



2nd September 2026

SPACE Data Platform

enabling high throughput spatial data studies at scale

Dedicated to the memory of Prof Greg Hannon

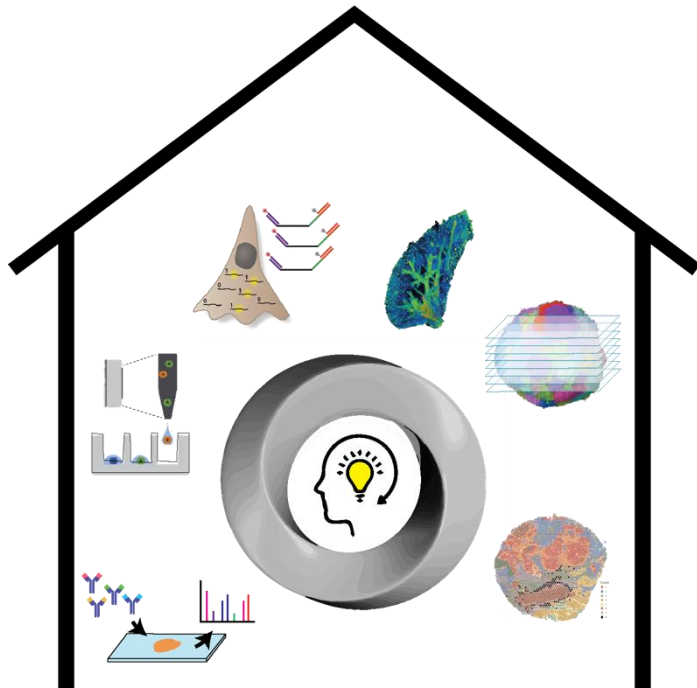
Nicholas Walton

Institute of Astronomy, University of Cambridge

Funded by

SPACE

Spatial Profiling and Annotation
Centre of Excellence



Spatial omics technologies are extremely powerful and can be **transformative** for cancer research

Can we identify mutant clones in large 3D organs and profile their growth during promotion?

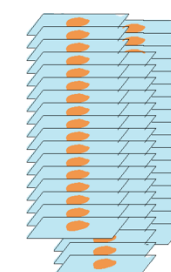
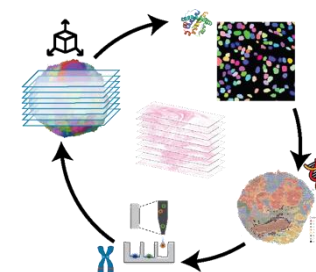
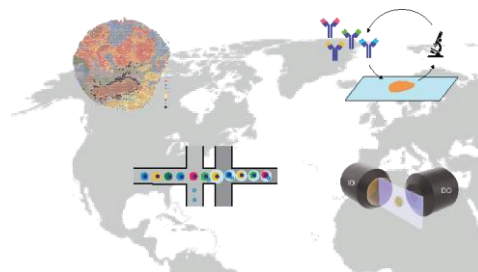
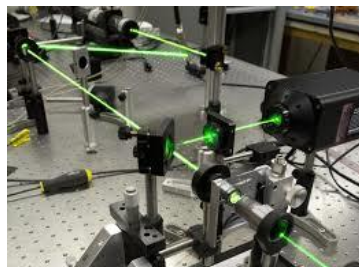
How heterogeneously expressed are the drug targets that I'm evaluating, and how could the local environment affect therapy?

Are there spatial predictors for drug resistance or relapse?

How do different tumour clones drive the formation of different tumour micro-environments?

SPACE evolved from the earlier Cancer Grand Challenge IMAXT project (2017-2023)

Spatial biology can be difficult: data is a bottleneck



Method
choice and
tech
availability

Technology
development
and optimization

Fragmentation

Multi-modal
integration

Throughput

Data
management,
visualization,
analysis

SPACE is a "one stop shop" for applying spatial biology to cancer research

Project ideation

Choice of technology

Sample identification
experimental design

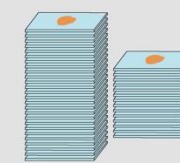
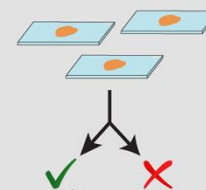
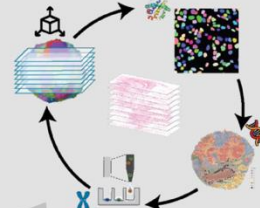
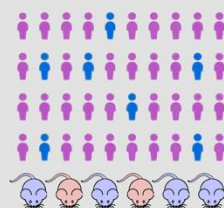
Technology optimization
and development

Proof of concept

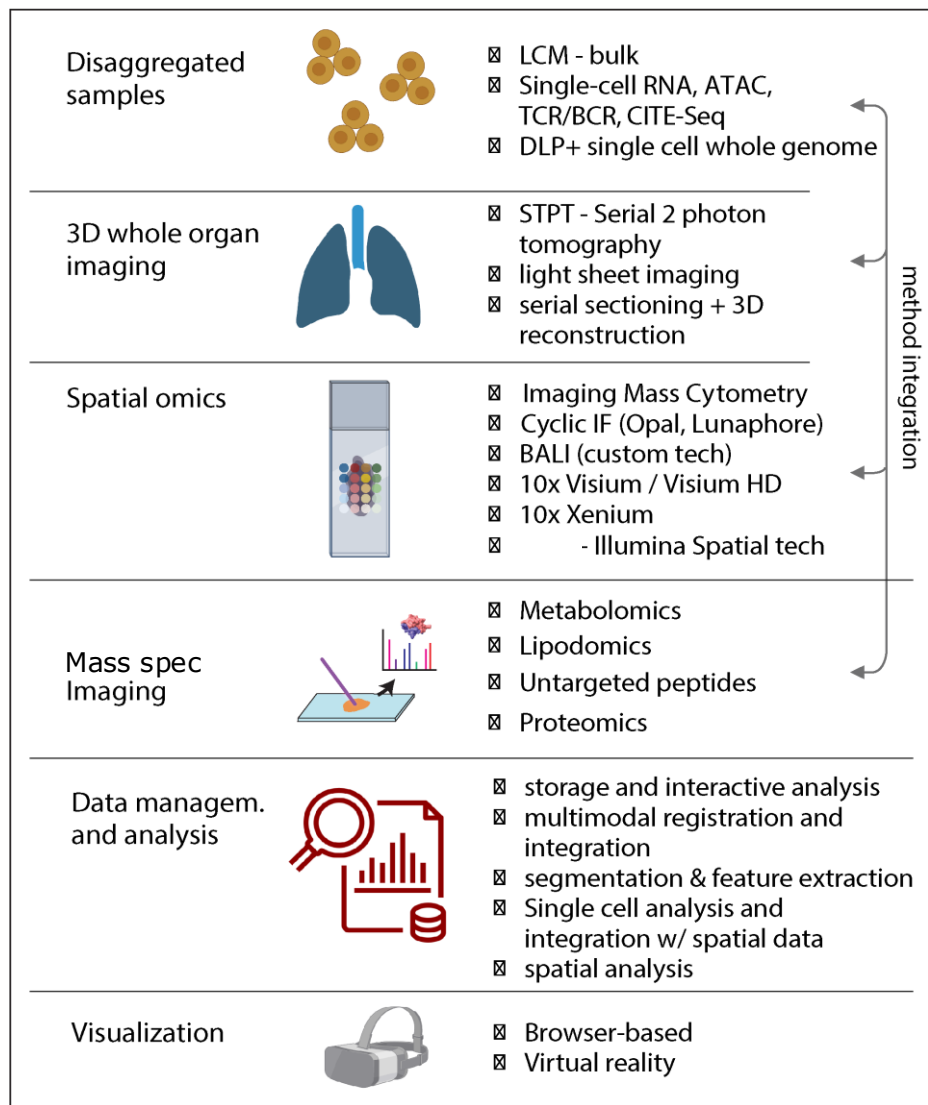
Data generation

Data analysis
and management

Conclusions
and new insights



Technology platform: samples to data



- High throughput 3D whole-organ imaging with integrated sectioning and section collection for downstream profiling
- Multiple options for spatial transcriptomics profiling
- Large capacity for multiplexed immunohistochemistry (up to 40 markers per run), with a large library of pre-optimized markers
- Integration of multiple technologies on same slide (multiplexed IHC + spatial transcriptomics + H&E)
- Mass spec imaging: metabolomics/lipidomics and peptides
- SPACE Data centre

A brief aside: looking up and out rather than down and in ...

In astronomy: telescopes rather than microscopes

<https://camcead.ast.cam.ac.uk/>

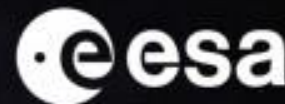
SPACE

Spatial Profiling and Annotation
Centre of Excellence



SKY-SCANNING COMPLETE FOR ESA'S MILKY WAY MAPPER GAIA

SPACE leverages expertise from 'big' space missions



From 24 July 2014 to 15 January 2025, Gaia made more than three trillion observations of two billion stars and other objects, which revolutionised the view of our home galaxy and cosmic neighbourhood.

