/gkn851.
102. Csardi G, Nepusz T. The igraph software package for complex network research. InterJ Complex Syst. 2006;1695:1–9.
103. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. https://doi.org/10.1038/nature15393.
104. Hart T, Moffat J. BAGEL: a computational framework for identifying essential genes from pooled library screens. BMC Bioinformatics. 2016;17(1):164. https://doi.org/10.1186/s12859-016-1015-8.
105. Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569(7757):503–8. https://doi.org/10.1038/s41586-019-1186-3.
106. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2016;45(D1):D353–D61. https://doi.org/10.1093/nar/gkw1092.
107. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074–d82. https://doi.org/10.1093/nar/gkx1037.

108. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of pharmacogenomic interactions in cancer. Cell. 2016;166(3):740–54. https://doi.org/10.1016/j.cell.2016.06.017.
109. Wagner AH, Walsh B, Mayfield G, Tamborero D, Sonkin D, Krysiak K, et al. A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. Nat Genet. 2020;52(4):448–57. https://doi.org/10.1038/s41588-020-0603-8.
110. MySQL 8.0 reference manual. https://dev.mysql.com/doc/refman/8.0/en/.
111. Bakken S, Suraski Z, Schmid E. PHP manual; 2020.
112. Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, et al. shiny: web application framework for R. 2021.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Fig. S1. Literature search, review and annotation workflow.



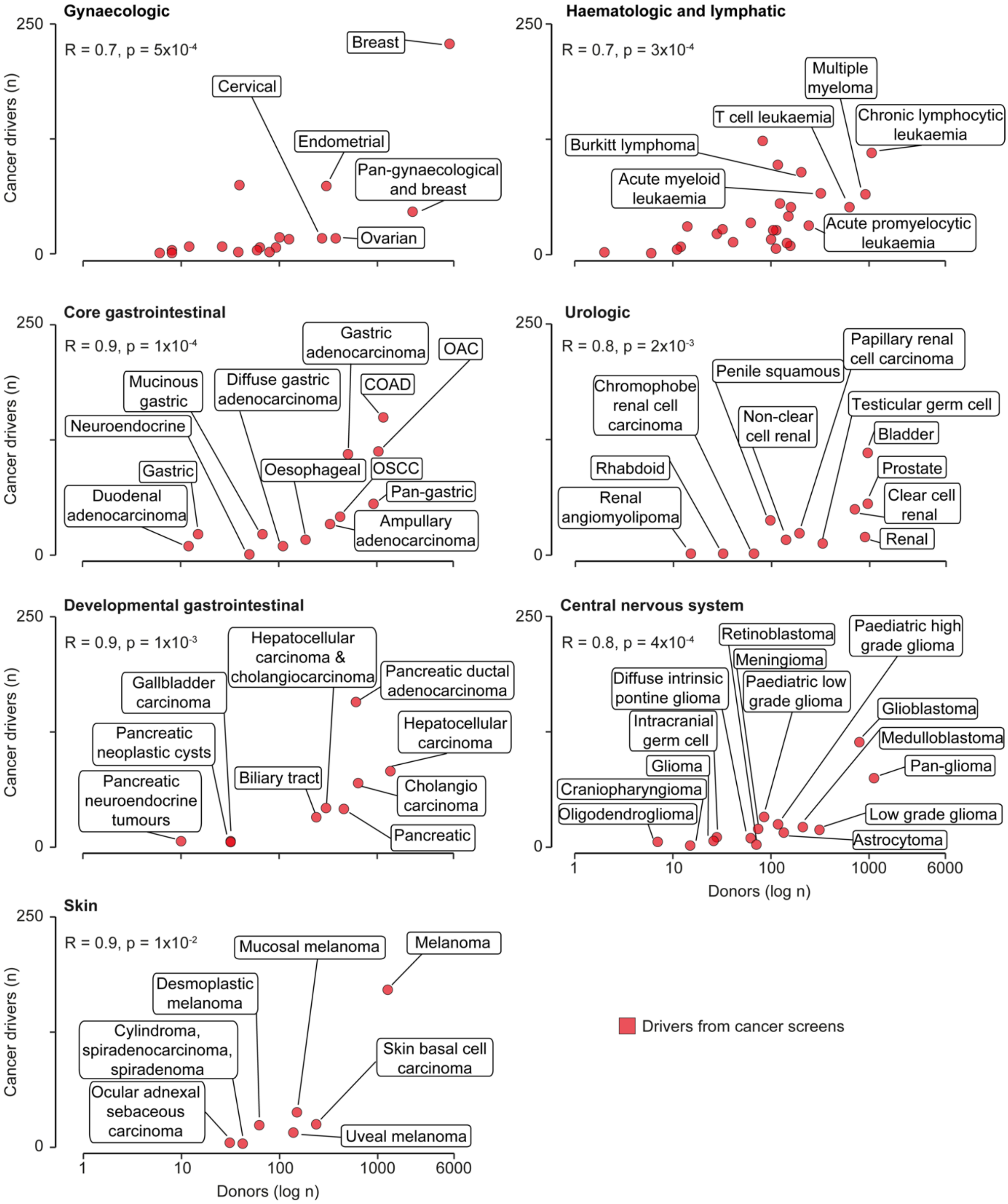
a. Publications reporting cancer and noncancer sequencing screens from the initial literature search in Pubmed, TCGA and ICGC were reviewed at the abstract and full text levels and added to three sources of canonical cancer genes and previously curated papers in NCG 6.0 [1] for a total of 331 publications.

b. Two experts reviewed each publication independently and conflicting annotations were further discussed. Lists of canonical cancer drivers, drivers from cancer and noncancer screens, cancer types and noncancer tissues, and methods used to detect drivers were annotated. Additionally, the number of cancer and noncancer donors were extracted.

c. The resulting lists of drivers were filtered out for possible false positives (**Additional File 3, Table S2**) and intersected to annotate the canonical drivers in cancer screens, candidate cancer drivers (remaining drivers in cancer screens), canonical healthy drivers, candidate healthy drivers, and remaining healthy drivers. Cancer types and noncancer tissues were mapped to organ systems [2]. The full workflow is explained in the Methods.

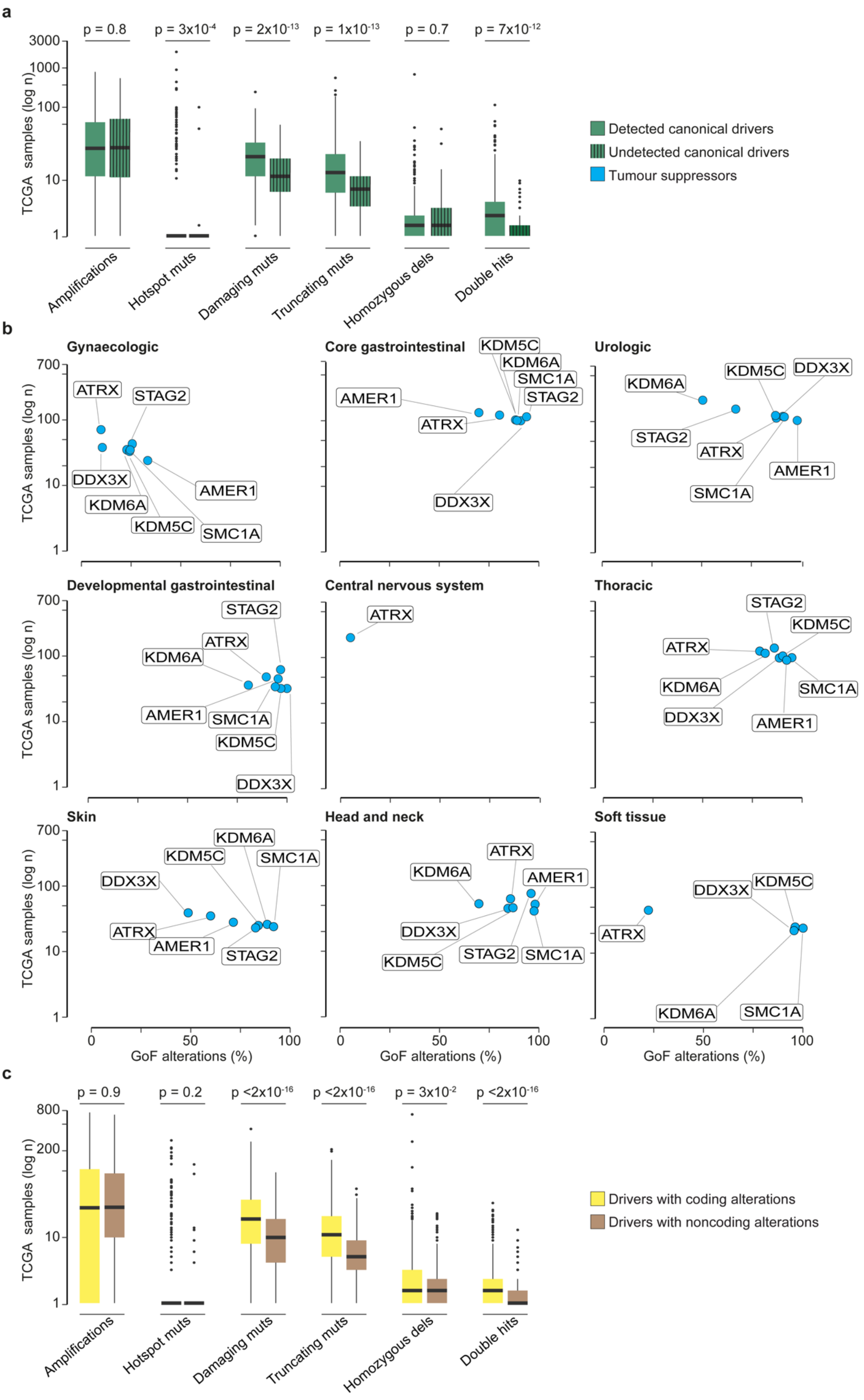
Usage of driver detection methods across (**d**) cancer and (**e**) noncancer screens. If a screen used several methods, it was counted multiple times. Multiple versions of the same method as well as methods used in less than 8 studies were aggregated. The full list is available in **Additional File 1, Table S1**.

Fig. S2. Correlation between numbers of donors and cancer drivers in individual organ systems



Correlations between numbers of sequenced donors and identified cancer drivers in individual cancer types mapping to each organ system. Only organ systems with significant correlations are reported. Spearman correlation coefficient R and associated p-value are shown. OAC: oesophageal adenocarcinoma. OSCC: oesophageal squamous cell carcinoma. COAD: colorectal adenocarcinoma.

Fig. S3: Patterns of driver damaging alterations in TCGA samples.



a. Number of TCGA samples with damaging alterations (all, LoF, GoF) in canonical drivers that were detected (421) or undetected (170) by cancer detection methods, divided by type of damaging alterations.

b. Proportion of gain of function (GoF) alterations affecting seven frequently damaged (>500 samples) canonical tumour suppressors. All these genes had an organ system-specific prevalence of GoF and loss of function (LoF) alterations.

c. Number of TCGA samples with damaging alterations in genes previously reported to have coding or only noncoding driver alterations, divided by type of damaging alterations.

All distributions were compared using a two-sided Wilcoxon rank-sum test.

References

1. Repana D, Nulsen J, Dressler L, Bortolomeazzi M, Venkata SK, Tourna A, Yakovleva A, Palmieri T, Ciccarelli FD: **The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens.** *Genome Biology* 2019, **20:**1.
2. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al: **Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer.** *Cell* 2018, **173:**291-304.e296.

Chapter 7. Discussion

Recent scientific and clinical advances have led to the widespread adoption of ICI to treat several tumour types, including late-stage CRC. Despite the vast progress made in this field, many patients still receive no durable clinical benefit from these therapies. This is due to innate resistance if the treatment has no effect from the beginning or to acquired resistance if, after an initial benefit from treatment, the tumour stops responding and resumes its unchecked growth. The mechanisms of innate and acquired resistance are still not fully understood due to the complex nature of the network that regulates the physiological and pathological function of the immune system as well as its interactions with the other cell populations constituting the TME[112].

In addition to the complexity of the TME, the other factor impairing a complete understanding of the mechanisms underlying resistance to immunotherapy is cancer heterogeneity[112]. Both innate and acquired resistance are driven by heterogeneity, which provides the pre-existing genetic alterations enabling immune escape[42]. Some patients will have innate resistance because of the specific genetic, epigenetic and TME features of their tumour, even if they would be expected to respond to therapy given their histologic and tumour features[42]. Acquired resistance is also driven by heterogeneity, but within the tumour. Genetic alterations, either pre-existing or developed during treatment, can confer a fitness advantage to one or more clones under the new selective pressures introduced by immunotherapy[2, 26]. Thus, ITH constitutes the primary factor driving the emergence of acquired resistance.

