nature reviews gastroenterology & hepatology

https://doi.org/10.1038/s41575-023-00784-1

Roadmap

Check for updates

A Roadmap for the Human Gut Cell Atlas

Matthias Zilbauer[1,2,3,37] ✉, Kylie R. James[4,5,37], Mandeep Kaur[6,37], Sebastian Pott[7,37], Zhixin Li[8,37], Albert Burger[9,37], Jay R. Thiagarajah[10], Joseph Burclaff[11,12], Frode L. Jahnsen[13], Francesca Perrone[1,2], Alexander D. Ross[1,2,14], Gianluca Matteoli[15], Nathalie Stakenborg[15], Tomohisa Sujino[16], Andreas Moor[17], Raquel Bartolome-Casado[13,18], Espen S. Bækkevold[13], Ran Zhou[7], Bingqing Xie[19], Ken S. Lau[20], Shahida Din[21], Scott T. Magness[11,12], Qiuming Yao[22], Semir Beyaz[23], Mark Arends[24], Alexandre Denadai-Souza[25], Lori A. Coburn[26,27], Jellert T. Gaublomme[28], Richard Baldock[29], Irene Papatheodorou[30], Jose Ordovas-Montanes[10], Guy Boeckxstaens[15], Anna Hupalowska[31], Sarah A. Teichmann[18,32], Aviv Regev[31,33], Ramnik J. Xavier[34], Alison Simmons[35], Michael P. Snyder[36], Keith T. Wilson[26,27], Gut Cell Atlas Consortium* & Human Cell Atlas Gut Biological Network Consortium*

Abstract

The number of studies investigating the human gastrointestinal tract using various single-cell profiling methods has increased substantially in the past few years. Although this increase provides a unique opportunity for the generation of the first comprehensive Human Gut Cell Atlas (HGCA), there remains a range of major challenges ahead. Above all, the ultimate success will largely depend on a structured and coordinated approach that aligns global efforts undertaken by a large number of research groups. In this Roadmap, we discuss a comprehensive forward-thinking direction for the generation of the HGCA on behalf of the Gut Biological Network of the Human Cell Atlas. Based on the consensus opinion of experts from across the globe, we outline the main requirements for the first complete HGCA by summarizing existing data sets and highlighting anatomical regions and/or tissues with limited coverage. We provide recommendations for future studies and discuss key methodologies and the importance of integrating the healthy gut atlas with related diseases and gut organoids. Importantly, we critically overview the computational tools available and provide recommendations to overcome key challenges.

Sections

Introduction

Mapping the human gastrointestinal tract

Sampling strategies

Documenting metadata

Protecting donor identity and engaging communities

Available methodologies and their application

Computational challenges and opportunities

Development of a CCF for the HGCA

Summary of relevant existing data sets, studies and portals

Future opportunities, including incorporating data sets

Conclusions

A full list of affiliations appears at the end of the paper. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: mz304@cam.ac.uk

Roadmap

Key points

- The number of studies applying single-cell sequencing methods to human intestinal tissue has been rapidly increasing, providing a unique opportunity to generate a complete map of the human intestine.
- Generation of the Human Gut Cell Atlas (HGCA) requires the coordinated efforts of groups across the globe and the integration of various data sets followed by their computational analyses.
- This article provides a roadmap for the generation of the HGCA based on the expertise and recommendations of the Gut Biological Network of the Human Cell Atlas.
- The HGCA will provide a unique and highly valuable reference map enhancing research in intestinal health and disease.

Introduction

The intestinal tract is one of the most complex organs in the human body and serves a wide range of functions, including the digestion and absorption of nutrients, and represents a major site of immune interactions. Different anatomical sections of the intestinal tract have specific roles in the digestive process, requiring the presence of various cell types and complex interactions between them. Additionally, the interaction with trillions of nearby microbes has been shown to be of critical importance, adding further complexity to this finely tuned symbiosis. A detailed knowledge of gut physiology and cellular function in health is a prerequisite for investigating related diseases such as bowel cancer and inflammatory conditions.

The development of methodologies enabling genome-wide molecular profiling on a single-cell level has opened unprecedented opportunities to generate detailed anatomical maps of the human body. Although the number of studies investigating the human intestine using various single-cell profiling methods has increased substantially in the past few years, generating the first comprehensive Human Gut Cell Atlas (HGCA) is associated with major challenges. Above all, the ultimate success will largely depend on a structured and coordinated approach that aligns global efforts. The Gut Biological Network of the Human Cell Atlas (HCA) connects leading experts in a wide range of related areas, providing an ideal platform to lead the development of the first single-cell HGCA. Here, we provide a detailed roadmap, including recommendations on how to combine existing data and requirements for future studies. We discuss key methodologies and the importance of integrating the healthy gut atlas with related diseases and gut organoids. Importantly, we critically reflect on available computational tools, highlight existing limitations and provide recommendations to overcome key challenges. Essentially, generation of the first comprehensive HGCA will provide researchers, scientists and clinicians with greater scope and power to enable novel discoveries into intestinal biology and pathophysiology with the ultimate goal of improving human health.

Mapping the human gastrointestinal tract

The human digestive tract reaches from the oral cavity to the rectum, including the oesophagus, stomach, and small and large intestines. Generating a complete map requires the inclusion of all organs associated with the digestive process and detailed sampling of each gut segment to capture distinct anatomical changes along the cephalad–caudal axis (Fig. 1). A unique feature of the human intestine is the presence of trillions of microbiota that exist in a finely balanced symbiosis and are thought to be of critical importance to the host[1]. As such, profiling microbial communities will provide an opportunity to interrogate the crosstalk between microbes and host cells. However, this task is associated with the major challenges of profiling the human gut microbiome, including the vast inter-individual variability, differences in its composition along the intestinal tract and the dynamic changes of microbiota over time[2,3]. As a result, integrating the gut microbiome in the HGCA is likely to be part of future efforts and will require close interaction with expert researchers in this area. Another important aspect unique to the gastrointestinal tract is the exposure to and interaction with a wide range of food and nutritional components as well as with antigens or potential toxins present in the daily diet. Robust evidence in mice and humans suggests that gastrointestinal development, function and predisposition to related diseases are strongly correlated with dietary habits[4,5]. Therefore, documenting details on dietary habits will add critical information and further increase the value of the generated molecular profiles. Importantly, epigenetic mechanisms can mediate the effect of exposure to environmental factors on stable cellular phenotypes[6]. The possibility of simultaneously profiling multiple molecular layers at single-cell resolution, including epigenetic, transcriptomic and proteomic signatures, will support us in unravelling the interactions between intestinal host cells and our environment.

Sampling strategies

Acquiring human tissue in sufficient quantities and from a large number of donors is one of the major challenges associated with the generation of the HGCA. There are several considerations that apply to the gastrointestinal tract. In general, there are three major strategies for sampling the alimentary tract, including mucosal biopsy samples, live surgical resections and resections from deceased donors[7]. Each strategy and tissue source have unique advantages and disadvantages (Fig. 2). Mucosal biopsy samples can be obtained during a routine endoscopy but access is restricted to the upper and lower gastrointestinal tract, with the jejunum and proximal ileum rarely sampled[8]. Multiple biopsy samples can be taken from the same patient, enabling comparisons of different anatomical regions, which can be accessed during routine upper endoscopy and colonoscopy[8]. Contrary to mucosal biopsy samples, which can be obtained from healthy individuals, live resections are derived from patients with major gastrointestinal pathologies. Work published in 2019 identified transcriptional differences, for example, in antimicrobial defence pathways and mucin biosynthesis, between healthy intestines and normal-appearing intestinal tissue proximal to chronic inflammatory regions in patients with ulcerative colitis[9]. Hence, resections from patients with gastrointestinal disease should be used with caution when mapping 'healthy' cells. Advanced endoscopic imaging technology, such as confocal laser endomicroscopy, can provide additional information and guide sampling strategies[10]. Tissue from deceased organ donors enables the sampling of entire organs, providing the opportunity for widescale comparisons between regions. Another major advantage of gut resections (either surgical or from deceased donors) is the provision of all layers of the intestinal tube from luminal contents through the musculature and serosa. By contrast, forceps biopsy samples obtained during clinical endoscopy only capture the mucosa, including the epithelium and immediate subepithelial cells[11].

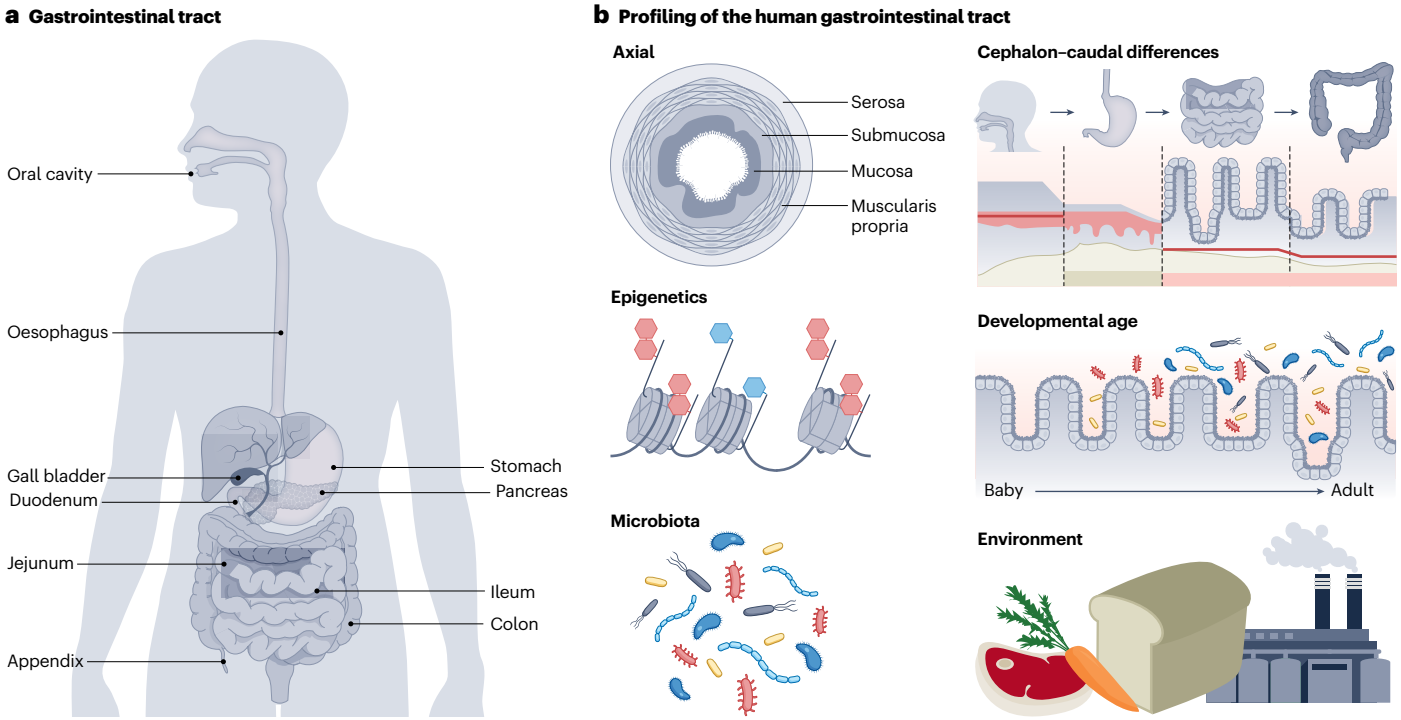

Fig. 1 | Profiling the human gastrointestinal tract. a, A complete map of the human gastrointestinal tract requires the inclusion of the entire intestinal tube and associated organs. **b**, Profiling must also consider the effect of the developmental stage, gut microbiome and potential effects of environmental factors such as diet, toxins and medication. In addition to transcriptional profiling, capturing the underlying genome sequence and the epigenetic programme will yield critical information.

Size also varies drastically between biopsy samples and resections, affecting the number and type of downstream analyses that can be performed. Finally, tissue quality might differ between strategies, with the time from sampling to processing of tissue being of critical importance[12].

In summary, compared to other human organs, the gastrointestinal tract is frequently sampled during routine clinical procedures, providing unique opportunities to obtain tissue from healthy individuals as well as from many related gut diseases. However, for complete coverage of all anatomical regions and layers of the intestinal tube, the use of surgical resection materials as well as deceased donor tissue is essential.