3 TRILLION
Observations

2 BILLION
Stars & other objects observed

938 MILLION
Camera pixels on board

15 300
Spacecraft 'pirouettes'

55 KG
Cold nitrogen gas consumed

580 MILLION
Accesses of Gaia catalogue so far

3827
Days in science operations

13 000
Refereed scientific publications so far

2.8 MILLION
Commands sent to spacecraft

142 TB
Downlinked data (compressed)

500 TB
Volume of data release 4
(5.5 years of observations)

50 000 HOURS
Ground station time used



Gaia: Currently at Release Gaia DR3

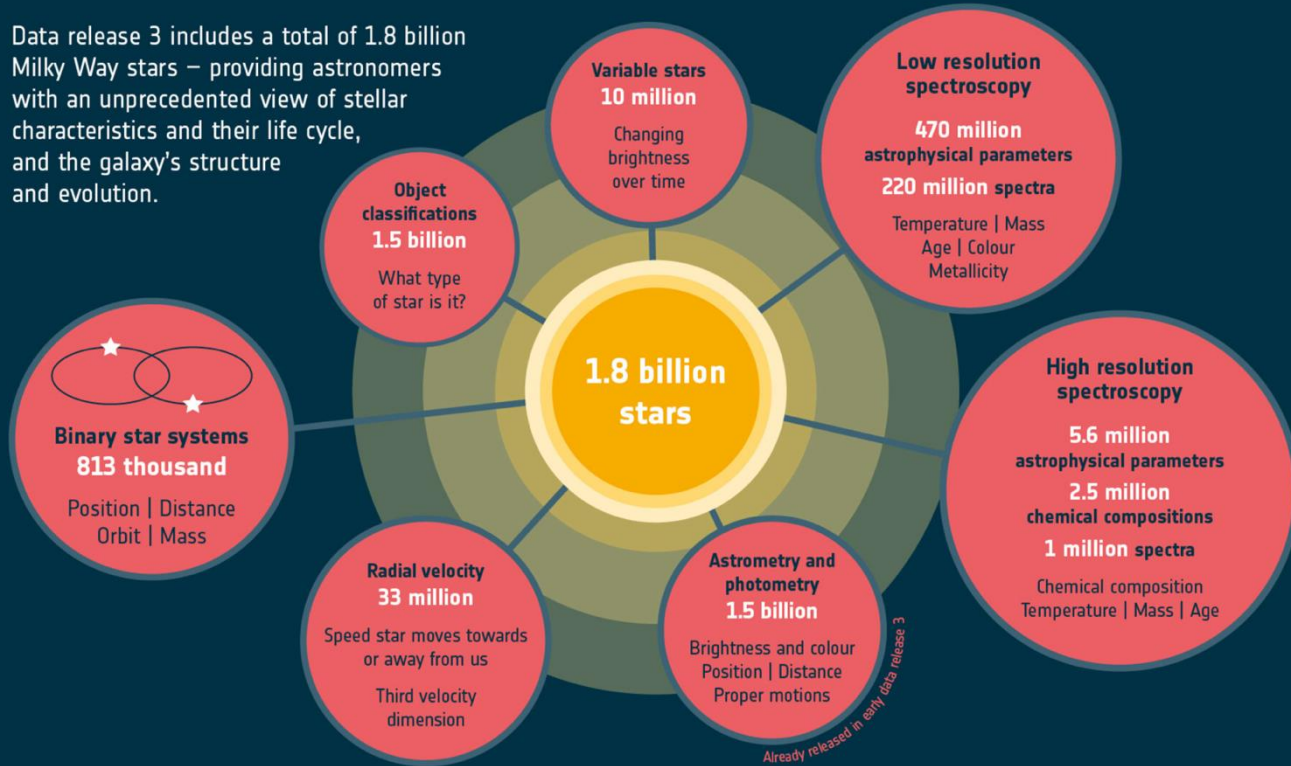
SPACE draws on Cambridge expertise in Gaia, where we lead (spectro)photometric analysis



MILKY WAY STARS

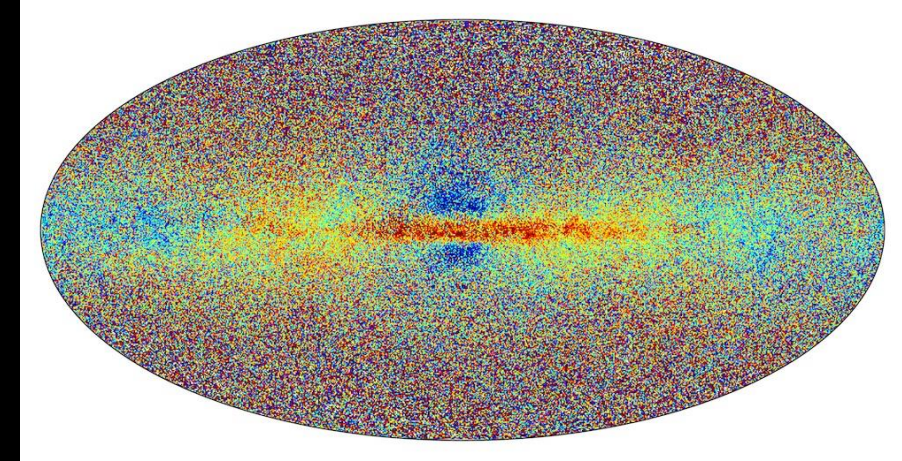


Data release 3 includes a total of 1.8 billion Milky Way stars – providing astronomers with an unprecedented view of stellar characteristics and their life cycle, and the galaxy's structure and evolution.



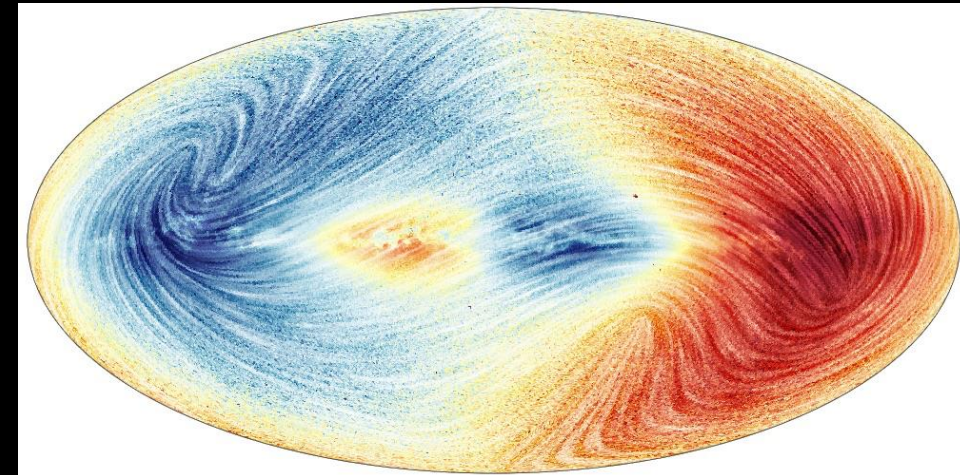
Gaia EDR3/ DR3: with larger to come

Next release Gaia DR4, ~2B high quality (~2.8B total) stars, epoch everything, 2nd Dec 2026



