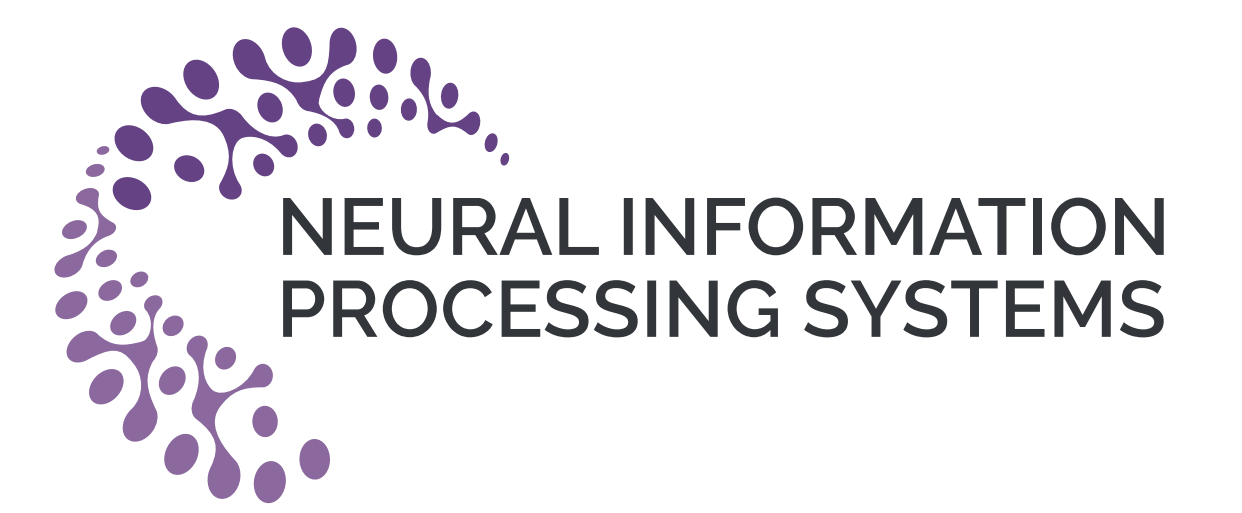# STimage-1K4M: A histopathology image–gene expression dataset for spatial transcriptomics

Jiawen Chen, Muqing Zhou, Wenrong Wu, Jinwei Zhang, Yun Li, Didong Li. Department of Biostatistics, UNC–Chapel Hill