Documenting metadata

Generation of the HGCA requires combining a vast number of different data sets generated by research groups around the globe[13,14]. A fundamental aspect that determines the integration and comparability of datasets is the documentation of detailed metadata[15]. The HCA (The Human Cell Atlas – Metadata) provides extensive guidance on this topic[16]. In the following section, we briefly describe the basic information required to enable future studies to be included in the HGCA and provide a template metadata table (Fig. 3). In line with recommendations provided by the HCA, required metadata can be divided into the following main aspects: (1) study design, (2) donor information, (3) sample information, (4) sample processing and (5) data generation[16]. In brief, a detailed description of the study design, including patient inclusion criteria and sampling strategy, is of major importance. Donor information should include baseline demographic data, details on any known medical conditions, particularly those affecting the intestine, and medications that can be of major benefit. Similarly, the effect of dietary habits on gut physiology is well established, and information on dietary habits could be used to identify novel aspects of dietary factors involved in gut health and disease[4,5]. Baseline information on sample type includes the method by which it was obtained, sample area covered and anatomic location[7,12,17]. Depending on the sample type, the exact anatomical location from which the sample was obtained can be difficult to determine. For example, during routine endoscopy, the exact location of mucosal biopsies can often only be estimated relative to anatomical landmarks such as the terminal ileum[17]. However, the level of detail provided has a major effect on the comparability of studies and the interpretation of data generated. If available, representative haematoxylin and eosin and/or other staining on tissue sections can be directly linked to samples processed for single-cell studies[17,18]. Information on sampling procedures should include details on the length of time between sampling and processing, storage duration, and storage type (for example, −20 °C or −80 °C, liquid nitrogen). Furthermore, providing a detailed protocol for sample processing, including tissue dissociation, cell viability, possible enrichment of individual cell types and equipment used, is important. The development of standardized protocols is subject to ongoing work within the HCA. Active engagement of the scientific community and the use of available resources, including recommended protocols, will increase the comparability of data generated by different groups.

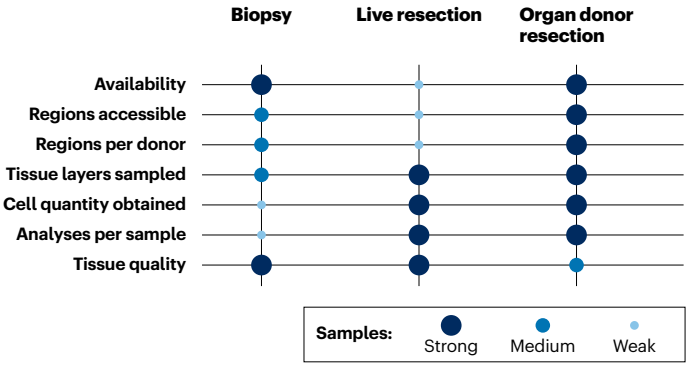

Fig. 2 | Summary of main tissue types and sampling strategies available. The main advantages and limitations of the tissue types and sampling strategies are illustrated.

Protecting donor identity and engaging communities

Documentation, storage and sharing of patient and donor information raise important ethical considerations[19]. Regulations vary considerably between different institutions and territories, further complicating the sharing of tissues and data. In the EU, regulations concerning the protection of personal data and the responsible sharing of such data have been formulated in the General Data Protection Regulation (GDPR) 2016/679 (ref. 20). In the USA, there are several federal regulations that must be considered when using human tissue and data in the context of research, including the Department of Health and Human Services and the FDA[21]. In addition to major differences in the rules and regulations between territories across the globe, exchanging tissue and data between regions poses further complexities and barriers. However, given the critical importance of metadata for the integration of different data sets, obtaining appropriate donor consent for sharing information, data and/or tissue is a prerequisite. In addition to explicit informed consent, pseudonymization of data and donor privacy must be ensured. A critical aspect of achieving a broad utility for the HGCA is the inclusion and participation of ancestrally and geographically diverse populations. This requires close partnerships between researchers, funders, and potential patient and donor communities, particularly in the case of historically neglected or mistreated populations in research[22,23]. Ideally, studies providing data for the HGCA should include as many community stakeholders as feasible, with the long-term goal of having research questions formulated by researchers and participating populations working in partnership[24].

Available methodologies and their application

A wide range of single-cell and single-nuclei genomics tools have been developed and have enabled the creation of high-resolution cellular atlases from human organs and tissues[13,25,26]. By and large, the methodologies that will be used to generate the HGCA are the same as those used to profile other organs or systems (Fig. 4). In addition to single-cell transcriptional profiling, key methods include spatial transcriptomics and single-cell profiling of other molecular layers such as the genotype[27] and epigenotype, including chromatin states (single-cell assay for transposase-accessible chromatin sequencing (ATAC-seq))[28] and DNA methylation (whole-genome bisulfite sequencing)[29]. Furthermore, several methods have been developed that enable simultaneous profiling of multiple molecular layers such as the 10x Genomics Multiome kit, which combines single-nuclei ATAC-seq and single-nuclei RNA sequencing (RNA-seq)[30]. Single-cell nucleosome, occupancy and methylome sequencing can measure DNA methylation and chromatin accessibility within single cells[31], whereas single-cell nucleosome, methylation and transcription sequencing capture transcript levels in addition to chromatin accessibility and DNA methylation[32]. Furthermore, the use of methods that enable enrichment for specific and/or rare cell types will be key to achieving complete coverage. For example, mining rare cells sequencing (MIRACL-seq) enables label-free enrichment of rare cell types[33], whereas the Chromium Single Cell Immune Profiling assay provided by 10x Genomics enables detailed immune cell profiling, including full-length, paired B cell or T cell receptor sequences, surface protein expression, antigen specificity, and 5′ gene expression[34]. Another key challenge for the HGCA is the timely processing of fresh tissue samples. Isolation of tightly connected epithelial cells is associated with damage[12], highlighting the need for suitable dissociation protocols as well as computational tools to exclude dead or damaged cells[12]. Furthermore, dissociation of cells at 37 °C for a prolonged time (that is, more than 1 h) induces the expression of immediate early genes, thereby disrupting nuanced cellular states. Current advances have led to the development of single-nuclei RNA-seq protocols, which can be applied to frozen tissue samples, thereby easing the burden of sample processing. High-quality nuclei can be readily isolated from frozen samples and subjected to single-nuclei RNA-seq[35]. Although fewer transcripts are recovered compared with fresh, non-frozen samples, these data capture most cell populations and vastly increase the availability of tissue sources[35]. The decision on which tissue processing method is most appropriate will be largely determined by the individual setup, including tissue availability. However, a detailed description of experimental procedures is of critical importance for the value of generated data for the HGCA.

As a result of methods requiring tissue dissociation prior to single-cell profiling, information on the spatial arrangement of cells, which is crucial for their function, is lost entirely. Multiple spatial methods are used to capture cells in their anatomical context, including next-generation sequencing-based (for example, 10x Genomics Visium)[36] and imaging-based (for example, multiplexed error-robust fluorescence in situ hybridization) assays[37,38]. Typically, frozen or formalin-fixed, paraffin-embedded tissue sections are used as starting material and, ideally, are obtained from the same area from which tissue for single-cell assays was taken[18,36]. Integration of spatial approaches with single-cell genomics provides both the cellular resolution as well as spatial organization of cell combinations and states (functional tissue units) as an essential framework for a comprehensive atlas of the human gut. Several computational tools have been developed and will be discussed later.

In summary, existing methodologies offer researchers a wide range of opportunities to address their research goals. Although integrating data derived from studies using different methodologies poses a challenge for the HGCA, the key factor determining the potential value is the quality of data generated combined with the documentation of detailed metadata. Additionally, we encourage researchers across the globe to engage with the HCA Gut Bionetwork prior to and during their single-cell studies to further maximize the use of generated data.

Computational challenges and opportunities

The HGCA aims to generate a resource for the scientific community that is reliable, easy to access and user-friendly. Given the vast number of diverse data sets generated by a wide range of research groups using

different methodological approaches (as outlined later), the usage of existing computational tools and development of novel ones are of critical importance. The main challenges include adequate measures to identify studies with sufficient quality and combining and mapping diverse data sets. Furthermore, the development of a user-friendly interface enabling the data to be explored by the scientific community is a key requirement. Although many of these tasks also apply to most, if not all, tissue mapping studies[39–41], there are several additional challenges and opportunities that are specific to the intestine. These include the integration of host cell molecular signatures with mucosa-associated microbial profiles, mapping of data sets to specific anatomical locations along the craniocaudal axis of the gut, and alignment of healthy gut-derived data sets with disease state and intestinal organoid models. With the rapid development of single-cell RNA-seq technology, numerous computational tools and pipelines have been developed to analyse single-cell data[42–44]. These include programmes and/or packages enabling pre-processing of data such as quality control checks, normalization and batch correction, sequence alignment, and detection and removal of cell doublets[45–47]. A summary of relevant existing computational tools and/or packages is provided in Table 1. In the following section, we briefly discuss the main computational challenges and opportunities related to the development of the HCGA.

Sample ID _____ **Patient ID** _____

Study design

Summary	
Patient number	
Sample number	
Patient inclusion criteria	
Sampling strategy	

Donor information

Age	
Sex	
Ethnicity	
Medical condition (gastrointestinal related)	
Medical condition (non-gastrointestinal related)	
Medication (gastrointestinal related)	
Medication (non-gastrointestinal related)	
Dietary habits (if applicable)	

Sample information

Sample type	Mucosal biopsy	
	Surgical resection	
	Deceased donor	
Macroscopical appearance of sample	Normal	
	Pathological (provide details)	
Microscopical appearance of sample	Normal	
	Pathological (provide details)	
	Representative H&E stain available (provide details of H&E and other stains)	**YES/NO**

Anatomical sample location

Oral cavity		
Oesophagus	Proximal	
	Mid	
	Distal	
Stomach	Antrum	
	Body	
	Fundus	
	Pylorus	
Small bowel[a]	Duodenum	
	D1	
	D2	
	D3	
	D4	
	Jejunum	
	Proximal	
	Mid	
	Distal	
	Ileum	
	Proximal	
	Mid	
	Terminal ileum	
Large bowel[a]	Appendix	
	Caecum	
	Ascending	
	Transverse	
	Descending	
	Sigmoid	
	Rectum	

CCF sample location

		Distance (mm/cm)
Prominal landmark		
Distal landmark		
Sample storage	4˚C	
	−80˚C	
	Liquid nitrogen	
Storage solution		
Starting material	Fresh tissue	
	Frozen tissue	
Tissue dissociation	Whole tissue	
	Cell type purification	
	Cell viability (%)	
Downstream applications	Cell culture	
	Organoid generation	
	Omics processing	

Data generation

Omics approach used	RNA-seq	
	scRNA-seq	
	ATAC-seq	
	Proteomics	
	Whole-genome sequencing	
Library preparation		
Sequencing platform		

Fig. 3 | Template metadata for gut-related single-cell studies. Template table for the collection of basic information required to enable future studies to be included in the HGCA. ATAC-seq, single-cell assay for transposase-accessible chromatin sequencing; CCF, common coordinate framework; H&E, haematoxylin and eosin; RNA-seq, RNA sequencing; scRNA-seq, single-cell RNA-seq. [a]If available, provide distance (in mm/cm) from the nearest anatomical landmark: gastroduodenal junction, ileocaecal valve, hepatic flexure, splenic flexure or anus.

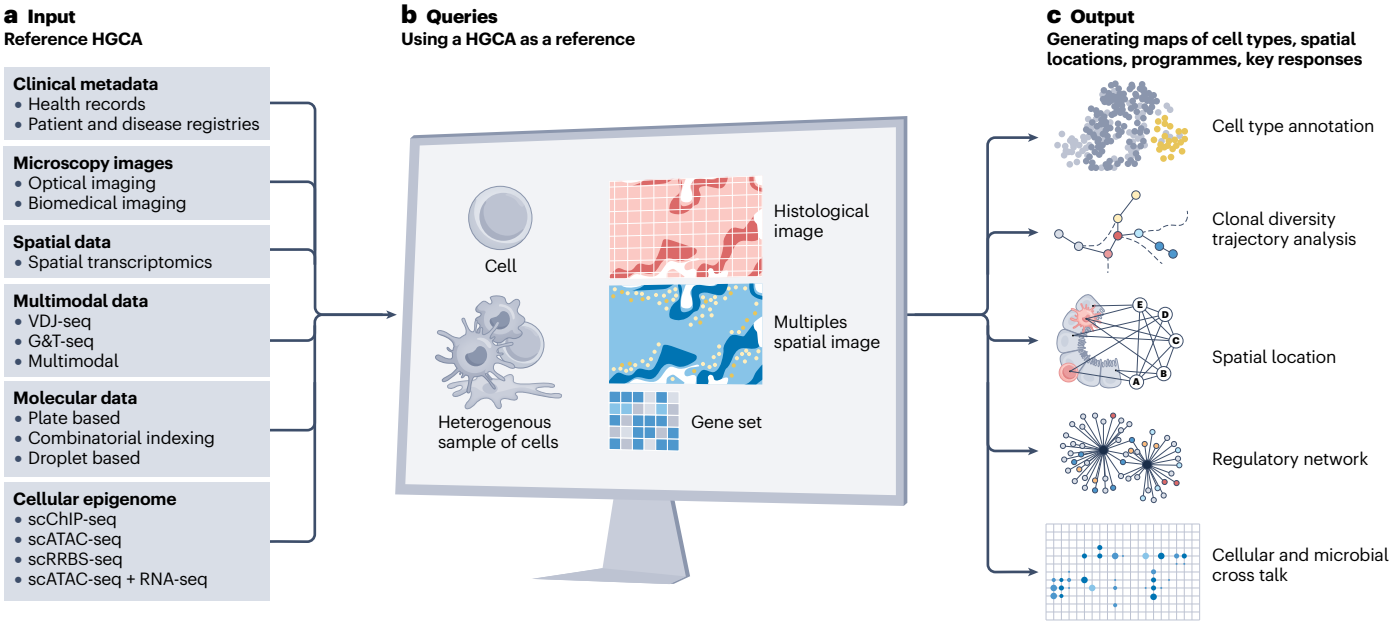

Fig. 4 | Data integration, processing and analysing strategies. a, Generation of the Human Gut Cell Atlas (HGCA) requires successful integration of various data sets, data types, and associated patient and donor metadata. **b**,**c**, Successful integration will allow a range of queries to be performed and outputs generated. Simultaneous profiling of epigenetic mechanisms on a single-cell level will allow the identification of regulatory cellular networks and help to define cellular identities. G&T-seq, genome and transcriptome sequencing; RNA-seq, RNA sequencing; scATAC-seq, single-cell assay for transposase-accessible chromatin; scChiP-seq, single-cell chromatin immunoprecipitation sequencing; scRRBS-seq, single-cell reduced representation bisulfite sequencing; VDJ-seq, variability, diversity and joining sequencing.

