A Roadmap for the Human Gut Cell Atlas

Title: A Roadmap for the Human Gut Cell Atlas

Author(s): Matthias Zilbauer*#‡[1,2,3], Kylie R James#‡[4,5], Mandeep Kaur#[6], Sebastian Pott‡#[7], Zhixin Li#[8], Albert Burger‡#[9], 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‡[19], Bingqing Xie‡[20], Ken S Lau‡[21], Shahida Din‡[22], Scott T Magness[12,23], Qiuming Yao[24], Semir Beyaz[25], Mark Arends‡[26], Alexandre Denadai-Souza‡[27], Lori A. Coburn‡[28,29], Jellert T Gaublomme‡[30], Richard Baldock‡[31], Irene Papatheodorou‡[32], Jose Ordovas-Montanes[10], Guy Boeckxstaens‡[15], Anna Hupalowska[33], Sarah A Teichmann‡[18,34], Aviv Regev[33,35], Ramnik J Xavier‡[36], Alison Simmons[37], Michael P Snyder[38], Keith T. Wilson‡[28,29], Gut Cell Atlas‡ & Human Cell Atlas Gut Biological Network

*Corresponding author

#Authors that acted as section leads.

‡Members of the Gut Cell Atlas, an Initiative Supported by the Helmsley Charitable Trust.

Author affiliations:

[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, NSW, Australia,

[5]School of Biomedical Sciences, University of New South Wales, 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, USA

[8]Dana-Farber Cancer Institute, Boston, 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, USA.

[11]Joint Department of Biomedical Engineering, University of North Carolina at Chapel Hill, North Carolina State University, Chapel Hill, North Carolina, USA
[12]Center for Gastrointestinal Biology and Disease, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 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, Cambridge, UK
[19]Section of Genetic Medicine, University of Chicago, Chicago, USA
[20]Department of medicine, University of Chicago, Chicago, USA
[21]Epithelial Biology Center and Department of Cell and Developmental Biology, Vanderbilt University School of Medicine, Nashville TN, USA
[22]Edinburgh IBD Unit, Western General Hospital, NHS Lothian
[23]Joint Department of Biomedical Engineering, University of North Carolina at Chapel Hill/North Carolina State University, Chapel Hill, North Carolina, USA
[24]Department of Computer Science and Engineering, University of Nebraska Lincoln, Lincoln, Nebraska, USA
[25]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 11724 USA
[26]Edinburgh Pathology, University of Edinburgh, Institute of Genetics and Cancer, Edinburgh, UK
[27]Laboratory of Mucosal Biology, Department of Chronic Diseases, Metabolism and Ageing, KU Leuven, Belgium
[28]Vanderbilt University Medical Center, Nashville, USA
[29]Veterans Affairs Tennessee Valley Healthcare System, Nashville, USA
[30]Department of Biological Sciences, Columbia University, New York, USA
[31]University of Edinburgh, UK
[32]European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Wellcome Genome Campus, Hinxton, UK.
[33]Genentech, San Francisco, CA, USA

[34]Theory of Condensed Matter Group, Cavendish Laboratory/Department of Physics, University of Cambridge, Cambridge, UK
[35]Klarman Cell Observatory, Broad Institute of Harvard and MIT, Cambridge, MA, USA.
[36]Broad Institute and Department of Molecular Biology, Massachusetts General Hospital, Harvard Medical School, Boston, USA
[37]MRC Human Immunology Unit, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK.
[38]Stanford University School of Medicine, Stanford, USA

Corresponding author:
Matthias Zilbauer
Email: mz304@cam.ac.uk

Abstract:

The number of studies investigating the human gastrointestinal tract using various single cell profiling methods have increased substantially in recent years. Although this provides a unique opportunity for the generation of a first comprehensive Human Gut Cell Atlas (HGCA), there are still 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. Here we present a detailed roadmap for the generation of the HGCA on behalf of the Human Cell Atlas (HCA) Gut Biological Network. Based on the consensus opinion of experts from across the globe, we outline the main requirements for a first complete Gut Atlas by summarising existing datasets and highlighting anatomical regions and/or tissues with limited coverage. We provide recommendations for future studies, discuss key methodologies as well as the importance of integrating the healthy gut atlas with related diseases and gut organoids. Importantly, we critically discuss computational tools available and provide recommendations to overcome key challenges.

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 as well as representing a major site of immune interactions. Different anatomical sections of the intestinal tract play

specific roles in the digestive process requiring the presence and complex interaction of various cell types. 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 allowing genome wide molecular profiling on a single cell (sc) 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 sc profiling methods have increased substantially in recent years, generating a 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 Human Cell Atlas (HCA) Gut Biological Network connects leading experts in a wide range of related areas providing an ideal platform to lead the development of the first sc HGCA. Here we provide a detailed roadmap including recommendations on how to combine existing data and requirements for future studies. We discuss key methodologies as well as the importance of integrating the healthy gut atlas with related diseases and gut organoids. Importantly, we will critically reflect on computational tools available, highlight any existing limitations and provide recommendations to overcome key challenges. Importantly, generating a first comprehensive HGCA will provide researchers, scientist 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 and includes the oesophagus, stomach, and the small and large intestine. 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 (Figure 1). A unique feature of the human intestine is the presence of trillions of microbes 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 cross talk 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 as well as the dynamic changes of microbes over

time. 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 researches in this area. Another important aspect unique to the gastrointestinal (GI) tract is the exposure to and interaction with a wide range of food/nutritional components as well as antigens or potential toxins present in the daily diet. Robust evidence suggests that GI development, function, and predisposition to related disease are strongly correlated with dietary habits[2,3]. Therefore, documenting details on dietary habits will add critical information and further increase the value of molecular profiles generated. Importantly, epigenetic mechanisms can mediate the impact of exposure to environmental factors into stable cellular phenotypes. The possibility of simultaneously profiling multiple molecular layers at sc resolution including epigenetic, transcriptomic, and proteomic signatures will us to unravel 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 GI tract. In general, there are three major strategies for sampling the alimentary tract including mucosal biopsies, live surgical resections, and resections from deceased donors. Each strategy and tissue source has unique advantages and disadvantages (Figure 2). Mucosal biopsies can be obtained during routine endoscopy, but access is restricted to the upper and lower GI tract, with jejunum and proximal ileum rarely sampled. Multiple biopsies can be taken from the same patient allowing comparisons of different anatomical regions. In contrast to mucosal biopsies, which can be obtained from healthy individuals, live resections are derived from patients with major GI pathology. Recent work identified transcriptional differences between healthy intestines and normal-appearing intestine proximal to chronic inflammation[4]. Hence, resections from patients with GI 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[5]. Tissue from deceased organ donors allows sampling of entire organs, providing the opportunity for widescale comparisons between regions. Another major advantage of gut resections is the provision of all tissue layers from luminal contents through to musculature and serosa. In contrast, forceps biopsies only capture the mucosa including the epithelium and immediate subepithelial cells. Size also varies drastically between biopsies and resections impacting on

the number and type of downstream analyses that can be performed. Finally, tissue quality may differ between strategies with the time from sampling to processing of tissue being of critical importance.

