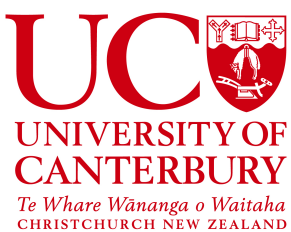
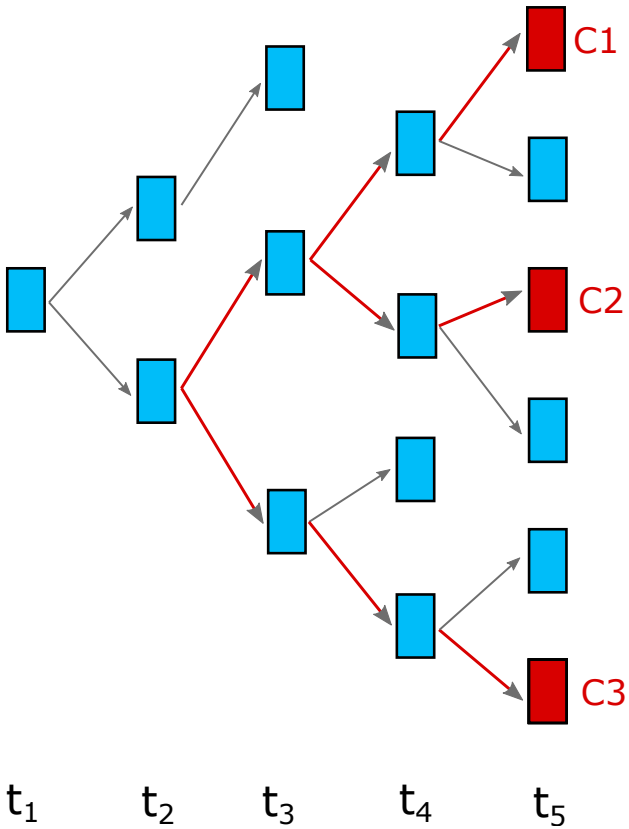
Inferring phylogenetic trees from single cell expression data of cancer

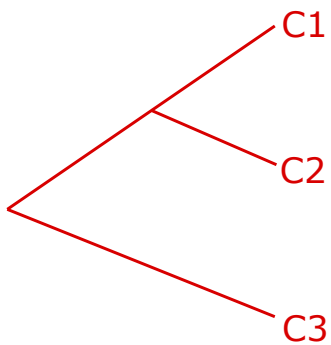
Alexandra "Sasha" Gavryushkina, Holley Pinkney, Sarah Diermeier, Alex Gavryushkin

Waiheke 2024
The 27th Annual New Zealand Phylogenomics Meeting



Single-cell phylogenetic tree of cancer



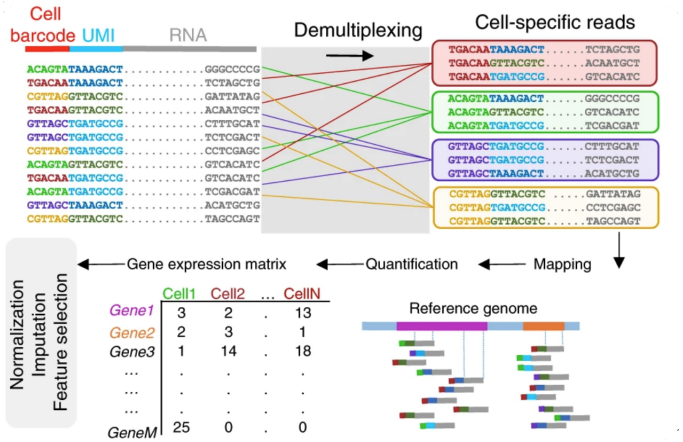


Inferring cell phylogenies from scRNA-seq data

- Many studies have used scDNA-seq data
- scRNA sequencing is cost-effective
- scRNA sequencing captures other evolutionary processes (e.g., methylation, copy number variation, etc.)

scRNA data:

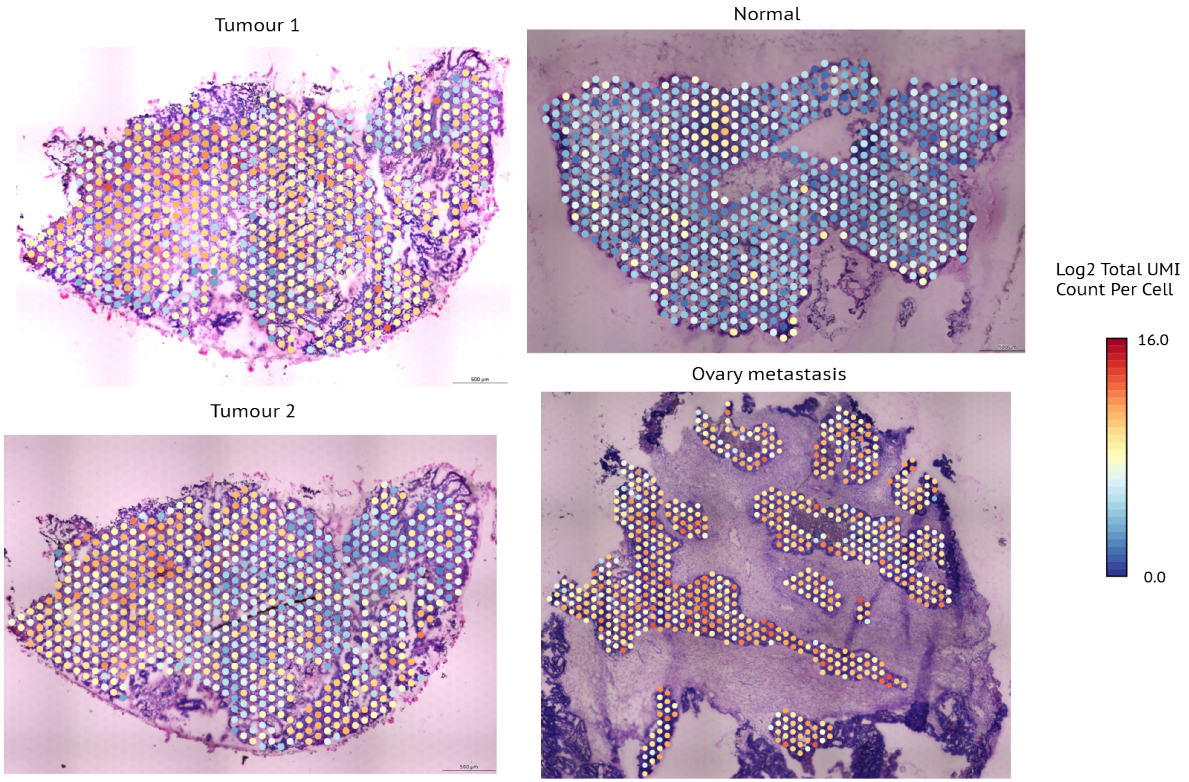
- Single nucleotide variants (SNVs)
- Expression counts