Combining and integrating data sets across modalities

Combining and integrating studies performed by different groups using either the same or different methodological approaches poses a major challenge. An important first step is the application of stringent selection criteria and quality control measures to select high-quality data sets. In addition to data quality, the provision of detailed metadata is a prerequisite for successful integration into the HGCA[17]. Another major challenge is the removal of potentially confounding batch effects that refer to differences caused by technical variation rather than reflecting true biological differences[48]. Available approaches and packages include Seurat[49], LIGER[50] and Harmony[51], which use a variety of different computational approaches[52]. The integration of different molecular profiles (for example, transcriptomes and epigenomes) forms another key computational challenge and analytical frameworks have been developed to integrate multiple data types in the same cells, including GLUE[53], MOFA+[54] and Cobolt[55]. Additionally, a statistical regression framework (MIRACL-seq) has been developed to integrate clinical metadata and genetic information in single-cell profiling studies[33].

Cell type annotation

Abundant and well-characterized cell types can be reliably identified through unsupervised clustering algorithms followed by the comparison of key marker gene expression profiles[56]. Generating a list of known marker genes based on existing literature forms a key aspect in the process of performing reliable cell annotation and requires the contribution of experts in the field. Indeed, combining expertise and large data sets provides a unique opportunity to develop an extensive cell marker gene list using both automated and supervised cell identification approaches (for example, MACA[57], singleR[58], ScType[59]). For each confidently identified gut cell type, marker genes can then be inferred by distinguishing the known clusters (for example, COMET[60], COSG[61]), ultimately leading to a fully automated annotation procedure. Applying automated annotation approaches to large data sets will also yield unknown and novel cell clusters, which must be subjected to additional validation studies and functional characterizations, for example, by using human gut organoid models, as will be discussed later.

Mapping cellular location

Mapping individual cell types to their specific anatomical location within the human intestine is of critical importance. Combining single-cell sequencing technologies with spatial transcriptomics enables accurate profiling of cell type topography and several tools have been developed to address the computational aspects involved[62], including ASAP[63], novoSpaRc[64] and Cell2location[65]. Multiomic spatial profiling efforts result in a gene plus a protein-by-cell matrix, annotated with spatial coordinates for the centroid position of each cell[66]. This matrix can then be used as input to the currently developed open source for spatial analysis pipelines[67], enabling an unbiased identification of spatial patterns based on gene expression, distinct cellular neighbourhoods and cell-to-cell interactions. Mapping the spatial network of cells in the human intestine will enable the identification of genes with coherent spatial expression patterns using methods such as BinSpect and SpatialDE[68,69]. Spatial domains with coherent gene expression patterns can be identified with hidden Markov random field models that find spatial domains by comparing the gene expression patterns of

each cell with its neighbourhood to search for coherent patterns[67,70]. The prevalence of specific cell-to-cell interactions in the human intestine could be evaluated by the frequency that each pair of cell types is proximal to each other[71]. Finally, spatially interacting cell types can be analysed to identify which known ligand–receptor pairs show increased or decreased co-expression, which could serve as a proxy for signalling activity, by creating a background distribution through spatially aware permutations[67], increasing predictive power when compared to spatially unaware permutations. The inclusion of imaging data in the HGCA provides an additional approach towards mapping cellular location and will be discussed in more detail below.

Cellular and microbial crosstalk

Investigating cellular crosstalk through the analyses of single-cell transcriptomic data sets is an active research field[72]. Amongst the most compelling existing approaches are CellPhoneDB, which enables interrogation of context-specific crosstalk between different cell types based on an extensive data base of known receptors and ligands[73]. Applying these algorithms to large and/or combined data sets is likely to yield novel insights into fundamental aspects of human intestinal physiology[18]. Extending such analyses to include the crosstalk between human host cells and the gut microbiome represents another substantial challenge. To this end, several computational strategies have been developed to identify cellular and microbial crosstalk in the gut utilizing single-cell data from human gut and microbial samples[74–76]. Current studies have provided evidence for the major value of combining in situ spatial profiling technologies with single-cell sequencing to interrogate host–microbial interactions[77].

Regulatory network inference

Successful integration of data sets and/or molecular profiles has the potential to unravel both known and novel cell type-specific molecular networks. Several tools are available to enable the prediction of gene regulatory networks that control fundamental biological functions such as cellular differentiation and cell state transitions (for example, SCENIC[78], GRNBoost2 (ref. 79), PIDC[80]). Identifying gene regulatory networks contributes towards our understanding of how coordinated expression of transcription factor networks drives expression of their respective target genes and ultimately shapes and maintains gut cell identity[81]. Examples include the identification of transcriptional networks that are present in the human fetal intestinal epithelium and reactivated in patients with Crohn's disease[18] and the coregulation of ulcerative colitis risk genes by a limited number of gene modules in disease-relevant cell types[82].

HGCA portal

HGCA portal is a user-friendly and interactive web portal, essential for making the HGCA accessible to a broad research community[83]. There are a number of existing portals of single-cell data that provide a wide range of tools to interrogate data sets, and a summary is provided later[84–86]. These portals provide an excellent starting point and we envision complete integration of the HGCA into the larger portals, which also enables gut-specific tools to be developed.

Development of a CCF for the HGCA

A critical part of the HGCA is its ability to capture the location of cells in their anatomical and physiological context[13]. The concept of a common coordinate framework (CCF) has been introduced to facilitate this capability. Rood et al. provide the following definition[87]: "An underlying reference map of organs, tissues, or cells that allows new individual samples to be mapped to determine the relative location of structural regions between samples". Like any other computational model, a CCF provides an abstract representation of a real-world item. CCFs have been developed for several human organs, including the lung[88] and brain[89], and the vasculature has been used as a CCF of the entire human body[90]. A first step towards the development of an HGCA CCF requires the generation of one-dimensional, two-dimensional and three-dimensional models of the intestinal tract. The one-dimensional conceptual model is based on the clinical view and biological organization of the digestive tract, in which distance from anatomical landmarks provides the critical location metric[17]. 'Anatomograms' are two-dimensional graphical models that include a simplified view of the gut for representational purposes, thereby providing a basic framework for the accurate capture of the anatomical location of the tissue component and cell types[91]. Three-dimensional models can be generated by integrating computerized tomography or magnetic resonance imaging data. Generating a framework also involves developing a system of associated coordinates that provide a detailed specification of anatomical location, tissue and cell type[17]. Hence, providing a detailed anatomical location of sampling sites is of critical importance. However, there are limitations to the accuracy of obtaining the exact anatomical location during surgery or intestinal endoscopy[17]. Stating proportional lengths and/or distances from reliable anatomical landmarks, such as the splenic flexure or ileocaecal valve, could improve accuracy and help facilitate the integration of generated data sets into

Table 1 | Summary of currently available computational packages

Main categories	Task of analysis	Available tools and/or algorithms
Combining and integrating data sets across modalities	General workflow	Seurat, Scanpy
	Trajectory inference	Monocle, PAGA, Slingshot, Velocyto, scVelo
	Imputation	MAGIC, scVI, SAVERX
	Batch effect correction	LIGER, Harmony, BBKNN, scVI
	Cell type matching and searching	FR-Match, Cell BLAST
	Multi-modal integration	GLUE, MOFA+, Cobolt, MultiMAP, ArchR, MultiVI
	Metadata integration	Drokhlyansky et al. (ref. 33)
Cell type annotation	Automatic cell type annotation	MACA, singleR, ScType, Celltypist
	Novel markers	COMET, COSG
Mapping spatial location	Spatial location inference	ASAP, novoSpaRc, Cell2location, BinSpect, SpatialDE
Cellular and microbial crosstalk	Cellular crosstalk	CellChat, scConnect, CellComm, CellPhoneDB
	Microbial crosstalk	Kang et al., 2020 (ref. 74), Chattopadhyay et al., 2018 (ref. 75), Liao et al., 2020 (ref. 76)
Regulatory network inference	Transcription factor regulatory network	SCENIC, GRNBoost2, PIDC

the HGCA CCF. Another complementary strategy to localize cell types in two-dimensional space is the combined use of single-cell RNA-seq with spatial transcriptomics, as discussed earlier. This strategy requires the development of computational tools and efforts to allow the integration of single-cell RNA-seq and spatial data into a CCF, and work in this area is ongoing.

Several CCFs have been developed and are accessible through existing data portals. For example, as part of the Human BioMolecular Atlas Program (HuBMAP), three-dimensional models of both the human large and small intestine have been developed[17]. The output is a series of surface models of the gut representing the outline within a three-dimensional context enabling the placement of tissue or cell types at the macro-level. This data can be accessed online through a CCF Registration User Interface (Fig. 5a), and mapped data can be explored using the Exploration User Interface. Furthermore, as part of a collaborative project funded by the Leona M. and Harry B. Helmsley Charitable Trust (Gut Cell Atlas), researchers from The European Bioinformatics Institute (EBI Cambridge) and Edinburgh have developed a one-dimensional linear conceptual CCF model for the human large and small intestine that is linked to the two-dimensional anatomogram[17,91]. These mappings, coupled with the inverse transform from the three-dimensional and two-dimensional spaces, back to the one-dimensional linear model and enable spatial interoperability between all representations and, therefore, the capability to compare and query data registered to any CCF using a web-based visualization platform[17] (Fig. 5b–d).

In addition to providing a framework for the accurate mapping of single data sets, CCFs are also capable of integrating other related data sets and can be managed across geographically dispersed data repositories[87]. Importantly, the value of CCFs is directly related to the level of standardization achieved as it determines the degree of potential interoperability across the data sets[17]. Amongst the ways to increase standardization is the development of consensus guidelines on how anatomical location can be most accurately documented for the purpose of integrating single-cell studies into the HGCA CCF. A starting point can be the use of a standard metadata template (Fig. 3) and the adherence to the minimum information standard for the description of cell and tissue sources in the gut. Ultimately, the successful development of a CCF will greatly enhance the value and broad application of the HGCA.