In summary, compared to other human organs, the GI tract is frequently being sampled during routine clinical procedures providing unique opportunities to obtain tissue from healthy individuals as well as 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

The generation of a HGCA requires combining a vast number of different datasets generated by groups around the globe[6,7]. A fundamental aspect that determines the integration and comparability of datasets, is the documentation of detailed metadata[8]. The HCA provides extensive guidance on this topic[9,10]. In the following we briefly describe basic information required to allow future studies to be included in the HGCA and provide a template metadata table (Table 1). In line with recommendations provided by the HCA, required metadata can be divided into the following main aspects: a) study design, b) donor information, c) sample information, d) sample processing and e) data generation. In brief, a detailed description of study design including patient inclusion criteria as well as 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 can be of major benefit. Similarly, the impact 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[2,3]. Baseline information on sample type include the method by which it was obtained, sample area covered and anatomic location. Depending on sample type, the exact anatomical location from which the sample was obtained can be difficult to determine. However, the level of detail provided has a major impact on the comparability of studies and interpretation of data generated. If available, representative haematoxylin & eosin (H&E) and/or other staining can be directly linked to samples processed for sc studies. Information on sampling procedures should include details on the time between sampling and processing, as well as storage duration. Furthermore, providing a detailed protocol on sample processing including tissue dissociation, cell viability, possible enrichment of individual cell types as well as equipment used is important. The development of standardised protocols is subject of

ongoing work within the Human Cell Atlas. Active engagement of the scientific community and use of available resources including recommended protocols will increase comparability of data generated by different groups.

Protecting donor identity and engaging communities

Documentation, storage and sharing of patient/donor information raises important ethical considerations[11]. Regulations vary considerably between different institutions and countries further complicating the sharing of tissues and data. In the European Union, 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[12]. In the United States 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 (HHS) and the Food and Drug Administration (FDA)[13]. In addition to major differences in the rules and regulations between countries across the globe, exchanging tissue and data between countries poses further complexities and barriers. However, given the critical importance of metadata for the integration of different datasets, appropriate donor consent for sharing information, data and/or tissue is a prerequisite. In addition to explicit informed consent, pseudonymisation 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/donor communities, particularly in the case of historically neglected or mistreated populations in research[14]. Ideally, studies providing data for the HGCA should include as many community stake-holders as feasible, with the long-term goal of having research questions formulated by researchers and participating populations working in partnership[15].

Available methodologies and their application

A wide range of sc and single nuclei (sn) genomics tools have been developed and enabled the creation of high-resolution cellular atlases from human organs and tissues[6 16 17]. By and large, methodologies that will be used to generate the HGCA are the same as those used to profile other organs or systems (Figure 3). In addition to sc transcriptional profiling, key methods include spatial transcriptomics and sc profiling of other molecular layers such as the genotype (G&T-seq)[18] and epigenotype including chromatin states (sc Assay for Transposase-Accessible Chromatin sequencing (ATAC)-seq)[19] and DNA methylation (whole genome bisulfite

sequencing (scGBseq))[20]. Furthermore, several methods have been developed that allow simultaneous profiling of multiple molecular layers such as the 10x Genomics Multiome kit, which combines snATACseq and snRNAseq[21]. ScNOMe-seq is capable of measuring DNA methylation and chromatin accessibility within single cells[22], whilst scNMTseq captures transcript levels in addition to chromatin accessibility and DNA methylation[23]. Furthermore, the use of methods that enable enrichment for specific and/or rare cell types will be key to achieving complete coverage. For example, MIRACL-seq allows label free enrichment of rare cell types[24] whilst the Chromium Single Cell Immune Profiling assay provided by 10x Genomics allows detailed immune cell profiling including full-length, paired B-cell or T-cell receptor (BCR/TCR) sequences, surface protein expression, antigen specificity, and 5' gene expression. Another key challenge for the HGCA is the timely processing of fresh tissue samples. Isolation of tightly connected epithelial cells is associated with damage, highlighting the need for suitable dissociation protocols as well as computational tools to exclude dead or damaged cells. Furthermore, dissociation of cells at 37°C for prolonged time induces the expression of early-immediate genes, thereby disrupting nuanced cellular states. Recent advances have led to the development of snRNAseq protocols which can be applied to frozen tissue samples therefore easing the burden of sample processing. High quality nuclei can be readily isolated from frozen samples and subjected to snRNAseq[25]. Although fewer transcripts are recovered, these data capture most cell populations and vastly increase the availability of tissue sources. 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 on 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 sc profiling, information on spatial arrangement of cells that are crucial for their function is lost entirely. Multiple spatial methods are used to capture cells in their anatomical context including Next Generation Sequencing (NGS)-based (e.g., 10x Genomics' Visium)[26] and imaging-based (e.g., MERFISH) assays[27,28]. Typically, frozen or Formalin-Fixed Paraffin-Embedded (FFPE) tissue sections are used as starting material and ideally are obtained from the same area from which tissue for sc assays was taken. Integration of spatial approaches with sc genomics provides both the cellular resolution as well as spatial organisation 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 below.

In summary, existing methodologies offer researchers a wide range of opportunities to address their research goals. Whilst 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 as well as during their sc studies in order to further maximise 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 amount of diverse datasets generated by a wide range of research groups using different methodological approaches, the usage of existing as well as development of novel computational tools is of critical importance. Main challenges include adequate measures to identify studies with sufficient quality as well as combining and mapping of diverse datasets. Furthermore, the development of a user-friendly interface allowing the data to be explored by the scientific community is a key requirement. Whilst many of these tasks also apply to most if not all tissue mapping studies, 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 datasets to specific anatomical locations along the cranio-caudal axis of the gut and alignment of healthy gut derived datasets with disease state and intestinal organoid models. With the rapid development of scRNAseq technology, numerous computational tools and pipelines have been developed to analyse sc data[29-31]. These include programmes/packages allowing pre-processing of data such as quality control (QC) checks, normalisation and batch correction, sequence alignment as well as detection and removal of cell doublets. A summary of relevant existing computational tools/packages is provided in Table 2. In the following, we briefly discuss the main computational challenges and opportunities related to the development of the HCGA.

Combining and integrating datasets 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 datasets. In addition to data quality, the provision of detailed metadata is a prerequisite for successful integration into the HGCA. Another major challenge is the removal of potentially confounding batch effects[32]. Available approaches including Seurat[33], LIGER[34], and Harmony[35]. The integration of different

molecular profiles forms another key computational challenge and analytical frameworks have been developed to integrate multiple data types in the same cells including GLUE[36], MOFA+[37] and Cobolt[38]. Additionally, a statistical regression framework has been developed to integrate clinical metadata in sc profiling studies[24].