The impact of inter-tumour heterogeneity in clinical practice is evidenced by the large percentage of patients who do not respond to ICI agents despite showing features associated with response, such as high TMB or PDL1 expression in the tumour[47, 48]. The genetic alterations responsible for primary immune escape are likely to have accumulated over months or years while being subject to the selective pressure exerted by the endogenous host immune response[113]. For this reason, these alterations are generally present in pre-existing tumour samples, from which they can be identified through sequencing the cancer genome. However, there are still significant differences in mechanism across primary sites and cancer types[113].

Besides inter-tumour heterogeneity, genetic ITH also shapes the TME and response to immunotherapy, as each clone interacts with the local environment according to its specific epigenetic and genetic makeup [24, 25]. For this reason, studies have taken into account inter-tumour genetic heterogeneity when investigating the TME across different tumours. For instance, Ock et al. [114] have shown how the potential benefit of anti-CTLA4 immunotherapy is highly variable across cancer types and how genomic alterations can significantly affect the efficacy of an immunotherapy regime[114]. Thus, resources like the TCGA[3] and COSMIC[6] database have been instrumental in providing the scientific community with a vast repository of genomic and transcriptomic data as well as the functional annotation of mutations instrumental in quantifying and interpreting inter-tumour heterogeneity. Another of these resources is the NCG database[7], whose most recent updates are described in chapters five and six.

NCG[7], differs from TCGA[3], COSMIC[6] and similar resources because instead of reporting mutations, it provides a comprehensive repository of genes that have been identified as proven or predicted cancer drivers. In our update of the NCG database, we investigated driver heterogeneity across primary sites and cancer types, showing that the number of cancer genes, as well as the ratio of candidate to known cancer genes, varies significantly across primary sites[100]. This variability is likely due not only to differences in the number of studies and overall sample sizes of the analysed cohorts, but also to variable levels of genomic instability and other factors, such as exposure to different agents like tobacco smoke[10].

We updated the NCG database to its next iteration NCGHD [7] by not only updating the repertoire of cancer genes and their system-level properties but also extending the analysis to genes that drive non-malignant clone formation in non-cancer tissues[7]. Additionally, we expanded our annotations to include genes whose driver role depends on mutations in non-coding elements, thus further expanding the range of genetic inter-tumour heterogeneity captured by NCG[7]. As part of the database update, we added to our annotations a collection of drug-gene interaction data, including both genes, which are targets of FDA-approved oncological drugs, and genes that are biomarkers of antineoplastic drug resistance or response in clinical studies and cancer cell lines[7].

Our analysis did not include cancer drivers originating from chromosomal rearrangements or epigenetic alterations in the literature review. This was because of the still limited availability of their annotations in the literature and the inherent complexity of mapping their properties to specific genomic locations[7].

Adding these genes to the repertoire when their knowledge is mature will help close the gaps in the knowledge of the genetic drivers of tumorigenesis. The annotations we produced for NCG have frequently employed the studies described in this thesis. First, a list of CRC drivers from NCG was employed for the analysis of somatic mutation and gene expression data in our study on the impact of ITH on the response to anti-PD1 immunotherapy in CRC[107]. Then, NCG was also a resource for designing all the antibody panels for IMC and mIF employed in this study.

Genetic Inter-tumour heterogeneity on its own does not account for the full spectrum of response to ICI treatment as this can only be understood by considering the whole TME, including its tumoral, stromal and immune components[14]. Only such an analysis will enable the development of a set of biomarkers sufficient to guide the clinical application of ICI therapies. This challenge will need to be tackled through approaches like those described in section 1.3. While spatial and single-cell omics enable the investigation of ITH, highly-multiplexed image analysis will complement these approaches by providing an ensemble view of the tissue composition and structure[55, 60]. This is because only highly-multiplexed image data allows the spatially resolved characterisation of the cell-cell interaction network underlying all pathological functions of the TME[56].

Despite the significant development of highly-multiplexed imaging technologies, the computational analysis of their output was still challenging with the available software tools[94]. These tools have significant drawbacks, including low scalability and reproducibility, which are compounded by the lack of portability of most tools [94], while many image analysis projects require resources available only in HPC environments [60, 89]. imcyto[87] and MCMICRO[89], published at the same time as SIMPLI, have partially addressed these issues. However, these tools have fixed workflows, which offer very little control on which analysis steps are executed. This limits their adaptability to the many challenges posed by the analysis of the TME, whose many cell types and levels of organisation require highly flexible analysis approaches. We developed SIMPLI, a pipeline that provides high portability, reproducibility and scalability while still leaving the user in control of its workflow[99]. This flexibility was not obtained at the expense of ease of use, as in our experience, new users can learn to run their analysis after less than a day of training.

The user-friendliness and flexibility will need to be maintained in future iterations of SIMPLI for it to remain relevant in the rapidly developing field of highly-multiplexed image analysis. The containerised implementation of SIMPLI makes the inclusion of additional tools and features easy to implement. This adaptability is crucial, as no software tool for pixel classification, cell segmentation, or phenotyping has yet reached the status of a gold standard in the field, and new approaches are constantly being developed. This is particularly the case for the cell-segmentation step, where machine learning-based approaches are constantly evolving, and many new open-source tools are being developed[77]. Additionally, SIMPLI would benefit from implementing a pixel classification tool to increase its flexibility in the pixel-level analysis step. This would enable users to adopt a supervised or an unsupervised approach or both, as is the case for the cell-level analysis.

These new developments should also conform to the latest standards for software pipelines and image data and metadata sharing. For instance, SIMPLI would benefit from adopting the nf-core[95] framework for pipeline development. Implementing this framework would make SIMPLI immediately accessible to a broad community of users and developers. This framework would also mandate the adoption of scientific-computing best practices like automated continuous-integration tests and consistent version control and Digital Object Identifier (DOI) linking. Apart from nf-core, SIMPLI would benefit from the implementation of an interface to OMERO[115] to automatically manage image data and metadata and retrieve them from remote servers. Finally, all updates to SIMPLI will need to be promptly documented and shared with the user base through SIMPLI's extensive wiki (https://github.com/ciccalab/SIMPLI/wiki), which already provides a detailed user manual for the current version of the software.

We applied SIMPLI to the analysis of IMC data in a multiregional multiomic study to characterise the genetic and TME-related factors driving the tumour-immune interactions determining anti-PD1 ICI response in CRC[107]. SIMPLI's unique cell masking and phenotyping process allowed us to first identify cells belonging to the main immune cell populations in the TME and then phenotype these populations by unsupervised clustering and expression thresholding in a single run. This allowed the in-depth characterisation of the heterogeneity of several immune cell populations across hypermutated and non-hypermutated CRCs, as well as DB and nDB CRCs. We then validated all these observations through SIMPLI's highly flexible pixel analysis approach, which enabled us to quantify the combinations of markers better describing each cell phenotype.

This cell– and pixel-level analysis found that hypermutated CRCs are enriched in proliferating and cytotoxic CD8^{+} T cells compared to non-hypermutated tumours[107]. This enrichment was probably linked to the low levels of Wnt activity in these tumours, as we observed genes of the Wnt pathway to be more frequently downregulated and subjected to damaging mutations in hypermutated compared with non-hypermutated CRCs. Additionally, Wnt downstream targets were significantly downregulated in hypermutated CRCs both in our cohort and in TCGA[3]. Our analysis confirms previous observations of CD8^{+} cell depletion in CRCs[116, 117] with high WNT pathway activity and provides a more detailed characterisation of these cells, phenotyping them as CD8^{+}GzB^{+} or CD8^{+}Ki67^{+} [107].

Across hypermutated CRCs, we observed the same levels of enrichment in T cells and their subpopulations between DB- and nDB tumours[107]. Additionally, there was no difference in PD1 and PDL1 expression and complex formation, and we identified no differences in T cell composition or activation that segregated with response. However, we observed variability in the level of disruption of antigen presentation mechanisms and the distribution of activated subpopulations of antigen-presenting cells[107].

We found that DB tumours presented more frequently transcriptional dysregulation or damaging mutations in multiple immune-related processes, including the interferon-gamma pathway and MHCI antigen presentation[107]. For instance, B2M protein expression derived from the analysis of IMC images with SIMPLI was significantly lower in the tumour but not in the stroma of DB tumours. Moreover, the analysis of IMC images enabled us to verify the absence of B2M expression from two CRCs harbouring clonal truncating mutations in B2M. These results, not observable through bulk RNA-seq only, highlight the importance of highly-multiplexed image analysis in multiomic studies[107]. While B2M loss has been previously observed in DB CRCs[118], the mechanisms allowing response in its absence require further studies.

We identified a single immune cell population, which differed between DB and nDB tumours: CD74$^+$ macrophages[107]. From the spatial analysis of IMC images with SIMPLI, we observed that these are positive for PDL1 and in DB patients have a higher tendency to form clusters close to PD1$^+$ CD8$^+$ T cells[107]. Thus, these could be the cells whose PD1/PDL1 interaction is disrupted by anti-PD1 ICI. Then the CD8$^+$ T cells are freed from inhibition and can perform their cytotoxic antitumor function. While the molecular mechanisms employed by hypermutated CRCs to avoid elimination by CD8$^+$ T cells require further investigation, alterations in antigen presentation mechanisms could represent a possible explanation. Since this observation remained valid across two cohorts showing genetic and immune inter- and intra- tumour heterogeneity and with varied clinical history, it suggests that CD74$^+$ macrophage infiltration has the potential to be further researched as a predictor of anti-PD1 response in CRC.

Our study of the inter- and intra- tumour heterogeneity of CRC in relation to response to anti-PD1 immunotherapy highlighted the need to identify cancer-specific markers of response. We showed the expression of PDL1 and PD1 and PD1-PDL1 complex formation were not associated with response[107]. Additionally, we observed that showed that in hypermutated CRC, TMB does no longer positively correlate to response, and even CRCs with very high TMB can be nDB tumours[107]. TMB was also not correlated with T cell infiltration. Thus, these biomarkers are not universal predictors of response, and the spectrum of variability in tumour genetics and immune features of the TME needs to be considered.

The lack of predictive power of the TMB above the 12 mutations/megabase pair is because the TMB t provides little information about the immunogenicity of the tumour. After predicting the immunogenicity of mutations in each patient, we observed a higher number of clonal immunogenic mutations in responders. Thus, DB-CRCs had a larger fraction of tumour cells with the same potential immune targets compared to nDB-CRCs. This difference was reflected in a higher clonality of the productive TCR repertoire in DB-CRCs, showing that while responders do not have more mutation, they have significantly larger clones harbouring immunogenic mutations and expanded T cell clones. These observations are in agreement with a recent study which showed that the TMB was associated with response only in cancer types whose neoantigen burden was positively correlated with CD8$^+$ T cell infiltration[119] and this was not the case for the CRCs in our analysis. These results show that the TMB is only weakly coupled to ICB response, and other factors, such as WNT pathway activation levels and availability of CD8$^+$ T Cells and active antigen-presenting cells, could represent better predictors of response to ICI in CRC.

A multiregional and multiomic approach like the one we adopted for our study in CRC is essential for analysing immunotherapy response when considering the significant impact of intra-tumour heterogeneity on acquired resistance. The study of acquired resistance to ICI is challenging because of the limited availability of longitudinal samples from before treatment and after progression. Moreover, the mechanisms of acquired resistance to immunotherapy could be highly variable both across cancer types and within CRC[113]. Studies of adoptive T cell therapy in CRC and melanoma have identified loss of neoantigen-specific HLA haplotypes[120] and loss of B2M[121] as possible mechanisms of resistance.

Additionally, because of intra-tumour heterogeneity, the number of sampled regions and the time of sampling in the treatment course could significantly impact biomarker positivity[113]. This would lead to an underrepresentation of the tumour's immune and genetic status because of sampling bias from a few or single small biopsies or the dynamic nature of immunity in the TME.

For these reasons, multiregional analyses, including highly-multiplexed imaging technologies, are necessary to progress beyond a broad categorisation of tumours and to identify the specific drivers, pathways and cells responsible for ICI resistance in a given patient. Multiplexed methodologies and image analysis strategies will enable the investigation of these features and drive the studies aiming to bring the identified biomarkers into clinical practice[56, 60]. These studies will require the concerted efforts of multidisciplinary teams, including engineers, bioinformaticians, pathologists, oncologists, and immunologists, to develop the technologies, analytical software and experimental frameworks to investigate intra-tumour heterogeneity in the TME and its impact on immunotherapy response.

References

1. Hausser J, Alon U. Tumour heterogeneity and the evolutionary trade-offs of cancer. *Nature Reviews Cancer* **20**, 247-257 (2020).

2. Marusyk A, Janiszewska M, Polyak K. Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance. *Cancer Cell* **37**, 471-484 (2020).

3. Chang K*, et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* **45**, 1113-1120 (2013).

4. Zhang J*, et al.* The International Cancer Genome Consortium Data Portal. *Nature Biotechnology* **37**, 367-369 (2019).