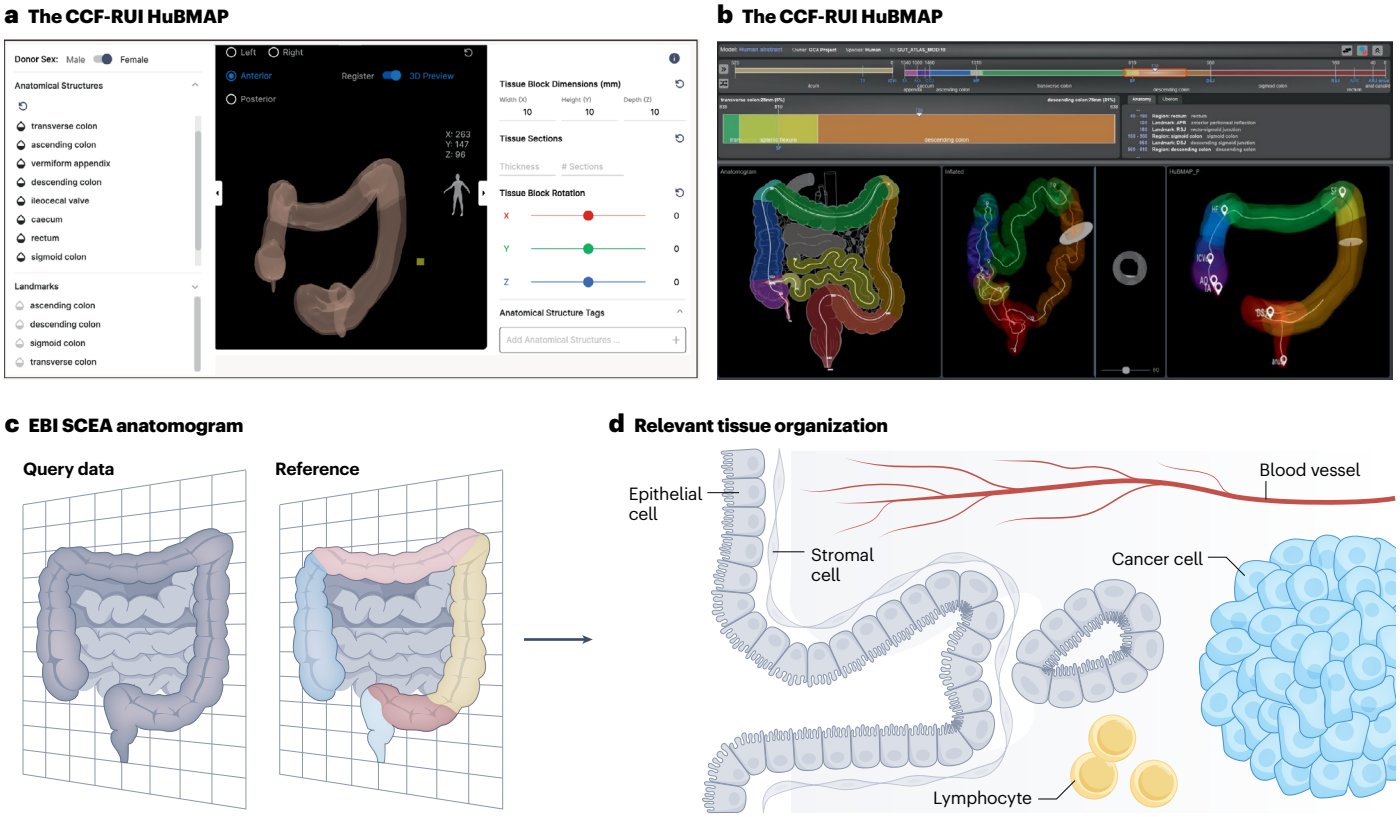

Fig. 5 | Current applications for the Human Gut Cell Atlas Common Coordinate Framework. a, Human BioMolecular Atlas Program (HuBMAP) Registration User Interface (CCF-RUI) showing the gut visibility controls in the left panel, interactive block registration interface in the centre panel and block specification controls in the right panel. **b**, Gut Atlas CCF browser interface shows the one-dimensional conceptual model for the full large and small intestines with a zoom panel and ontology listings at the top, with two-dimensional and three-dimensional interactive displays for all of the available models. The position of the current region of interest is displayed in each common coordinate framework (CCF) view and is fully interoperable in the sense that position selection in any view will be updated immediately in all the other views. **c**, European Bioinformatics Institute (EBI) Single-Cell Expression Atlas (SCEA) anatomogram for the large and small intestines enables the selection of any structure identified in the anatomical structures, cell types and biomarker (ASCT) tables. **d**, Expanded view of the anal canal to show relevant cell types and tissue organization of that region.

Summary of relevant existing data sets, studies and portals

The number of published studies reporting on single-cell profiling of the human gut has increased substantially since the founding of the HCA in 2016. Table 2 provides a summary of some of the major studies that have profiled primary tissue obtained from individual gut segments, including the oral cavity[92–94], oesophagus[95,96], stomach[96–100], and small[18,25,101–108] and large[18,25,33,82,104,105,107,109–113] intestine. Successful integration of these studies provides a strong foundation for the development of the HGCA. Indeed, several publicly accessible portals have already been established that contain intestinal single-cell data sets and enable access in a user-friendly way. Furthermore, a substantial number of studies have profiled intestinal tissues obtained from patients with gut diseases[82,107,108,114–127] and/or human intestinal organoids[18,128,129] (Fig. 6 and Table 2). In this section, we provide a summary of the main existing data sets and portals, highlight examples of novel findings derived from single-cell research, and provide recommendations for future work.

In utero development and healthy gut data sets

The high throughput scalability intrinsic to most single-cell technologies has provided unprecedented advances in the field. One area that benefited from this aspect is gastrointestinal organogenesis as its progress has been hindered by the scarcity of human samples. Indeed, data sets derived from single-cell RNA-seq spanning from 6 to 25 post-conception weeks have yielded several novel insights. For example, during early fetal development, the crypt–villus axis begins to emerge in the small intestine, concomitant to the appearance of *FOXL1*$^+$ mesenchymal cells co-expressing *PDGFRA* and *F3* (ref. 18). In other single-cell studies, a population of mesenchymal cells displaying a similar transcriptional signature were found to co-express *NRG1*, which has been demonstrated to support the differentiation of *LGR5*$^+$ stem cells into mature intestinal epithelial cells[130,131]. Additionally, distinct clusters of single-cell transcriptomes identified during the early stages of human intestinal development provided novel insight into the processes of regionalization during early intestinal development[131,132]. There are numerous examples of novel findings based on single-cell RNA-seq data sets derived from healthy adult gut samples. For example, an assessment of epithelial cells from the ileum, colon and rectum revealed a high degree of functional diversity between the small and large intestines as reflected by different nutrient absorption preferences[105]. The proposed existence of Paneth-like cells in the large intestine based on a cluster of colorectal epithelial cells co-expressing *LYZ*, *CA4*, *CA7* and *SPIB* was later attributed to a new absorptive epithelial cell type expressing *BEST4*$^+$ (refs. 25,95,111). In 2022, *BEST4*$^+$ epithelial cells were also reported to vary in abundance and transcriptional signature across different regions of the gut[104]. Together, these findings illustrate the major benefit of combining and comparing data sets from different studies, ultimately reaching reliable insight into healthy gut physiology and cellular function. Besides overcoming the challenges inherent to scarcity of material, data sets generated by single-cell RNA-seq have aided the characterization of the cellular diversity and transcriptional signatures of rare cell types. For instance, the characteristics of the human enteric nervous system remained elusive until a few years ago. In 2020, by employing MIRACL-seq (a novel method designed to enrich samples for rare cell types), 1,445 enteric neurons were recovered from the human colon and found to cluster into 14 subsets based on their transcriptional signatures[33]. Interstitial cells of Cajal (ICCs) are rare entities critical for gastrointestinal peristalsis through both the generation of slow-wave pacemaker activity to smooth muscle cells and the mediation of neurotransmission from enteric neurons. By applying the same strategy, transcriptional signatures of 1,103 ICCs from the human colon were generated[33]. Another single-cell data base, generated by immunophenotyping and fluorescence-activated cell sorting, enriched ICCs from gastric resections and provided a comprehensive characterization of pathways and channels participating in their pacemaker activity[99]. Chemosensory cells, such as tuft cells and enteroendocrine cells (EECs), are rare intestinal epithelial cells operating as an interface for signal transduction between the intestinal lumen and the body, relaying diet-derived and microbiota-derived signals through the release of numerous peptide hormones, neurotransmitters and cytokines[133]. In addition, tuft cells are essential for mounting T helper 2 immune responses against parasites[134]. Based on single-cell studies, tuft cells were demonstrated to interact with the innate and adaptive immune systems through previously unreported receptors[104], including immunoglobulin G receptors[25]. EECs sense intestinal content and release hormones to regulate gastrointestinal activity, systemic metabolism and food intake[133]. By using an organoid-based platform wherein EEC differentiation was induced by transient expression of *NEUROG3* and hormones were tagged with gene reporters, the researchers generated a comprehensive data set of EEC subtypes derived from the small intestine and colon[128].

Intestinal diseases and gut organoids

Generating a complete HGCA in health provides unique opportunities to study the pathogenesis of related diseases. Furthermore, single-cell transcriptional profiles of primary tissue samples can be utilized as a valuable reference map allowing validation of existing and future organoid models. It is, therefore, of critical importance to take steps towards ensuring that data sets generated from related samples and patient cohorts can be integrated into the HGCA (Fig. 6). Amongst the related gut diseases that have been investigated using single-cell profiling approaches are colorectal cancer (CRC)[33,105,114,116–125], inflammatory bowel diseases (IBD)[135], Crohn's disease[18,106,136–138], ulcerative colitis[82,112,115,139,140] and coeliac disease[101].

Examples of major findings in IBD include the identification of distinct immune cell signatures in ulcerative colitis and Crohn's disease[141], a pathogenic cellular module associated with resistance to anti-TNF therapy[137], inference of genetic risk genes to single-cell function[82] and the reactivation of fetal intestinal epithelial transcriptional profiles in childhood-onset Crohn's disease[18]. Similarly, the application of single-cell molecular profiling methods to colonic tissue obtained from patients with CRC has led to major advances in our understanding of disease pathogenesis. Specifically, current single-cell insights into the stem and metaplastic origins of human pre-cancers[117] have led to the reclassification of the consensus of molecular subtypes of CRC by their intrinsic features[126]. The transition of benign lesions into malignancy is accompanied by tumour cell acquisition of stem characteristics[117,125] and reorganization of the microenvironment into suppressive immune-stromal hubs that can potentially be therapeutically targeted[116,142]. Although a comprehensive summary of all relevant available single-cell studies in IBD and CRC is beyond the scope of this manuscript, it is important to highlight that the ability to integrate and compare studies performed on disease tissues at different stages of progression is of major benefit. Hence, ensuring compatibility with the HGCA remains a key priority for every study. Although integration of common gut-related conditions for which extensive data sets are already available will be prioritized in the first phase of data integration, it is important to emphasize the major value of investigating

Table 2 | Summary of existing studies that have used single-cell profiling methods for human intestinal tissue samples obtained from healthy individuals, patients with colorectal cancer or inflammatory bowel diseases, and patient-derived intestinal organoids

Study	Donors and patients	Gut segment sampled	Tissue/sample type	Number of cells and main/key cell types
Williams et al., *Cell*, 2021 (ref. 92)	Healthy individuals	Oral cavity	Biopsy samples of the buccal and gingival mucosa	88,000 cells; epithelial and endothelial cells, fibroblasts, and immune lineage cells
Zhao et al., *J. Cell. Biochem.*, 2022 (ref. 93)	Healthy individuals	Oral cavity	Buccal mucosa	26,398 cells; fibroblasts, immune and endothelial cells, melanocytes, myofibroblasts, epithelial and neuron-like cells
Caetano et al., *eLife*, 2021 (ref. 94)	Healthy individuals and patients with periodontitis	Oral cavity	Resections of gingival tissue from healthy individuals or patients with periodontitis	12,411 cells; epithelial, stromal, immune, endothelial and perivascular cells
Busslinger et al., *Cell Rep.*, 2021 (ref. 95)	Healthy individuals	Oesophagus, stomach, duodenum	Mucosal biopsy samples	4,581 cells; oesophageal squamous, gastric glandular and duodenal crypt, and villus epithelial cells
Nowicki-Osuch et al., *Science*, 2021 (ref. 96)	Deceased organ donors and patients with Barrett oesophagus	Oesophagus, stomach, duodenum	Mucosal biopsy samples and surgical resection material	43,000 cells; epithelial cells, including squamous basal and superficial cells, gastric foveolar, endocrine, parietal and chief cells
Zhang et al., *Gut*, 2021 (ref. 97)	Patients diagnosed with gastric adenocarcinoma and healthy controls	Stomach	Mucosal biopsy samples	27,667 cells; epithelial, stromal and immune cells
Sorini et al., *JCI Insight*, 2023 (ref. 98)	Patients undergoing gastric resection for obesity (with and without *Helicobacter pylori* infection)	Stomach	Surgical resection	22,033 cells; T cells, B cells, ILCs and myeloid cells
Foong et al., *Neurogastroenterol. Motil.*, 2022 (ref. 99)	Patients undergoing gastric resection for obesity	Stomach	Surgical resection	424 cells; intestinal cells of Cajal and haematopoietic cells
Kumar et al., *Cancer Discov.*, 2022 (ref. 100)	Patients with gastric adenocarcinoma	Stomach	Mucosal biopsy samples and surgical resection	152,423 cells; epithelial, stromal and immune cells
Atlasy et al., *Nat. Comm.*, 2022 (ref. 101)	Patients with coeliac disease and healthy controls	Small intestine (duodenum)	Mucosal biopsy samples	6,248 cells; immune cells, including B and T cells, macrophages, and dendritic cells
Egozi et al., *bioRxiv*, 2022 (ref. 102)	Infants with necrotizing enterocolitis and controls	Small intestine	Surgical resection	11,308 cells; myeloid, B cells, T and NK cells, lymphatic and blood endothelial cells, fibroblasts, and enterocyte populations
Domínguez Conde et al., *Science*, 2022 (ref. 103)	Deceased donors	Small and large intestine (also lung, spleen, bone marrow and lymphoid tissue)	Resection	360,000 cells, mucosal immune cells
Burclaff et al., *Cell. Mol. Gastroenterol. Hepatol.*, 2022 (ref. 104)	Transplant donors	Small intestine (duodenum, jejunum, ileum), large intestine (ascending, transverse and descending colon)	Surgical resection	428,000 cells; immune, epithelial, mesenchymal, endothelial, neural and red blood cells
Elmentaite et al., *Nature*, 2021 (ref. 25)	Deceased donors, healthy individuals, patients with IBD, human fetal gut	Small intestine (duodenum, jejunum, ileum), appendix, large intestine (caecum, ascending colon, transverse colon, descending colon, sigmoid colon, rectum)	Surgical resection and mucosal biopsy samples	428,000 cells; immune, epithelial, mesenchymal, endothelial, neural and red blood cells
Beumer et al., *Cell*, 2020 (ref. 128)	Patients with colorectal cancer, healthy individuals	Small intestine (duodenum, ileum), large intestine (ascending colon)	Intestinal organoids generated from surgical resection or mucosal biopsy samples	4,281 cells; enteroendocrine cells, enterocytes, stem, goblet, and Paneth cells

Table 2 (continued) | Summary of existing studies that have used single-cell profiling methods for human intestinal tissue samples obtained from healthy individuals, patients with colorectal cancer or inflammatory bowel diseases, and patient-derived intestinal organoids