Cell type annotation: Abundant and well characterised cell types can be reliably identified through unsupervised clustering algorithms followed by comparison of key marker gene expression profiles[39]. 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 datasets provides a unique opportunity to develop an extensive cell marker gene list using both automated and supervised cell identification approaches (e.g. MACA[40], singleR[41], ScType[42]). For each confidently identified gut cell type, marker genes can then be inferred by distinguishing the known clusters (e.g. COMET[43], COSG[44]) ultimately leading to a fully automated annotation procedure. Applying automated annotation approaches to larger datasets will also yield unknown/novel cell clusters, which must be subjected to additional validation studies and functional characterisations for example by using human gut organoid models as will be discussed below.

Mapping cellular location: Mapping individual cell types to their specific anatomical location within the human intestine is of critical importance. Combining sc sequencing technologies with spatial transcriptomics enables accurate profiling of cell type topography and several tools have been developed to address the computational aspects involved[45] including ASAP[46], novoSpaRc[47], and Cell2location[48]. Multi-omic spatial profiling efforts result in a gene, plus protein-by-cell matrix, annotated with spatial coordinates for the centroid position of each cell. This matrix can then be used as input to the recently developed open source for spatial analysis pipelines[49] allowing an unbiased identification of spatial patterns based on gene expression, distinct cellular neighborhoods 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[50]. Spatial domains with coherent gene expression patterns can be identified with a Hidden Markov random field models, that find spatial domains by comparing the gene expression patterns of each cell with its neighborhood to search for coherent patterns[49]. The prevalence of specific cell-to-cell interactions in the human intestine can be evaluated by the frequency that each pair of cell types is proximal to each other[51]. Finally, spatially interacting cell types can be analyzed to identify which known ligand-receptor pairs show significantly increased or decreased co-expression,

which could serve as a proxy for signaling activity, by creating a background distribution through spatially aware permutations[49], which increases 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 cross talk: Investigating cellular cross-talk through the analyses of sc transcriptomic datasets is subject to active research in the field[52]. Amongst the most compelling existing approaches are CellPhoneDB, which enables interrogation of context specific crosstalk between different cell types based on an extensive database of known receptors and ligands[53]. Applying these algorithms to large and/or combined datasets is likely to yield novel insight into fundamental aspects of human intestinal physiology. Extending such analyses to include the crosstalk between human host cells and the gut microbiome represents another significant challenge. To this end, several computational strategies have been developed to identify cellular and microbial cross talk in the gut utilizing sc data from human gut and microbial samples[54 55]. Recent studies have provided evidence for the major value of combining in situ spatial-profiling technologies with single-cell sequencing to interrogate host-microbial interactions[56].

Regulatory network inference: Successful integration of datasets 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 (GRNs) that control fundamental biological functions such as cellular differentiation and cell state transitions (e.g. SCENIC[57], GRNBoost2[58], PIDC[59]). Identifying the GRNs contributes towards our understanding of how coordinated expression of transcription factor networks drives the expression of their respective target genes and ultimately to shape and maintain gut cell identity[60].

HGCA portal: A user-friendly and interactive web portal is essential for making the HGCA accessible to a broad research community[61]. There are a number of existing sc data portals providing a wide range of tools to interrogate datasets and a summary is provided below[62-64]. 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 Common Coordinate Framework (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[6]. The concept of a Common Coordinate Framework (CCF) has been introduced to facilitate this capability. Rood et al. provide the following definition[65]: *"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[66], the brain[67] and the use of vasculature as a CCF of the entire human body[68]. A first step towards the development of a HGCA CCF requires the generation of one-, two- and three-dimensional models of the intestinal tract. The one-dimensional conceptual model is based on the clinical view and biological organisation of the digestive tract in which distance from anatomical landmarks provides the critical location metric[69]. 'Anatomograms' are two-dimensional graphical models that include a simplified view of the gut for representational purposes thereby providing a basic framework for accurately capturing the anatomical location of the tissue component and cell types[70]. Three-dimensional models can be generated by integrating Computerized Tomography (CT) or Magnetic Resonance Imaging (MRI) data. Generating a framework also involves developing a system of associated coordinates that provide detailed specification of anatomical location as well as tissue and cell type. Hence, providing a detailed anatomical location of sampling sites is of critical importance. However, there are limitations to the accuracy of obtaining exact anatomical location during surgery or intestinal endoscopy. Stating proportional lengths and/or distances from reliable anatomical landmarks could improve accuracy and help to facilitate integration of generated datasets into the HGCA CCF. Another complementary strategy to localise cell types in 2D space is the combined use of scRNAseq with spatial transcriptomics as discussed above. This requires the development of computational tools and efforts to allow the integration of scRNAseq 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), 3D models of both the human large and small intestine have been developed. The output is a series of surface models of the gut representing the outline within a 3D context allowing placement of tissue or cell types at the macro-level. This data can be accessed online through a CCF Registration User Interface (CCF-RUI, Figure 4A) and mapped data can be explored using the Exploration User Interface (CCF-EUI). Furthermore, as part of a collaborative project funded by the Leona M. and Harry B. Helmsley Charitable Trust (Gut Cell Atlas)[71], researchers from The European

Bioinformatics Institute (EBI Cambridge) and Edinburgh have developed a 1D linear conceptual CCF model for the human large and small intestine that is linked to the 2D anatomogram[69,70]. These mappings coupled with the inverse transform from the 3D and 2D spaces back to the 1D linear model allow spatial interoperability between all representations and therefore the capability to compare and query data registered to any CCF using a web-based visualisation platforms (Figure 4B, C and D)[69].

In addition to providing a framework for accurately mapping single datasets, CCFs are also capable of integrating other related datasets and can be managed across geographically dispersed data repositories. Importantly, the value of CCFs is directly related to the level of standardisation achieved as it determines the degree of potential interoperability across the datasets. Amongst the ways to increase standardisation are the development of consensus guidelines on how anatomical location can be most accurately documented for the purpose of integrating sc studies into the HGCA CCF. A starting point can be the use of a standard metadata template (Table 1) and the adherence to the minimum information standard for the description of cell/tissue sources in the gut. Ultimately, successful development of a CCF will greatly enhance the value and broad application of the HGCA.

Summary of relevant existing datasets, studies and portals

The number of published studies reporting on sc profiling of the human gut has increased substantially since the founding of the HCA in 2016 (Figure 5 and Supplementary Table 1). 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 sc datasets and enable access in a user-friendly way. Furthermore, a substantial number of studies have profiled intestinal tissues obtained from patients with gut diseases and/or human intestinal organoids (Figure 5). In this section, we provide a summary of the main existing datasets and portals, highlight examples of novel findings derived from sc research, and provide recommendations for future work.

