Title: Comparative Workbench for Atlas Data

Summary

The Human Cell Atlas resource will deliver single cell transcriptome data spatially organised in terms of gross anatomy, tissue location and with images of cellular histology. This will enable the application of bioinformatics analysis, machine learning and data-mining revealing an atlas of cell types and sub-types and ultimately disease conditions. However, to obtain an understanding of specific conditions, pathologies and histopathological phenotypes with their spatial relationships and dependencies we need a *spatial descriptive framework* to enable integration and analysis in spatial terms. In addition, tools are needed for the clinical domain expert to retrieve, analyse, visualise, and annotate the data and in particular *sharing* for expert-led collaboration, annotation and analysis. This is particularly important in the context of the histology of normal and diseased tissues within and across species to fully understand animal models of disease and for integration with human data.

Project Aims

We propose to extend tools and standards developed in the context of the Edinburgh Mouse Atlas with two specific aims:

1. To implement a focussed data visual-analytics "workbench" (Comparative Workbench for Atlas data or CWA) to provide the capability for clinical pathologists to analyse, annotate and compare complex cell-level imaging in conjunction with sequencing data analysis of the same or related tissues. This will provide access to the atlas data in tabular views enabling comparison between individuals, time-course data and cross-species coupled with management of configuration and "state" for tracking, saving and sharing the analysis process.
2. To develop a conceptual coordinate model of the application domain (small and large intestine) for spatial annotation of data samples and to develop the spatial descriptors for data retrieval, analysis and visualisation. The model will encode biologically and clinically relevant locations and regions and when a full HCA atlas framework becomes available the biological conceptual model will be mapped to the real-space coordinate frame. This mapping from biologically natural coordinates to real-space image coordinates has been demonstrated for the mouse embryo and the same techniques will apply here.

For specific focus, the CWA context will be normal small intestine and colon and pathology of Familial Adenomatous Polyposis (FAP) due to inherited *APC* mutations. For this we have tissue and data across four species: Min mice, Pirc rats, *APC*-mutant pigs and FAP humans. We will extend our OMERO-based prototype to provide a configurable tabular viewer of multiple "view-cells" (Figure 1). The CWA enables a structured visual-analysis that utilises both human and animal data to interrogate and understand normal physiological and pathological mechanisms. The data-samples and image-data will be annotated in spatial terms using the conceptual atlas framework and ontologies to develop the spatial-description standards needed to annotate both sample-based and image-based data and enable intra- and inter-species query and analysis. The gut exemplar provides a conceptually "simple" biological model with a complex mapping to adult anatomy that will test the requirements for spatial annotation and semantic integration needed for all tissues to deliver a fully integrated atlas framework enabling machine learning to cross the divide from sequence analysis to spatial organisation.

Prior Contributions

The Edinburgh University Division of Pathology (Professor Arends) will deliver the core datasets for the project and the workbench development. Heriot Watt University (Professor Burger) will develop the conceptual gut model and deliver the required atlas integration and

interoperability. Professor Baldock will contribute to both groups with his long-term expertise in digital atlases, coordinates systems and atlasing software.

The Human Cell Atlas DCP will be a multi-level, multi-modality, spatio-temporal data framework for terabytes of data. Our prior work on logic-based spatial descriptions of anatomy, specifically the '**straight mouse**' offers support for such a framework[1]. By defining natural coordinate axes in images and atlases that correspond to biologically relevant (anterior–posterior etc.) directions (fig 2), we are able to map between semantically rich descriptions of location (using spatial relations such as *adjacent to*, *lateral/dorsal to*, *overlapping with*, etc.) and the corresponding pixel/voxel sets in images/atlases and associated data to deliver robust data integration and interoperability. It will also provide a rich environment for NLP research on mappings between logic-based spatial descriptions and textual references facilitating data mining opportunities from location-based cell information to the scientific literature and large-scale image resources.

We (Edinburgh & Heriot-Watt) played key roles in the International Neuroinformatics Coordination Facility (INCF – www.incf.org) Atlasing Program culminating in the Waxholm Space Atlas and the Digital Atlasing Infrastructure[2]. Registration into a common spatial framework and query across distributed atlas resources were addressed, and data modelling and system architecture lessons from the DAI will be applicable to DCP.

Professors Baldock and Burger are involved in ELIXIR (www.elixir-europe.org), specifically the special interest group in atlases and images. ELIXIR is Europe's primary bioinformatics infrastructure programme and interoperability between the Human Cell Atlas DCP and the ELIXIR infrastructure will be enhanced by spatial annotation standards.

Professor Arends directs the Centre for Comparative Pathology and leads the colorectal cancer pathology research group, working on animal models of adenomatous polyposis pathology using well-characterised mouse[3], rat[4] and pig[5] mutant-*APC* models and human FAP for cross-species comparison. We have generated standardised datasets of histopathology images including gene expression data with RNA In-Situ Hybridisation (ISH) and Immunohistochemistry (IHC) and some DNAseq of mutational landscapes across these species that will be used to explore the potential for co-visualisation and linkage between multimodal datasets. For example, mutant-*APC* driven tumour burden and intestinal location are species-specific and can be quantified through quantitative analysis of gross and histopathological images, using intestinal atlas coordinates. Visualization of within-tumour distribution of cell proliferation indices (e.g. Ki67 IHC) and stem cell transcriptional markers (e.g. Lgr5 ISH) may reveal new correlations that might explain the architectural pattern differences.