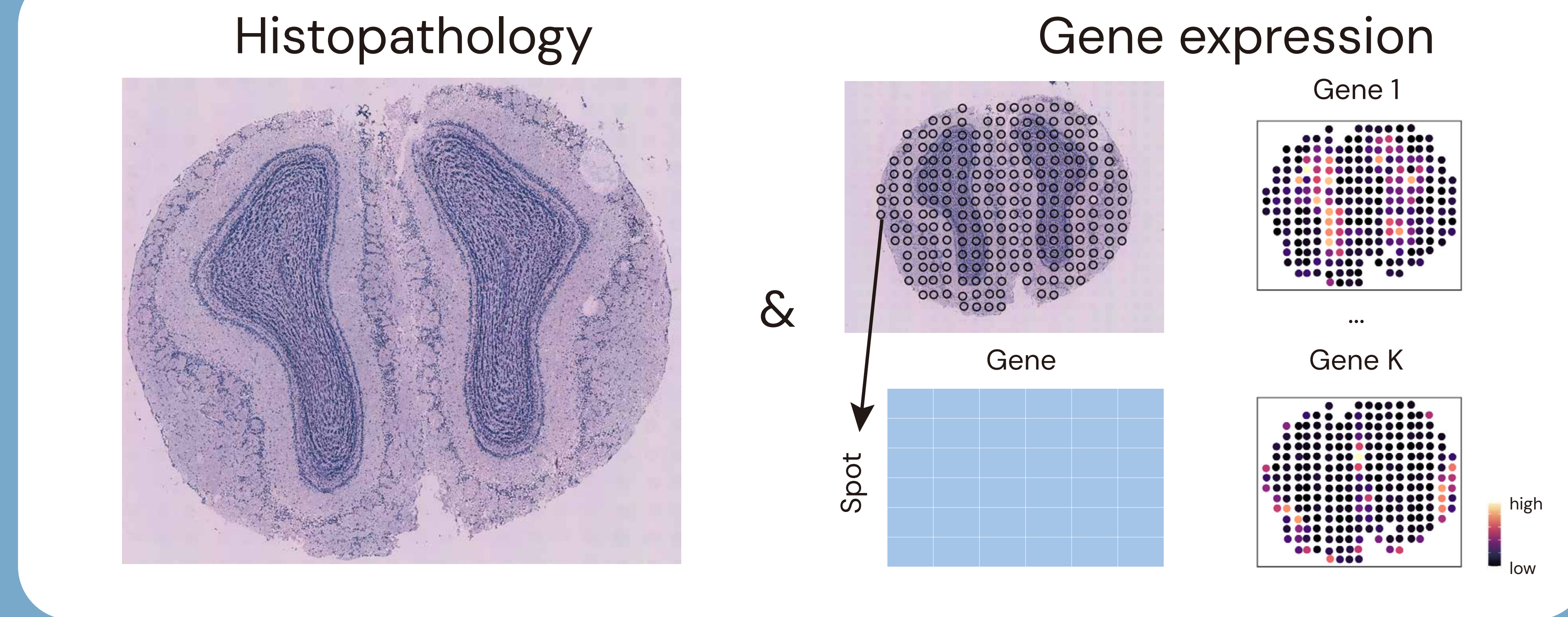THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