In utero development and healthy gut datasets

The high throughput scalability intrinsic to most sc technologies has provided unprecedented advances to the field. One area that benefited from this aspect is GI organogenesis, since its progress has been hindered by the scarcity of samples. Indeed, datasets derived from scRNAseq 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*$^+$ mesenchyme cells co-expressing *PDGFRA* and *F3*[72]. In other sc studies a population of mesenchymal cells displaying a similar transcriptional signature were found to co-express *NRG1*, which has been demonstrated to support differentiation of *LGR5*$^+$ stem cells into mature intestinal epithelial cells[73,74]. Additionally, distinct clusters of sc transcriptomes identified during early stages of human intestinal development provided novel insight into the processes of regionalization during early intestinal development[75 74]. There are numerous examples of novel findings based on scRNAseq datasets derived from healthy adult gut samples. An assessment of epithelial cells from ileum, colon, and rectum revealed a high degree of functional diversity between the small and large intestine, as reflected by different nutrient absorption preferences[76]. 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*$^+$ [16,77,78]. Recently, *BEST4*$^+$ epithelial cells have also been reported to vary in abundance and transcriptional signature across different regions of the gut[79]. Together, these findings illustrate the major benefit of combining and comparing datasets from different studies ultimately reaching reliable insight into healthy gut physiology and cellular function. Besides overcoming the challenges inherent to scarcity of material, datasets generated by scRNAseq 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 has remained elusive until recently. 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[24]. Interstitial Cells of Cajal (ICC) are rare entities critical for GI peristalsis through both the generation of slow wave pacemaker activity to smooth muscle cells, and mediation of neurotransmission from enteric neurons. By applying the same strategy, transcriptional signatures of 1,103 ICCs from the human colon have been generated[24]. Another sc database, 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[80]. Chemosensory cells such as tuft cells (TF) and enteroendocrine cells (EEC) are rare intestinal epithelial cells operating as an interface for signal transduction between the intestinal lumen and the body, relaying diet- and microbiota-derived signals through the release of numerous peptide hormones, neurotransmitters, and cytokines. In addition, TF cells are essential for mounting Th2 immune responses against parasites. Based on sc studies, TF cells

were demonstrated to interact with the innate and adaptive immune systems through previously unreported receptors[79], including IgG receptors[16]. EECs sense intestinal content and release hormones to regulate gastrointestinal activity, systemic metabolism, and food intake. By using an organoid-based platform wherein EEC differentiation was induced by transient expression of *NEUROG3*, and hormones were tagged with gene reporters, the authors generated a comprehensive dataset of EEC subtypes derived from the small intestine and colon[81].

Intestinal diseases and gut organoids

Generating a complete HGCA in health provides unique opportunities to study pathogenesis of related diseases. Furthermore, sc transcriptional profiles of primary tissue samples can be utilised 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 datasets generated from related sample and patient cohorts can be integrated into the HGCA (Figure 5). Amongst the related gut diseases that have been investigated using sc profiling approaches are colorectal cancer (CRC)[24,76,82-92], the inflammatory bowel diseases (IBD)[93], Crohn's disease (CD)[72,94-97] and ulcerative colitis (UC)[98-102] as well as celiac disease[103].

Examples of major findings in IBD include the identification of distinct immune cell signatures in both UC and CD[104], a pathogenic cellular module associated with resistance to anti-TNF therapy[95], inference of genetic risk genes to sc function[98] as well as the reactivation of fetal intestinal epithelial transcriptional profiles in childhood onset CD[72]. Similarly, application of sc molecular profiling methods to colonic tissue obtained from CRC patients has led to major advances in our understanding of disease pathogenesis. Specifically, recent sc insights into the stem and metaplastic origins of human pre-cancers[84] have led to reclassification of the consensus molecular subtypes of CRC by their intrinsic features[105]. Transition of benign lesions into malignancy is accompanied by tumor cell acquisition of stem characteristics[84,92], and reorganization of the microenvironment into suppressive immune-stromal hubs that can potentially be therapeutically targeted[83,106]. Although a comprehensive summary of all relevant available sc 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 datasets are already available will be prioritised in the first phase of data integration, it is important to emphasise the major value of investigating and ultimately integrating rarer conditions. Indeed, combining datasets from less commonly profiled conditions represents another unique opportunity of the HGCA by increasing

computational power and allowing validation of key findings. Furthermore, applying a variety of methodologies (e.g. single multi-ome profiling) to the same condition will further improve the value of the HGCA and the chances of gaining novel insight in diseases pathogenesis.

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[107]. Performing transcriptional profiling of patient derived intestinal organoids on a sc level is of great value as reflected in the increasing numbers of studies reporting novel findings. Main benefits include the ability to validate cellular composition of organoids to further improve culture conditions as well as evaluating to what extent disease associated cellular alterations are retained in vitro. A recent example is a study published by He and colleagues, who applied scRNAseq to human small intestinal organoids[108]. Exposure of organoids to IL-22 resulted in increased expression of antimicrobial peptides suggesting the presence of Paneth cells[108]. Furthermore, sc profiling of primary human fetal gut and organoids derived from the same tissue sample revealed in vitro maturation of fetal organoids highlighting their value to investigate early stages of human intestinal epithelial cell development[72]. Work by Ishikawa and colleagues combined scRNAseq of human colonic epithelium cells with genetically modified human intestinal organoids leading to the identification of quiescent LGR5+ stem cells in the human colon[109]. Similarly, combining cancer organoid assays and scRNAseq of biopsies from patients with CRC identified a novel role of a β-hydroxybutyrate (BHB)-triggered pathway in regulating intestinal tumorigenesis[110]. As for studies based on profiling primary tissue, the provision of metadata including clinical donor details, sampling sites, biobanking as well as experimental procedures are 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 sc datasets that have been generated in the recent past, several web-based data repositories have emerged to provide access in a user-friendly format. While most portals share common features of storing data, curating datasets 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[70,111]: The Single Cell Expression Atlas (SCEA) is the sc component of EMBL-EBI's Expression Atlas. This is an added-value resource that enables simple gene and meta-data queries, allowing users to answer questions such as "where is my gene of interest expressed", "how does its expression level change in a disease". SCEA collects expression data from all species, annotates their metatada with appropriate ontology terms to allow nested across studies and crucially, re-analyses all datasets using standardised analysis pipelines to enable comparison across studies. The SCEA works in close collaboration with the HGCA, releasing the datasets as they become publicly available and generating the first full gut 2D 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[112]: The Human BioMolecular Atlas Program (HuBMAP) is a NIH Common Fund effort to integrate and map diverse biological data across the healthy human body. Three main features enable: 1) analysis of scRNAseq experiments with *Azimuth*, a web application that uses reference datasets to automate annotation and interpretation of data; 2) spatial sc data visualization with *Vitessce*; and 3) navigation of the healthy human cells with the CCF to interact with the virtual human body to focus on anatomical structures, cell types, and biomarkers.

Tabula Sapiens[113]: Like HuBMAP, the Tabula Sapiens focuses on healthy human subjects, and has created a first draft HCA of 24 organs from 15 different human subjects. Tabula Sapiens is funded by the Chan Zuckerberg Initiative (CZI) and is unique in that the sc data sets derived for each organ are from the same human subject, controlling for inter-individual factors. This also allows comparison of cell types that are shared between different organs. The web portal offers the ability to peruse the all the sc data combined (currently at 500,000 cells) or curated cell subsets (*i.e. endothelial, epithelial, immune, and stromal)* in an easy-to-navigate graphical user interface.