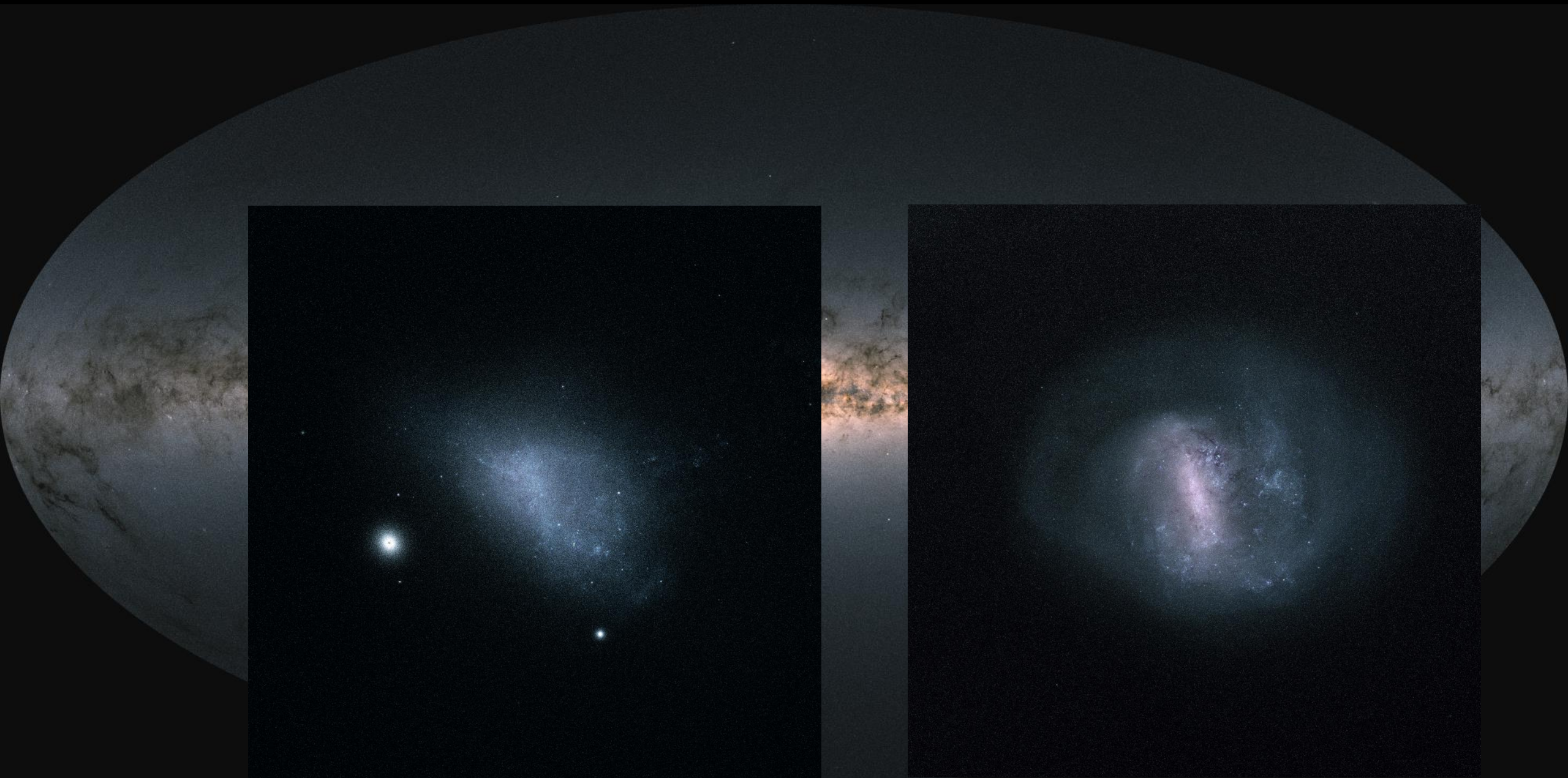
Credit: ESA/ Gaia/ DPAC

Stellar abundances



Stellar velocities ...

Credit: ESA/ Gaia/ DPAC



Gaia EDR3 flux colour (~1.5 Billion stars) /Credit ESA/Gaia/DPAC

Gaia Science

the most scientifically productive ESA science mission, in part because the data and data flow is described and accessible to all

μarcsec-level
astrometry:
direct distance
measurements
across the
Galaxy (20 kpc
($\sim 6 \times 10^{17}$ km))
reaching ~ 10
μas precision for
bright stars.

Gaia science spans most of astrophysics from
solar system asteroids, stars, the Milky Way, dark matter, to fundamental physics
Science highlights: <https://www.cosmos.esa.int/web/gaia/highlights-of-gaia-dr3>

DATA IN ASTRONOMY AS AN EXAMPLE: FULL LIFE CYCLE PLANNING FROM THE START

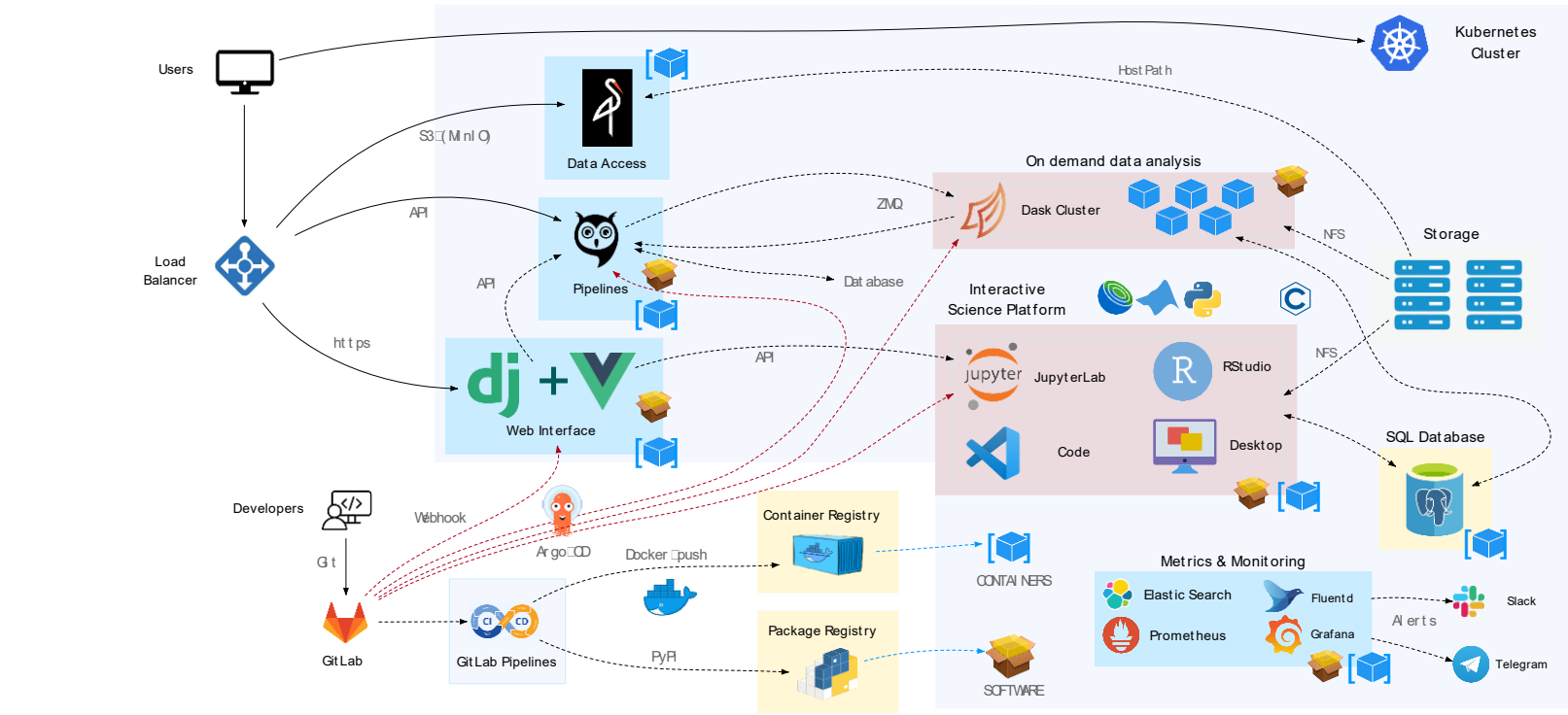
- Strong focus in planning new experiments from science requirements to full end data and science data delivery
- Requirements → Design → Implementation → Operations → Archive
- Essentially all data from astronomy is, or rapidly, becomes openly available to all
 - Facilities and experiments can have proprietary periods (typically 1 year) but then the data is made accessible globally
 - FAIR access to science data products
- Requires that data conforms to standard formats, tools are available to read/write the data, the data is discoverable and accessible via an archive or programmatic API, documented, help available
- Data policies apply to new experimental (and increasingly theory) facilities both large and small.

All images: Credit: ESO

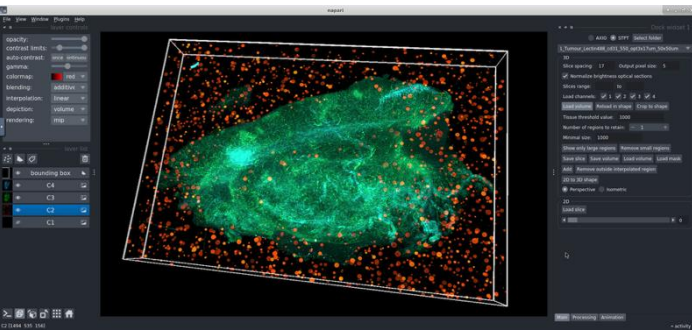
Facilities such as the ESO Extremely Large Telescope (39-m diameter mirror) are expensive! All data is precious.

THE SPACE ARCHITECTURE AND DATA FLOW

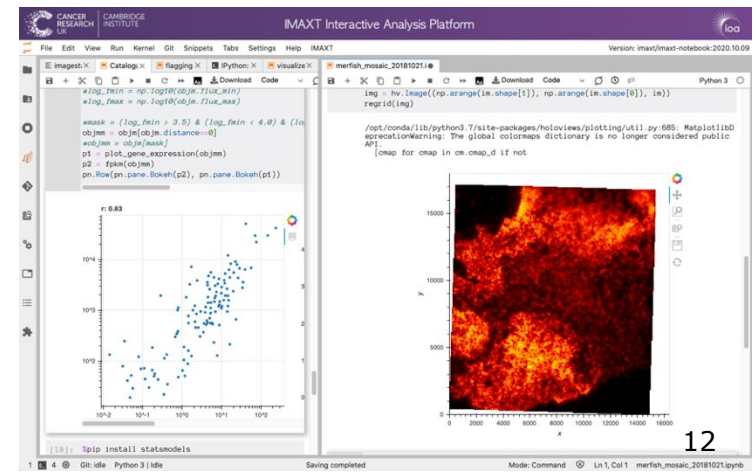
A SCALABLE SCIENCE END TO END SYSTEM



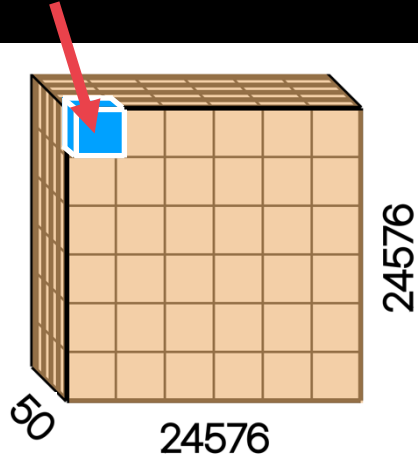
- Data storage: Local storage with remote access using S3 (Simple Storage Service) protocol
- Kubernetes cluster: Automates the deployment, scaling and management of containerised applications
- Data processing: Data pipelines that run on the datasets to make them scientifically usable
- Data analysis: Run analysis in the cluster using Jupyter Notebooks, R (-Studio), Fiji, or any other tool.