5. Campbell PJ*, et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93 (2020).

6. Tate JG*, et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* **47**, D941-D947 (2019).

7. Dressler L*, et al.* Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the Network of Cancer Genes (NCG) resource. *Genome Biology* **23**, 35 (2022).

8. Clark K*, et al.* The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging* **26**, 1045-1057 (2013).

9. Vandin F. Computational Methods for Characterizing Cancer Mutational Heterogeneity. *Frontiers in Genetics* **8**, 83 (2017).

10. Alexandrov LB*, et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).

11. Hanahan D, Weinberg Robert A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646-674 (2011).

12. Andrews MC, Reuben A, Gopalakrishnan V, Wargo JA. Concepts Collide: Genomic, Immune, and Microbial Influences on the Tumor Microenvironment and Response to Cancer Therapy. *Frontiers in Immunology* **9**, 946 (2018).

13. Sautès-Fridman C, Petitprez F, Calderaro J, Fridman WH. Tertiary lymphoid structures in the era of cancer immunotherapy. *Nature Reviews Cancer* **19**, 307-325 (2019).

14. Binnewies M*, et al.* Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nature Medicine* **24**, 541-550 (2018).

15. Chen DS, Mellman I. Elements of cancer immunity and the cancer–immune set point. *Nature* **541**, 321-330 (2017).

16. Sautès-Fridman C*, et al.* Tertiary lymphoid structures in cancers: prognostic value, regulation, and manipulation for therapeutic intervention. *Frontiers in immunology* **7**, 407 (2016).

17. Rodriguez-Salas N*, et al.* Clinical relevance of colorectal cancer molecular subtypes. *Critical Reviews in Oncology/Hematology* **109**, 9-19 (2017).

18. Singh MP, Rai S, Pandey A, Singh NK, Srivastava S. Molecular subtypes of colorectal cancer: An emerging therapeutic opportunity for personalized medicine. *Genes & Diseases* **8**, 133-145 (2021).

19. Wang W*, et al.* Molecular subtyping of colorectal cancer: Recent progress, new challenges and emerging opportunities. *Seminars in Cancer Biology* **55**, 37-52 (2019).

20. Guinney J*, et al.* The consensus molecular subtypes of colorectal cancer. *Nature Medicine* **21**, 1350-1356 (2015).

21. Eide PW*, et al.* Metastatic heterogeneity of the consensus molecular subtypes of colorectal cancer. *npj Genomic Medicine* **6**, 59 (2021).

22. Sottoriva A, Barnes CP, Graham TA. Catch my drift? Making sense of genomic intra-tumour heterogeneity. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1867**, 95-100 (2017).

23. Bakhoum SF, Cantley LC. The Multifaceted Role of Chromosomal Instability in Cancer and Its Microenvironment. *Cell* **174**, 1347-1360 (2018).

24. Biswas A, De S. Drivers of dynamic intratumor heterogeneity and phenotypic plasticity. *American Journal of Physiology-Cell Physiology* **320**, C750-C760 (2021).

25. Janiszewska M. The microcosmos of intratumor heterogeneity: the space-time of cancer evolution. *Oncogene* **39**, 2031-2039 (2020).

26. Ramón y Cajal S*, et al.* Clinical implications of intratumor heterogeneity: challenges and opportunities. *Journal of Molecular Medicine* **98**, 161-177 (2020).

27. Stanta G, Bonin S. Overview on Clinical Relevance of Intra-Tumor Heterogeneity. *Frontiers in Medicine* **5**, (2018).

28. Korenchan DE, Flavell RR. Spatiotemporal pH heterogeneity as a promoter of cancer progression and therapeutic resistance. *Cancers* **11**, 1026 (2019).

29. Jiménez-Sánchez A*, et al.* Heterogeneous Tumor-Immune Microenvironments among Differentially Growing Metastases in an Ovarian Cancer Patient. *Cell* **170**, 927-938.e920 (2017).

30. Zheng Z, Yu T, Zhao X, Gao X, Zhao Y, Liu G. Intratumor heterogeneity: A new perspective on colorectal cancer research. *Cancer Medicine* **9**, 7637-7645 (2020).

31. Sagaert X, Vanstapel A, Verbeek S. Tumor Heterogeneity in Colorectal Cancer: What Do We Know So Far? *Pathobiology* **85**, 72-84 (2018).

32. Jones HG*, et al.* Genetic and epigenetic intra-tumour heterogeneity in colorectal cancer. *World journal of surgery* **41**, 1375-1383 (2017).

33. Jeantet M*, et al.* High intra-and inter-tumoral heterogeneity of RAS mutations in colorectal cancer. *International journal of molecular sciences* **17**, 2015 (2016).

34. Kogita A*, et al.* Inter-and intra-tumor profiling of multi-regional colon cancer and metastasis. *Biochemical and biophysical research communications* **458**, 52-56 (2015).

35. Normanno N*, et al.* Heterogeneity of KRAS, NRAS, BRAF and PIK3CA mutations in metastatic colorectal cancer and potential effects on therapy in the CAPRI GOIM trial. *Annals of oncology* **26**, 1710-1714 (2015).

36. Losi L, Baisse B, Bouzourene H, Benhattar J. Evolution of intratumoral genetic heterogeneity during colorectal cancer progression. *Carcinogenesis* **26**, 916-922 (2005).

37. De Smedt L*, et al.* Microsatellite instable vs stable colon carcinomas: analysis of tumour heterogeneity, inflammation and angiogenesis. *British Journal of Cancer* **113**, 500-509 (2015).

38. Joung J-G*, et al.* Tumor heterogeneity predicts metastatic potential in colorectal cancer. *Clinical Cancer Research* **23**, 7209-7216 (2017).

39. Sveen A*, et al.* Intra-patient inter-metastatic genetic heterogeneity in colorectal cancer as a key determinant of survival after curative liver resection. *PLoS genetics* **12**, e1006225 (2016).

40. Hobor S, Van Emburgh BO, Crowley E, Misale S, Di Nicolantonio F,

Bardelli A. TGFα and amphiregulin paracrine network promotes

resistance

to EGFR blockade in colorectal cancer cells. *Clinical Cancer Research* **20**, 6429-6438 (2014).

41. Li X, Shao C, Shi Y, Han W. Lessons learned from the blockade of immune checkpoints in cancer immunotherapy. *Journal of Hematology & Oncology* **11**, 31 (2018).

42. Galluzzi L, Chan Timothy A, Kroemer G, Wolchok Jedd D, López-Soto A. The hallmarks of successful anticancer immunotherapy. *Science Translational Medicine* **10**, eaat7807 (2018).

43. Sahin IH*, et al.* Immune checkpoint inhibitors for the treatment of MSI-H/MMR-D colorectal cancer and a perspective on resistance mechanisms. *British Journal of Cancer* **121**, 809-818 (2019).

44. Bagchi S, Yuan R, Engleman EG. Immune Checkpoint Inhibitors for the Treatment of Cancer: Clinical Impact and Mechanisms of Response and Resistance. *Annual Review of Pathology: Mechanisms of Disease* **16**, 223-249 (2021).

45. Ott PA*, et al.* T-cell–inflamed gene-expression profile, programmed death ligand 1 expression, and tumor mutational burden predict efficacy in patients treated with pembrolizumab across 20 cancers: KEYNOTE-028. *Journal of Clinical Oncology* **37**, 318-327 (2019).

46. Jenkins RW, Barbie DA, Flaherty KT. Mechanisms of resistance to immune checkpoint inhibitors. *British Journal of Cancer* **118**, 9-16 (2018).

47. Alsaab HO*, et al.* PD-1 and PD-L1 Checkpoint Signaling Inhibition for Cancer Immunotherapy: Mechanism, Combinations, and Clinical Outcome. *Frontiers in Pharmacology* **8**, (2017).

48. Salmaninejad A*, et al.* PD-1/PD-L1 pathway: Basic biology and role in cancer immunotherapy. *Journal of Cellular Physiology* **234**, 16824-16837 (2019).

49. Sharpe AH, Pauken KE. The diverse functions of the PD1 inhibitory pathway. *Nature Reviews Immunology* **18**, 153-167 (2018).

50. Ganesh K*, et al.* Immunotherapy in colorectal cancer: rationale, challenges and potential. *Nature Reviews Gastroenterology & Hepatology* **16**, 361-375 (2019).

51. Oliveira AF, Bretes L, Furtado I. Review of PD-1/PD-L1 Inhibitors in Metastatic dMMR/MSI-H Colorectal Cancer. *Frontiers in Oncology* **9**, (2019).

52. André T*, et al.* Pembrolizumab in Microsatellite-Instability–High Advanced Colorectal Cancer. *New England Journal of Medicine* **383**, 2207-2218 (2020).

53. Lenz H-J*, et al.* First-Line Nivolumab Plus Low-Dose Ipilimumab for Microsatellite Instability-High/Mismatch Repair-Deficient Metastatic Colorectal Cancer: The Phase II CheckMate 142 Study. *Journal of Clinical Oncology* **40**, 161-170 (2021).

54. Galon J*, et al.* MSI status plus immunoscore to select metastatic colorectal cancer patients for immunotherapies. *Annals of Oncology* **29**, x4 (2018).

55. Bodenmiller B. Multiplexed Epitope-Based Tissue Imaging for Discovery and Healthcare Applications. *Cell Systems* **2**, 225-238 (2016).

56. Lewis SM*, et al.* Spatial omics and multiplexed imaging to explore cancer biology. *Nature Methods* **18**, 997-1012 (2021).

57. Fan J, Slowikowski K, Zhang F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Experimental & Molecular Medicine* **52**, 1452-1465 (2020).

58. Lee J, Hyeon DY, Hwang D. Single-cell multiomics: technologies and data analysis methods. *Experimental & Molecular Medicine* **52**, 1428-1442 (2020).

59. Bortolomeazzi M, Keddar MR, Ciccarelli FD, Benedetti L. Identification of non-cancer cells from cancer transcriptomic data. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1863**, 194445 (2020).

60. Parra ER, Francisco-Cruz A, Wistuba II. State-of-the-Art of Profiling Immune Contexture in the Era of Multiplexed Staining and Digital Analysis to Study Paraffin Tumor Tissues. *Cancers* **11**, (2019).

61. Tsujikawa T*, et al.* Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis. *Cell Reports* **19**, 203-217 (2017).

62. Tan WCC*, et al.* Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Communications* **40**, 135-153 (2020).

63. Morrison LE*, et al.* Brightfield multiplex immunohistochemistry with multispectral imaging. *Laboratory Investigation* **100**, 1124-1136 (2020).

64. Stack EC, Wang C, Roman KA, Hoyt CC. Multiplexed immunohistochemistry, imaging, and quantitation: A review, with an

assessment of Tyramide signal amplification, multispectral imaging and multiplex analysis. *Methods* **70**, 46-58 (2014).

65. Gerdes MJ*, et al.* Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proceedings of the National Academy of Sciences* **110**, 11982 (2013).

66. Lin J-R*, et al.* Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *eLife* **7**, e31657 (2018).

67. Jungmann R, Avendaño MS, Woehrstein JB, Dai M, Shih WM, Yin P. Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and Exchange-PAINT. *Nature methods* **11**, 313-318 (2014).

68. Goltsev Y*, et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**, 968-981.e915 (2018).

69. Black S*, et al.* CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nature Protocols* **16**, 3802-3835 (2021).

70. Angelo M*, et al.* Multiplexed ion beam imaging of human breast tumors. *Nature Medicine* **20**, 436-442 (2014).

71. Giesen C*, et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods* **11**, 417-422 (2014).

72. Vicar T*, et al.* Cell segmentation methods for label-free contrast microscopy: review and comprehensive comparison. *BMC Bioinformatics* **20**, 360 (2019).

73. Schmidt U, Weigert M, Broaddus C, Myers G. Cell Detection with Star-Convex Polygons.). Springer International Publishing (2018).

74. Hollandi R*, et al.* nucleAIzer: A Parameter-free Deep Learning Framework for Nucleus Segmentation Using Image Style Transfer. *Cell Systems* **10**, 453-458.e456 (2020).

75. Stringer C, Wang T, Michaelos M, Pachitariu M. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods* **18**, 100-106 (2021).

76. Greenwald NF*, et al.* Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *bioRxiv*, 2021.2003.2001.431313 (2021).

77. Durkee MS, Abraham R, Clark MR, Giger ML. Artificial Intelligence and Cellular Segmentation in Tissue Microscopy Images. *The American Journal of Pathology* **191**, 1693-1701 (2021).

78. Levine Jacob H*, et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184-197 (2015).

79. Van Gassen S*, et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A* **87**, 636-645 (2015).

80. Hao Y*, et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e3529 (2021).

81. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *Journal of machine learning research* **9**, (2008).

82. Leland M, John H, Nathaniel S, Lukas G. UMAP: uniform manifold approximation and projection. *Journal of Open Source Software* **3**, 861 (2018).