UCSC Cell Browser[114]: The University of California at Santa Cruz (UCSC) *Cell Browser* is a sub-project under 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 sc 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 datasets generated from human, mouse, fly, and sponges, and curated as conventional Atlases or analysed in the context of development and evolution. The datasets 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[115]: The Single Cell Portal is hosted by the Broad Institute and was created with the idea to make sc 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 datasets. The site is easy to use and has succinct overviews describing 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[116]: This resource provides a detailed gut cell survey by combining single cell data generated from a range of human gut tissues. Datasets include studies that profiled various gut segments obtained from fetal, pediatric and adult donors as well as patients diagnosed with CD. In addition to direct access to raw datasets, the site provides an interactive viewer that allows basic analyses and interrogation of datasets such as exploring single cell expression profiles of individual cell lineages, gut regions, or comparisons between age groups.

Future opportunities including incorporating datasets

In many aspects, the current collection of high quality published sc 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 scRNAseq technology include intestinal development[72,75,117], profiling of cell types, states and tissue composition[81], mapping regional differences across intestinal tissues and along the intestinal tube[81,118], as well as comparisons between healthy individuals and IBD[79,119], CRC[89,90], or other related diseases[92,120]. 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 underrepresented 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. However, such cross-tissue studies are of major value as they enable the identification of common/distinct cell types across intestinal regions as well as studying migratory patterns of specific immune cell populations[16]. Furthermore, greater emphasis should be placed on profiling rare cell types particularly in less frequently sampled gut regions. This can be achieved by applying methods designed to enrich for such cell types prior to scRNAseq 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[24]. However, advantages and disadvantages of these methods must be carefully considered in the context of each particular research study[121,122]. Finally, there has been a major bias with regards to anatomical sampling sites with a large proportion focussing on profiling the colon and distal small bowel. This is likely to be partly caused by ease of access during endoscopic procedures as well as the major relevance to related GI diseases such as IBD and CRC. Future studies should also include less frequently sampled gut segments, such as the mid small intestine, oesophagus, stomach as well as 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 in the pathophysiology of related diseases including those occurring in specific age groups.

Conclusions

Generating a complete map of the human intestine at the sc level will improve our understanding of gut health and disease. The inherent complexities and scale of this task requires 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 is essential to achieving our goal the generation of a HGCA.

References:

1 Rooks, M. G. & Garrett, W. S. Gut microbiota, metabolites and host immunity. *Nature Reviews Immunology* **16**, 341-352, doi:10.1038/nri.2016.42 (2016).

2 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, doi:10.1016/j.chom.2018.05.012 (2018).

3 Khalili, H. *et al.* The role of diet in the aetiopathogenesis of inflammatory bowel disease. *Nat Rev Gastroenterol Hepatol* **15**, 525-535, doi:10.1038/s41575-018-0022-9 (2018).

4 Smillie, C. S. *et al.* Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178**, 714-730.e722, doi:10.1016/j.cell.2019.06.029 (2019).

5 Pilonis, N. D., Januszewicz, W. & di Pietro, M. Confocal laser endomicroscopy in gastro-intestinal endoscopy: technical aspects and clinical applications. *Transl Gastroenterol Hepatol* **7**, 7, doi:10.21037/tgh.2020.04.02 (2022).

6 Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, doi:ARTN e27041 10.7554/eLife.27041 (2017).
7 Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451-453, doi:DOI 10.1038/550451a (2017).
8 Skinnider, M. A., Squair, J. W. & Courtine, G. Enabling reproducible re-analysis of single-cell data. *Genome Biol* **22**, doi:ARTN 215 10.1186/s13059-021-02422-y (2021).
9 The Human Cell Atlas - Metadata.
10 Füllgrabe, A. *et al.* Guidelines for reporting single-cell RNA-seq experiments. *Nat Biotechnol* **38**, 1384-1386, doi:10.1038/s41587-020-00744-z (2020).
11 Lee, S. S. J. The Ethics of Consent in a Shifting Genomic Ecosystem. *Annual Review of Biomedical Data Science, Vol 4* **4**, 145-164, doi:10.1146/annurev-biodatasci-030221-125715 (2021).
12 EUR-Lex, T. E. P. a. t. C. o. t. E. U. *General Data Protection Regulation (EU) 2016/679 (GDPR)*, <https://eur-lex.europa.eu/eli/reg/2016/679/oj.> (2016).
13 Bledsoe, M. J. & Grizzle, W. E. Use of human specimens in research: the evolving United States regulatory, policy, and scientific landscape. *Diagn Histopathol (Oxf)* **19**, 322-330, doi:10.1016/j.mpdhp.2013.06.015 (2013).
14 Shore, N. *et al.* Understanding community-based processes for research ethics review: a national study. *Am J Public Health* **101 Suppl 1**, S359-364, doi:10.2105/ajph.2010.194340 (2011).
15 Mikesell, L., Bromley, E. & Khodyakov, D. Ethical community-engaged research: a literature review. *Am J Public Health* **103**, e7-e14, doi:10.2105/ajph.2013.301605 (2013).
16 Elmentaite, R. *et al.* Cells of the human intestinal tract mapped across space and time. *Nature* **597**, 250-+, doi:10.1038/s41586-021-03852-1 (2021).
17 Camp, J. G., Platt, R. & Treutlein, B. Mapping human cell phenotypes to genotypes with single-cell genomics. *Science* **365**, 1401-+, doi:10.1126/science.aax6648 (2019).
18 Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods* **12**, 519-522, doi:10.1038/nmeth.3370 (2015).
19 Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat Biotechnol* **37**, 916-924, doi:10.1038/s41587-019-0147-6 (2019).
20 Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* **11**, 817-820, doi:10.1038/nmeth.3035 (2014).
21 Hao, Y. H. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-+, doi:10.1016/j.cell.2021.04.048 (2021).
22 Pott, S. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *Elife* **6**, doi:10.7554/eLife.23203 (2017).
23 Clark, S. J. *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun* **9**, doi:ARTN 781 10.1038/s41467-018-03149-4 (2018).
24 Drokhlyansky, E. *et al.* The Human and Mouse Enteric Nervous System at Single-Cell Resolution. *Cell* **182**, 1606-+, doi:10.1016/j.cell.2020.08.003 (2020).

25 Slyper, M. *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat Med* **26**, 792-802, doi:10.1038/s41591-020-0844-1 (2020).

26 Wu, S. Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet* **53**, 1334-1347, doi:10.1038/s41588-021-00911-1 (2021).

27 Xia, C., Babcock, H. P., Moffitt, J. R. & Zhuang, X. Multiplexed detection of RNA using MERFISH and branched DNA amplification. *Scientific Reports* **9**, 7721, doi:10.1038/s41598-019-43943-8 (2019).

28 Black, S. *et al.* CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nat Protoc* **16**, 3802-3835, doi:10.1038/s41596-021-00556-8 (2021).