Professor Baldock has led the Mouse Atlas programme[6] (www.emouseatlas.org) at the IGMM and developed the core concepts used by many spatial atlas resources internationally, anatomy ontologies and the 3D spatial frameworks and techniques for spatial data mapping[7] and visualisation[8]. These techniques led to the concepts and implementation of natural coordinates for spatial data analysis and integration[1]. In addition, he developed techniques for functional tissue unit analysis of cellular arrangement for physiological modelling[9].

Professor Albert Burger leads the Biomedical Informatics Systems Engineering Lab (BISEL) at Heriot- Watt University. He has closely worked with the Mouse Atlas program for over 15 years. His research expertise focuses on distributed bioinformatics systems, including integration of distributed, heterogeneous databases and web service infrastructures (specifically the integration of spatio-temporal biomedical atlases).

Proposed Work and Deliverables

There are two interlocked components to this proposal, a demonstrator interface for data analysis and the atlas model and associated spatial annotation framework for data analysis

across samples and between species coupled with visualisation in a spatial context, and analysis.

Edinburgh University lead: The Comparative Workbench for Atlas data is a configurable user-interface with a tabular view analogous to a “spreadsheet”, where each “cell” visualises a selected data type. The key properties of the workbench are user-control of the configuration, saving state and the ability to integrate the visual-analysis with collaboration. It is a digital work-bench for the biomedical scientist using existing techniques for data visualisation and interaction. Figure 1 illustrates the concept. Here we will implement the interface with the FAP data delivering an application demonstrator. The deliverable will be software of a functioning prototype, developed using best web-application standards of Javascript/Ajax with access to data/databases via web-services to exemplar spatially annotated datasets.

Heriot-Watt University lead: Develop and implement the conceptual model for the small and large intestine as a generalised cylinder with all layering to match the recognised histology. This is analogous to the straight-mouse embryo (fig 2) already developed. The model will include significant landmarks and regional organisation as required by anatomists and clinical pathologists and will be mapped onto the HCA spatial framework when available using transform techniques already in place[7]. This conceptual model in *biological space* will be used for spatial annotation of the data: tissue biopsies, imaging and transcriptome sequences across the exemplar-species. The deliverable is a mammalian spatial model for small and large intestines and an analysis of the minimum annotation requirements to correctly locate and query tissue and cellular samples across multiple samples and species. In the context of the HCA we will map this model onto the image-based framework to provide the natural coordinate mapping in image-space. This will be an example of tissue and system mapping that will be critical in capturing, querying and analysing the different data modalities contributing to HCA.

Evaluation and Dissemination

Evaluation of the workbench: 1) technical evaluation that the required functionality is delivered in terms of image and data visualisation and annotation, configuration management and saving, storing and collaborative of sharing specific views, and 2) UX testing with a small cohort of domain experts to test that it provides a useful mechanism for collaborative analysis.

Key evaluation of the spatial annotation in the conceptual framework is the *expressivity* of the annotation standards, i.e. can the locations as defined by experts be captured and the *accuracy* i.e. can the descriptions correctly locate the data so that spatial queries return data appropriately and spatial organisation (e.g. relative colonic position) is retained.

Dissemination via usual academic route of conference and meeting presentations and journal publication coupled with sharing within the HCA and CZI groups and meetings. The data will be archived and freely available following FAIR guidelines and all software open-source and available from GitHub.

Sharing Statement

All data used to develop the workbench and spatial annotations will be made freely available both directly and via the University of Edinburgh DataShare archive under a CC by 4 licence. All software will be open-source and maintained in a GitHub repository.