Study	Donors and patients	Gut segment sampled	Tissue/sample type	Number of cells and main/key cell types
Wang et al., *J. Exp. Med.*, 2020 (ref. 105)	Patients with intestinal tumours	Small intestine (ileum), large bowel (colon and rectum)	Mucosal resection (sampled 10 cm away from the tumour)	14,537 cells; enterocytes, goblet cells, Paneth cells, enteroendocrine cells, progenitor cells, transit amplifying and stem cells
Huang et al., *Cell*, 2019 (ref. 109)	Paediatric patients with Crohn's disease, ulcerative colitis and healthy controls	Appendix, large intestine (caecum, ascending colon, transverse colon, descending colon, sigmoid colon, rectum)	Mucosal biopsy samples	73,165 cells; epithelial cells, stromal cells and immune cells, including myeloid cells, B cell subsets, plasma cells, and T and NK cells
Jaeger et al., *Nat. Comm.*, 2021 (ref. 106)	Patients with Crohn's disease and non-IBD controls (bowel cancer)	Small intestine (terminal ileum)	Surgical resection	15,731 intraepithelial lymphocytes and 29,247 lamina propria cells; T cells
Elmentaite et al., *Dev. Cell*, 2020 (ref. 18)	Paediatric patients with Crohn's disease and healthy controls, human fetal gut	Small intestine (terminal ileum), colon (fetal colon)	Mucosal biopsy samples (paediatric patients), resection (fetal gut)	62,854 (fetal) and 11,302 (paediatric) cells; immune, epithelial, mesenchymal, endothelial, neural and red blood cells
Drokhlyansky et al., *Cell*, 2020 (ref. 33)	Patients with colorectal cancer	Large intestine (colon)	Surgical resection	436,202 cells; adipose tissue, epithelial, glial, mesenchymal, and endothelial cells, T cells, fibroblasts, macrophages, myocytes, and neuronal subsets from muscularis propria
Kinchen et al., *Cell*, 2018 (ref. 110)	Patients with ulcerative colitis and healthy controls	Large bowel (colon)	Mucosal biopsy samples	4,378 cells; stromal cell subsets in healthy human colon and ulcerative colitis colon
Parikh et al., *Nature*, 2019 (ref. 111)	Patients with ulcerative colitis and healthy controls	Large bowel (colon)	Mucosal biopsy samples	11,175 cells; intestinal epithelial cells, including progenitor cells, colonocytes and goblet cells
Corridoni et al., *Nat. Med.*, 2020 (ref. 112)	Patients diagnosed with ulcerative colitis and healthy controls	Large bowel (colon)	Mucosal biopsy samples	8,581 cells; $CD8^+$ T cells
James et al., *Nat. Immunol.*, 2020 (ref. 113)	Transplant donors	Large bowel (caecum, transverse colon, sigmoid colon and mesenteric lymph nodes)	Surgical resection	40,108 cells; B cells, T cells, ILCs and myeloid cells; includes microbiota data
Kong et al., *Immunity*, 2023 (ref. 107)	Patients with Crohn's disease and healthy controls	Small bowel (terminal ileum), large bowel (sigmoid colon)	Mucosal biopsy samples	720,633 cells; intestinal, epithelium, myofibroblasts, immune and stromal cells
Smillie et al., *Cell*, 2019 (ref. 82)	Patients with ulcerative colitis and healthy controls	Large bowel (colon)	Mucosal biopsy samples	107,784 cells; epithelial, stromal and immune cells
Kondo et al., *Gastroenterology* 2021 (ref. 108)	Patients with ulcerative colitis, Crohn's disease and healthy controls	Small bowel (terminal ileum), large bowel (colon)	Mucosal biopsy samples	>300,000 cells; epithelial, stromal, T cells, B cells, plasma, macrophage and myeloid cells from 31 protein markers
Lee et al., *Nat. Genet.*, 2020 (ref. 114)	Patients with colorectal cancer	Large bowel (colon)	Surgical resection	93,333 cells; goblet cells, stem-like transit amplifying cells, colonocytes, fibroblasts, glia cells, endothelial cells, macrophages, dendritic cells, T and NK cells, plasma cells and B cells
Uzzan et al., *Nat. Med.*, 2022 (ref. 115)	Patients with ulcerative colitis and healthy controls	Large bowel (left colon)	Mucosal biopsy samples	18,720 cells; T cells, B cells, macrophages, plasma cells and stromal cells
Pelka et al., *Cell*, 2021 (ref. 116)	Patients with colorectal cancer	Large bowel (colon)	Surgical resection	371,223 cells; immune, endothelial, fibroblasts and epithelial cells
Chen et al., *Cell*, 2021 (ref. 117)	Patients diagnosed with colorectal cancer, intestinal polyps and healthy controls	Large bowel (colon)	Surgical resection, intestinal polyps	142,065 cells; intestinal epithelium, and immune cells and fibroblasts
Li et al., *Nat. Genet.*, 2017 (ref. 118)	Patients with colorectal cancer	Large bowel (colon)	Surgical resection	622 cells; epithelial, fibroblasts, endothelial, mast and immune cells

Table 2 (continued) | Summary of existing studies that have used single-cell profiling methods for human intestinal tissue samples obtained from healthy individuals, patients with colorectal cancer or inflammatory bowel diseases, and patient-derived intestinal organoids

Study	Donors and patients	Gut segment sampled	Tissue/sample type	Number of cells and main/key cell types
Qian et al., *Cell Res.*, 2020 (ref. 119)	Patients with colorectal cancer	Large bowel (colon)	Surgical resection	44,685 cells; T cells, B cells, macrophages, dendritic cells, fibroblasts, endothelial cells, epithelial cells and enteric glia
Zhang et al., *Cell*, 2020 (ref. 120)	Patients with colorectal cancer	Large bowel (colon)	Surgical resection	43,817 cells (10X Genomics), 10,468 cells (Smart-seq2); T cells, B cells, ILCs, macrophages, plasma cells, fibroblasts and endothelial cells
Domanska et al., *J. Exp. Med.*, 2022 (ref. 121)	Patients with colorectal cancer	Large bowel (colon)	Surgical resection	63,917 cells; colonic macrophages
Uhlitz et al., *EMBO Mol. Med.*, 2021 (ref. 122)	Patients with colorectal cancer	Large bowel (colon)	Surgical resection	>30,000 cells; epithelial, stromal and immune cells
Zhang et al., *Nature*, 2018 (ref. 123)	Patients with colorectal cancer	Large bowel (colon)	Surgical resection	8,530 cells; T cell subsets
Che et al., *Cell Discov.*, 2021 (ref. 124)	Patients with colorectal cancer	Large bowel (colon), also liver metastases and blood	Surgical resection	111,292 cells; T and NK cells, B and plasma cells, cancer-associated fibroblasts, endothelial and myeloid cells
Becker et al., *Nat. Genet.*, 2022 (ref. 125)	Patients with familial adenomatous polyposis, colon precancer and colon cancer	Large bowel (colon)	Surgical resection	201,884 (snRNA-seq), 447,829 (scATAC-seq) cells; immune, endothelial, fibroblasts, glia and epithelial cells
Joanito et al., *Nat. Genet.*, 2022 (ref. 126)	Patients with colorectal cancer	Large bowel (colon)	Surgical resection	373,058, focus on epithelial cells (49,155 cells)
Luoma et al., *Cell*, 2020 (ref. 127)	Patients diagnosed with colorectal cancer and healthy controls	Large bowel (descending colon, sigmoid colon, rectum)	Mucosal biopsy samples	51,652 cells; $CD45^+$ mononuclear cells, $CD3^+$ T cells
Holloway et al., *Cell Stem Cell*, 2021 (ref. 130)	Human fetal samples	Intestine, liver and kidney	Resection	24,783 cells; mesenchymal, epithelium, endothelium and neuronal cells
Fawkner-Corbett et al., *Cell*, 2021 (ref. 149)	Human fetal samples	Small and large bowel	Resection	76,592 cells; epithelial, fibroblast, endothelial, pericytes, neural, muscularis, mesothelium, myofibroblast and immune cells
Yu et al., *Cell*, 2021 (ref. 131)	Human fetal samples	Oesophagus, stomach, small and large bowel, liver	Resection	155,232 cells; epithelial, mesenchymal, immune, endothelial and neuronal cells, and erythroid lineages
Cao et al., *Science*, 2020 (ref. 150)	Human fetal samples	Stomach, small and large bowel (also adrenal gland, cerebellum, cerebrum, eye, heart, intestine, kidney, liver, lung, muscle, pancreas, placenta, spleen, thymus and sentinel tissue)	Resection	4,062,980 cells; stromal, endothelial, myeloid, lymphoid, epithelial, mesothelial, chromaffin, erythroblasts, ENS neurons and ENS glial cells
Gao et al., *Nat. Cell Biol.*, 2018 (ref. 132)	Human fetal samples	Oesophagus, stomach, small and large bowel	Resection	5,227 cells; epithelium, mesenchymal, endothelial, immune and neuronal cells
Li et al., *Nat. Immunol.*, 2019 (ref. 151)	Human fetal samples	Small and large intestine	Resection	1,804 cells; T cell subsets
He et al., *Cell Stem Cell*, 2022 (ref. 129)	Intestinal organoids	Small bowel	Mucosal biopsy-derived organoids	1,283 cells; intestinal epithelium
Ishikawa et al., *Gastroeneterology*, 2022 (ref. 147)	Intestinal organoids	Large intestine	Resection derived organoids	7,103 cells; intestinal epithelium

ENS, enteric nervous system; IBD, inflammatory bowel disease; ILCs, innate lymphoid cells; NK, natural killer; scATAC-seq, single-cell sequencing assay for transposase-accessible chromatin; snRNA-seq, single-nucleus RNA sequencing.

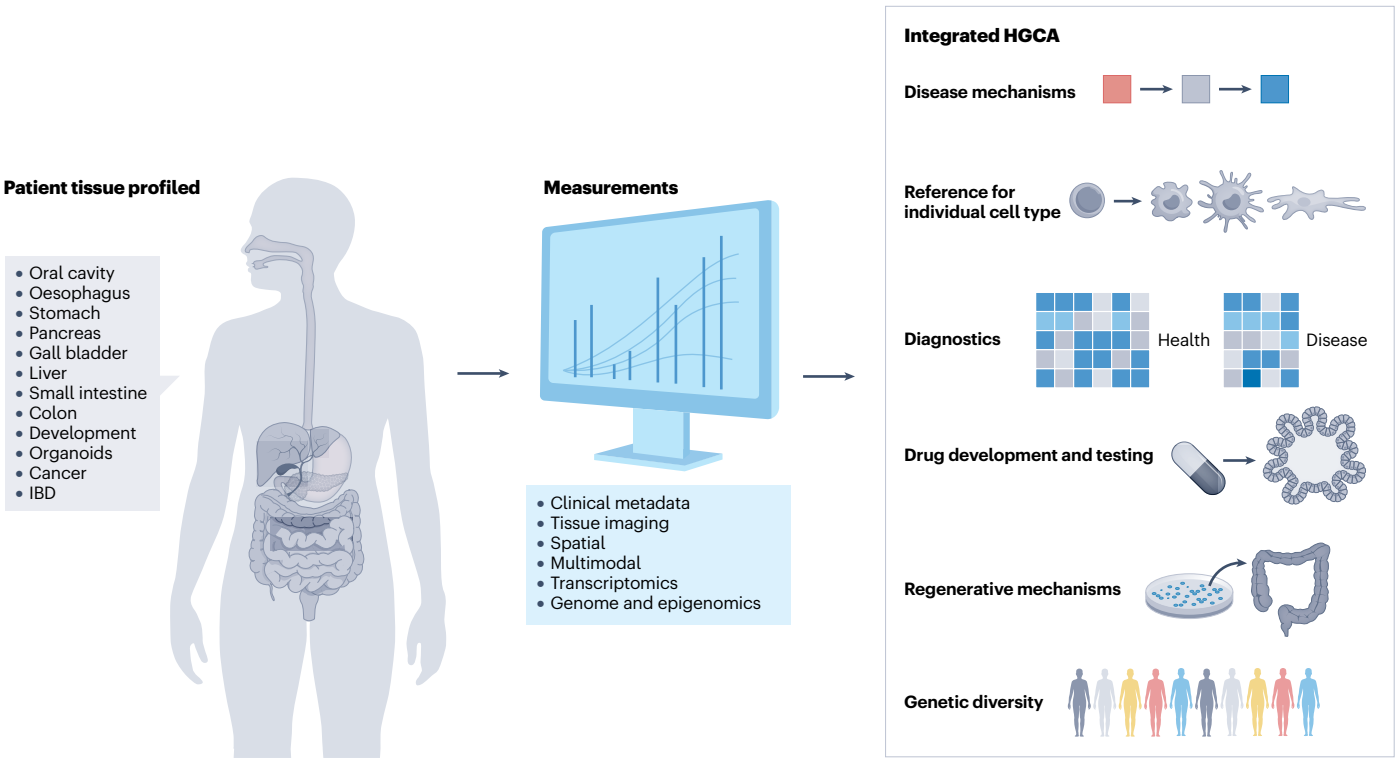

Fig. 6 | HGCA in health and disease. Integration of data sets generated from intestinal tissues obtained from healthy individuals and patients with related diseases, including inflammatory bowel disease (IBD) and colorectal cancer, will enhance the future value of the Human Gut Cell Atlas (HGCA). Single-cell profiling of intestinal organoids, followed by their integration into the HGCA, will provide unique opportunities for translational research, including regenerative medicine, drug testing and development.

and ultimately integrating rarer conditions, for example, tufting enteropathy[143], early-onset and monogenic forms of IBD[144] as well as a congenital disease affecting the enteric nervous system such as Hirschsprung disease[145]. Indeed, combining data sets from less commonly profiled conditions represents another unique opportunity for the HGCA by increasing computational power and allowing validation of key findings. Furthermore, applying a variety of methodologies (for example, single-cell multiomics profiling) to the same condition will further improve the value of the HGCA and the chances of gaining novel insight into disease pathogenesis.