29 Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformaticspipelines. *Experimental & Molecular Medicine* **50**, 1-14, doi:10.1038/s12276-018-0071-8 (2018).

30 Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLOS Computational Biology* **14**, e1006245, doi:10.1371/journal.pcbi.1006245 (2018).

31 Moreno, P. *et al.* User-friendly, scalable tools and workflows for single-cell RNA-seq analysis. *Nat. Methods* **18**, 327-328, doi:10.1038/s41592-021-01102-w (2021).

32 Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology* **21**, 12, doi:10.1186/s13059-019-1850-9 (2020).

33 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e3529, doi:10.1016/j.cell.2021.04.048 (2021).

34 Liu, J. *et al.* Jointly defining cell types from multiple single-cell datasets using LIGER. *Nature protocols* **15**, 3632-3662, doi:10.1038/s41596-020-0391-8 (2020).

35 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289-1296, doi:10.1038/s41592-019-0619-0 (2019).

36 Cao, Z. J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol*, doi:10.1038/s41587-022-01284-4 (2022).

37 Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology* **21**, 111, doi:10.1186/s13059-020-02015-1 (2020).

38 Gong, B., Zhou, Y. & Purdom, E. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome Biology* **22**, 351, doi:10.1186/s13059-021-02556-z (2021).

39 Jones, R. C. *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, 711-+, doi:10.1126/science.abl4896 (2022).

40 Xu, Y., Baumgart, S. J., Stegmann, C. M. & Hayat, S. MACA: Marker-based automatic cell-type annotation for single cell expression data. *Bioinformatics*, doi:10.1093/bioinformatics/btab840 (2021).

41 Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology* **20**, 163-172, doi:10.1038/s41590-018-0276-y (2019).

42 Ianevski, A., Giri, A. K. & Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nature Communications* **13**, 1246, doi:10.1038/s41467-022-28803-w (2022).

43 Delaney, C. *et al.* Combinatorial prediction of marker panels from single-cell transcriptomic data. *Mol Syst Biol* **15**, e9005, doi:10.15252/msb.20199005 (2019).

44 Dai, M., Pei, X. & Wang, X.-J. Accurate and fast cell marker gene identification with COSG. *Briefings in Bioinformatics* **23**, doi:10.1093/bib/bbab579 (2022).

45 Andersson, A. *et al.* Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications Biology* **3**, 565, doi:10.1038/s42003-020-01247-y (2020).

46 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 Research* **48**, W403-W414, doi:10.1093/nar/gkaa412 (2020).

47 Moriel, N. *et al.* NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature protocols* **16**, 4177-4200, doi:10.1038/s41596-021-00573-7 (2021).

48 Kleshchevnikov, V. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.*, doi:10.1038/s41587-021-01139-4 (2022).

49 Dries, R. *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol* **22**, 78, doi:10.1186/s13059-021-02286-2 (2021).

50 Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat Methods* **15**, 343-346, doi:10.1038/nmeth.4636 (2018).

51 Goltsev, Y. *et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**, 968-981 e915, doi:10.1016/j.cell.2018.07.010 (2018).

52 Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell-cell interactions and communication from gene expression. *Nature Reviews Genetics* **22**, 71-88, doi:10.1038/s41576-020-00292-x (2021).

53 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, doi:10.1038/s41596-020-0292-x (2020).

54 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. *Sensors and Actuators B-Chemical* **309**, doi:ARTN 127789 10.1016/j.snb.2020.127789 (2020).

55 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**, doi:ARTN 4638 10.1038/s41467-018-06214-0 (2018).

56 Galeano Niño, J. L. *et al.* Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer. *Nature* **611**, 810-817, doi:10.1038/s41586-022-05435-0 (2022).

57 Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083-1086, doi:10.1038/nmeth.4463 (2017).

58 Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159-2161, doi:10.1093/bioinformatics/bty916 (2019).

59 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.e253, doi:10.1016/j.cels.2017.08.014 (2017).

60 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, doi:10.1038/s41592-019-0690-6 (2020).
61 Cakir, B. *et al.* Comparison of visualization tools for single-cell RNAseq data. *NAR Genomics and Bioinformatics* **2**, doi:10.1093/nargab/lqaa052 (2020).
62 Megill, C. *et al.* chanzuckerberg/cellxgene: Release 0.15.0. doi:10.5281/ZENODO.3710410 (2020).
63 Elmentaite, R. *et al.* Single-Cell Sequencing of Developing Human Gut Reveals Transcriptional Links to Childhood Crohn's Disease. *Developmental Cell* **55**, 771-783.e775, doi:https://doi.org/10.1016/j.devcel.2020.11.010 (2020).
64 Moreno, P. *et al.* Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Research* **50**, D129-D140, doi:10.1093/nar/gkab1030 (2021).
65 Rood, J. E. *et al.* Toward a Common Coordinate Framework for the Human Body. *Cell* **179**, 1455-1467, doi:10.1016/j.cell.2019.11.019 (2019).
66 Schiller, H. B. *et al.* The Human Lung Cell Atlas: A High-Resolution Reference Map of the Human Lung in Health and Disease. *American Journal of Respiratory Cell and Molecular Biology* **61**, 31-41, doi:10.1165/rcmb.2018-0416TR (2019).
67 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. *Nature Neuroscience* **24**, 584-594, doi:10.1038/s41593-020-00794-1 (2021).
68 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. *Frontiers in Cardiovascular Medicine* **7**, doi:ARTN 29 10.3389/fcvm.2020.00029 (2020).
69 Burger, A. *et al.* Towards a Clinically-based Common Coordinate Framework for the Human Gut Cell Atlas - The Gut Models. *bioRxiv*, 2022.2012.2008.519665, doi:10.1101/2022.12.08.519665 (2022).
70 Moreno, P. *et al.* Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Res* **50**, D129-D140, doi:10.1093/nar/gkab1030 (2022).
71 *Gut Cell Atlas, an Initative Supported by the Helmsley Charitable Trust* <https://helmsleytrust.org/our-focus-areas/crohns-disease/crohns-disease-therapeutics/gut-cell-atlas/> (
72 Elmentaite, R. *et al.* Single-Cell Sequencing of Developing Human Gut Reveals Transcriptional Links to Childhood Crohn's Disease. *Developmental Cell* **55**, 771-+, doi:10.1016/j.devcel.2020.11.010 (2020).
73 Holloway, E. M. *et al.* Mapping Development of the Human Intestinal Niche at Single-Cell Resolution. *Cell Stem Cell* **28**, 568-+, doi:10.1016/j.stem.2020.11.008 (2021).
74 Yu, Q. H. *et al.* Charting human development using a multi-endodermal organ atlas and organoid models. *Cell* **184**, 3281-+, doi:10.1016/j.cell.2021.04.028 (2021).
75 Gao, S. *et al.* Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell RNA-sequencing (vol 20, pg 721, 2018). *Nature Cell Biology* **20**, 1227-1227, doi:10.1038/s41556-018-0165-5 (2018).
76 Wang, Y. L. *et al.* Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine. *Journal of Experimental Medicine* **217**, doi:10.1084/jem.20191130 (2020).