83. Eling N, Hoch T, Zanotelli V, Fischer J, Schulz D. imcRtools: Methods for imaging mass cytometry data analysis. R package version 1.0.0. https://githubcom/BodenmillerGroup/imcRtools, (2021).

84. Stoltzfus CR*, et al.* CytoMAP: A Spatial Analysis Toolbox Reveals Features of Myeloid Cell Organization in Lymphoid Tissues. *Cell Reports* **31**, 107523 (2020).

85. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Kdd*) (1996).

86. Eling N, Damond N, Hoch T, Bodenmiller B. cytomapper: an R/Bioconductor package for visualization of highly multiplexed imaging data. *Bioinformatics* **36**, 5706-5708 (2020).

87. van Maldegem F*, et al.* Characterisation of tumour microenvironment remodelling following oncogene inhibition in preclinical studies with imaging mass cytometry. *Nature Communications* **12**, 5906 (2021).

88. Zanotelli V, Bodenmiller B. ImcSegmentationPipeline: A pixelclassification based multiplexed image segmentation pipeline. https://githubcom/BodenmillerGroup/ImcSegmentationPipeline, (2017).

89. Schapiro D*, et al.* MCMICRO: A scalable, modular image-processing pipeline for multiplexed tissue imaging. *bioRxiv*, 2021.2003.2015.435473 (2021).

90. McQuin C*, et al.* CellProfiler 3.0: Next-generation image processing for biology. *PLOS Biology* **16**, e2005970 (2018).

91. Berg S*, et al.* ilastik: interactive machine learning for (bio)image analysis. *Nature Methods* **16**, 1226-1232 (2019).

92. Bankhead P*, et al.* QuPath: Open source software for digital pathology image analysis. *Scientific Reports* **7**, 16878 (2017).

93. Catena R, Montuenga LM, Bodenmiller B. Ruthenium counterstaining for imaging mass cytometry. *The Journal of pathology* **244**, 479-484 (2018).

94. Wiesmann V, Franz D, Held C, MÜNzenmayer C, Palmisano R, Wittenberg T. Review of free software tools for image analysis of fluorescence cell micrographs. *Journal of Microscopy* **257**, 39-53 (2015).

95. Ewels PA*, et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology* **38**, 276-278 (2020).

96. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35**, 316-319 (2017).

97. Afgan E*, et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research* **46**, W537-W544 (2018).

98. Grüning B*, et al.* Practical Computational Reproducibility in the Life Sciences. *Cell Systems* **6**, 631-635 (2018).

99. Bortolomeazzi M*, et al.* A SIMPLI (Single-cell Identification from MultiPLexed Images) approach for spatially-resolved tissue phenotyping at single-cell resolution. *Nature Communications* **13**, 781 (2022).

100. Repana D*, et al.* The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biology* **20**, 1 (2019).

101. Sochat VV, Prybol CJ, Kurtzer GM. Enhancing reproducibility in scientific computing: Metrics and registry for Singularity containers. *PLOS ONE* **12**, e0188511 (2017).

102. imctools. *https://githubcom/BodenmillerGroup/imctools*, (2017).

103. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411-420 (2018).

104. Henning C. fpc. https://cranr-projectorg/web/packages/fpc/indexhtml, (2020).

105. Melville JL, Aaron, Djekidel MN, Hao Y. uwot. https://cranr-projectorg/package=uwot, (2018).

106. Pau G, Fuchs F, Sklyar O, Boutros M, Huber W. EBImage—an R package for image processing with applications to cellular phenotypes. *Bioinformatics* **26**, 979-981 (2010).

107. Bortolomeazzi M*, et al.* Immunogenomics of Colorectal Cancer Response to Checkpoint Blockade: Analysis of the KEYNOTE 177 Trial and Validation Cohorts. *Gastroenterology* **161**, 1179-1193 (2021).

108. Eisenhauer EA*, et al.* New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer* **45**, 228-247 (2009).

109. Otsu N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics* **9**, 62-66 (1979).

110. Kramer AS*, et al.* InForm software: a semi-automated research tool to identify presumptive human hepatic progenitor cells, and other histological features of pathological significance. *Scientific Reports* **8**, 3418 (2018).

111. Strokotov D*, et al.* Is there a difference between T- and B-lymphocyte morphology? *Journal of Biomedical Optics* **14**, 064036 (2009).

112. Kumar A, Swain CA, Shevde LA. Informing the new developments and future of cancer immunotherapy. *Cancer and Metastasis Reviews* **40**, 549-562 (2021).

113. Hegde PS, Chen DS. Top 10 Challenges in Cancer Immunotherapy. *Immunity* **52**, 17-35 (2020).

114. Ock C-Y*, et al.* Genomic landscape associated with potential response to anti-CTLA-4 treatment in cancers. *Nature Communications* **8**, 1050 (2017).

115. Allan C*, et al.* OMERO: flexible, model-driven data management for experimental biology. *Nature Methods* **9**, 245-253 (2012).

116. Feng M*, et al.* Pharmacological inhibition of β-catenin/BCL9 interaction overcomes resistance to immune checkpoint blockades by modulating Treg cells. *Science Advances* **5**, eaau5240.

117. Xue J, Yu X, Xue L, Ge X, Zhao W, Peng W. Intrinsic β-catenin signaling suppresses CD8+ T-cell infiltration in colorectal cancer. *Biomedicine & Pharmacotherapy* **115**, 108921 (2019).

118. Middha S*, et al.* Majority of B2M-mutant and-deficient colorectal carcinomas achieve clinical benefit from immune checkpoint inhibitor therapy and are microsatellite instability-high. *JCO precision oncology* **3**, 1-14 (2019).

119. McGrail DJ*, et al.* High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types. *Annals of Oncology* **32**, 661-672 (2021).

120. Tran E*, et al.* T-cell transfer therapy targeting mutant KRAS in cancer. *New England Journal of Medicine* **375**, 2255-2262 (2016).

121. Restifo NP, Marincola FM, Kawakami Y, Taubenberger J, Yannelli JR, Rosenberg SA. Loss of functional beta2-microglobulin in metastatic melanomas from five patients receiving immunotherapy. *JNCI: Journal of the National Cancer Institute* **88**, 100-108 (1996).

Appendix A: Supplementary Tables

Supplementary Table1_S. Samples used in the study

Sample ID	Tissue	Anatomical site	Source	Imaging approach	T cells (%)	Macrophages (%)	B cells (%)	IgA+ cells (%)	Epithelial cells (%)	Dendritic cells (%)	CD8+ PD1+ cells (%)	CD8+ PD1- cells (%)	PLD1+ cells (%)	CD8+ T cells (%)	CD4+ T cells (%)	Treg (%)
CLN1	Human colon mucosa	Colon	UCLH	IMC	27.84	29.96	NA	35.19	NA	7.00	NA	NA	NA	NA	NA	NA
CLN2	Human colon mucosa	Transverse colon	UCLH	IMC	20.15	26.79	NA	13.90	NA	39.16	NA	NA	NA	NA	NA	NA
CLN3	Human colon mucosa	Ascending colon	UCLH	IMC	20.56	24.69	NA	44.13	NA	10.63	NA	NA	NA	NA	NA	NA
CLN4	Human colon mucosa	Colon	UCLH	IMC	56.33	27.98	NA	4.22	NA	11.46	NA	NA	NA	NA	NA	NA
CLN5	Human colon mucosa	Ascending colon	ICH	IMC	41.71	33.70	NA	18.65	NA	5.94	NA	NA	NA	NA	NA	NA
CLN6	Human colon mucosa	Descending colon	ICH	IMC	38.59	23.39	NA	29.60	NA	8.42	NA	NA	NA	NA	NA	NA
APP1	Human appendix	Appendix	UCLH	IMC	27.04	5.74	36.30	NA	25.65	5.27	NA	NA	NA	NA	NA	NA
CRC1	Human rectal cancer	Rectum	UCLH	mIF	NA	NA	NA	NA	NA	NA	0.17	5.85	3.73	NA	NA	NA
CRC_01_02_B	Human CRC	Rectum	Schürch et al., 2020	CODEX	NA	61.63	6.76	NA	NA	0.00	NA	NA	NA	7.79	1.72	0.23
CRC_02_04_B	Human CRC	Ascending colon	Schürch et al., 2020	CODEX	NA	20.93	12.56	NA	NA	0.18	NA	NA	NA	19.50	7.30	2.05
CRC_03_05_A	Human CRC	Descending colon	Schürch et al., 2020	CODEX	NA	31.44	9.83	NA	NA	1.59	NA	NA	NA	9.74	17.54	0.97
CRC_04_08_B	Human CRC	Sigma	Schürch et al., 2020	CODEX	NA	27.65	24.05	NA	NA	0.72	NA	NA	NA	10.92	7.52	1.63
CRC_05_09_A	Human CRC	Transversum	Schürch et al.,	CODEX	NA	24.98	3.83	NA	NA	0.87	NA	NA	NA	17.06	11.08	2.42

			2020													
CRC_06_11_A	Human CRC	Ascending colon	Schürch et al., 2020	CODEX	NA	22.03	12.96	NA	NA	0.13	NA	NA	NA	45.82	13.03	0.65
CRC_07_13_B	Human CRC	Ascending colon	Schürch et al., 2020	CODEX	NA	32.24	21.10	NA	NA	0.68	NA	NA	NA	7.68	6.24	1.01
CRC_08_16_B	Human CRC	Rectum	Schürch et al., 2020	CODEX	NA	23.53	8.55	NA	NA	0.07	NA	NA	NA	20.08	33.66	0.66
CRC_09_17_B	Human CRC	Cecum	Schürch et al., 2020	CODEX	NA	23.44	6.70	NA	NA	0.57	NA	NA	NA	14.56	11.86	0.74
CRC_10_20_B	Human CRC	Rectum	Schürch et al., 2020	CODEX	NA	31.82	4.91	NA	NA	0.50	NA	NA	NA	19.08	12.31	1.06
CRC_11_21_A	Human CRC	Descending colon	Schürch et al., 2020	CODEX	NA	27.35	4.66	NA	NA	0.39	NA	NA	NA	49.66	12.90	1.45
CRC_12_24_B	Human CRC	Sigma	Schürch et al., 2020	CODEX	NA	22.99	25.58	NA	NA	0.43	NA	NA	NA	6.48	24.11	0.35
CRC_13_26_B	Human CRC	Cecum	Schürch et al., 2020	CODEX	NA	35.34	6.71	NA	NA	2.04	NA	NA	NA	14.34	10.20	2.51
CRC_14_28_B	Human CRC	Cecum	Schürch et al., 2020	CODEX	NA	47.74	3.88	NA	NA	0.49	NA	NA	NA	14.90	15.75	0.28
CRC_15_29_A	Human CRC	Sigma	Schürch et al., 2020	CODEX	NA	25.49	16.43	NA	NA	0.89	NA	NA	NA	23.45	14.48	1.87
CRC_16_32_A	Human CRC	Cecum	Schürch et al., 2020	CODEX	NA	33.62	4.79	NA	NA	0.93	NA	NA	NA	15.15	9.89	0.77
CRC_17_34_B	Human CRC	Cecum	Schürch et al., 2020	CODEX	NA	54.65	6.34	NA	NA	0.59	NA	NA	NA	17.92	15.12	0.66
CRC_18_36_B	Human CRC	Sigma	Schürch et al., 2020	CODEX	NA	40.84	6.25	NA	NA	4.71	NA	NA	NA	9.59	12.16	0.17
CRC_19_38_	Human CRC	Cecum	Schürch	CODEX	NA	18.70	16.36	NA	NA	0.61	NA	NA	NA	27.10	8.57	1.13