The development of human intestinal organoid culture models has transformed many aspects of gut-related research providing researchers with unprecedented opportunities to study fundamental aspects of intestinal biology[146]. Performing transcriptional profiling of patient-derived intestinal organoids at the single-cell level is of great value as reflected in the increasing number of studies reporting novel findings[128]. The main benefits include the ability to validate the cellular composition of organoids to further improve culture conditions and evaluate to what extent disease-associated cellular alterations are retained in vitro. An example is a study published by He et al. in 2022, wherein the researchers applied single-cell RNA-seq to human small intestinal organoids[129]. Exposure of organoids to IL-22 resulted in increased expression of antimicrobial peptides suggesting the presence of Paneth cells[129]. Furthermore, single-cell profiling of primary human fetal gut and organoids derived from the same tissue sample revealed in vitro maturation of fetal organoids, highlighting their value in investigating the early stages of human intestinal epithelial cell development[18]. Work by Ishikawa et al. combined single-cell RNA-seq of human colonic epithelium cells with genetically modified human intestinal organoids, leading to the identification of quiescent LGR5$^+$ stem cells in the human colon[147]. Similarly, combining cancer organoid assays and single-cell RNA-seq of biopsy samples from patients with CRC identified the novel role of a β-hydroxybutyrate-triggered pathway in regulating intestinal tumorigenesis[148]. As for studies based on profiling primary tissue, the provision of metadata, including clinical donor details, sampling sites, biobanking and experimental procedures, is equally important to organoid-related studies as they will determine their future compatibility and ensure successful integration into the HGCA.

Existing data portals

In an effort to consolidate the growing number of single-cell data sets that have been generated in the past few years, several web-based data repositories have emerged to provide access in a user-friendly format. Although most portals share common features of storing data, curating data sets by different levels of metadata, and offering a variety of analysis and visualization tools, distinct aspects render each portal valuable. Here, we summarize key features of the main existing data portals relevant to the HGCA.

Single-Cell Expression Atlas. The Single-Cell Expression Atlas (SCEA)[91] is the single-cell component of the EMBL-EBI expression atlas.

This is an added-value resource that enables simple gene and metadata queries, allowing users to answer questions such as 'Where is my gene of interest expressed?' or 'How does its expression level change in a disease?'. SCEA collects expression data from all species and annotates their metadata with appropriate search (ontology) terms to enable standardized analysis pipelines across studies. The SCEA works in close collaboration with the HGCA, releasing the data sets as they become publicly available and generating the first full gut two-dimensional anatomogram that will enable easy visual exploration of gene and marker gene expression across the cell types in the different anatomical sub-structures of the gut.

HubMAP. The HuBMAP Portal is a National Institutes of Health Common Fund effort to integrate and map diverse biological data across the healthy human body. Three main features enable (1) analysis of single-cell RNA-seq experiments with Azimuth, a web application that uses reference data sets to automate annotation and interpretation of data, (2) spatial single-cell data visualization with Vitessce, and (3) navigation of healthy human cells with the CCF to interact with the virtual human body and focus on anatomical structures, cell types, and biomarkers.

Tabula Sapiens. Like HuBMAP, Tabula Sapiens focuses on healthy human participants and has created a first draft HCA of 24 organs from 15 different human subjects. Tabula Sapiens is funded by the Chan Zuckerberg Initiative and is unique in that the single-cell data sets derived for each organ are from the same human participant, controlling for inter-individual factors. This also enables the comparison of cell types that are shared between different organs. The web portal offers the ability to peruse all the single-cell data combined (currently at 500,000 cells) or curated cell subsets (that is, endothelial, epithelial, immune and stromal) in an easy-to-navigate graphical user interface.

University of California at Santa Cruz Cell Browser. The University of California at Santa Cruz (UCSC) Cell Browser is a sub-project under the supervision of the UCSC Genome Browser project that was developed and is maintained by a cross-departmental team in the UCSC Genomics Institute. The Cell Browser is an interactive viewer where users can interrogate single-cell data sets from a wide variety of species, organs and tissues from a menu list. A unique feature is the broad scientific focus of the data sets generated from human, mouse, fly, and sponge and curated as conventional Atlases or analysed in the context of development and evolution. The data sets are converted for compatibility by the UCSC Cell Browser Group and placed in an open-source portal. The analysis and viewer package can also be downloaded and installed locally.

Broad Institute Single Cell Portal. The Single Cell Portal is hosted by the Broad Institute and was created with the idea of making single-cell data easy to share and access. The portal is a deposit site where investigators can create their own collection as a body of work or contribute to the growing list of curated data sets. The site is easy to use and has succinct overviews describing the study design and experimental conditions. Like other portals, the Single Cell Portal has conventional visualization tools that are interactive in a simple graphical user interface that allows the user to filter several parameters in the data sets and sub-sampling data.

Gut Cell Atlas, Wellcome Sanger Institute. Gut Cell Atlas provides a detailed gut cell survey by combining single-cell data generated from a range of human gut tissues. Data sets include studies that profiled various gut segments obtained from fetal, paediatric and adult donors as well as patients diagnosed with Crohn's disease. In addition to direct access to raw data sets, the site provides an interactive viewer that enables basic analyses and interrogation of data sets such as exploring single-cell expression profiles of individual cell lineages, gut regions or comparisons between age groups.

Future opportunities, including incorporating data sets

In many aspects, the current collection of high-quality published single-cell studies achieve good coverage of the human intestinal tract and therefore provide a solid foundation for the generation of the HGCA. Areas that have been explored using single-cell RNA-seq technology include intestinal development[18,132,149–151], profiling of cell types, states and tissue composition[128], mapping regional differences across intestinal tissues and along the intestinal tube[113,128], and comparisons between healthy individuals and people with IBD[104,110], CRC[122,123], or other related diseases[96,125]. However, there are several less well-explored areas that require careful consideration in future studies. For example, to achieve a comprehensive and truly global HGCA, greater effort must be made to include samples obtained from under-represented and ethnically diverse communities to reflect potential cellular differences across human populations. Currently, only a few studies have sampled multiple intestinal regions within the same individual[25,107]. However, such cross-tissue studies are of major value as they enable the identification of common and/or distinct cell types across intestinal regions as well as studying migratory patterns of specific immune cell populations[25]. Furthermore, greater emphasis should be placed on profiling rare cell types, particularly in less frequently sampled gut regions such as the jejunum and proximal ileum. This objective can be achieved by applying methods designed to enrich such cell types prior to single-cell RNA-seq and/or the application of spatial transcriptomics. Additionally, analysing single nuclei rather than single cells is a useful strategy to profile cells that cannot be readily recovered using standard dissociation protocols[33]. However, the advantages and disadvantages of these methods must be carefully considered in the context of each particular research study[152,153]. Finally, there has been a major bias regarding the anatomical sampling sites, with a large proportion of studies focusing on profiling the colon and distal small bowel. This is likely to be partly caused by the ease of access during endoscopic procedures as well as the major relevance to related gastrointestinal diseases such as IBD and CRC. Future studies should also include less frequently sampled gut segments such as the mid-small intestine, oesophagus, stomach and other organs involved in the digestive process, including the liver and pancreas. Last but not least, profiling the intestine at different developmental stages by obtaining tissue from donors of all age groups will provide further insight into the pathophysiology of related diseases, including those occurring in specific age groups.

Conclusions

Generating a complete map of the human intestine at the single-cell level will improve our understanding of gut health and disease. The inherent complexities and scale of this task require a coordinated effort led by experts in this rapidly evolving field. This manuscript forms a key part of our strategy by providing a detailed Roadmap to the scientific community. Broad distribution and constructive discussion of our proposal are essential to achieving our goal of the generation of an HGCA.

Published online: 31 May 2023

References

1. Rooks, M. G. & Garrett, W. S. Gut microbiota, metabolites and host immunity. *Nat. Rev. Immunol.* **16**, 341–352 (2016).
2. Kolodziejczyk, A. A., Zheng, D. & Elinav, E. Diet-microbiota interactions and personalized nutrition. *Nat. Rev. Microbiol.* **17**, 742–753 (2019).
3. Knight, R. et al. Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* **16**, 410–422 (2018).
4. Makki, K., Deehan, E. C., Walter, J. & Backhed, F. The impact of dietary fiber on gut microbiota in host health and disease. *Cell Host Microbe* **23**, 705–715 (2018).
5. Khalili, H. et al. The role of diet in the aetiopathogenesis of inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol.* **15**, 525–535 (2018).
6. Safi-Stibler, S. & Gabory, A. Epigenetics and the developmental origins of health and disease: parental environment signalling to the epigenome, critical time windows and sculpting the adult phenotype. *Semin. Cell Dev. Biol.* **97**, 172–180 (2020).
7. Perrone, F. & Zilbauer, M. Biobanking of human gut organoids for translational research. *Exp. Mol. Med.* **53**, 1451–1458 (2021).
8. Enns, R. A. et al. Clinical practice guidelines for the use of video capsule endoscopy. *Gastroenterology* **152**, 497–514 (2017).
9. Smillie, C. S. et al. Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* **178**, 714–730.e22 (2019).
10. Pilonis, N. D., Januszewicz, W. & di Pietro, M. Confocal laser endomicroscopy in gastrointestinal endoscopy: technical aspects and clinical applications. *Transl. Gastroenterol. Hepatol.* **7**, 7 (2022).
11. Chen, W. C. & Wallace, M. B. Endoscopic management of mucosal lesions in the gastrointestinal tract. *Expert. Rev. Gastroenterol. Hepatol.* **10**, 481–495 (2016).
12. Pensold, D. & Zimmer-Bensch, G. Methods for single-cell isolation and preparation. *Adv. Exp. Med. Biol.* **1255**, 7–27 (2020).
13. Regev, A. et al. The human cell atlas. *eLife* **6**, e27041 (2017).
14. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The human cell atlas: from vision to reality. *Nature* **550**, 451–453 (2017).
15. Skinnider, M. A., Squair, J. W. & Courtine, G. Enabling reproducible re-analysis of single-cell data. *Genome Biol.* **22**, 215 (2021).
16. Füllgrabe, A. et al. Guidelines for reporting single-cell RNA-seq experiments. *Nat. Biotechnol.* **38**, 1384–1386 (2020).
17. Burger, A. et al. Towards a clinically-based common coordinate framework for the human gut cell atlas: the gut models. *BMC Med. Inform. Decis. Mak.* **23**, 36 (2023).
18. Elmentaite, R. et al. Single-cell sequencing of developing human gut reveals transcriptional links to childhood Crohn's disease. *Dev. Cell* **55**, 771–783.e5 (2020).
19. Lee, S. S. J. The ethics of consent in a shifting genomic ecosystem. *Annu. Rev. Biomed. Data Sci.* **4**, 145–164 (2021).
20. EUR-Lex, The European Parliament and the Council of the European Union. *General Data Protection Regulation (EU) 2016/679 (GDPR)* https://eur-lex.europa.eu/eli/reg/2016/679/oj (2016).
21. Bledsoe, M. J. & Grizzle, W. E. Use of human specimens in research: the evolving United States regulatory, policy, and scientific landscape. *Diagn. Histopathol.* **19**, 322–330 (2013).
22. Shore, N. et al. Understanding community-based processes for research ethics review: a national study. *Am. J. Public Health* **101**, S359–S364 (2011).
23. Editorial. Tackling helicopter research. *Nat. Geosci.* **15**, 597 (2022).
24. Mikesell, L., Bromley, E. & Khodyakov, D. Ethical community-engaged research: a literature review. *Am. J. Public Health* **103**, e7–e14 (2013).
25. Elmentaite, R. et al. Cells of the human intestinal tract mapped across space and time. *Nature* **597**, 250–255 (2021).
26. Camp, J. G., Platt, R. & Treutlein, B. Mapping human cell phenotypes to genotypes with single-cell genomics. *Science* **365**, 1401–1405 (2019).
27. Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
28. Lareau, C. A. et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).
29. Smallwood, S. A. et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).
30. Hao, Y. H. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–357 (2021).
31. Pott, S. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *eLife* https://doi.org/10.7554/eLife.23203 (2017).
32. Clark, S. J. et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).
33. Drokhlyansky, E. et al. The human and mouse enteric nervous system at single-cell resolution. *Cell* **182**, 1606–1622.e23 (2020).
34. Setliff, I. et al. High-throughput mapping of B cell receptor sequences to antigen specificity. *Cell* **179**, 1636–1646.e15 (2019).
35. Slyper, M. et al. A single-cell and single-nucleus RNA-seq toolbox for fresh and frozen human tumors. *Nat. Med.* **26**, 792–802 (2020).
36. Wu, S. Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).
37. Xia, C., Babcock, H. P., Moffitt, J. R. & Zhuang, X. Multiplexed detection of RNA using MERFISH and branched DNA amplification. *Sci. Rep.* **9**, 7721 (2019).
38. Black, S. et al. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nat. Protoc.* **16**, 3802–3835 (2021).
39. Haniffa, M. et al. A roadmap for the human developmental cell atlas. *Nature* **597**, 196–205 (2021).
40. Travaglini, K. J. et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
41. Zou, Z. et al. A single-cell transcriptomic atlas of human skin aging. *Dev. Cell* **56**, 383–397.e8 (2021).
42. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 1–14 (2018).
43. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.* **14**, e1006245 (2018).
44. Moreno, P. et al. User-friendly, scalable tools and workflows for single-cell RNA-seq analysis. *Nat. Methods* **18**, 327–328 (2021).
45. Wilbrey-Clark, A., Roberts, K. & Teichmann, S. A. Cell Atlas technologies and insights into tissue architecture. *Biochem. J.* **477**, 1427–1442 (2020).
46. Ke, M., Elshenawy, B., Sheldon, H., Arora, A. & Buffa, F. M. Single cell RNA-sequencing: a powerful yet still challenging technology to study cellular heterogeneity. *Bioessays* **44**, e2200084 (2022).
47. Su, M. et al. Data analysis guidelines for single-cell RNA-seq in biomedical studies and clinical applications. *Mil. Med. Res.* **9**, 68 (2022).
48. Tran, H. T. N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
49. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
50. Liu, J. et al. Jointly defining cell types from multiple single-cell datasets using LIGER. *Nat. Protoc.* **15**, 3632–3662 (2020).
51. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
52. Ryu, Y., Han, G. H., Jung, E. & Hwang, D. Integration of single-cell RNA-seq datasets: a review of computational methods. *Mol. Cell* **46**, 106–119 (2023).
53. Cao, Z. J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-022-01284-4 (2022).
54. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
55. Gong, B., Zhou, Y. & Purdom, E. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome Biol.* **22**, 351 (2021).
56. Tabula Sapiens Consortium et al. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
57. Xu, Y., Baumgart, S. J., Stegmann, C. M. & Hayat, S. MACA: marker-based automatic cell-type annotation for single cell expression data. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btab840 (2021).
58. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
59. Ianevski, A., Giri, A. K. & Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat. Commun.* **13**, 1246 (2022).
60. Delaney, C. et al. Combinatorial prediction of marker panels from single-cell transcriptomic data. *Mol. Syst. Biol.* **15**, e9005 (2019).
61. Dai, M., Pei, X. & Wang, X.-J. Accurate and fast cell marker gene identification with COSG. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbab579 (2022).
62. Andersson, A. et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun. Biol.* **3**, 565 (2020).
63. David, F. P. A., Litovchenko, M., Deplancke, B. & Gardeux, V. ASAP 2020 update: an open, scalable and interactive web-based portal for (single-cell) omics analyses. *Nucleic Acids Res.* **48**, W403–W414 (2020).
64. Moriel, N. et al. NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nat. Protoc.* **16**, 4177–4200 (2021).
65. Kleshchevnikov, V. et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-01139-4 (2022).
66. Longo, S. K., Guo, M. G., Ji, A. L. & Khavari, P. A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat. Rev. Genet.* **22**, 627–644 (2021).
67. Dries, R. et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22**, 78 (2021).
68. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).
69. Dries, R. et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22**, 78 (2021).
70. Zhu, Q., Shah, S., Dries, R., Cai, L. & Yuan, G.-C. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nat. Biotechnol.* **36**, 1183–1190 (2018).
71. Goltsev, Y. et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* **174**, 968–981.e15 (2018).
72. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell-cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22**, 71–88 (2021).
73. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).