77 Parikh, K. *et al.* Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature* **567**, 49-+, doi:10.1038/s41586-019-0992-y (2019).
78 Busslinger, G. A. *et al.* Human gastrointestinal epithelia of the esophagus, stomach, and duodenum resolved at single-cell resolution. *Cell Reports* **34**, doi:ARTN 108819 10.1016/j.celrep.2021.108819 (2021).
79 Burclaff, J. *et al.* A Proximal-to-Distal Survey of Healthy Adult Human small Intestine and Colon Epithelium by Single-Cell Transcriptomics. *Cellular and Molecular Gastroenterology and Hepatology* **13**, 1554-1589, doi:10.1016/j.jcmgh.2022.02.007 (2022).
80 Foong, D. *et al.* Single-cell RNA sequencing predicts motility networks in purified human gastric interstitial cells of Cajal. *Neurogastroenterology and Motility* **34**, doi:ARTN e14303 10.1111/nmo.14303 (2022).
81 Beumer, J. *et al.* High-Resolution mRNA and Secretome Atlas of Human Enteroendocrine Cells. *Cell* **181**, 1291-+, doi:10.1016/j.cell.2020.04.036 (2020).
82 Lee, H. O. *et al.* Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat Genet* **52**, 594-603, doi:10.1038/s41588-020-0636-z (2020).
83 Pelka, K. *et al.* Spatially organized multicellular immune hubs in human colorectal cancer. *Cell* **184**, 4734-4752.e4720, doi:10.1016/j.cell.2021.08.003 (2021).
84 Chen, B. *et al.* Differential pre-malignant programs and microenvironment chart distinct paths to malignancy in human colorectal polyps. *Cell* **184**, 6262-6280.e6226, doi:10.1016/j.cell.2021.11.031 (2021).
85 Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet* **49**, 708-718, doi:10.1038/ng.3818 (2017).
86 Qian, J. *et al.* A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Res* **30**, 745-762, doi:10.1038/s41422-020-0355-0 (2020).
87 Zhang, L. *et al.* Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer. *Cell* **181**, 442-459.e429, doi:10.1016/j.cell.2020.03.048 (2020).
88 Domanska, D. *et al.* Single-cell transcriptomic analysis of human colonic macrophages reveals niche-specific subsets. *J Exp Med* **219**, doi:10.1084/jem.20211846 (2022).
89 Uhlitz, F. *et al.* Mitogen-activated protein kinase activity drives cell trajectories in colorectal cancer. *Embo Molecular Medicine* **13**, doi:ARTN e14123 10.15252/emmm.202114123 (2021).
90 Zhang, L. *et al.* Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature* **564**, 268-+, doi:10.1038/s41586-018-0694-x (2018).
91 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, doi:10.1038/s41421-021-00312-y (2021).
92 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. *Nature Genetics* **54**, 985-+, doi:10.1038/s41588-022-01088-x (2022).

93 Bolton, C. *et al.* An Integrated Taxonomy for Monogenic Inflammatory Bowel Disease. *Gastroenterology* **162**, 859-876, doi:10.1053/j.gastro.2021.11.014 (2022).
94 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, doi:10.1016/j.jcmgh.2022.02.005 (2022).
95 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 e1420, doi:10.1016/j.cell.2019.08.008 (2019).
96 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, doi:10.1038/s41467-021-22164-6 (2021).
97 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 e822, doi:10.1053/j.gastro.2018.10.046 (2019).
98 Smillie, C. S. *et al.* Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178**, 714-+, doi:10.1016/j.cell.2019.06.029 (2019).
99 Uzzan, M. *et al.* Ulcerative colitis is characterized by a plasmablast-skewed humoral response associated with disease activity. *Nat Med* **28**, 766-779, doi:10.1038/s41591-022-01680-y (2022).
100 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, doi:10.1053/j.gastro.2020.12.076 (2021).
101 Devlin, J. C. *et al.* Single-Cell Transcriptional Survey of Ileal-Anal Pouch Immune Cells From Ulcerative Colitis Patients. *Gastroenterology* **160**, 1679-1693, doi:10.1053/j.gastro.2020.12.030 (2021).
102 Corridoni, D. *et al.* Single-cell atlas of colonic CD8(+) T cells in ulcerative colitis. *Nat Med* **26**, 1480-1490, doi:10.1038/s41591-020-1003-4 (2020).
103 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, doi:10.1038/s41467-022-32691-5 (2022).
104 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 e510, doi:10.1053/j.gastro.2020.04.074 (2020).
105 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, doi:10.1038/s41588-022-01100-4 (2022).
106 Qi, J. J. *et al.* Single-cell and spatial analysis reveal interaction of FAP(+) fibroblasts and SPP1(+) macrophages in colorectal cancer. *Nat Commun* **13**, doi:ARTN 1742 10.1038/s41467-022-29366-6 (2022).
107 Gunther, C., Winner, B., Neurath, M. F. & Stappenbeck, T. S. Organoids in gastrointestinal diseases: from experimental models to clinical translation. *Gut* **71**, 1892-1908, doi:10.1136/gutjnl-2021-326560 (2022).

108 He, G. W. *et al.* Optimized human intestinal organoid model reveals interleukin-22-dependency of paneth cell formation. *Cell Stem Cell* **29**, 1333-1345 e1336, doi:10.1016/j.stem.2022.08.002 (2022).
109 Ishikawa, K. *et al.* Identification of Quiescent LGR5(+) Stem Cells in the Human Colon. *Gastroenterology*, doi:10.1053/j.gastro.2022.07.081 (2022).
110 Dmitrieva-Posocco, O. *et al.* beta-Hydroxybutyrate suppresses colorectal cancer. *Nature* **605**, 160-165, doi:10.1038/s41586-022-04649-6 (2022).
111 *Single Cell Expression Atlas*, <https://www.ebi.ac.uk/gxa/sc/home> (
112 *Human BioMolecular Atlas Program (HuBMAP) Data Portal*, <https://portal.hubmapconsortium.org> (
113 *Tabula Sapiens Data Portal*, <https://tabula-sapiens-portal.ds.czbiohub.org> (
114 *UCSC Cell Browser*, <https://cells.ucsc.edu> (
115 *Broad Institute Single Cell Portal*, <https://singlecell.broadinstitute.org/single_cell> (
116 *Gut Cell Survey*, <https://www.gutcellatlas.org> (
117 Fawkner-Corbett, D. *et al.* Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* **184**, 810-+, doi:10.1016/j.cell.2020.12.016 (2021).
118 James, K. R. *et al.* Distinct microbial and immune niches of the human colon. *Nat Immunol* **21**, 343-+, doi:10.1038/s41590-020-0602-z (2020).
119 Kinchen, J. *et al.* Structural Remodeling of the Human Colonic Mesenchyme in Inflammatory Bowel Disease. *Cell* **175**, 372-+, doi:10.1016/j.cell.2018.08.067 (2018).
120 Nowicki-Osuch, K. *et al.* Molecular phenotyping reveals the identity of Barrett's esophagus and its malignant transition. *Science* **373**, 760-+, doi:10.1126/science.abd1449 (2021).
121 Nguyen, Q. H., Pervolarakis, N., Nee, K. & Kessenbrock, K. Experimental Considerations for Single-Cell RNA Sequencing Approaches. *Frontiers in Cell and Developmental Biology* **6**, doi:ARTN 108
10.3389/fcell.2018.00108 (2018).
122 Haque, A., Engel, J., Teichmann, S. A. & Lonnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine* **9**, doi:ARTN 75
10.1186/s13073-017-0467-4 (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 organised by The Leona M. and Harry B. Helmsley Charitable Trust and provides funding for members in form of project grants. MZ 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); KSL was supported by NIDDK R01DK103831, and The Helmsley Trust - G-1903-03793. STM received funding from (STM) National Institutes of Health USA R01DK115806 & P30DK034987. TS was supported by the Japanese Science and Technology (JST) forest, and the Japanese Society for the Promotion of Science (JSPS) (21K18272). LAC and KTW. were supported by The Helmsley Charitable Trust - G-1903-03793. KTW was also supported by NIDDK R01DK128200. LAC was supported by a Veterans Affairs Merit Award 1I01BX004366. MK was supported by the National Research Foundation, South Africa grant no: 129356.