SPACE analysis platform hardware:
12 PB disk, 2000 vCPUs, 11 TB RAM, 14 GPUs
Full user documentation
(Gonzalez-Solares et al. 2023, Biological
Imaging 3:e11, DOI
10.1017/S2633903X23000090)

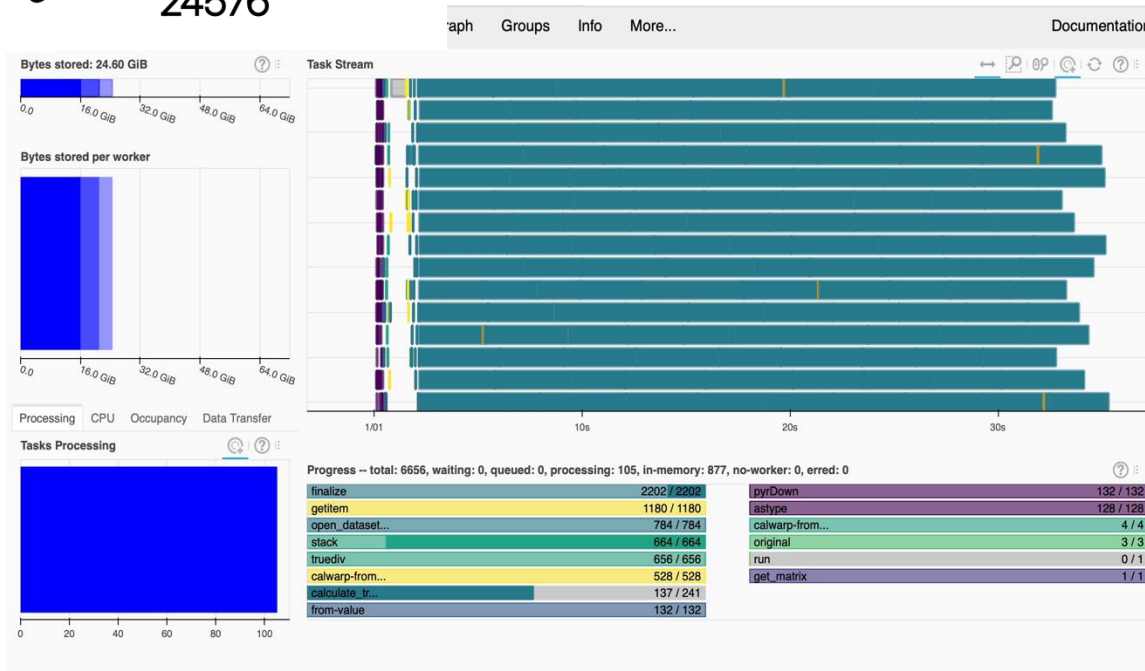


SPACE PLATFORM COMPONENTS



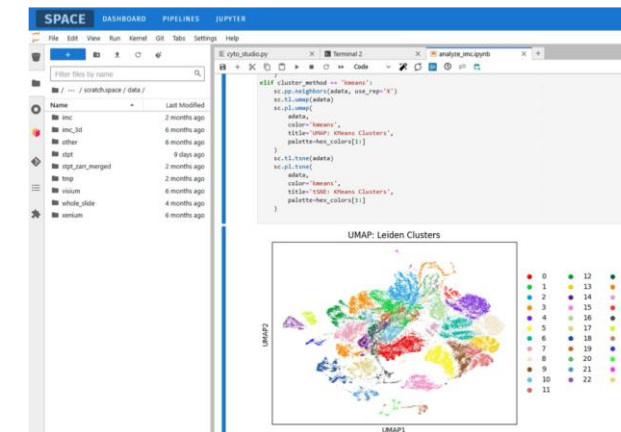
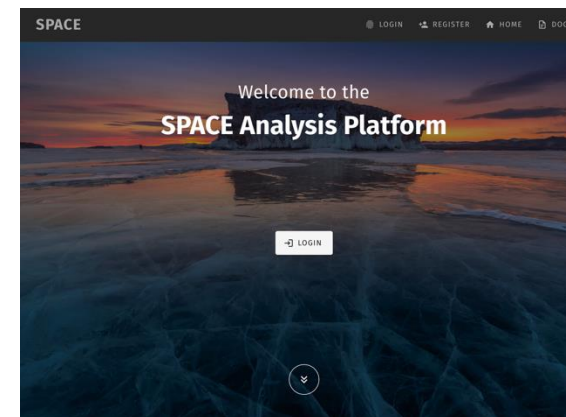
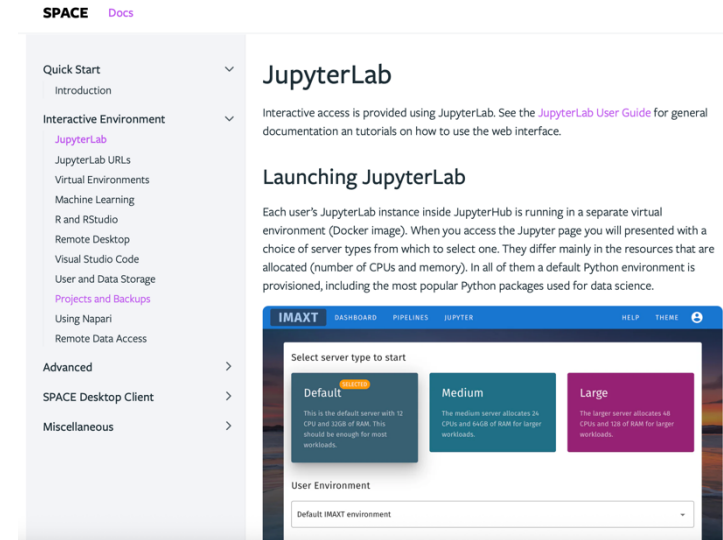
Data Formats: Use of Zarr to handle large multidimensional arrays

Data Pipelines: OWL scheduler for submission of pipelines to the cluster



Parallel and Distributed Computing to give performance and scalability

Documentation: provides guidance and help – along with a ticket system to support end user queries.



User Access: interactive processing. 'slurm' jobs via OWL.

SPACE DATA UPLOADER AND ARCHIVE END TO END METADATA CAPTURE AND QUERY

SPACE DESKTOP CLIENT

User Login

Username

Password

2FA Code

Log in

Provenance: every dataset traceable from sample to result.

Data upload from end user experiment to the server. Data model developed for key experiments (e.g. antibodies)
Data available in queryable archive. Performance ensured as all via SPACE cloud

Centre CEN-0001
CRUK Cambridge Institute
Cambridge CB2 0RE, UK

Instrument INS-0001
STPT
STPT instrument at CI

Sample GL1-0006
PAR1221X2GFP-X1_AN20-005271_FFPE1

Mouse ID: AN20-005271
Mouse Strain: NSG-EGFP
Sample type: FFPE
Sample source: Mouse PDX
Embedding method: Paraffin
Extraction method: Necropsy
Fixation method: 10% NBF
Fiducials type: Unknown
Cryoprotection type: None
Induction method: Subcutaneous fragment implantation

Study STUD-0002
ClonexSeq

Scan STPT-190516-1317
20190516_411_T_tc_day20_Dil_Lungs_100x1Sum_set1
Sample information
METADATA
Channels: 4
Sections: 100
Optical layers: 1
Thickness: 15µm
Tiles: 8 x 6
Original data: s3://imaxt/incoming/
Raw Zarr data: s3://imaxt/raw/
Related entries
STPT-190516-1317
STPT-190513-1251

Experiment EXP-0001
200918_BcNSG_411ElvinParental
Injected 100k 411 parental cells from Elvin 2013 (p15) into 3 Balb/c (1R, 1L, 2R) and 3 NSG (NM, 1R, 1L). Let grow for 21 days.

Analysis ANA-0020
JobID: 741
Status: FINISHED
Pipeline: stpt-mosaic (v0.6.4)
User: daemon
1/2 of 2

SPACE Antibody Portal

Welcome to SPACE Antibody Portal

ARCHIVE

Antibody Validations

+ Add ✎ Change

Antigens

+ Add ✎ Change

Clones

+ Add ✎ Change

+ Add ✎ Change

+ Add ✎ Change

View

View

Home > Archive > Antibody Validations

Start typing to filter...