A			et al., 2020													
CRC_20_39_A	Human CRC	Sigma	Schürch et al., 2020	CODEX	NA	42.11	4.06	NA	NA	0.25	NA	NA	NA	19.28	3.30	0.38
CRC_21_41_A	Human CRC	Sigma	Schürch et al., 2020	CODEX	NA	27.88	7.29	NA	NA	1.58	NA	NA	NA	16.85	11.13	2.66
CRC_22_43_A	Human CRC	Descending colon	Schürch et al., 2020	CODEX	NA	39.09	11.36	NA	NA	0.23	NA	NA	NA	14.32	20.23	1.93
CRC_23_46_B	Human CRC	Cecum	Schürch et al., 2020	CODEX	NA	32.84	7.99	NA	NA	3.29	NA	NA	NA	11.29	25.55	2.66
CRC_24_47_A	Human CRC	Cecum	Schürch et al., 2020	CODEX	NA	63.32	7.19	NA	NA	0.16	NA	NA	NA	14.94	4.74	0.87
CRC_25_49_A	Human CRC	Rectum	Schürch et al., 2020	CODEX	NA	29.20	26.05	NA	NA	0.08	NA	NA	NA	10.81	24.92	2.33
CRC_26_52_A	Human CRC	Rectum	Schürch et al., 2020	CODEX	NA	43.92	6.49	NA	NA	1.24	NA	NA	NA	21.65	17.73	4.23
CRC_27_54_A	Human CRC	Sigma	Schürch et al., 2020	CODEX	NA	32.36	6.43	NA	NA	0.19	NA	NA	NA	19.30	25.45	1.14
CRC_28_56_B	Human CRC	Rectum	Schürch et al., 2020	CODEX	NA	41.91	5.62	NA	NA	0.79	NA	NA	NA	4.44	31.76	0.20
CRC_29_58_A	Human CRC	Cecum	Schürch et al., 2020	CODEX	NA	48.56	10.50	NA	NA	0.26	NA	NA	NA	13.65	21.52	0.00
CRC_30_59_A	Human CRC	Sigma	Schürch et al., 2020	CODEX	NA	33.59	15.36	NA	NA	2.69	NA	NA	NA	22.22	15.10	2.17
CRC_31_61_B	Human CRC	Ascending colon	Schürch et al., 2020	CODEX	NA	44.17	7.00	NA	NA	5.61	NA	NA	NA	10.71	10.71	7.07
CRC_32_64_B	Human CRC	Sigma	Schürch et al., 2020	CODEX	NA	36.44	14.43	NA	NA	6.20	NA	NA	NA	7.94	23.40	0.00

CRC_33_65_A	Human CRC	Sigma	Schürch et al., 2020	CODEX	NA	37.43	15.08	NA	NA	0.43	NA	NA	NA	11.55	8.34	1.60
CRC_34_68_A	Human CRC	Descending colon	Schürch et al., 2020	CODEX	NA	46.34	10.99	NA	NA	1.70	NA	NA	NA	12.96	8.64	2.09
CRC_35_69_A	Human CRC	Descending colon	Schürch et al., 2020	CODEX	NA	39.07	7.55	NA	NA	0.79	NA	NA	NA	4.13	5.71	0.09

For each sample reported are the ID, the tissue and anatomical site of origin, the experimental approach, and the percentage of cells identified by SIMPLI and used in the analysis and the corresponding figure of the original paper where these data are shown. In Fig. 2g, the percentages of cells were measured over the total immune cells; In Figs. 3c and 4d they were measured over the total cells. Data shown in Fig. 5b were obtained from a previously published CODEX dataset (1). IMC: Imaging Mass Cytometry, mIF: multiplexed immunofluorescence, UCLH: University College London Hospital, ICH: Istituto Clinico Humanitas, NA: not applicable. This table was adapted from supplementary data 1 of:

Bortolomeazzi M*, et al.* A SIMPLI (Single-cell Identification from MultiPLexed Images) approach for spatially-resolved tissue phenotyping at single-cell resolution. *Nature Communications* **13**, 781 (2022)

1. Schürch, C. M. et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. Cell 182, 1341–1359.e1319 (2020).

Supplementary Table 2_S. Antibodies used in the study

Cell population	Antibody Specificity	Vendor	Catalogue Number	Metal or Opal Tag	Antibody Dilution	Application
All leukocytes	CD45	Fluidigm	3152016D	152Sm	1 in 500	IMC
B cells	CD20	Fluidigm	3161029D	161Dy	1 in 250	IMC
B cells	IgA	NovusBio	NB500-469	142Nd	1 in 100	IMC
B cells	IgM	NovusBio	NBP2-34254	169Tm	1 in 200	IMC
B cells / T cells	CD27	Fluidigm	3171024D	171Yb	1 in 300	IMC
T cells	CD45RA	Fluidigm	3166028D	166Er	1 in 2000	IMC
T cells	CD45RO	Fluidigm	3173016D	173Yb	1 in 500	IMC
T cells / macrophages	CD4	Fluidigm	3156033D	156Gd	1 in 200	IMC
T cells	CD8	Fluidigm	3162035D	162Dy	1 in 800	IMC
T cells	PD1	Fluidigm	3165039D	165Ho	1 in 50	IMC
T cells	CD3	Fluidigm	3170019D	170Er	1 in 800	IMC
T cells	FOXP3	Fluidigm	3155016D	155Gd	1 in 200	IMC
Macrophages	CD68	Fluidigm	3159035D	159Tb	1 in 400	IMC
Macrophages	CD16	Fluidigm	3146020D	146Nd	1 in 200	IMC
Macrophages and dendritic cells	CD11c	Abcam	ab216655	175 Lu	1 in 400	IMC
Macrophages, dendritic cells, tumour cells	PDL1	RnD System	MAB1561	150Nd	1 in 70	IMC
Endothelial cells	CD34	Abcam	ab213058	164Dy	1 in 150	IMC
Epithelial cells	Pan keratin	Fluidigm	3148020D	148Nd	1 in 3000	IMC
Epithelial cells	E-Cadherin	Fluidigm	3158029D	158Gd	1 in 3000	IMC
Basement membrane cells	Collagen type IV	NovusBio	NBP1-97716	176Yb	1 in 30	IMC
Proliferating cells	Ki67	Fluidigm	3168022D	168Er	1 in 400	IMC
Stromal cells	Vimentin	Fluidigm	3143029D	143Nd	1 in 8000	IMC
Stromal cells	SMA	Fluidigm	3141017D	141Pr	1 in 4000	IMC
Various	CAMK4	NovusBio	NBP2-37428	174Yb	1 in 250	IMC
Various	IFNA5	CloudClone	MAG975Hu22	147Sm	1 in 100	IMC
Various	VEGFC	Abcam	ab191274	154Sm	1 in 600	IMC
T cells	CD3	Dako	A0452	NA	1 in 200	IHC
NA	Anti-Rabbit-HRP	Dako	P0448	NA	1 in 200	IHC
T cells	CD8	Cell Signaling Technologies	85336	Opal 780 (1:75)	1 in 200	mIF
T cells	GzB	Abcam	ab208586	Opal 480 (1:600)	1 in 100	mIF
T cells	PD1	Abcam	ab137132	Opal 650 (1:100)	1 in 300	mIF
Proliferating cells	ki67	BD Biosciences	550609	Opal 690 (1:150)	1 in 200	mIF
Macrophages	CD68	BioLegend	916104	Opal 620 (1:700)	1 in 1000	mIF
Macrophages, dendritic cells, tumour cells	PDL1	RnD System	MAB1561	Opal 520 (1:150)	1 in 450	mIF
Stromal cells	CD44	NA	NA	NA	NA	CODEX
T cells	FOXP3	NA	NA	NA	NA	CODEX
Epithelial cells	CDX2	NA	NA	NA	NA	CODEX
T cells	CD8	NA	NA	NA	NA	CODEX
Various	p53	NA	NA	NA	NA	CODEX
T cells	GATA3	NA	NA	NA	NA	CODEX
All leukocytes	CD45	NA	NA	NA	NA	CODEX
T cells	Tbet	NA	NA	NA	NA	CODEX
Various	Beta catenin	NA	NA	NA	NA	CODEX
Macrophages and dendritic cells	HLADR	NA	NA	NA	NA	CODEX

Macrophages, dendritic cells, tumour cells	PDL1	NA	NA	NA	NA	CODEX
Proliferating cells	Ki67	NA	NA	NA	NA	CODEX
T cells	CD45RA	NA	NA	NA	NA	CODEX
T cells	CD4	NA	NA	NA	NA	CODEX
Dendritic cells	CD21	NA	NA	NA	NA	CODEX
Epithelial cells	MUC1	NA	NA	NA	NA	CODEX
Various	CD30	NA	NA	NA	NA	CODEX
T cells	CD2	NA	NA	NA	NA	CODEX
Stromal cells	Vimentin	NA	NA	NA	NA	CODEX
B cells	CD20	NA	NA	NA	NA	CODEX
Various	LAG3	NA	NA	NA	NA	CODEX
Various	NaKATPase	NA	NA	NA	NA	CODEX
T cells	CD5	NA	NA	NA	NA	CODEX
Various	IDO1	NA	NA	NA	NA	CODEX
Epithelial cells	Cytokeratin	NA	NA	NA	NA	CODEX
Macrophages and dendritic cells	CD11b	NA	NA	NA	NA	CODEX
NK cells	CD56	NA	NA	NA	NA	CODEX
Stromal cells	aSMA	NA	NA	NA	NA	CODEX
Various	BCL2	NA	NA	NA	NA	CODEX
Various	CD25	NA	NA	NA	NA	CODEX
Basement membrane cells	Collagen type IV	NA	NA	NA	NA	CODEX
Macrophages and dendritic cells	CD11c	NA	NA	NA	NA	CODEX
T cells	PD1	NA	NA	NA	NA	CODEX
T cells	GzB	NA	NA	NA	NA	CODEX
Various	EGFR	NA	NA	NA	NA	CODEX
Various	VISTA	NA	NA	NA	NA	CODEX
Granulocytes	CD15	NA	NA	NA	NA	CODEX
Various	CD194	NA	NA	NA	NA	CODEX
Various	ICOS	NA	NA	NA	NA	CODEX
Various	MMP9	NA	NA	NA	NA	CODEX
Neuroendocrine cells	Synaptophysin	NA	NA	NA	NA	CODEX
Various	CD71	NA	NA	NA	NA	CODEX
Nervous cells	GFAP	NA	NA	NA	NA	CODEX
T cells	CD7	NA	NA	NA	NA	CODEX
T cells	CD3	NA	NA	NA	NA	CODEX
Neuroendocrine cells	ChromograninA	NA	NA	NA	NA	CODEX
Macrophages	CD163	NA	NA	NA	NA	CODEX
NK cells	CD57	NA	NA	NA	NA	CODEX
T cells	CD45RO	NA	NA	NA	NA	CODEX
Macrophages	CD68	NA	NA	NA	NA	CODEX
Endothelial cells	CD31	NA	NA	NA	NA	CODEX
Lymphatic endothelial cells	Podoplanin	NA	NA	NA	NA	CODEX
Endothelial cells	CD34	NA	NA	NA	NA	CODEX
B cells	CD38	NA	NA	NA	NA	CODEX
Plasma cells	CD138	NA	NA	NA	NA	CODEX
Various	MMP12	NA	NA	NA	NA	CODEX

For each antibody reported are the associated cell population, the catalogue number, the vendor, the tag, the dilution used in the staining, the experimental application and the corresponding reference figure. Data shown in

Fig. 5b were obtained from a previously published CODEX dataset (1). IMC: Imaging Mass Cytometry, IHC: Immunohistochemistry, mIF: multiplexed immunofluorescence, CODEX: co-detection by indexing, NA: not applicable. This table was adapted from supplementary data 2 of:

Bortolomeazzi M*, et al.* A SIMPLI (Single-cell Identification from MultiPLexed Images) approach for spatially-resolved tissue phenotyping at single-cell resolution. *Nature Communications* **13**, 781 (2022)

1. Schürch, C. M. et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. Cell 182, 1341–1359.e1319 (2020).

Appendix A: Supplementary Tables

Supplementary Table 1_I. Patients and samples used in the study

Patient ID	Treatment	Cohort	Prior lines of chemotherapy (n)	Treatment response	PFS (days)	Benefit	Tumor source	Histological type	Anatomical site	Tumour diameter (mm)	Tumour staging	MMR protein loss (IHC)	MMR germline mutation	Lynch syndrome	MSI status (Biocartis)	TMB (FM1)	TMB (WES)	Hypermutated phenotype
UH1	Pembrolizumab	Discovery	0	CR	1223	DB	Primary	Adenocarcinoma	Rectum	NA	pT4b N2 Mx	MLH1, PMS2	Unknown	Y	Y	76	46	Y
UH2	Pembrolizumab	Discovery	0	PD	56	nDB	Primary	Adenocarcinoma	Sigmoid	50	pT4b N2 Mx	MLH1, PMS2	Unknown	N	N	2	4	N
UH3	Pembrolizumab	Discovery	0	SD then PD	155	nDB	Primary	Adenocarcinoma	Caecum	60	pT4b N1 M0	MLH1, PMS2	Unknown	N	N	4	5	N
UH4	Pembrolizumab	Discovery	1	PR	719	DB	Primary	Mucinous adenocarcinoma	Sigmoid	55	pT4 N0 Mx	Not performed	MSH2	Y	Y	56	38	Y
UH5	Pembrolizumab	Discovery	0	SD then PD	113	nDB	Primary	Adenocarcinoma	Caecum	40	pT4b N0 Mx	MLH1, PMS2	Unknown	N	Y	77	56	Y
UH6	Pembrolizumab	Discovery	0	PR then PD	292	nDB	Primary	Adenocarcinoma	Descending colon	25	pT4b N1 M0	MLH1, PMS2	Unknown	N	N	9	4	N
UH7	Pembrolizumab	Discovery	0	CR	802	DB	Primary	Mucinous adenocarcinoma	Transverse colon	115	pT4b pN2 pM1	MLH1, PMS2	Unknown	N	Y	28	43	Y
UH8	Pembrolizumab	Discovery	0	CR	1223	DB	Primary	Adenocarcinoma	Caecum	55	pT4b N1 MX	MLH1, PMS2	Unknown	Y	Y	33	33	Y
UH9	Pembrolizumab	Discovery	0	CR	1103	DB	Primary	Adenocarcinoma	Caecum	45	pT4 pN0	MLH1, PMS2	Unknown	N	Y	53	44	Y
UH10	Pembrolizumab	Discovery	0	SD then PD	194	nDB	Anastomotic recurrence	Adenocarcinoma	Caecum	10	pT4 N2 M0	Not performed	MSH2	Y	Y	48	55	Y
UH11	Nivolumab	Discovery	3	SD then PD	179	nDB	Primary	Mucinous adenocarcinoma	Caecum	20	pT30 pN1 pM1	MLH1, PMS2	Unknown	Y	NA	NA	122	Y
UH12	Nivolumab	Discovery	0	PD	84	nDB	Primary	Poorly differentiat	Caecum	52	pT3 pN0	MLH1, PMS2	Unknown	N	NA	NA	65	Y