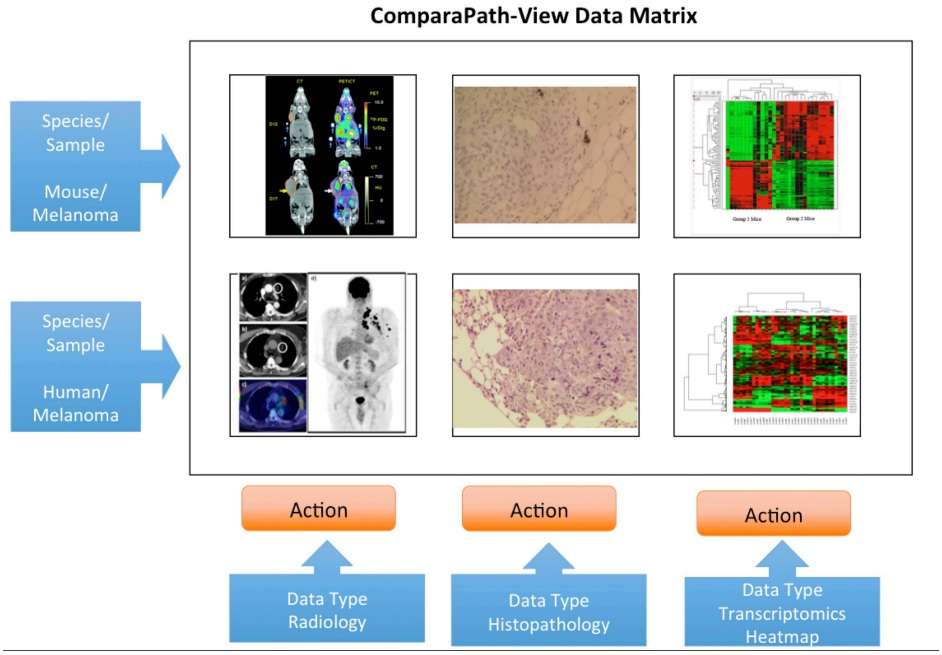

Figure 1. The concept of the ComparaPath-View interface is a configurable tabular view of selected visualisations of the data. Each "cell" of the interface can provide a range of visualisations including high-resolution (zoomable) images, 3D morphology data, confocal images, 3D anatomy, movie sequences, heatmaps, tabular data, networks interaction graphs and standard bioinformatic dataviews. For example using the open-source BioJS javascript software library specifically designed for browser integration. These will be complemented by the Edinburgh developments for 3D images and tile-based image visualisation. It is envisaged each "cell" view will allow drill-down by popup of a new window for higher resolution with full controls for the particular visualisation. For example view controls and annotation tools for 2D and 3D images and selection, filtering and analysis options for bioinformatics dataviews (heatmaps, networks etc). Data selection, view layout configurations and annotations will all be held centrally to enable recovery and sharing of specific views and sharing of annotations and comments. This is an annotation and visual analysis tool for domain experts to collaborate on and share the data analysis and interpretation.

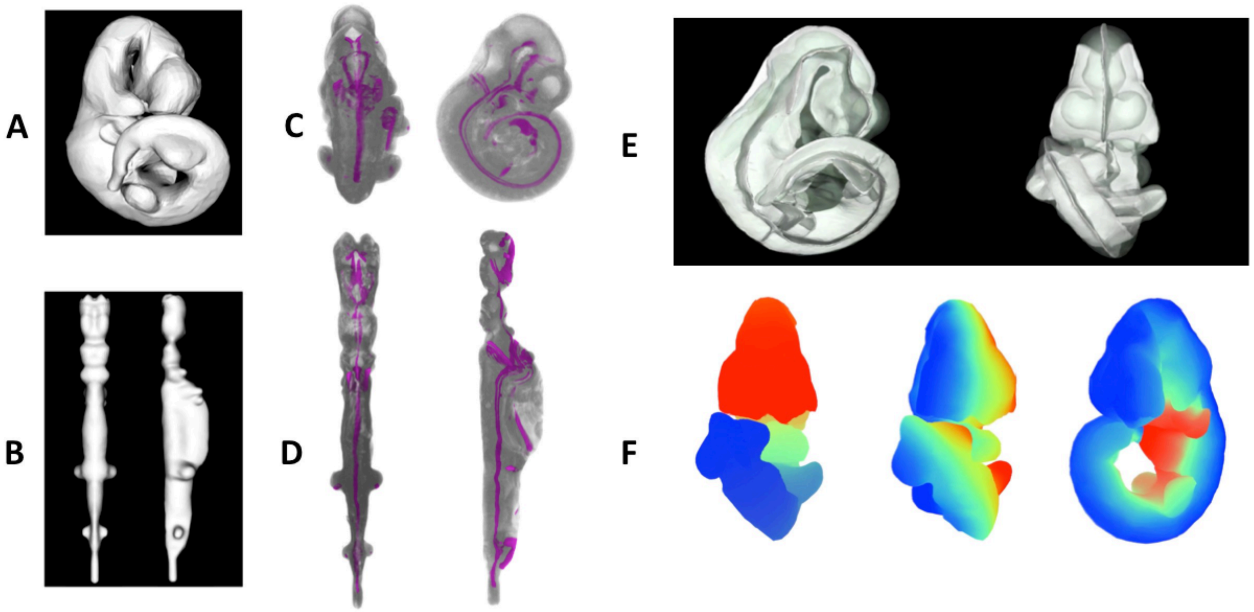

Figure 2. Biological coordinates of anterior-posterior (AP), left-right (LR) and dorsal-ventral (DV) for the developing mouse mapped to the complex 3D shape of a real embryo. These *natural* coordinates provide a spatial reference based on the underlying biology and therefore are a robust framework for cross data modalities, cross time courses and cross-species interoperability. The cartesian "straight" mouse embryo (B) is mapped onto a real embryo model (A) using the constrained distance transform. This allows expression patterns (C) to be mapped onto the biological coordinate frame (D). The midline, mid-dorsal-ventral and mid left-right planes are depicted in (F) and the coloured images (F) show AP, LR and DV coordinate values in the spatial context of a standard embryo model. This shows the principle of mapping between "real-world" image coordinates and the intrinsic biological or natural coordinates to enable robust spatial descriptions in terms of biology for comparison and interoperability.