ARCHIVE

Antibody Validations + Add

Antigens + Add

Clones + Add

Metals + Add

TOOLS

Choices

Search

Select Antibody Validation to change

Q

Action: -----

☐	NAME	TISSUE
☐	Clone1/143Nd/1.2/lung	lung
☐	Clone1/143Nd/2.0/lung	lung
☐	Clone1/143Nd/3.0/lung	lung
☐	Clone2/143Nd/1.2/lung	lung
☐	Clone3/143Nd/2.0/lung	lung

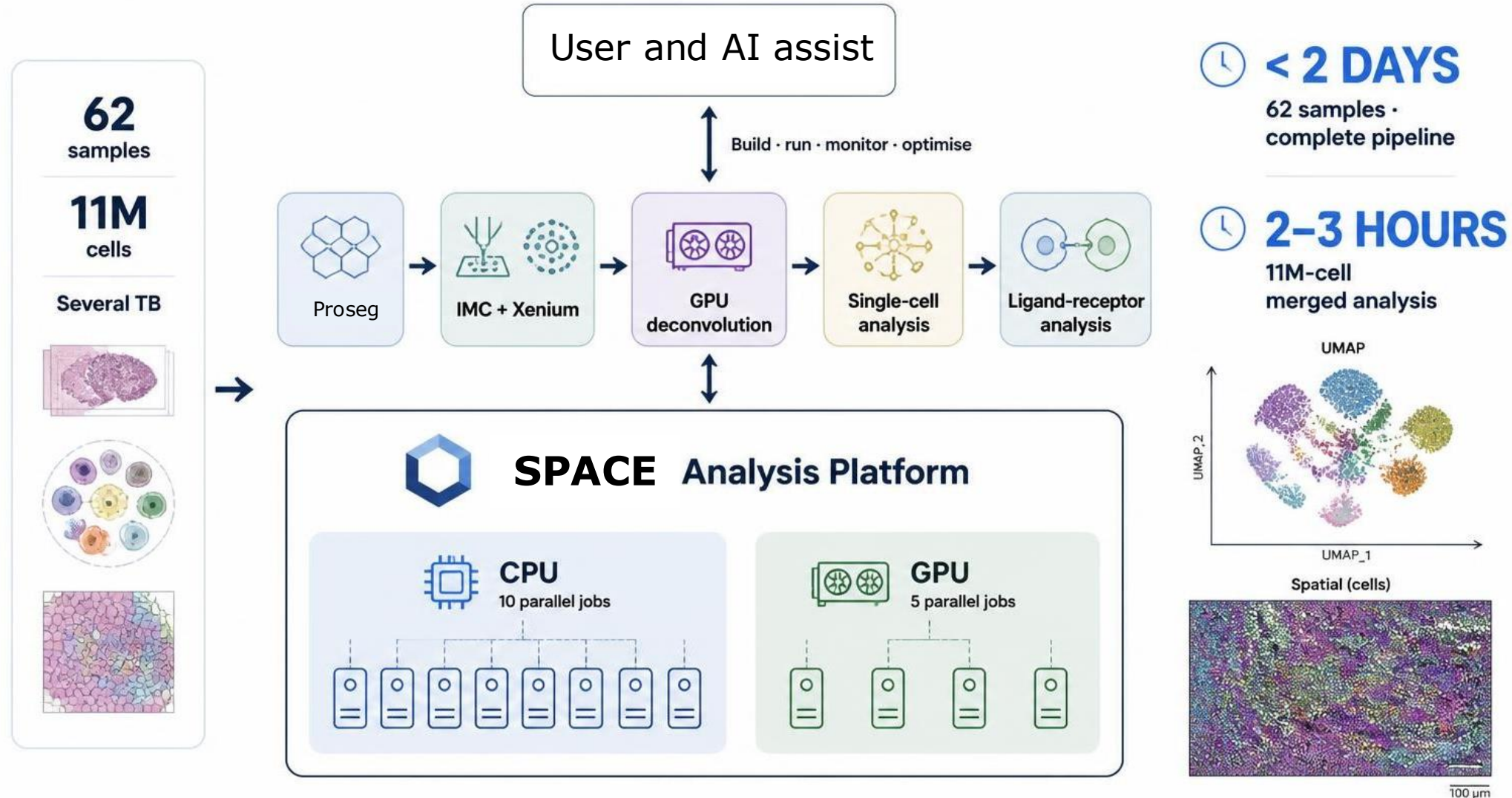
5 Antibody Validations

All

By metal

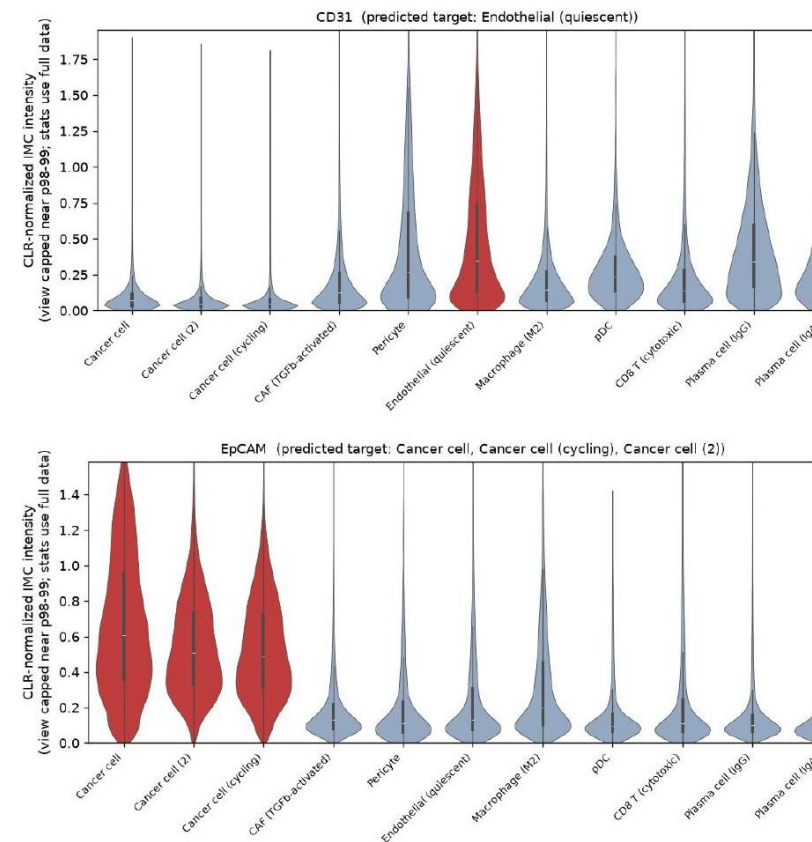
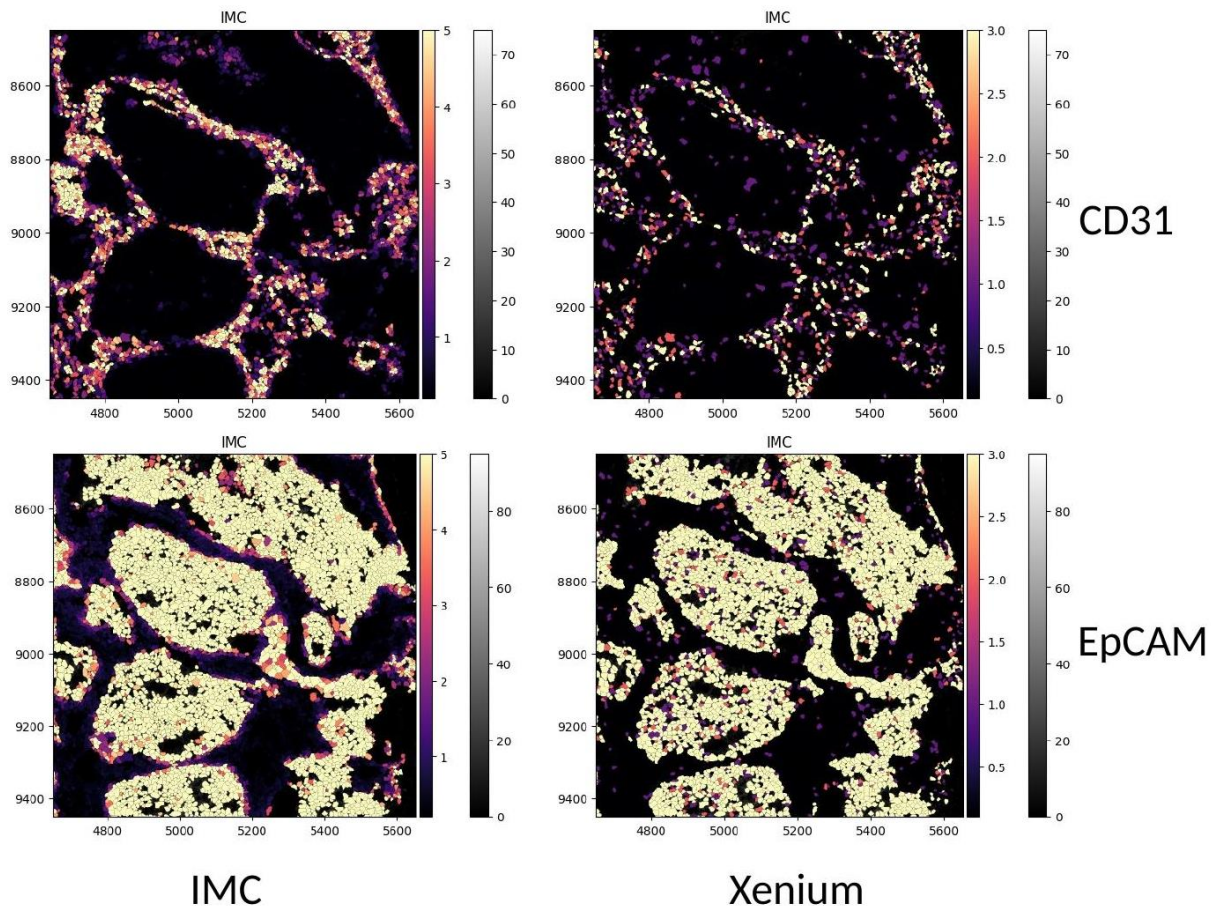
Initial Biology: KS, I