								ed adenocarcinoma	ascending colon									
UH13	Nivolumab	Discovery	3	SD	365	DB	Primary	Mucinous adenocarcinoma	Transverse colon	75	pT3 pN1 pMX	MLH1, PMS2	Unknown	Y	NA	20	27	Y
UH14	Nivolumab	Discovery	0	CR	392	DB	Primary	Adenocarcinoma	Caecum	60	pT4b N1b	MLH1, PMS2	Unknown	N	NA	NA	53	Y
UH15	Nivolumab	Discovery	1	SD	533	DB	Primary	Adenocarcinoma	Appendix	60	pT4b N2 Mx	MSH2, MSH6	Unknown	N	NA	NA	16	Y
UH16	Nivolumab	Discovery	0	SD	578	DB	Primary	Mucinous adenocarcinoma	Splenic flexure	34	pT4a N1 Mx	MLH1	Unknown	Y	NA	15	10	Y
UH17	Pembrolizumab	Validation	1	PR	653	DB	Primary	Poorly differentiated adenocarcinoma	Ascending colon	40	pT3 pN0 pMx	PMS2	PSM2	Y	NA	21	NA	Y
UH18	Pembrolizumab	Validation	1	PR	613	DB	Primary	Poorly differentiated adenocarcinoma	Hepatic flexure	55	pT4b N1 pMX	MLH1, PMS2	Unknown	Y	NA	32	NA	Y
UH19	Pembrolizumab	Validation	1	SD then PD	111	nDB	Primary	Poorly differentiated adenocarcinoma	Transverse colon	80	pT4a pN2b pMx	MSH2, MSH6	MSH2	Y	NA	38	NA	Y
UH20	Nivolumab	Validation	1	PR	508	DB	Primary	Mucinous adenocarcinoma	Transverse colon	70	pT4b pN1b pMx	MLH1, PMS2	Unknown	N	NA	NA	NA	Y
UH21	Nivolumab	Validation	2	SD	878	DB	Primary	Adenocarcinoma	Transverse colon	30	pT4b pN1 pM1	MLH1, PMS2	Unknown	N	NA	NA	NA	Y
UH22	Nivolumab	Validation	1	PR then PD	970	DB	Primary	Adenocarcinoma	Sigmoid	70	pT3 pN1 pM1	MSH2, MSH6	Unknown	Y	NA	NA	NA	Y
UH23	Nivolumab	Validation	1	PD	45	nDB	Primary	Mucinous adenocarcinoma	Sigmoid	130	pT4b N1 MX	MSH2, MSH6	MSH2	Y	NA	NA	NA	Y

UH24	Nivolumab	Validation	0	PD	55	nDB	Biopsy from primary	Adenocarcinoma	Ascending colon	NA	pT3N2 M1b	MLH1, PMS2	MSH2	Y	NA	NA	NA	Y	
UH25	Nivolumab	Validation	1	PR	436	DB	Primary	Adenocarcinoma	Ascending colon	20	pT3 N2 Mx	MSH2, MSH6	Unknown	Y	NA	39	NA	Y	
UH26	Pembrolizumab	Validation	3	PR	1581	DB	Primary	Adenocarcinoma	Transverse colon	75	pT4 N2 Mx	MLH1, PMS2	Unknown	Y	NA	29	NA	Y	
UH27	Nivolumab + Ipilimumab followed by Nivolumab	Validation	1	PR	583	DB	Primary	Adenocarcinoma	Ascending colon	30	pT4a pN2a pM0	MLH1, PMS2	Unknown	N	NA	44	NA	Y	
UH28	Nivolumab followed by Nivolumab + Ipilimumab	Validation	1	PR	334	DB	Biopsy from primary	Adenocarcinoma	Rectum	NA	pT4bN2 M1b	MSH2, MSH6	Unknown	N	NA	NA	NA	Y	

Treatment response was assessed with RECIST 1.1 and progression free survival (PFS) was used to divide patients into durable benefit (DB, PFS >12 months; censor date: 31/12/2019 for the discovery cohort and 30/10/2020 for the validation cohort) or non-durable benefit (nDB, progressive disease or benefit <12 months), with the exception of UH28 (Methods). Prior lines of chemotherapy refer to the metastatic setting before receiving anti-PD1 treatment. Tumour histological type, anatomical site, maximum diameter and staging, and MMR deficiency were derived from the patient pathology report. The MSI status was

assessed using Idylla MSI assay (Biocartis) (1) and the tumour mutational burden (TMB) was assessed using the FoundationOne (FM1) test (2) and whole exome sequencing (WES). For WES, shown is the mean of TMB between the two tumour regions sequenced. Hypermutated status was defined as TMB higher than 12 muts/Mbps in at least one of the regions sequenced. CR: complete response; PR: partial response; PD: progressive disease; SD: stable disease; MMR: mismatch repair; MSI: microsatellite instability; NA: not available.
This table was adapted from supplementary table 1 of:

Bortolomeazzi, M. et al. Immunogenomics of Colorectal Cancer Response to Checkpoint Blockade: Analysis of the KEYNOTE 177 Trial and Validation Cohorts. Gastroenterology 161, 1179-1193 (2021).

1. Craene, B. D. et al. Detection of microsatellite instability (MSI) in colorectal cancer samples with a novel set of highly sensitive markers by means of the Idylla MSI Test prototype. Journal of Clinical Oncology 36, e15639-e15639, doi:10.1200/JCO.2018.36.15_suppl.e15639 (2018).

2. Frampton, G. M. et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol 31, 1023-1031, doi:10.1038/nbt.2696 (2013).

Supplementary Table 2_I. Description of tumour regions

Patient ID	Treatment	Cohort	Sample ID	CD3 staining (slide A; 90) Both cohorts	Tumour content (%; slide A; 55) Both cohorts	CD3 cells/mm2 (median; slide B; 76) Discovery cohort	CD3 cells/mm2 (median; slide F; 59) Discovery cohort	CD3 cells/mm2 (median; slide H; 59) Discovery cohort	CD3 cells/mm2 (median; slide J; 75) Discovery cohort	HE staining (slide B; 13) Validation cohort	CD3 cells/mm2 (median; slide E; 29) Validation cohort	CD3 cells/mm2 (median; slide G; 30) Validation cohort	IMC (slide C; 77) Both cohorts	mIF (slide D; 24) Both cohorts	WES (slides E1-5; 32) Discovery cohort	RNAseq (slides G1-5; 58) Discovery cohort	RNAseq (slides F1-5; 30) Validation cohort	TCRseq (slides I1-5; 28) Discovery cohort	PD1-PDL1 A-FRET (median % efficiency; slides K1-2; 58) Discovery cohort
UH1	Pembrolizumab	Discovery	UH1_TOTAL	NA	NA	421.6	NA	NA	299.55	NA	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH1	Pembrolizumab	Discovery	UH1_34	Y	NA	364.6	505.1	524.1	399.4	NA	NA	NA	N	NA	N	Y	NA	N	0.0
UH1	Pembrolizumab	Discovery	UH1_35	Y	50.0	518.5	521.4	588.0	421.6	NA	NA	NA	Y (2)	NA	Y	Y	NA	N	0.0
UH1	Pembrolizumab	Discovery	UH1_36	Y	60.0	247.3	288.5	255.2	222.0	NA	NA	NA	Y (2)	NA	Y	Y	NA	N	0.0
UH1	Pembrolizumab	Discovery	UH1_37	Y	NA	188.3	NA	292.21	244.08	NA	NA	NA	N	NA	N	Y	NA	N	0.0
UH2	Pembrolizumab	Discovery	UH2_TOTAL	NA	NA	66.6	NA	NA	99.85	NA	NA	NA	NA	N	NA	NA	NA	NA	NA
UH2	Pembrolizumab	Discovery	UH2_39	Y	70.0	63.3	122.0	77.7	110.9	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH2	Pembrolizumab	Discovery	UH2_40	Y	NA	46.0	55.5	199.7	94.7	NA	NA	NA	N	NA	N	Y	NA	N	0.0
UH2	Pembrolizumab	Discovery	UH2_41	Y	NA	99.9	88.8	156.5	122.0	NA	NA	NA	N	NA	N	Y	NA	N	0.0
UH2	Pembrolizumab	Discovery	UH2_42	Y	65.0	55.5	99.9	44.4	69.9	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH3	Pembrolizumab	Discovery	UH3_TOTAL	NA	NA	166.4	NA	NA	277.36	NA	NA	NA	NA	N	NA	NA	NA	NA	NA
UH3	Pembrolizumab	Discovery	UH3_12	Y	80.0	433.8	654.6	310.6	420.8	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH3	Pembrolizumab	Discovery	UH3_13	Y	NA	615.7	776.6	510.3	965.2	NA	NA	NA	N	NA	N	Y	NA	N	3.8
UH3	Pembrolizumab	Discovery	UH3_15	Y	NA	377.2	499.3	266.3	299.6	NA	NA	NA	N	NA	N	Y	NA	N	2.6
UH3	Pembrolizumab	Discovery	UH3_14	Y	50.0	33.3	33.3	22.8	149.2	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH3	Pembrolizumab	Discovery	UH3_16	Y	NA	160.9	366.1	119.4	55.5	NA	NA	NA	N	NA	N	Y	NA	N	2.0
UH4	Pembrolizumab	Discovery	UH4_TOTAL	NA	NA	44.4	NA	NA	44.56	NA	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH4	Pembrolizumab	Discovery	UH4_4	Y	20.0	116.5	180.6	70.7	66.6	NA	NA	NA	Y (overlapping with UH4_5)	NA	Y (merged with UH4_5)	Y (merged with UH4_5)	NA	N	0.0 (merged with UH4_H5)