74. Kang, R., Park, B., Eady, M., Ouyang, Q. & Chen, K. J. Single-cell classification of foodborne pathogens using hyperspectral microscope imaging coupled with deep learning frameworks. *Sens. Actuators B Chem.* **309**, 127789 (2020).
75. Chattopadhyay, P. K., Roederer, M. & Bolton, D. L. A deadly dance: the choreography of host-pathogen interactions, as revealed by single-cell technologies. *Nat. Commun.* **9**, 4638 (2018).
76. Liao, C., Wang, T., Maslov, S. & Xavier, J. B. Modeling microbial cross-feeding at intermediate scale portrays community dynamics and species coexistence. *PLoS Comput. Biol.* **16**, e1008135 (2020).
77. Galeano Niño, J. L. et al. Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer. *Nature* **611**, 810–817 (2022).
78. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
79. Moerman, T. et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159–2161 (2019).
80. Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.* **5**, 251–267.e3 (2017).
81. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **17**, 147–154 (2020).
82. Smillie, C. S. et al. Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* **178**, 714–71 (2019).
83. Cakir, B. et al. Comparison of visualization tools for single-cell RNAseq data. *NAR Genom. Bioinform.* **2**, lqaa052 (2020).
84. Megill, C. et al. chanzuckerberg/cellxgene: Release 0.15.0. https://doi.org/10.5281/ZENODO.3710410 (2020).
85. Elmentaite, R. et al. Single-cell sequencing of developing human gut reveals transcriptional links to childhood Crohn's disease. *Dev. Cell* **55**, 771–783.e5 (2020).
86. Moreno, P. et al. Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Res.* **50**, D129–D140 (2021).
87. Rood, J. E. et al. Toward a common coordinate framework for the human body. *Cell* **179**, 1455–1467 (2019).
88. Schiller, H. B. et al. The human lung cell atlas: a high-resolution reference map of the human lung in health and disease. *Am. J. Respir. Cell Mol. Biol.* **61**, 31–41 (2019).
89. Eze, U. C., Bhaduri, A., Haeussler, M., Nowakowski, T. J. & Kriegstein, A. R. Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nat. Neurosci.* **24**, 584–594 (2021).
90. Weber, G. M., Ju, Y. N. & Borner, K. Considerations for using the vasculature as a coordinate system to map all the cells in the human body. *Front. Cardiovasc. Med.* **7**, 29 (2020).
91. Moreno, P. et al. Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Res.* **50**, D129–D140 (2022).
92. Williams, D. W. et al. Human oral mucosa cell atlas reveals a stromal-neutrophil axis regulating tissue immunity. *Cell* **184**, 4090–4104.e15 (2021).
93. Zhao, R. W. et al. Function of normal oral mucosa revealed by single-cell RNA sequencing. *J. Cell. Biochem.* **123**, 1481–1494 (2022).
94. Caetano, A. J. et al. Defining human mesenchymal and epithelial heterogeneity in response to oral inflammatory disease. *eLife* **10**, e62810 (2021).
95. Busslinger, G. A. et al. Human gastrointestinal epithelia of the esophagus, stomach, and duodenum resolved at single-cell resolution. *Cell Rep.* **34**, 108819 (2021).
96. Nowicki-Osuch, K. et al. Molecular phenotyping reveals the identity of Barrett's esophagus and its malignant transition. *Science* **373**, 760–767 (2021).
97. Zhang, M. et al. Dissecting transcriptional heterogeneity in primary gastric adenocarcinoma by single cell RNA sequencing. *Gut* **70**, 464–475 (2021).
98. Sorini, C. et al. Metagenomic and single-cell RNA-Seq survey of the Helicobacter pylori-infected stomach in asymptomatic individuals. *JCI Insight* https://doi.org/10.1172/jci.insight.161042 (2023).
99. Foong, D. et al. Single-cell RNA sequencing predicts motility networks in purified human gastric interstitial cells of Cajal. *Neurogastroenterol. Motil.* **34**, e14303 (2022).
100. Kumar, V. et al. Single-cell atlas of lineage states, tumor microenvironment, and subtype-specific expression programs in gastric cancer. *Cancer Discov.* **12**, 670–691 (2022).
101. Atlasy, N. et al. Single cell transcriptomic analysis of the immune cell compartment in the human small intestine and in Celiac disease. *Nat. Commun.* **13**, 4920 (2022).
102. Egozi, A. et al. Single cell atlas of the neonatal small intestine with necrotizing enterocolitis. *bioRxiv* https://doi.org/10.1101/2022.03.01.482508 (2022).
103. Domínguez Conde, C. et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
104. Burclaff, J. et al. A proximal-to-distal survey of healthy adult human small intestine and colon epithelium by single-cell transcriptomics. *Cell. Mol. Gastroenterol. Hepatol.* **13**, 1554–1589 (2022).
105. Wang, Y. L. et al. Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine. *J. Exp. Med.* https://doi.org/10.1084/jem.20191130 (2020).
106. Jaeger, N. et al. Single-cell analyses of Crohn's disease tissues reveal intestinal intraepithelial T cells heterogeneity and altered subset distributions. *Nat. Commun.* **12**, 1921 (2021).
107. Kong, L. et al. The landscape of immune dysregulation in Crohn's disease revealed through single-cell transcriptomic profiling in the ileum and colon. *Immunity* **56**, 444–458.e5 (2023).
108. Kondo, A. et al. Highly multiplexed image analysis of intestinal tissue sections in patients with inflammatory bowel disease. *Gastroenterology* **161**, 1940–1952 (2021).
109. Huang, B. et al. Mucosal profiling of pediatric-onset colitis and IBD reveals common pathogenics and therapeutic pathways. *Cell* **179**, 1160–1176.e24 (2019).
110. Kinchen, J. et al. Structural remodeling of the human colonic mesenchyme in inflammatory bowel disease. *Cell* **175**, 372–386.e17 (2018).
111. Parikh, K. et al. Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature* **567**, 49–55 (2019).
112. Corridoni, D. et al. Single-cell atlas of colonic $CD8^+$ T cells in ulcerative colitis. *Nat. Med.* **26**, 1480–1490 (2020).
113. James, K. R. et al. Distinct microbial and immune niches of the human colon. *Nat. Immunol.* **21**, 343–353 (2020).
114. Lee, H. O. et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat. Genet.* **52**, 594–603 (2020).
115. Uzzan, M. et al. Ulcerative colitis is characterized by a plasmablast-skewed humoral response associated with disease activity. *Nat. Med.* **28**, 766–779 (2022).
116. Pelka, K. et al. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell* **184**, 4734–4752.e20 (2021).
117. Chen, B. et al. Differential pre-malignant programs and microenvironment chart distinct paths to malignancy in human colorectal polyps. *Cell* **184**, 6262–6280.e26 (2021).
118. Li, H. et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49**, 708–718 (2017).
119. Qian, J. et al. A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Res.* **30**, 745–762 (2020).
120. Zhang, L. et al. Single-cell analyses inform mechanisms of myeloid-targeted therapies in colon cancer. *Cell* **181**, 442–459.e29 (2020).
121. Domanska, D. et al. Single-cell transcriptomic analysis of human colonic macrophages reveals niche-specific subsets. *J. Exp. Med.* https://doi.org/10.1084/jem.20211846 (2022).
122. Uhlitz, F. et al. Mitogen-activated protein kinase activity drives cell trajectories in colorectal cancer. *EMBO Mol. Med.* **13**, e14123 (2021).
123. Zhang, L. et al. Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature* **564**, 268–272 (2018).
124. Che, L. H. et al. A single-cell atlas of liver metastases of colorectal cancer reveals reprogramming of the tumor microenvironment in response to preoperative chemotherapy. *Cell Discov.* **7**, 80 (2021).
125. Becker, W. R. et al. Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. *Nat. Genet.* **54**, 985–995 (2022).
126. Joanito, I. et al. Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. *Nat. Genet.* **54**, 963–975 (2022).
127. Luoma, A. M. et al. Molecular pathways of colon inflammation induced by cancer immunotherapy. *Cell* **182**, 655–671.e22 (2020).
128. Beumer, J. et al. High-resolution mRNA and secretome atlas of human enteroendocrine cells. *Cell* **181**, 1291–1306.e19 (2020).
129. He, G. W. et al. Optimized human intestinal organoid model reveals interleukin-22-dependency of paneth cell formation. *Cell Stem Cell* **29**, 1333–1345.e6 (2022).
130. Holloway, E. M. et al. Mapping development of the human intestinal niche at single-cell resolution. *Cell Stem Cell* **28**, 568–580.e4 (2021).
131. Yu, Q. H. et al. Charting human development using a multi-endodermal organ atlas and organoid models. *Cell* **184**, 3281–3298.e22 (2021).
132. Gao, S. et al. Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell RNA-sequencing. *Nat. Cell Biol.* **20**, 1227 (2018).
133. Sanchez, J. G., Enriquez, J. R. & Wells, J. M. Enteroendocrine cell differentiation and function in the intestine. *Curr. Opin. Endocrinol. Diabetes Obes.* **29**, 169–176 (2022).
134. Du, Y. et al. An update on the biological characteristics and functions of tuft cells in the gut. *Front. Cell Dev. Biol.* **10**, 1102978 (2022).
135. Bolton, C. et al. An integrated taxonomy for monogenic inflammatory bowel disease. *Gastroenterology* **162**, 859–876 (2022).
136. Kanke, M. et al. Single-cell analysis reveals unexpected cellular changes and transposon expression signatures in the colonic epithelium of treatment-naive adult Crohn's disease patients. *Cell Mol. Gastroenterol. Hepatol.* **13**, 1717–1740 (2022).
137. Martin, J. C. et al. Single-cell analysis of Crohn's disease lesions identifies a pathogenic cellular module associated with resistance to anti-TNF therapy. *Cell* **178**, 1493–1508.e20 (2019).
138. Uniken Venema, W. T. et al. Single-cell RNA sequencing of blood and ileal T cells from patients with Crohn's disease reveals tissue-specific characteristics and drug targets. *Gastroenterology* **156**, 812–815.e22 (2019).
139. Chen, E. et al. Inflamed ulcerative colitis regions associated with MRGPRX2-mediated mast cell degranulation and cell activation modules, defining a new therapeutic target. *Gastroenterology* **160**, 1709–1724 (2021).
140. Devlin, J. C. et al. Single-cell transcriptional survey of ileal-anal pouch immune cells from ulcerative colitis patients. *Gastroenterology* **160**, 1679–1693 (2021).
141. Mitsialis, V. et al. Single-cell analyses of colon and blood reveal distinct immune cell signatures of ulcerative colitis and Crohn's disease. *Gastroenterology* **159**, 591–608.e10 (2020).
142. Qi, J. J. et al. Single-cell and spatial analysis reveal interaction of FAP^+ fibroblasts and $SPP1^+$ macrophages in colorectal cancer. *Nat. Commun.* **13**, 1742 (2022).