CASE STUDY: 62 SAMPLES, 11 MILLION CELLS, ONE INTEGRATED ANALYSIS



CASE STUDY QC: IMC AND XENIUM COREGISTRATION IS ACCURATE AT THE SINGLE-CELL LEVEL

Dario Bressan's talk yesterday



For most IMC markers, single-cell expression recapitulates the Xenium data across the panel — segmentation and coregistration validated at single-cell level

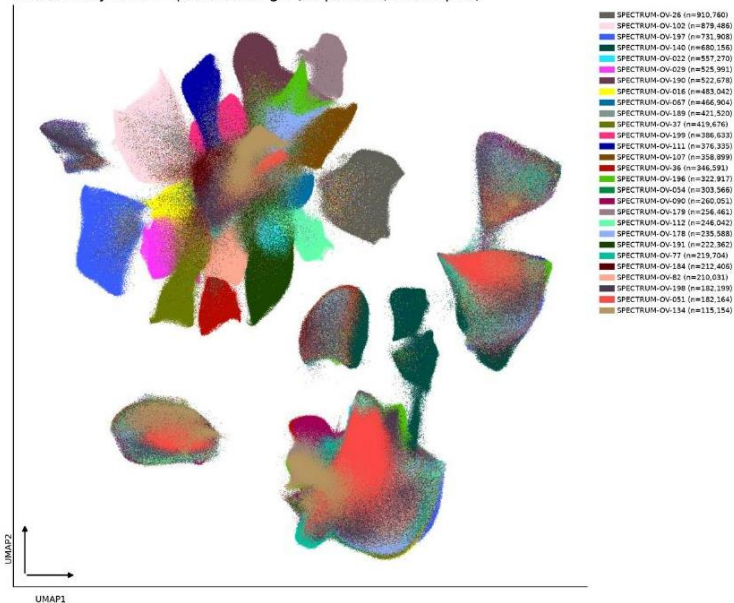
Unpublished data · Samples: MSK SPECTRUM cohort (collab. Shah lab, MSKCC)

2 Sep 2026

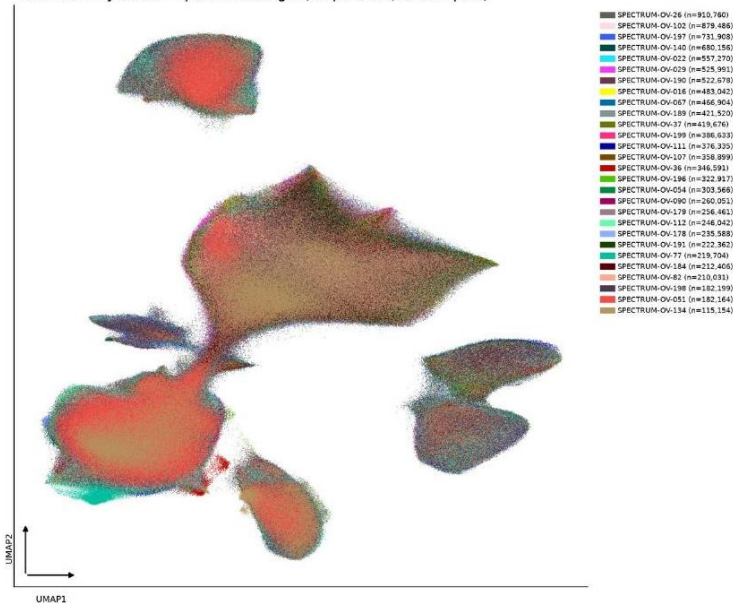
SPACE Data - Walton @ CGC Spatial Biology: RS, London, UK

CASE STUDY RESULTS: 11M CELLS, 62 SAMPLES, 28 PATIENTS

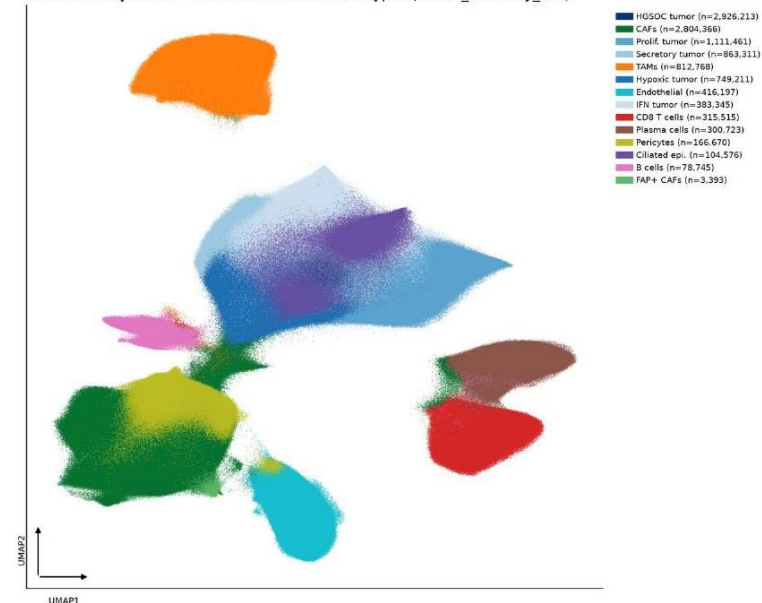
Pre-Harmony UMAP — patient of origin (28 patients, 62 samples)



Post-Harmony UMAP — patient of origin (28 patients, 62 samples)



Post-Harmony UMAP — de novo annotated cell types (leiden_harmony_r07)



Pre-integration (coloured by patient) → batch-corrected → de novo annotated cell types

<2 days end-to-end / 11M cells

2–3 hour 11M-cell merged analysis (10 CPU + 5 GPU parallel jobs)

Cell types called by LLM from marker expression, human-verified

What AI assisted automation does

The researcher describes the analysis in natural language; an AI assistant builds, runs, monitors and optimises the pipeline on the SPACE platform

The AI uses the same interfaces as human users, OWL scheduler, Dask cluster, archive queries, so every action is visible and auditable

This optimises the workflow generation process: pipeline construction, parameter tuning, job monitoring, failure recovery, results summarisation

Trust

Human-in-the-loop: the expert researcher reviews the plan, the generated code, and QC outputs at each stage

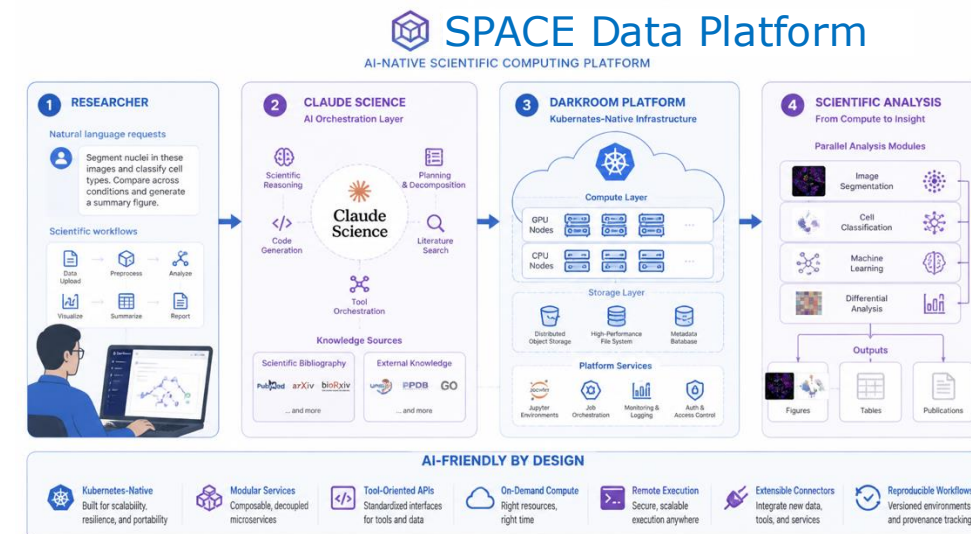
Every step versioned and logged in the data flow, full provenance from prompt to result

Outputs are benchmarked against manually built reference pipelines; the raw data is not modified