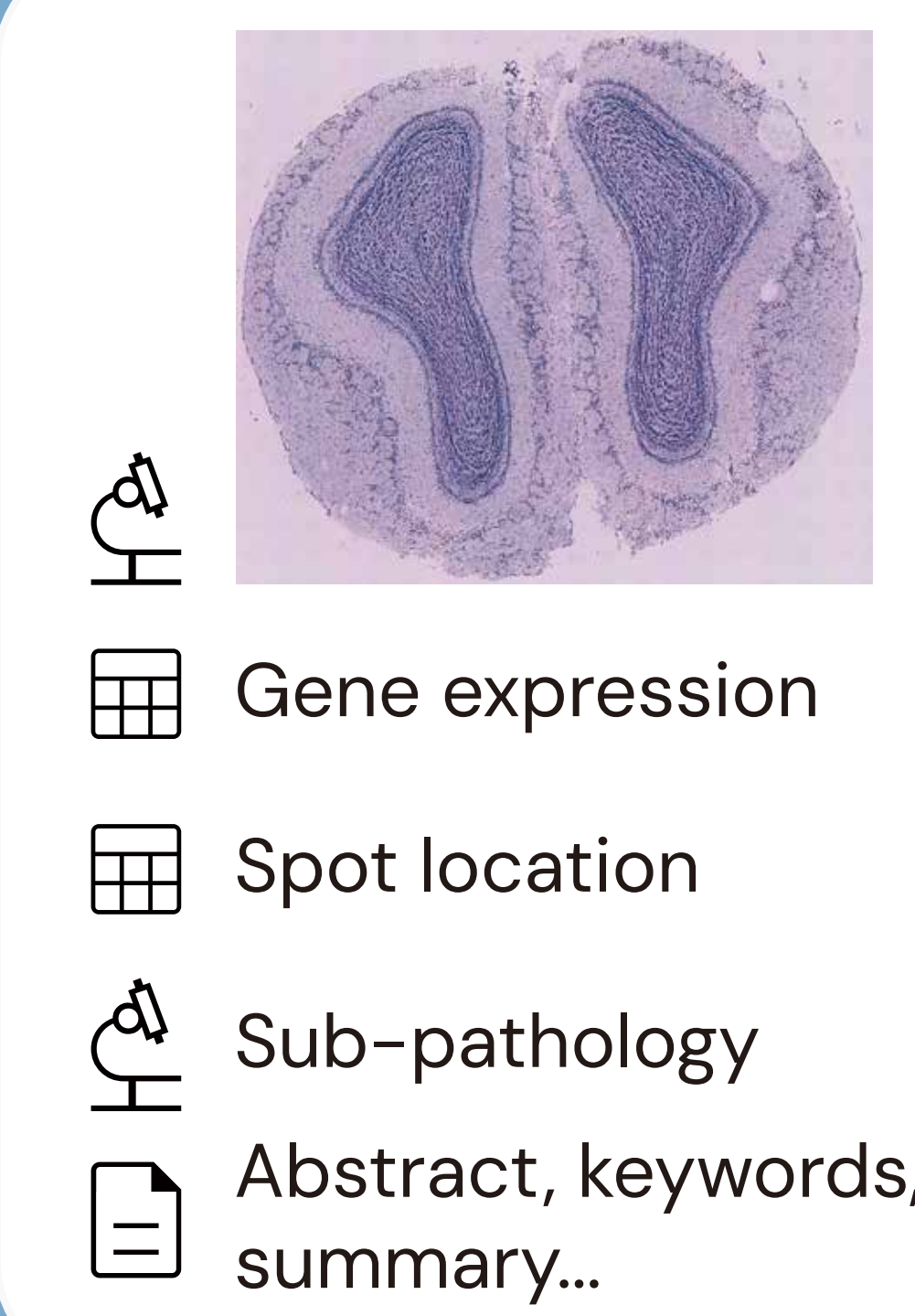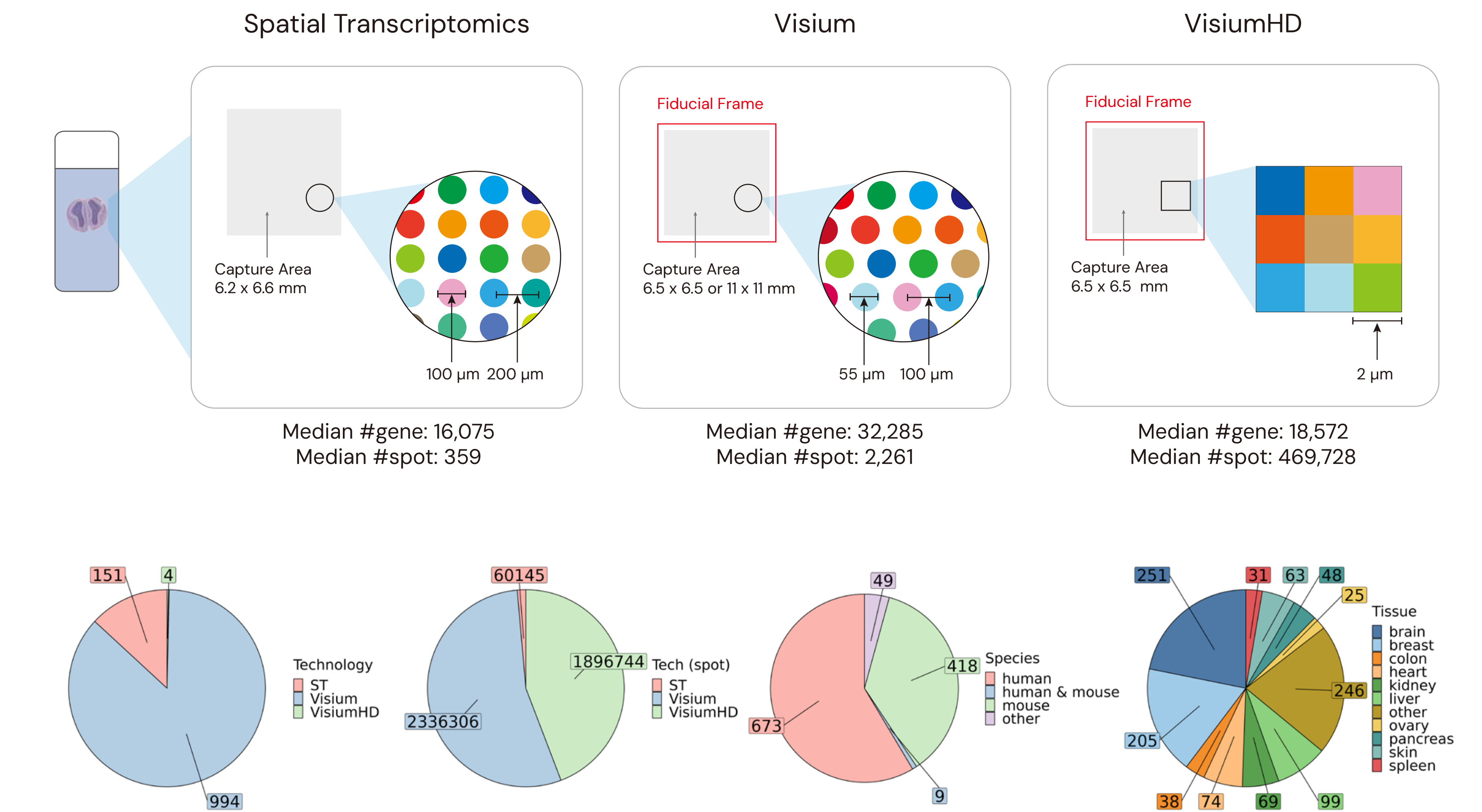NEURAL INFORMATION PROCESSING SYSTEMS