UH4	Pembrolizumab	Discovery	UH4_5	Y	20.0	110.9	194.7	110.9	110.9	NA	NA	NA	Y (overlapping with UH4_4)	NA	Y (merged with UH4_4)	Y (merged with UH4_4)	NA	N	0.0 (merged with UH4_H4)
UH4	Pembrolizumab	Discovery	UH4_2	Y	20.0	34.9	99.9	33.3	11.1	NA	NA	NA	Y (overlapping with UH4_3)	NA	Y (merged with UH4_3)	Y (merged with UH4_3)	NA	N	0.0 (merged with UH4_L3)
UH4	Pembrolizumab	Discovery	UH4_3	Y	20.0	23.0	69.5	27.7	20.7	NA	NA	NA	Y (overlapping with UH4_2)	NA	Y (merged with UH4_2)	Y (merged with UH4_2)	NA	N	0.0 (merged with UH4_L2)
UH5	Pembrolizumab	Discovery	UH5_TOTAL	NA	NA	463.8	NA	NA	355.02	NA	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH5	Pembrolizumab	Discovery	UH5_7	Y	60.0	1048.5	870.9	704.5	733.8	NA	NA	NA	Y	NA	Y	Y	NA	N	0.0
UH5	Pembrolizumab	Discovery	UH5_8	Y	NA	823.5	1027.6	632.4	976.3	NA	NA	NA	N	NA	N	Y	NA	N	0.0
UH5	Pembrolizumab	Discovery	UH5_10	Y	60.0	255.3	345.1	199.7	177.5	NA	NA	NA	Y	NA	Y	Y	NA	N	0.0
UH5	Pembrolizumab	Discovery	UH5_9	Y	NA	576.9	632.4	355.0	515.9	NA	NA	NA	N	NA	N	Y	NA	N	0.0
UH6	Pembrolizumab	Discovery	UH6_TOTAL	NA	NA	355.0	NA	NA	321.74	NA	NA	NA	NA	N	NA	NA	NA	NA	NA
UH6	Pembrolizumab	Discovery	UH6_29	Y	NA	521.4	360.6	466.0	288.5	NA	NA	NA	Y	NA	N	Y	NA	N	0.6
UH6	Pembrolizumab	Discovery	UH6_30	Y	NA	593.4	687.9	624.2	504.8	NA	NA	NA	Y	NA	N	Y	NA	N	8.3
UH6	Pembrolizumab	Discovery	UH6_31	Y	40.0	588.0	515.0	NA	NA	NA	NA	NA	Y	NA	Y	Y	NA	Y	4.6
UH6	Pembrolizumab	Discovery	UH6_28	Y	30.0	255.2	271.8	342.4	342.7	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.7
UH7	Pembrolizumab	Discovery	UH7_TOTAL	NA	NA	55.5	NA	NA	55.47	NA	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH7	Pembrolizumab	Discovery	UH7_52	Y	50.0	177.5	161.1	210.8	168.3	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH7	Pembrolizumab	Discovery	UH7_53	Y	NA	77.7	66.7	88.8	180.3	NA	NA	NA	N	NA	N	Y	NA	N	0.0
UH7	Pembrolizumab	Discovery	UH7_54	Y	NA	44.4	36.0	66.6	44.4	NA	NA	NA	N	NA	N	Y	NA	N	0.0
UH7	Pembrolizumab	Discovery	UH7_55	Y	40.0	55.5	55.6	44.4	77.8	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH8	Pembrolizumab	Discovery	UH8_TOTAL	NA	NA	133.1	NA	NA	133.13	NA	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH8	Pembrolizumab	Discovery	UH8_23	Y	NA	521.4	476.0	244.1	507.9	NA	NA	NA	N	NA	N	Y	NA	N	0.1
UH8	Pembrolizumab	Discovery	UH8_24	Y	70.0	490.4	476.0	459.9	593.6	NA	NA	NA	Y (2)	NA	Y	Y	NA	Y	0.0
UH8	Pembrolizumab	Discovery	UH8_25	Y	75.0	110.9	79.5	155.3	77.4	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.6
UH8	Pembrolizu	Discove	UH8_26	Y	NA	99.9	97.2	66.6	300.2	NA	NA	NA	N	NA	N	Y	NA	N	0.0

Appendix A: Supplementary Tables

	mab	ry																	
UH9	Pembrolizumab	Discovery	UH9_TOTAL	NA	NA	122.0	NA	NA	305.00	NA	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH9	Pembrolizumab	Discovery	UH9_18	Y	60.0	338.4	221.9	255.2	466.0	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.3
UH9	Pembrolizumab	Discovery	UH9_20	Y	NA	151.7	233.0	221.9	233.0	NA	NA	NA	Y (2)	NA	N	Y	NA	N	0.0
UH9	Pembrolizumab	Discovery	UH9_19	Y	NA	231.3	244.1	299.6	477.1	NA	NA	NA	Y (2)	NA	N	Y	NA	N	0.0
UH9	Pembrolizumab	Discovery	UH9_21	Y	50.0	22.2	144.2	64.5	110.9	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH10	Pembrolizumab	Discovery	UH10_TOTAL	NA	NA	931.9	NA	NA	650.01	NA	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH10	Pembrolizumab	Discovery	UH10_47	Y	30.0	1717.2	1852.8	1825.1	1531.0	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH10	Pembrolizumab	Discovery	UH10_48	Y	NA	703.0	672.0	718.4	654.6	NA	NA	NA	N	NA	N	Y	NA	N	0.0
UH10	Pembrolizumab	Discovery	UH10_49	Y	60.0	693.4	628.8	687.9	432.7	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH10	Pembrolizumab	Discovery	UH10_50	Y	NA	699.2	604.6	964.0	809.9	NA	NA	NA	N	NA	N	Y	NA	N	0.0
UH11	Nivolumab	Discovery	UH11_TOTAL	NA	NA	44.4	NA	NA	77.66	NA	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH11	Nivolumab	Discovery	UH11_73	Y	25.0	66.6	88.8	99.9	122.0	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH11	Nivolumab	Discovery	UH11_74	Y	28.0	77.7	77.7	67.1	77.7	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH12	Nivolumab	Discovery	UH12_TOTAL	NA	NA	209.2	NA	NA	403.36	NA	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH12	Nivolumab	Discovery	UH12_76	Y	80.0	865.4	931.9	887.6	931.9	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH12	Nivolumab	Discovery	UH12_78	Y	80.0	550.7	805.6	685.7	801.0	NA	NA	NA	Y	NA	N	Y	NA	Y	0.6
UH12	Nivolumab	Discovery	UH12_77	Y	80.0	288.5	310.6	366.1	283.1	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH12	Nivolumab	Discovery	UH12_79	Y	80.0	421.6	344.0	489.7	410.5	NA	NA	NA	Y	NA	N	Y	NA	Y	8.6
UH13	Nivolumab	Discovery	UH13_TOTAL	NA	NA	44.4	NA	NA	44.38	NA	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH13	Nivolumab	Discovery	UH13_68	Y	13.0	95.3	99.9	122.0	110.9	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH13	Nivolumab	Discovery	UH13_69	Y	15.0	24.8	19.9	33.3	11.1	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH14	Nivolumab	Discovery	UH14_TOTAL	NA	NA	321.7	NA	NA	520.34	NA	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH14	Nivolumab	Discovery	UH14_57	Y	NA	521.4	599.1	687.9	244.1	NA	NA	NA	N	NA	N	Y	NA	N	0.0
UH14	Nivolumab	Discovery	UH14_58	Y	80.0	1309.1	1342.4	1542.1	1331.3	NA	NA	NA	Y (2)	NA	Y	Y	NA	Y	0.0
UH14	Nivolumab	Discovery	UH14_56	Y	60.0	471.7	599.4	699.0	456.5	NA	NA	NA	Y (2)	NA	Y	Y	NA	Y	0.0
UH14	Nivolumab	Discovery	UH14_59	Y	NA	1126.1	1458.0	1708.5	1420.1	NA	NA	NA	N	NA	N	Y	NA	N	0.0
UH15	Nivolumab	Discove	UH15_TOT	NA	NA	1698.7	NA	NA	1921.35	NA	NA	NA	NA	Y	NA	NA	NA	NA	NA

		ry	AL																
UH15	Nivolumab	Discovery	UH15_60	Y	NA	1076.2	1497.7	1484.2	2019.2	NA	NA	NA	N	NA	N	Y	NA	N	0.0
UH15	Nivolumab	Discovery	UH15_61	Y	80.0	1830.6	1625.4	1739.1	1412.8	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH15	Nivolumab	Discovery	UH15_62	Y	60.0	1664.2	1669.8	1573.3	2336.1	NA	NA	NA	Y	NA	Y	Y	NA	Y	0.0
UH16	Nivolumab	Discovery	UH16_TOTAL	NA	NA	305.1	NA	NA	310.65	NA	NA	NA	NA	N	NA	NA	NA	NA	NA
UH16	Nivolumab	Discovery	UH16_65	Y	NA	386.5	349.5	416.7	277.4	NA	NA	NA	N	NA	N	Y	NA	Y (merged with UH16_H66)	0.0
UH16	Nivolumab	Discovery	UH16_66	Y	30.0	454.9	416.3	532.5	366.1	NA	NA	NA	Y	NA	Y	Y	NA	Y (merged with UH16_H65)	0.0
UH16	Nivolumab	Discovery	UH16_63	Y	NA	133.1	199.7	219.9	177.5	NA	NA	NA	N	NA	N	Y	NA	Y (merged with UH16_L62)	0.0
UH16	Nivolumab	Discovery	UH16_64	Y	20.0	33.3	156.1	255.2	110.9	NA	NA	NA	Y	NA	Y	Y	NA	Y (merged with UH16_L63)	0.0
UH17	Pembrolizumab	Validation	UH17_TOTAL	NA	NA	NA	NA	NA	NA	Y	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH17	Pembrolizumab	Validation	UH17_93	Y	65.0	NA	NA	NA	NA	NA	2637.5	1947.3	Y	NA	NA	NA	Y	NA	NA
UH17	Pembrolizumab	Validation	UH17_94	Y	65.0	NA	NA	NA	NA	NA	2371.0	2287.7	Y	NA	NA	NA	Y	NA	NA
UH18	Pembrolizumab	Validation	UH18_TOTAL	NA	NA	NA	NA	NA	NA	Y	NA	NA	NA	N	NA	NA	NA	NA	NA
UH18	Pembrolizumab	Validation	UH18_101	Y	NA	NA	NA	NA	NA	NA	688.5	527.5	N	NA	NA	NA	Y	NA	NA
UH18	Pembrolizumab	Validation	UH18_102	Y	NA	NA	NA	NA	NA	NA	322.1	507.3	N	NA	NA	NA	Y	NA	NA
UH18	Pembrolizumab	Validation	UH18_103	Y	90.0	NA	NA	NA	NA	NA	649.7	410.9	Y	NA	NA	NA	Y	NA	NA
UH18	Pembrolizumab	Validation	UH18_104	Y	NA	NA	NA	NA	NA	NA	669.3	702.7	N	NA	NA	NA	Y	NA	NA
UH19	Pembrolizumab	Validation	UH19_TOTAL	NA	NA	NA	NA	NA	NA	Y	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH19	Pembrolizumab	Validation	UH19_85	Y	NA	NA	NA	NA	NA	NA	888.4	699.6	N	NA	NA	NA	Y	NA	NA
UH19	Pembrolizumab	Validation	UH19_86	Y	NA	NA	NA	NA	NA	NA	798.0	798.8	N	NA	NA	NA	Y	NA	NA
UH19	Pembrolizumab	Validation	UH19_87	Y	80.0	NA	NA	NA	NA	NA	1849.1	1317.7	Y	NA	NA	NA	Y	NA	NA
UH19	Pembrolizu	Validati	UH19_88	Y	80.0	NA	NA	NA	NA	NA	1410.4	932.8	Y	NA	NA	NA	Y	NA	NA

	mab	on																	
UH20	Nivolumab	Validation	UH20_TOTAL	NA	NA	NA	NA	NA	NA	Y	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH20	Nivolumab	Validation	UH20_107	Y	NA	NA	NA	NA	NA	NA	844.0	723.7	N	NA	NA	NA	Y	NA	NA
UH20	Nivolumab	Validation	UH20_108	Y	65.0	NA	NA	NA	NA	NA	932.9	844.0	Y	NA	NA	NA	Y	NA	NA
UH20	Nivolumab	Validation	UH20_105	Y	NA	NA	NA	NA	NA	NA	721.8	789.7	N	NA	NA	NA	Y	NA	NA
UH20	Nivolumab	Validation	UH20_106	Y	65.0	NA	NA	NA	NA	NA	618.5	621.9	Y	NA	NA	NA	Y	NA	NA
UH21	Nivolumab	Validation	UH21_TOTAL	NA	NA	NA	NA	NA	NA	Y	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH21	Nivolumab	Validation	UH21_109	Y	80.0	NA	NA	NA	NA	NA	684.6	1278.7	Y	NA	NA	NA	Y	NA	NA
UH21	Nivolumab	Validation	UH21_110	Y	80.0	NA	NA	NA	NA	NA	521.9	276.3	Y	NA	NA	NA	Y	NA	NA
UH22	Nivolumab	Validation	UH22_TOTAL	NA	NA	NA	NA	NA	NA	Y	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH22	Nivolumab	Validation	UH22_111	Y	60.0	NA	NA	NA	NA	NA	207.8	187.8	Y (2)	NA	NA	NA	Y	NA	NA
UH22	Nivolumab	Validation	UH22_112	Y	60.0	NA	NA	NA	NA	NA	255.4	333.2	Y (2)	NA	NA	NA	Y	NA	NA
UH23	Nivolumab	Validation	UH23_TOTAL	NA	NA	NA	NA	NA	NA	Y	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH23	Nivolumab	Validation	UH23_81	Y	40.0	NA	NA	NA	NA	NA	344.3	155.5	Y	NA	NA	NA	Y	NA	NA
UH23	Nivolumab	Validation	UH23_82	Y	40.0	NA	NA	NA	NA	NA	885.9	1100.1	Y	NA	NA	NA	Y	NA	NA
UH23	Nivolumab	Validation	UH23_83	Y	NA	NA	NA	NA	NA	NA	NA	995.0	N	NA	NA	NA	Y	NA	NA
UH24	Nivolumab	Validation	UH24	NA	NA	NA	NA	NA	NA	Y	NA	NA	Y	Y	NA	NA	NA	NA	NA
UH25	Nivolumab	Validation	UH25_TOTAL	NA	NA	NA	NA	NA	NA	Y	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH25	Nivolumab	Validation	UH25_91	Y	60.0	NA	NA	NA	NA	NA	805.1	465.2	Y (2)	NA	NA	NA	Y	NA	NA
UH25	Nivolumab	Validation	UH25_92	Y	60.0	NA	NA	NA	NA	NA	1365.9	1010.6	Y (2)	NA	NA	NA	Y	NA	NA
UH26	Pembrolizumab	Validation	UH26_TOTAL	NA	NA	NA	NA	NA	NA	Y	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH26	Pembrolizumab	Validation	UH26_115	Y	50.0	NA	NA	NA	NA	NA	488.6	510.8	Y	NA	NA	NA	Y	NA	NA
UH26	Pembrolizumab	Validation	UH26_116	Y	50.0	NA	NA	NA	NA	NA	301.8	355.6	Y (2)	NA	NA	NA	Y	NA	NA
UH26	Pembrolizumab	Validation	UH26_117	Y	NA	NA	NA	NA	NA	NA	165.5	123.3	N	NA	NA	NA	Y	NA	NA
UH27	Nivolumab (Ipilimumab)	Validation	UH27_TOTAL	NA	NA	NA	NA	NA	NA	Y	NA	NA	NA	Y	NA	NA	NA	NA	NA
UH27	Nivolumab (Ipilimumab)	Validation	UH27_95	Y	NA	NA	NA	NA	NA	NA	621.9	566.4	N	NA	NA	NA	Y	NA	NA
UH27	Nivolumab (Ipilimumab)	Validation	UH27_96	Y	80.0	NA	NA	NA	NA	NA	577.5	460.0	Y (2)	NA	NA	NA	Y	NA	NA
UH27	Nivolumab	Validati	UH27_97	Y	80.0	NA	NA	NA	NA	NA	910.6	838.4	Y (2)	NA	NA	NA	Y	NA	NA