What AI improves

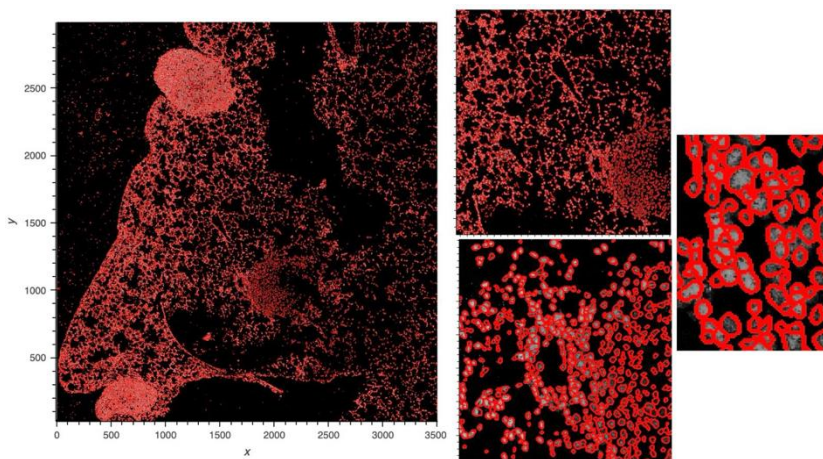
Case study (previous slides): the 62-sample pipeline was built and run this way, a complete analysis in < 2 days rather than weeks

Analyst time shifts from plumbing to interpretation and validation



SPACE TOOLS AND APPLICATIONS FULLY INTEGRATED IN THE DATA FLOW

- Range of analysis tools offered through the SPACE platform
- Includes advanced 'consensus' segmentation, UberSeg (Irfan et al, 2025 10.3389/fgene.2025.1547788)
- Creation of virtual H&E images from multiplexed data (e.g. IMC)
- Server based visualisation (e.g. Napari, QuPath)
- User provided python or R analysis
- Access to GPU to support AI/ML



Spatial correlations within multiplexed datasets

M. O. Irfan, E. A. González-Solares, T. Whitmarsh, A. Molaeinezhad, M. Al Sa'd, C. M. Mulvey, M. Páez Ribes, D. Bressan and N. A. Walton

ABSTRACT

Single cell segmentation followed by clustering is a common approach for identifying correlations in multiplexed imaging datasets. We investigate an alternative strategy for quantifying correlations across imaging channels exploiting 2D Fourier transforms. This technique requires only images and is capable of highlighting biologically significant regions across large multiplexed datasets.

METHOD

The 2D Fourier transform of an image reveals the image power distribution (P) over angular scales. The power correlation coefficient (-1 to 1):

$$r = \frac{P_{ij}}{\sqrt{P_i P_j}}$$

between two channels of the same image can be used to identify common features at different spatial scales. In Fig 1. we show the correlation between six different IMC channels for a BRCA2 breast cancer tissue sample¹. The largest spatial scales that can be probed are capped by the image size itself, as the power measurements approach these scales, the error bars increase. Strong correlation across the largest image scales are seen between the channels which target Progesterone Receptor (PR) A/B, Epithelial cell adhesion molecule (EpCAM) and ribosomal protein S6; the latter two being markers of epithelial tumour cells and cell proliferation whilst the presence of PR-A/B can be used as a prognostic tool.

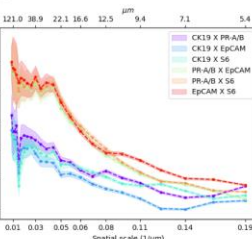


Fig 1. Correlation across spatial scales for a single breast cancer tissue sample as imaged by six IMC channels. The shaded regions denote the uncertainties.

To rapidly assess the strongest correlations within a large number of tissue samples, the average correlation coefficient (across all available spatial scales) can be used. For Fig 2. we calculated the average correlation between IMC channels for 230 BRCA2 images¹. We define our threshold for two signals being strongly correlated as 0.68 and in Fig 2. we plot the number of images where two channels display strong correlation with each other. We include structural markers such as the DNA1, DNA2 and Histone channels, which all highlight nuclear detection, to verify our technique. We expect each channel to be 100% correlated with itself as well as strong correlation between the nuclear marker channels. In addition to the expected, we see strong correlations between CK19 and CK8/18 and panCK, markers which identify breast cancer tissue as well as S6 and c-erbB-2-Her2 and Progesterone Receptor A/B, markers which define the tumour proteome.

Fig 2. A channel-channel matrix showing the number of images where two channels display a correlation coefficient greater than 0.68 with each other.

Larger images can be portioned into subsections in order for the correlation coefficient to highlight specific regions of the tissue which would merit further investigation, such as cell segmentation. In Fig 3. we show a single mouse-lung tissue sample as seen at 1 micron resolution by two IMC channels. The average (over available spatial scales) correlation coefficients within subsections of the images reveal notable correlation in the bottom right of the image between these two channels. The channels selected provide markers for S6 and CD68 which, respectively, highlight tumour regions and immune response cells.

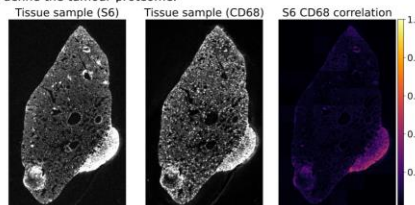


Fig 3. A single image of mouse-lung tissue seen in two IMC channels alongside the average correlation coefficient across patches within the images.

CONCLUSION

We have adapted the technique of cross-correlation power spectra, common within astrophysics, for application to biomedical images from multiplexed datasets. This Fourier-based analysis can quickly identify channels and regions of biological interest across large numbers of tissue samples and provides annotation insight without the need for cell segmentation.

1) Jackson et al. (2020)



A Toolkit for Generating Virtual Brightfield Images of Histological and Immunohistochemical Stains from Multiplexed Data with AI-Based Channel Selection and Image Enhancement

Tristan Whitmarsh, Mohammad Al Sa'd, Eduardo González Solares, Alireza Molaeinezhad, Melis O Irfan, Claire Mulvey, Marta Paez-Ribes, Atefeh Fatemi, Wei Cope, Kui Hua, Gregory Hannon, Dario Bressan, Nicholas Walton

BACKGROUND

Multiplex imaging reveals rich molecular and spatial information in tissue, but clinical brightfield imaging is still often required.

Acquiring additional stained sections can be impractical when tissue is limited or unavailable and may require additional registration steps.

We present a framework for generating virtual brightfield images directly from multiplexed imaging data, removing the need for additional staining or sectioning. Using a physics-based stain model, molecular markers are mapped to realistic stain colors, producing images that closely resemble conventional histology.

METHODS

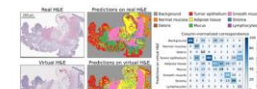
Multiplex channels are normalized and combined into stain classes to define virtual stain intensity maps. A physics-based light transmission model simulates brightfield image formation to generate virtual H&E, IHC, and custom stains. Marker to stain assignments are defined using curated presets or optionally selected using a large language model. Post-processing includes optional denoising and super-resolution enhancement to improve visual quality while preserving structural detail.

Imaging Mass Cytometry (IMC) image of a mouse embryo with virtual brightfield image regions of custom special stains.

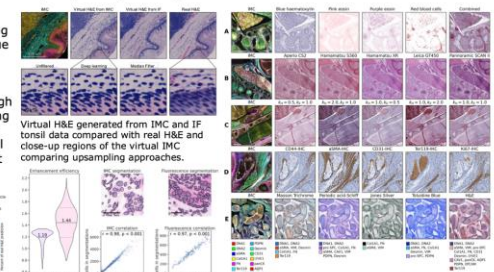
RESULTS

Quantitative comparisons showed strong agreement in pixel level similarity, tissue classification, and cell counts between virtual and real H&E.

Pathologist evaluation demonstrated high concordance with clinical tumour grading and deep learning-based enhancement improved noise reduction and structural detail. Channel selection was consistent between large language models.



Real and virtual H&E from CyCIF data with TIAtoolbox tissue-type predictions and aggregate tile-level correspondence across slides shown as a column-normalized confusion matrix.



Enhancement efficiency (edge preservation vs. noise reduction) comparing deep learning and median filtering, with StarDist segmentations and cell-count correlations of to hematoxylin (k_1) and eosin (k_2). D: Virtual IHC ground truth for IMC and fluorescence data. E: Special stains.

CONCLUSION

- Virtual brightfield images can be generated directly from multiplexed imaging data
- This eliminates the need for additional tissue staining or sectioning

Code available on GitHub at github.com/TristanWhitmarsh/multiplex2brightfield and as python package on PyPi (`pip install multiplex2brightfield`).

Accepted for publication in Frontiers in Bioinformatics



Irfan et al 2026 in prep

SPACE Data - Walton @ CGC Spatial Biology: RS, London, UK

Whitmarsh et al, 2026, Front. Bioinform. 6:1765143. doi: 10.3389/fbinf.2026.1765143

SPACE Poster and Talks at this meeting

- In search of the hidden "omics", Spatial analysis of the tumour micro-environment across scales and molecular domains (Dario Bressan – talk)
- BALI – Barcoding by activated ligation of indices (Giorgia Battistoni – talk)
- Investigation of low-to-high grade transition in glioma (Mirta Vojnovic)
- Breast cancer metastasis TME response to therapy (Claire Mulvey)
- Spatial investigation of neuroendocrine tumours (Marta Paez-Ribes)
- Opal Studio, a new tool for multiplexed ISH analysis (Tristan Whitmarsh)
- Spatialstudio, AI-assisted spatial omics investigation (Kui Hua)

2 Sep 2026

SPACE Data - Walton @ CGC Spatial Biology: RS, L

Opal Studio: An Interactive Spatial Biology Workbench for Exploring Cell Segmentation and Phenotyping Methods

Tristan Whitmarsh, Eduardo González Solares¹, Alireza Molaeinezhad¹, Mohammad Al Sa'di¹, Claire Mulvey², Marta Paez-Ribes², Atefeh Fatemi², Dario Bressan², Nicholas Walton²

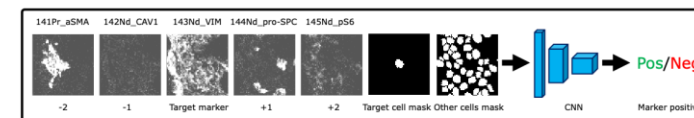
¹ Institute of Astronomy, University of Cambridge, Cambridge, UK. ² CRUK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge, UK

BACKGROUND

The analysis of highly multiplexed imaging data, such as imaging mass cytometry (IMC), requires repeated visual assessment of cell segmentation and phenotyping. These steps are often time consuming and distributed across separate tools. Opal Studio integrates these workflows in a single interactive environment for rapid exploration and comparison of spatial analysis methods.