Author contributions MZ, KTW, AS, MPS and RJX are coordinators of the HCA Gut Biological Network. MZ conceived the idea, co-ordinated the writing process, wrote parts of the paper and edited all sections. SP was section lead for methodologies. ZL was section lead for computational tools and challenges. MK was section lead of the integration of HGCA with gut diseases and organoids. KJ was section lead for the summary of existing datasets and portals. AGB was section lead for the Common Coordinate Framework (CCF). AH helped to design and create the figures. All other authors wrote parts of the paper, contributed to critical discussions and provided feedback on the entire manuscript.

Competing interests: SAT: In the past three years, SAT. 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. GM has received grant funding from Boehringer Ingelheim. A.R. is a co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas, and was an SAB member of ThermoFisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov until 31 July 2020. Since 1 August 2020, A.R. has been an employee of Genentech and has equity in Roche. A.R. is an inventor on patents and patent applications filed at the Broad related to single cell genomics.

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.
- The generation of a Human Gut Cell Atlas (HGCA) requires the coordinated efforts of groups across the globe and the integration of various datasets followed by their computational analyses.
- This article provides a roadmap for the generation of the HGCA based on the expertise and recommendations of the Human Cell Atlas (HCA) Gut Biological Network.
- The Human Gut Cell Atlas (HGCA) will provide a unique and highly valuable reference map enhancing research in intestinal health and disease.

Figure legends

Figure 1: Profiling the human digestive tract: A complete map of the human intestinal tract requires the inclusion of the entire intestinal tube and associated organs (left panel). Profiling must also consider the impact of developmental stage, the gut microbiome, and the potential impact of environmental factors such as diet, toxins and medication. In addition to transcriptional profiling, capturing the underlying genome sequence and epigenetic programme will yield critical information (right panel).

Figure 2: Summary of main tissue types and sampling strategies available. Main advantages and limitations are illustrated.

Figure 3: Data integration, processing and analysing strategies: 1) Generation of the HGCA requires successful integration of various datasets, datatypes, and associated patient/donor metadata. 2 and 3) Successful integration will allow a range of queries to be performed and outputs generated.

Figure 4: Current applications for the Human Gut Cell Atlas Common Coordinate Framework: A: HuBMAP Registration User Interface (CCF-RUI) showing the gut visibility controls on the LH panel, interactive block registration interface in the centre panel and the

block specification controls in the RH panel. **B:** Gut Atlas CCF browser interface showing the 1D conceptual model for the full large and small intestines with zoom panel and ontology listings at the top with 2D and 3D interactive displays for all of the available models. The position of the current ROI is displayed in each 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:** EBI SCEA anatomogram for the large and small intestines allowing selection of any structure identified in the ASCT tables. **D:** Expanded view of the anal canal to show relevant cell-types and tissue organisation of that region.

Figure 5: HGCA in health and disease: **A)** Integration of datasets generated from intestinal tissues obtained from healthy individuals and patients with related diseases including IBD and colorectal cancer will enhance the future value of the HGCA. Single cell profiling of intestinal organoids followed by their integration in the HGCA will provide unique opportunities for translational research including regenerative medicine, drug testing and development. **B)** Summary of existing studies that have used sc profiling methods to human tissue samples obtained from healthy gut, colorectal cancer, Inflammatory Bowel Diseases and patient derived intestinal organoids.

Table 1: Template metadata for gut-related single-cell studies

Sample ID	
Patient ID	

Study design	
Summary	*(brief study summary)*
Patient numbers	
Sample numbers	
Patient inclusion criteria	
Sampling strategy	

Donor information		
Age		
Gender		
Ethnicity		
Medical condition	Gastro-intestinal (GI) related	
	Non-GI related	
Medication	Gastro-intestinal (GI) related	
	Non-GI related	
Dietary habits (if applicable)		
Sample Information		
Sample type		
Mucosal biopsy		
Surgical resection		
Deceased donor		
Macroscopical appearance of sample		
Normal		
Pathological (*Provide details)*		
Microscopic appearance of sample		
Normal		
Pathological (*Provide details)*		
Representative H&E stain available	Yes/No	*(Provide details of H&E and other stains)*
Anatomical Sample location		
Oral cavity		

Oesophagus	Proximal		
	Mid		
	Distal		
Stomach	Antrum		
	Body		
	Fundus		
	Pylorus		
Small bowel*	Duodenum	D1	
		D2	
		D3	
		D4	
	Jejunum	Proximal	
		Mid	
		Distal	
	Ileum	Proximal	
		Mid	
		Terminal Ileum	
Large Bowel*	Appendix		
	Caecum		
	Ascending		
	Transverse		
	Descending		
	Sigmoid		
	Rectum		

CCF Sample Location*			
Proximal Landmark		Distance (mm/cm)	
Distal Landmark		Distance (mm/cm)	

Sample processing

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	
Organoids generation	
-omics processing	
Data generation	
omics approach used	
RNA-seq	
scRNA-seq	
ATAC-seq	
Proteomics	
Whole genome sequencing	
Library preparation	
Sequencing platform	

*If available provide distance (in mm/cm) from nearest anatomical landmark: gastroduodenal junction, ileocaecal valve, hepatic flexure, splenic flexure, or anus.

Table 2 Summary of currently available computational packages.

Main categories	Task of analysis	Available tools/algorithms
Combining and integrating datasets 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., 2020)
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 cross talk	Cellular cross talk	CellChat, scConnect, CellComm, CellPhoneDB
	Microbial cross talk	(Kang et al., 2020) (Chattopadhyay et al., 2018) (Liao et al., 2020)
Regulatory network inference	TF regulatory network	SCENIC, GRNBoost2, PIDC

1051

1052