Data source

SPATIAL research

GEO Gene Expression Omnibus

10x GENOMICS

Publications

Image-Gene expression pairs

Histopathology & Gene expression

Gene 1 ... Gene K

Gene / Spot

high / low

Data

- Gene expression
- Spot location
- Sub-pathology
- Abstract, keywords, summary...

We release the first **large scale histopathology image–gene expression dataset** for spatial transcriptomics (ST), containing 1,149 ST slides, encompassing 4,293,195 spots with paired gene expression information.

Data composition

Spatial Transcriptomics
Capture Area 6.2 x 6.6 mm
100 μm 200 μm
Median #gene: 16,075
Median #spot: 359

Visium
Fiducial Frame
Capture Area 6.5 x 6.5 or 11 x 11 mm
55 μm 100 μm
Median #gene: 32,285
Median #spot: 2,261

VisiumHD
Fiducial Frame
Capture Area 6.5 x 6.5 mm
2 μm
Median #gene: 18,572
Median #spot: 469,728

Technology: ST, Visium, VisiumHD — 151, 4, 994

Tech (spot): ST, Visium, VisiumHD — 60145, 1896744, 2336306

Species: human, human & mouse, mouse, other — 49, 418, 673, 9

Tissue: brain, breast, colon, heart, kidney, liver, other, ovary, pancreas, skin, spleen — 251, 31, 63, 48, 25, 246, 205, 38, 74, 69, 99