MARKER POSITIVITY METHOD

For each cell, the positivity for a marker is determined using a convolutional neural network (CNN) where the input includes the target marker image and the cell mask derived from the cell segmentation. Neighboring cell masks are included to account for signal from neighboring cells, similar to CellTune⁹, while adjacent mass channels are included to account for IMC-specific channel spillover. The current model was trained on manually annotated IMC cell-marker positivity labels augmented with artificial channel spillover.



EVALUATION

Eleven paired mouse-lung ROIs from two samples were analyzed using identical segmentations and phenotype gates. Marker positivity was determined using either the AI-based method or Otsu thresholding. Cell-type separation between manually delineated lesions and normal lung tissue was quantified using Fisher's criterion (J).

OPAL STUDIO FEATURES

- Multiplexed images can be loaded from OME-TIFF or SpatialData¹, with fully adjustable visualization and overlays for segmentations, marker positivity, and phenotype masks.
- Image processing tools include channel merging, denoising, histogram equalization, and virtual brightfield image generation².
- Regions of interest can be manually drawn for targeted analysis.
- Cell segmentation tools include default and IMC-trained models for **InstanSeg³**, **Mesmer⁴**, **StarDist⁵**, **Cellpose⁶**, **Omnipose⁷**, as well as a classical **Watershed** implementation.
- The segmentations can be applied to only the cells in the view area, or a manually defined region for rapid comparison, then processed and merged to create composite segmentations⁸.
- Phenotyping is performed using threshold- or AI-based marker positivity determination followed by rule-based gating.
- Projects can be saved as SpatialData for downstream analysis

RESULT

AI-based marker positivity produced stronger lesion-lung separation (higher J).

	Tumor epithelial cells		Mean of all cell types	
	Otsu	AI	Otsu	AI
Sample 1	0.196	0.570	0.050	0.077
Sample 2	0.369	0.542	0.046	0.082
Average	0.283	0.556	0.048	0.080

¹ Marconato et al., Nat. Methods (2025) ² Whitmarsh et al., Front. Bioinform. (2026) ³ Goldsborough et al., arXiv (2024) ⁴ Greenwald et al., Nat. Biotechnol. (2022) ⁵ Schmidt et al., MICCAI (2018) ⁶ Stringer et al., Nat. Methods (2021) ⁷ Cutler et al., Nat. Methods (2022) ⁸ Irfan et al., Front. Genet. (2025) ⁹ Bussi et al., bioRxiv (2025)

CONCLUSION

- Opal Studio provides an interactive environment for spatial image analysis across any multiplex imaging modality, from segmentation to cell phenotyping.
- A marker positivity classifier, trained on IMC data, achieved better separation of lesion and normal lung cell types than conventional Otsu thresholding.

Code available on GitHub at <https://github.com/TristanWhitmarsh/opal-studio>

Consortium portals

(data coordination at scale)

HTAN (US, cancer): ~8,500 biospecimens, ~2,050 participants, 20+ assays incl. multiplexed imaging (*Nat Methods* 2025)

HuBMAP (US, healthy tissue): ~9,250 public datasets, 25 data types, 29 organs (*June 2026*)

NCI Imaging Data Commons: ~100 TB cancer imaging incl. multiplexed fluorescence; >95% covered by the permissive CC-BY licence

Archives & standards

(the common language)

EBI BioImage Archive: >1.5 PB, accepts OME-Zarr and SpatialData submissions

OME-Zarr / NGFF v0.5: cloud-native imaging format. SPACE is Zarr-native throughout

scverse SpatialData (*Nat Methods* 2025): universal spatial-omics framework on OME-Zarr + AnnData

The SPACE position

A centre that couples **wet-lab data generation + PB-scale compute + a user science platform + a queryable archive** in one service

SPACE is standards-aligned: a natural upstream generator feeding the public archives

Scarce resource: provenance-tracked, **3D multimodal** raw + processed data — AI/training-ready

SPACE complements. It provides an end-to-end spatial data centre of excellence within an open, standards-based ecosystem

SPACE



<https://indico.ast.cam.ac.uk/e/CGCDDataWorkshopM2>
(access key: by request)

 naw@ast.cam.ac.uk

The outline timetable includes the current baseline set of focus topics per session block.

 There are no materials yet.

[See details >](#)

see more.

to.png

cgc-data2-attendees-photo.JPG
cgc-data2-attendees-zoom-photo.png

Best Practice for use of AI/ML techniques: Breakout 2

22

SPACE Data - Walton @ CGC Spatial Biology: RS, London, UK

SHARING EXPERTISE ACROSS THE CGC COMMUNITY

THE SECOND CGC DATA WORKSHOP: DEC 2025



SPACE

- Workshop brought together representatives from SPACE and nine Cancer Grand Challenge Teams and input from the Cancer Research UK data strategy team to focus on current issues in handling data both within CGC teams and across the CGC initiative more widely.
- Breakout sessions around three discussion areas:
 - Data Interoperability: data formats and metadata
 - Archiving and Data Access across the CGC Projects
 - Best Practice for use of AI/ML techniques
- Clear that there are significant concerns around topics such as: availability of data describing data, data standards, privacy, policy, archives, what to put in the archives, AI, use of AI, LargeLM's, etc.
- For **interoperability**, metadata standards and their implementation, the tension between standards and flexibility were recurring topics. For **archiving**, the need to implement FAIR principles balanced against patient privacy. For **AI**, creating benchmark datasets, trust, and computational demands came to the fore.

Core Challenge: Developing common metadata standards that streamline recording and access without creating excessive burden for researchers.

Key Outcomes:

- Identified five universal metadata categories essential for data reusability across modalities (sequencing, spatial imaging, microbiome, clinical): experimental context, sample provenance, technical parameters, data processing, and quality assessment
- Consensus favouring recommendations and best practices rather than strict mandates, recognising the tension between standardisation and research flexibility
- Emphasis on nomenclature standardisation for both human and machine readability to enable cross-project integration
- Concerns about data format longevity and whether current standards will persist

Initial Actions:

- Survey CGC projects to map the data landscape, develop tiered metadata guidelines starting with essential elements and maintain ongoing cross-project dialogue.

Challenge, outcomes and actions also identified for
AI+Applications and **Archive+Data Access**

Gaia works because its data products (generated by the project), are described, standard, and open to all. 16,000 papers to date, all from researchers who didn't make the observations at the telescope.

Some points for discussion, and will likely be addressed in the upcoming presentations

- **The standards question:** OME-Zarr and SpatialData give us the formats. Proprietary periods, then open data, works in astronomy. What's the equivalent for tissue data, and does patient privacy change the answer, or just the engineering?
- **The AI question:** Foundation models are being trained on 2D, dissociated data. What does "AI-ready" actually mean for 3D, multimodal, provenance-tracked tissue, and who decides the benchmarks?
- **The archive question:** What would it take to build one interoperable, queryable spatial-tissue archive network the whole field works from, and who funds archives, not just experiments?

ACKNOWLEDGEMENTS

In memory of
Prof Greg Hannon



SPACE

SPACE: <https://www.space-cgc.org/>

SPACE Team

- Nicholas Walton
- Dario Bressan
- Claire Mulvey
- Marta Paez-Ribes
- Atefeh Fatemi
- Kui Hua
- Qiuchen Meng
- Carolina Mendonca
- Phoebe Hicks
- Martina Alini
- Giorgia Battistoni
- Clare Rebbeck
- Eduardo Gonzales-Solares
- Mo Al Sa'd
- Tristan Whitmarsh
- Melis Irfan
- Alireza Molaeinezhad
- Leigh Smith
- Neil Millar

Hannon laboratory team

CRUK-CI Core facilities and Admin

Collaborators

- Sam Freeman, Sohrab Shah, Wesley Tansey (MSKCC)
- Jean Abraham (UCAM)
- Richard Mair (UCAM)
- Andrew Smith, Krista Rombouts (UCL)
- Louis Chesler, John Anderson, Massimo Giuliani (ICR/GOSH)
- Isidro Cortes-Ciriano (EBI)
- Charles Keller (cc-TDI)
- Allan Balmain (UCSF)

Industry partners

- Abcam
- Tissuevision inc.
- 10X genomics
- Oxford Nanopore
- Standard Biotoools