	(Ipilimumab)	on																	
UH27	Nivolumab (Ipilimumab)	Validation	UH27_98	Y	NA	NA	NA	NA	NA	NA	299.8	266.5	N	NA	NA	NA	Y	NA	NA
UH28	Nivolumab (Ipilimumab)	Validation	UH28	NA	NA	NA	NA	NA	NA	Y	NA	NA	Y (3)	Y	NA	NA	NA	NA	NA
UH29	Nivolumab	Validation	UH29	NA	NA	NA	NA	NA	NA	Y	NA	NA	Y (2)	Y	NA	NA	NA	NA	NA

For each experiment, the block slides (as reported in the Methods, Fig.1B and Extended Data Fig.1A) and the number of all analysed regions across samples are indicated in brackets, for a total of 738 profiled regions or samples. Tumour content was estimated by the pathologist in the regions of slide A that were subsequently projected in the sequential slides for IMC, WES and TCR-seq. QuPath (1) was used to calculate the median number of CD3 cells/mm2 in all selected regions of slides B, F, H and J for the discovery cohort and slides E and G for the validation cohort. For IMC, WES, and TCR-seq the two regions (one with low and one with high CD3 infiltration) with the highest tumour content were selected, whenever possible. This was possible for all samples except UH4, where DNA and RNA from the two CD3 high and low regions was merged. Where IMC was performed on the same region more than once, the total number of regions is reported in brackets. For TCR-seq, all four regions were sequenced in UH12, the two high and low CD3 regions were merged in UH16. UH1, UH4 and UH5 were excluded from TCR-Seq analysis because DNA was not sufficient. HE: Haematoxylin and Eosin; IMC: Imaging Mass Cytometry; mIF: Multiplexed Immunofluorescence; WES: Whole Exome Sequencing; A-FRET/; Amplified Förster Energy Transfer; NA = Not applicable.

This table was adapted from supplementary table 2 of:

Bortolomeazzi, M. et al. Immunogenomics of Colorectal Cancer Response to Checkpoint Blockade: Analysis of the KEYNOTE 177 Trial and Validation Cohorts. Gastroenterology 161, 1179-1193 (2021).

1. Bankhead, P. et al. QuPath: Open source software for digital pathology image analysis. Sci Rep 7, 16878, doi:10.1038/s41598-017-17204-5 (2017).

Supplementary Table 5_I. Antibodies used in the study

Cell population	Antibody Specificity	Cell clustering (IMC Panel I)	Cell clustering (IMC Panel III)	Vendor	Catalogue Number	Primary Antibody Dilution	Metal or Opal Tag	Application
1. T cells	CD3	T cells, PD1+ cells	NA	Fluidigm	3170019D	1:800	170Er	IMC Panel I, II, III
1. T cells	CD45RA	T cells, PD1+ cells, B cells	NA	Fluidigm	3166028D	1:2000	166Er	IMC Panel I and II
1. T cells	CD45RO	T cells, PD1+ cells, macrophages, dendritic cells, PDL1+ cells, B cells, neutrophils	CD74+ macrophages	Fluidigm	3173016D	1:500	173Yb	IMC Panel I, II, III
1. T cells	CD57	T cells, PD1+ cells	NA	Abcam	ab212405	1:300	174Yb	IMC Panel I and II
1. T cells	CD8	T cells, PD1+ cells	NA	Fluidigm	3162035D	1:800	162Dy	IMC Panel I, II, III
1. T cells	FOXP3	T cells, PD1+ cells	NA	Fluidigm	3155016D	1:200	155Gd	IMC Panel I and II
1. T cells	PD1	T cells, PD1+ cells	NA	Fluidigm	3165039D	1:50	165Ho	IMC Panel I, II, III
1. T cells	CD134	T cells	NA	Fluidigm	3151024D	1:100	151Eu	IMC Panel II
1. T cells	LAG3	T cells	NA	Fluidigm	3153028D	1:100	153Eu	IMC Panel II
1. T cells	TIM3	T cells	NA	Fluidigm	3154024D	1:100	154Sm	IMC Panel II
1. T cells	Vista	T cells	NA	Fluidigm	3160025D	1:100	160Gd	IMC Panel II
1. T cells	TCF7	T cells	NA	Cell Signalling Technology	2203	1:100	164Dy	IMC Panel II
1. T cells and Macrophages	CD4	T cells, PD1+ cells, macrophages, dendritic cells, PDL1+ cells	CD74+ macrophages	Fluidigm	3156033D	1:200	156Gd	IMC Panel I, II, III
11. Tumour cells	E-Cadherin	NA	NA	Fluidigm	3158029D	1:3000	158Gd	IMC Panel I, II, III
11. Tumour cells	Pan-keratin	NA	NA	Fluidigm	3148020D	1:3000	148Nd	IMC Panel I, II, III
12. Stromal cells	FAP	NA	NA	Abcam	ab207178	1:100	153Eu	IMC Panel I
12. Stromal cells	SMA	NA	NA	Fluidigm	3141017D	1:4000	141Pr	IMC Panel I
12. Stromal cells	Vimentin	Macrophages, dendritic cells, PDL1+ cells	NA	Fluidigm	3143029D	1:500	143Nd	IMC Panel I, II, III
2. B cells	CD20	B Cells	NA	Fluidigm	3161029D	1:250	161Dy	IMC Panel I
2. B cells	CD27	T cells, PD1+ Cells, B Cells	NA	Fluidigm	3171024D	1:300	171Yb	IMC Panel I and II
2. B cells	IgA	B Cells	NA	NovusBio	NB500-469	1:100	142Nd	IMC Panel I
2. B cells	IgM	B Cells	CD74+ macrophages	NovusBio	NBP2-34650	1:200	169Tm	IMC Panel I
3. Macrophages	CD16	Macrophages, dendritic cells, PDL1+ cells, neutrophils	CD74+ macrophages	Fluidigm	3146020D	1:200	146Nd	IMC Panel I, II, III
3. Macrophages	CD163	Macrophages, dendritic cells, PDL1+ cells, neutrophils	CD74+ macrophages	Fluidigm	3147021D	1:300	147Sm	IMC Panel I, II, III
3. Macrophages	CD68	Macrophages, dendritic cells, PDL1+ cells	CD74+ macrophages	Fluidigm	3159035D	1:400	159Tb	IMC Panel I, II, III
3. Macrophages	CD74	Macrophages, dendritic cells, PDL1+ cells, neutrophils, B cells	CD74+ macrophages	Biolegend	326802	1:100	144Nd	IMC Panel I, II, III
3. Macrophages	PDL1	Macrophages, dendritic cells, PDL1+ cells	CD74+ macrophages	RnD System	MAB1561	1:70	150Nd	IMC Panel I, II, III
3. Macrophages	CD40	NA	CD74+ macrophages	NovusBio	NBP2-34488	1:500	172Yb	IMC Panel III
3. Macrophages	CD206	NA	CD74+ macrophages	NovusBio	NBP2-52927	1:100	163Dy	IMC Panel III
3. Macrophages	FOLR2	NA	CD74+ macrophages	Origene	CF808026	1:100	153Eu	IMC Panel III
3. Macrophages	HLA-DR/DP/DQ	NA	CD74+ macrophages	Abcam	ab7856	1:500	141Pr	IMC Panel III

4. Dendritic cells and Macrophages	CD11b	NA	CD74+ macrophages	Cell Signaling Technology	D6X1N	1:1000	142Nd	IMC Panel III
4. Dendritic cells and Macrophages	CD11c	Macrophages, dendritic cells, PDL1+ cells	CD74+ macrophages	Abcam	ab216655	1:400	175Lu	IMC Panel I, II, III
4. Dendritic cells and Macrophages	CD103	NA	CD74+ macrophages	Abcam	ab254201	1:100	160Gd	IMC Panel III
5. Neutrophils	CD15	Neutrophils	NA	Fluidigm	3149026D	1:100	149Sm	IMC Panel I
6. Leukocytes	CD45	T cells, PD1+ cells, macrophages, dendritic cells, PDL1+ cells, B cells, neutrophils	NA	Fluidigm	3152016D	1:500	152Sm	IMC Panel I, II, III
7. Nucleated cells	Beta2Microglobulin (B2M)	Macrophages, dendritic cells, PDL1+ cells, B cells, neutrophils	CD74+ macrophages	Abcam	ab212756	1:100	176Yb	IMC Panel I, II, III
7. Nucleated cells	HLA-A/B/C	NA	CD74+ macrophages	BD Biosciences	565292	1:1500	161Dy	IMC Panel III
8. Proliferating cells	Ki67	T cells, PD1+ cells, B cells, neutrophils	NA	Fluidigm	3168022D	1:400	168Er	IMC Panel I, II, III
9. Endothelial cells Endothelium	CD31	NA	NA	Fluidigm	3151025D	1:300	151 Eu	IMC Panel I
9. Endothelial cells Endothelium	CD34	NA	NA	Abcam	ab213058	1:150	164Dy	IMC Panel I
10. Cytotoxic cells	Granzyme B (GzB)	T cells, PD1+ cells, neutrophils	NA	Fluidigm	3167021D	1:300	167Er	IMC Panel I, II, III
NA	CD3	NA	NA	Dako	A0452	1:200	NA	IHC
NA	Anti-Rabbit-HRP	NA	NA	Dako	P0448	1:200	NA	IHC
NA	Anti-Mouse-ATTO488	NA	NA	FASTBASE	in house produced	1:25	NA	A-FRET
NA	Anti-Rabbit-HRP	NA	NA	Jackson Laboratories	711-036-152	1:200	NA	A-FRET
NA	PD1	NA	NA	Abcam	ab52587	1:100	NA	A-FRET
NA	PDL1	NA	NA	Abcam	ab205921	1:500	NA	A-FRET
1. T cells	CD8	NA	NA	Cell Signaling Technology	85336	1:200	Opal 780 (1:75)	mIF
1. T cells	TCF7	NA	NA	Cell Signaling Technology	2203	1:50	Opal 540 (1:500)	mIF
1. T cells	PD1	NA	NA	Abcam	ab137132	1:300	Opal 650 (1:100)	mIF
1. T cells	GzB	NA	NA	Abcam	ab208586	1:100	Opal 480 (1:600)	mIF
1. T cells	Ki67	NA	NA	BD Biosciences	550609	1:200	Opal 690 (1:150)	mIF
3. Macrophages	CD68	NA	NA	Biolegend	916104	1:1000	Opal 620 (1:700)	mIF
3. Macrophages	CD74	NA	NA	Biolegend	326802	1:600	Opal 570 (1:700)	mIF
3. Macrophages	PDL1	NA	NA	RnD System	MAB1561	1:450	Opal 520 (1:150)	mIF

For each antibody, the cell population, target proteins, the target cell population in the phenotyping clustering with the IMC panels I or III and the application in the study are reported. IMC: Imaging Mass Cytometry; IHC: immunohistochemistry; A-FRET: Amplified Foster Resonant Energy Transfer; mIF: multiplexed Immunofluorescence; NA: not applicable.

This table was adapted from supplementary table 5 of:

Bortolomeazzi, M. et al. Immunogenomics of Colorectal Cancer Response to Checkpoint Blockade: Analysis of the KEYNOTE 177 Trial and Validation Cohorts. Gastroenterology 161, 1179-1193 (2021).