143. Goulet, O., Pigneur, B. & Charbit-Henrion, F. Congenital enteropathies involving defects in enterocyte structure or differentiation. *Best Pract. Res. Clin. Gastroenterol.* **56-57**, 101784 (2022).
144. Kelsen, J. R. & Baldassano, R. N. The role of monogenic disease in children with very early onset inflammatory bowel disease. *Curr. Opin. Pediatr.* **29**, 566–571 (2017).
145. Karim, A., Tang, C. S. & Tam, P. K. The emerging genetic landscape of Hirschsprung disease and its potential clinical applications. *Front. Pediatr.* **9**, 638093 (2021).
146. Gunther, C., Winner, B., Neurath, M. F. & Stappenbeck, T. S. Organoids in gastrointestinal diseases: from experimental models to clinical translation. *Gut* **71**, 1892–1908 (2022).
147. Ishikawa, K. et al. Identification of quiescent LGR5$^+$ stem cells in the human colon. *Gastroenterology* https://doi.org/10.1053/j.gastro.2022.07.081 (2022).
148. Dmitrieva-Posocco, O. et al. β-Hydroxybutyrate suppresses colorectal cancer. *Nature* **605**, 160–165 (2022).
149. Fawkner-Corbett, D. et al. Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* **184**, 810–81 (2021).
150. Cao, J. et al. A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020).
151. Li, N. et al. Memory CD4$^+$ T cells are generated in the human fetal intestine. *Nat. Immunol.* **20**, 301–312 (2019).
152. Nguyen, Q. H., Pervolarakis, N., Nee, K. & Kessenbrock, K. Experimental considerations for single-cell RNA sequencing approaches. *Front. Cell Dev. Biol.* **6**, 108 (2018).
153. Haque, A., Engel, J., Teichmann, S. A. & Lonnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 75 (2017).

Acknowledgements

This publication is part of the Human Cell Atlas (HCA) www.humancellatlas.org/publications. The HCA initiative receives funding from The Wellcome Trust, the UK Research and Innovation Medical Research Council, EU Horizon 2020, INSERM (HuDeCA), and the Knut and Alice Wallenberg and Erling-Persson foundations. We thank the HCA Executive Office for their support. The Gut Cell Atlas is organized by The Leona M. and Harry B. Helmsley Charitable Trust and provides funding for members in the form of project grants. M.Z. was supported by an MRC New Investigator research grant (MR/T001917/1) and a project grant from the Great Ormond Street Hospital Children's Charity, Sparks (V4519); K.S.L. was supported by NIDDK R01DK103831, and The Helmsley Trust — G-1903-03793. S.T.M. received funding from National Institutes of Health USA R01DK115806 and P30DK034987. T.S. was supported by the Japanese Science and Technology (JST) FOREST and the Japanese Society for the Promotion of Science (JSPS) (21K18272). L.A.C. and K.T.W. were supported by The Helmsley Charitable Trust — G-1903-03793. K.T.W. was also supported by NIDDK R01DK128200. L.A.C. was supported by a Veterans Affairs Merit Award 1I01BX004366. M.K. was supported by the National Research Foundation, South Africa grant no: 129356.

Author contributions

M.Z., K.R.J., M.K., S.P., Z.L., A.B., J.R.T., J.B., F.L.J., F.P., A. Ross, N.S., R.B.C., E.S.B., R.Z., B.X., K.L., S.D., S.T.M., Q.Y., S.B., M.J.A., A.D-.S., L.C., J.G., R.B., I.P., J.O-.M., G.E.B., A.H., S.A.T., A. Regev, R.J.X., M.P.S. and K.T.W. researched data for the article. M.Z., K.R.J., M.K., S.P., Z.L., A.B., J.R.T., J.B., F.L.J., F.P., A. Ross, G.M., N.S., T.S., A.M., R.B.C., E.S.B., R.Z., B.X., K.L., S.D., S.T.M., Q.Y., S.B., M.J.A., A.D-.S., L.C., J.G., R.B., I.P., J.O-.M., G.E.B., A.H., S.A.T., A. Regev, R.J.X., M.P.S. and K.T.W. contributed substantially to discussion of the content. M.Z., K.R.J., M.K., S.P., Z.L., A.B., J.R.T., J.B., F.L.J., F.P., A. Ross, N.S., T.S., R.B.C., E.S.B., R.Z., B.X., K.L., S.D., S.T.M., Q.Y., S.B., M.J.A., A.D-.S., L.C., J.G., R.B., I.P., G.E.B., S.A.T., M.P.S. and K.T.W. wrote the article. M.Z., K.R.J., M.K., S.P., Z.L., A.B., J.R.T., J.B., F.L.J., F.P., A. Ross, G.M., T.S., A.M., R.B.C., E.S.B., R.Z., B.X., K.L., S.D., S.T.M., Q.Y., S.B., M.J.A., A.D-.S., L.C., J.G., R.B., I.P., J.O-.M., G.E.B., A.H., S.A.T., A. Regev, R.J.X., A.S., M.P.S. and K.T.W reviewed and/or edited the manuscript before submission.

Competing interests

In the past 3 years, S.A.T. has consulted or been a member of scientific advisory boards at Roche, Genentech, Biogen, GlaxoSmithKline, Qiagen and ForeSite Labs and is an equity holder of Transition Bio. G.M. has received grant funding from Boehringer Ingelheim. A. Regev is a co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas, and was a SAB member of ThermoFisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov until 31 July 2020. Since 1 August 2020, A. Regev has been an employee of Genentech and has equity in Roche. A. Regev is an inventor on patents and patent applications filed at the Broad Institute related to single-cell genomics. The remaining authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Reviews Gastroenterology & Hepatology* thanks the anonymous reviewers for their contribution to the peer review of this work.

Related links

Gut Cell Atlas: https://www.gutcellatlas.org
HCA Gut Bionetwork: https://www.humancellatlas.org/biological-networks/
Helmsley Charitable Trust (Gut Cell Atlas): https://helmsleytrust.org/our-focus-areas/crohns-disease/crohns-disease-therapeutics/gut-cell-atlas/
HuBMAP Portal: https://portal.hubmapconsortium.org/
Human BioMolecular Atlas Program: https://hubmapconsortium.github.io/ccf/pages/ccf-3d-reference-library.html
Single Cell Expression Atlas: https://www.ebi.ac.uk/gxa/sc/home
Single Cell Portal: https://singlecell.broadinstitute.org/single_cell
Tabula Sapiens: https://tabula-sapiens-portal.ds.czbiohub.org
The Human Cell Atlas – Metadata: https://data.humancellatlas.org/metadata
University of California at Santa Cruz (UCSC) Cell Browser: https://cells.ucsc.edu

Gut Cell Atlas Consortium

Matthias Zilbauer[1,2,3,37], Kylie R. James[4,5,37], Sebastian Pott[7,37], Albert Burger[9,37], Nathalie Stakenborg[15], Andreas Moor[17], Ran Zhou[7], Bingqing Xie[19], Ken S. Lau[20], Shahida Din[21], Mark Arends[24], Alexandre Denadai-Souza[25], Lori A. Coburn[26,27], Jellert T. Gaublomme[28], Richard Baldock[29], Irene Papatheodorou[30], Guy Boeckxstaens[15], Sarah A. Teichmann[18,32], Ramnik J. Xavier[34] & Keith T. Wilson[26,27]

Human Cell Atlas Gut Biological Network Consortium

Matthias Zilbauer[1,2,3,37], Kylie R. James[4,5,37], Mandeep Kaur[6,37], Sebastian Pott[7,37], Zhixin Li[8,37], Albert Burger[9,37], Jay R. Thiagarajah[10], Joseph Burclaff[11,12], Frode L. Jahnsen[13], Francesca Perrone[1,2], Alexander D. Ross[1,2,14], Gianluca Matteoli[15], Nathalie Stakenborg[15], Tomohisa Sujino[16], Andreas Moor[17], Raquel Bartolome-Casado[13,18], Espen S. Bækkevold[13], Ran Zhou[7], Bingqing Xie[19], Ken S. Lau[20], Shahida Din[21], Scott T. Magness[11,12], Qiuming Yao[22], Semir Beyaz[23], Mark Arends[24], Alexandre Denadai-Souza[25], Lori A. Coburn[26,27], Jellert T. Gaublomme[28], Richard Baldock[29], Irene Papatheodorou[30], Jose Ordovas-Montanes[10], Guy Boeckxstaens[15], Anna Hupalowska[31], Sarah A. Teichmann[18,32], Aviv Regev[31,33], Ramnik J. Xavier[34], Alison Simmons[35], Michael P. Snyder[36] & Keith T. Wilson[26,27]

[1]Wellcome-MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. [2]University Department of Paediatrics, University of Cambridge, Cambridge, UK. [3]Department of Paediatric Gastroenterology, Hepatology and Nutrition, Cambridge University Hospitals, Cambridge, UK. [4]Garvan Institute of Medical Research, Darlinghurst, NSW, Australia. [5]School of Biomedical Sciences, University of New South Wales, Sydney, NSW, Australia. [6]School of Molecular and Cell Biology, University of the Witwatersrand, Johannesburg, South Africa. [7]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL, USA. [8]Dana-Farber Cancer Institute, Boston, MA, USA. [9]Department of Computer Science, Heriot-watt University, Edinburgh, UK. [10]Division of Gastroenterology, Hepatology and Nutrition, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. [11]Joint Department of Biomedical Engineering, University of North Carolina at Chapel Hill and North Carolina State University', Chapel Hill, NC, USA. [12]Center for Gastrointestinal Biology and Disease, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [13]Department of Pathology, Oslo University Hospital and University of Oslo, Oslo, Norway. [14]University Department of Medical Genetics, University of Cambridge, Cambridge, UK. [15]Translational Research Center for Gastrointestinal Disorders (TARGID), Department of Chronic Diseases, Metabolism and Ageing, KU Leuven, Leuven, Belgium. [16]Center for the Diagnostic and Therapeutic Endoscopy, School of Medicine, Keio University, Tokyo, Japan. [17]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland. [18]Wellcome Sanger Institute, Hinxton, UK. [19]Department of Medicine, University of Chicago, Chicago, IL, USA. [20]Epithelial Biology Center and Department of Cell and Developmental Biology, Vanderbilt University School of Medicine, Nashville, TN, USA. [21]Edinburgh IBD Unit, Western General Hospital, NHS Lothian, Edinburgh, UK. [22]Department of Computer Science and Engineering, University of Nebraska Lincoln, Lincoln, NE, USA. [23]Cold Spring Harbour Laboratory, Cold Spring Harbour, New York, NY, USA. [24]Division of Pathology, Centre for Comparative Pathology, Cancer Research UK Edinburgh Centre, Institute of Cancer and Genetics, University of Edinburgh, Edinburgh, UK. [25]Laboratory of Mucosal Biology, Department of Chronic Diseases, Metabolism and Ageing, KU Leuven, Leuven, Belgium. [26]Vanderbilt University Medical Center, Nashville, TN, USA. [27]Veterans Affairs Tennessee Valley Healthcare System, Nashville, TN, USA. [28]Department of Biological Sciences, Columbia University, New York, NY, USA. [29]University of Edinburgh, Edinburgh, UK. [30]European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Wellcome Genome Campus, Hinxton, UK. [31]Genentech, San Francisco, CA, USA. [32]Theory of Condensed Matter Group, Cavendish Laboratory/Department of Physics, University of Cambridge, Cambridge, UK. [33]Klarman Cell Observatory, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [34]Broad Institute and Department of Molecular Biology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [35]MRC Human Immunology Unit, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [36]Stanford University School of Medicine, Stanford, CA, USA. [37]These authors jointly supervised this work: Matthias Zilbauer, Kylie R. James, Mandeep Kaur, Sebastian Pott, Zhixin Li & Albert Burger.