+ pathologist annotation for 71 slides

Experiment #1: Gene expression prediction

ERBB2

CLIP   PLIP   UNI
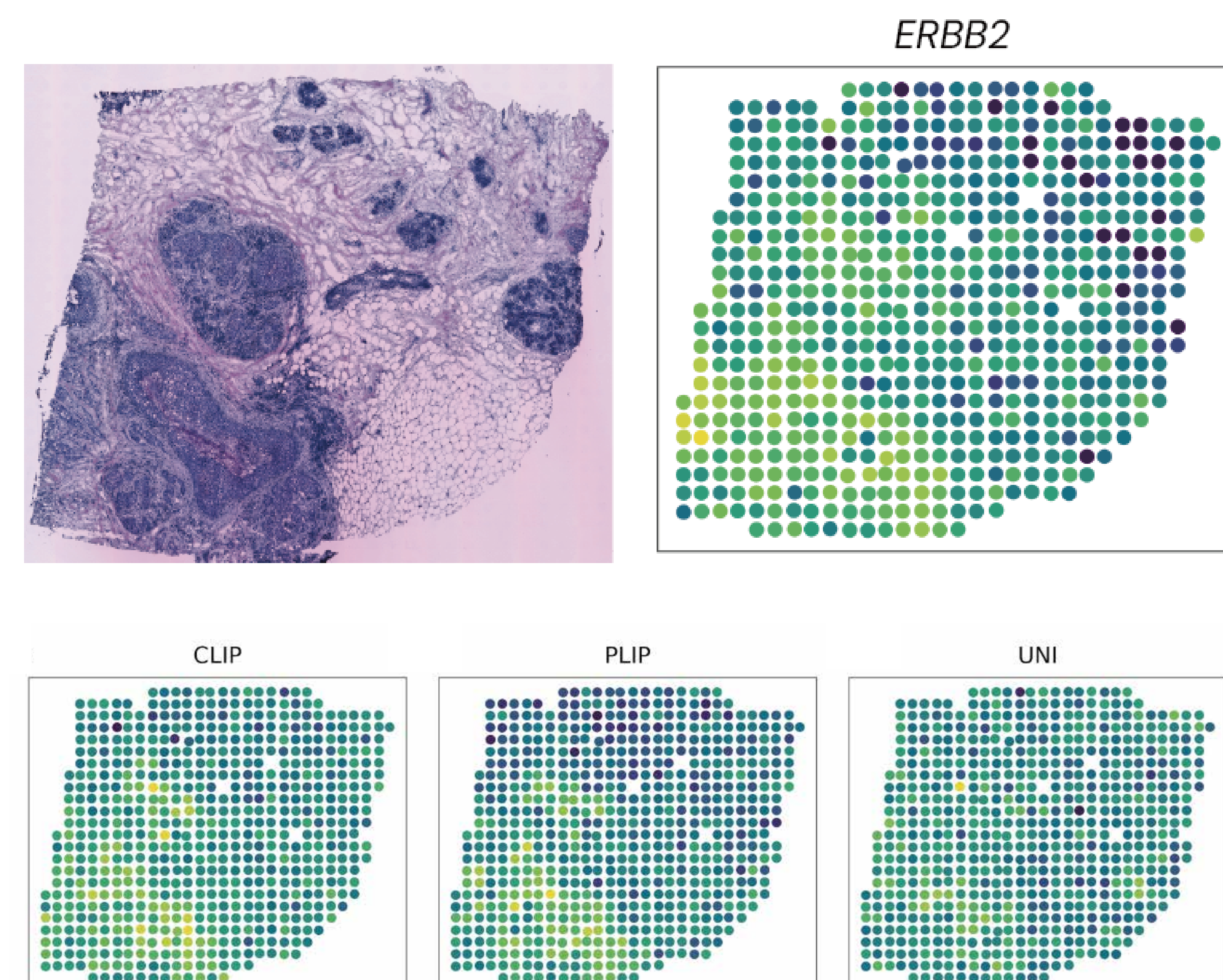
Experiment #2: Deconvolution

CellViT-SAM   CellViT-256

A1, A2, A3   B1, B2, B3

Neoplastic   Epithelial   Inflammatory   Connective   Dead

Cancer Epithelial   CellViT 256   CellViT SAM

0.2 0.4 0.6   0.00 0.25 0.50 0.75 1.00

Cancer Epithelial
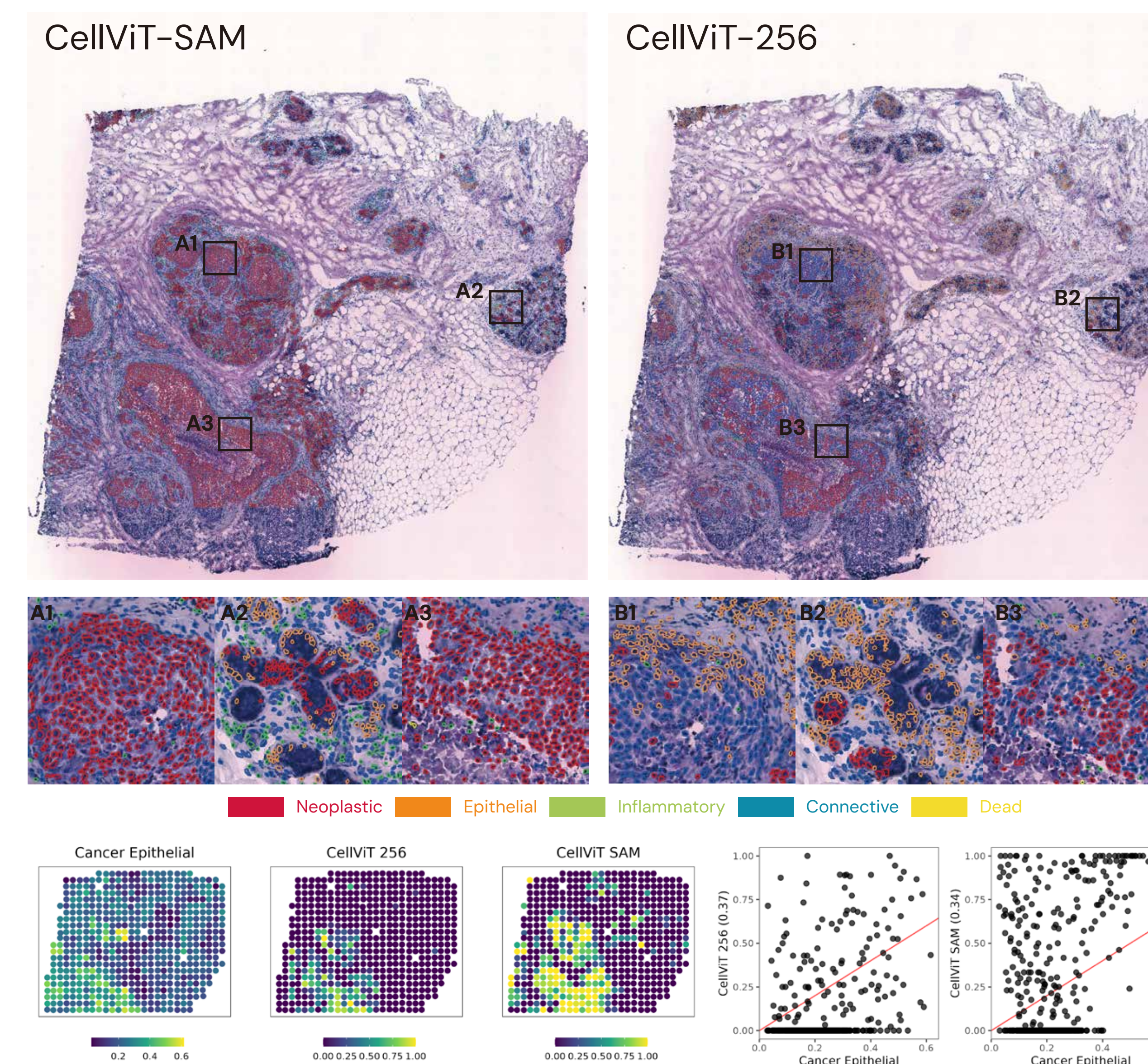
Experiment #3: Image representation learning

Fine-tune large vision model image encoder

Spot images → Image Encoder → Dim=512 Dim=32 → $I_1, I_2, ..., I_n$ → Refine ← $G_1, G_2, ..., G_n$ Dim=32 Dim=K ← Gene expression Encoder ← Spot ← K Genes

Fully connected layer

Human DLPFC

CLIP   finetune CLIP hvg   finetune CLIP overlap-hvg   PLIP   finetune PLIP hvg   finetune PLIP overlap-hvg   UNI

Brain layer: L1, L2, L3, L4, L5, L6, WM

Paper

Website