from Lafzi *et al.* 2018

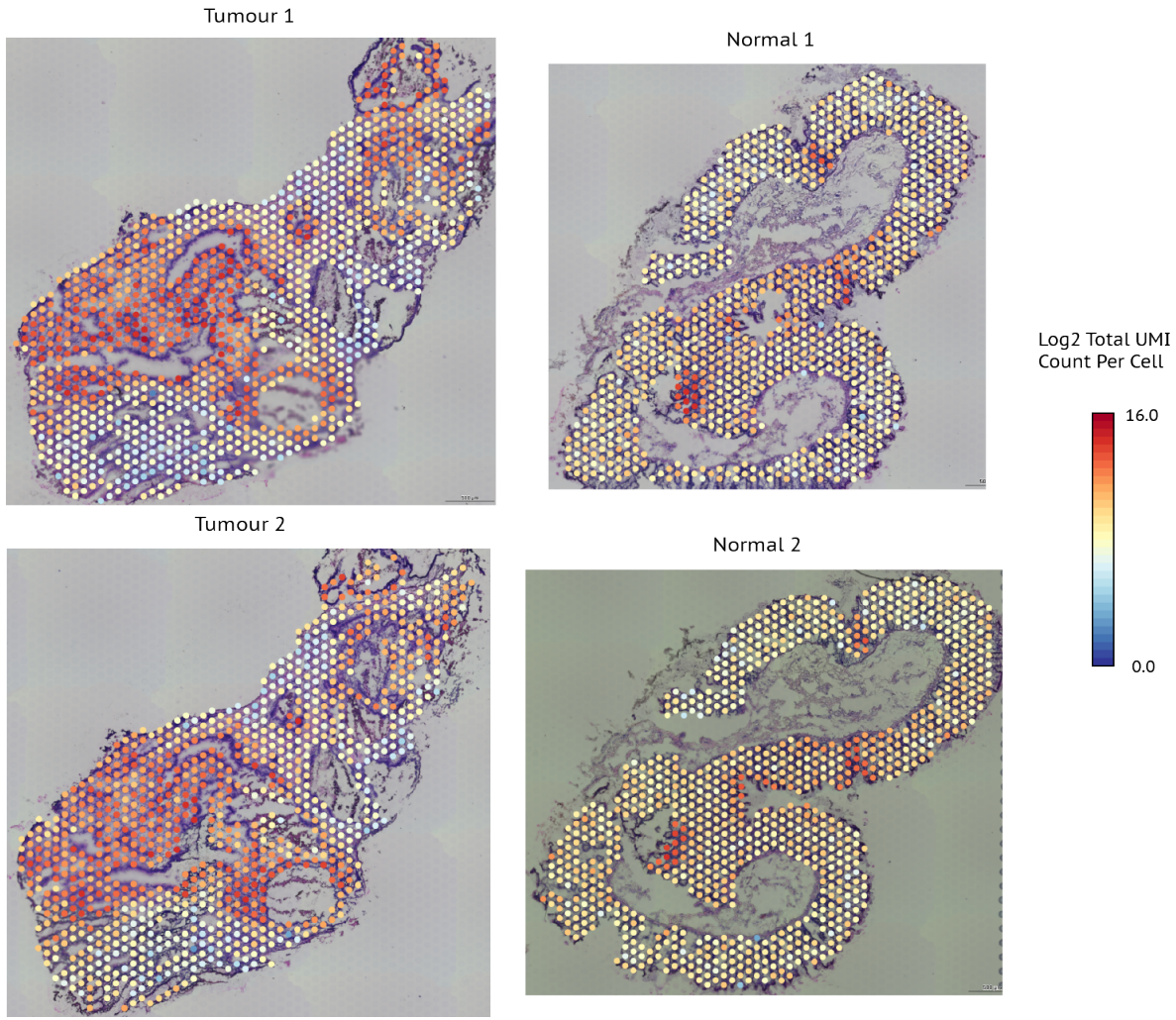
Visium Spatial Gene Expression data: Patient 1 (H Pinkney, S Diermeier)

Serial colon tumour sections, an adjacent normal and an ovary metastasis:



Visium Spatial Gene Expression data: Patient 2 (H Pinkney, S Diermeier)

Serial colon tumour and adjacent normal sections:



Visium Spatial Transcriptomics data phylogenetic analyses challenges

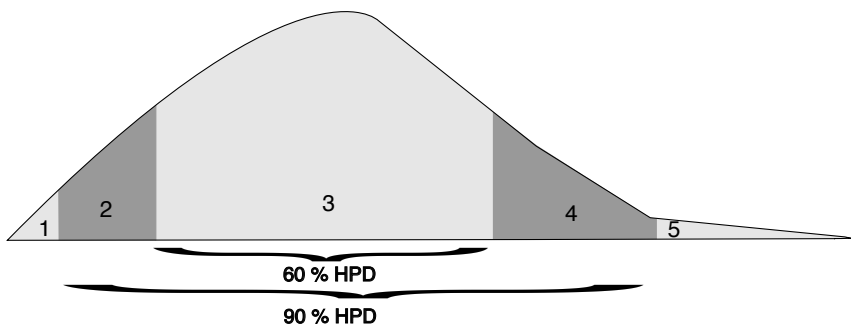
- Relatively low detection efficiency (up to 15%)
- Uneven RNA capture across spots of the same section
- Uneven RNA capture across sections
- Multicellular resolution (each spot covers up to 15 cells)
- Phylogenetically uninformative genes
  - Genes under selection
  - Cycle dependant genes
  - *Environmentally expressed* genes
- How to model expression evolution?

Visium Spatial Transcriptomics data phylogenetic analyses challenges

- Relatively low detection efficiency (up to 15%)
- Uneven RNA capture across spots of the same section
- Uneven RNA capture across sections
- Multicellular resolution (each spot covers up to 15 cells)
- Technical zeros: genes that are not identified due to not being sampled, ambiguously aligned to the reference, lost during the library preparation and sequencing
- Phylogenetically uninformative genes
  - Genes under selection
  - Cycle dependant genes
  - *Environmentally expressed* genes
- How to model expression evolution?

Modeling expression evolution

- Continuous counts are too computationally expensive to analyse for thousands of genes
- Discretisation approach (Moravec *et al.* 2021):
  - **0** is a separate category
  - Category **1, 2, 3, 4, 5** obtained from standardised by gene data:

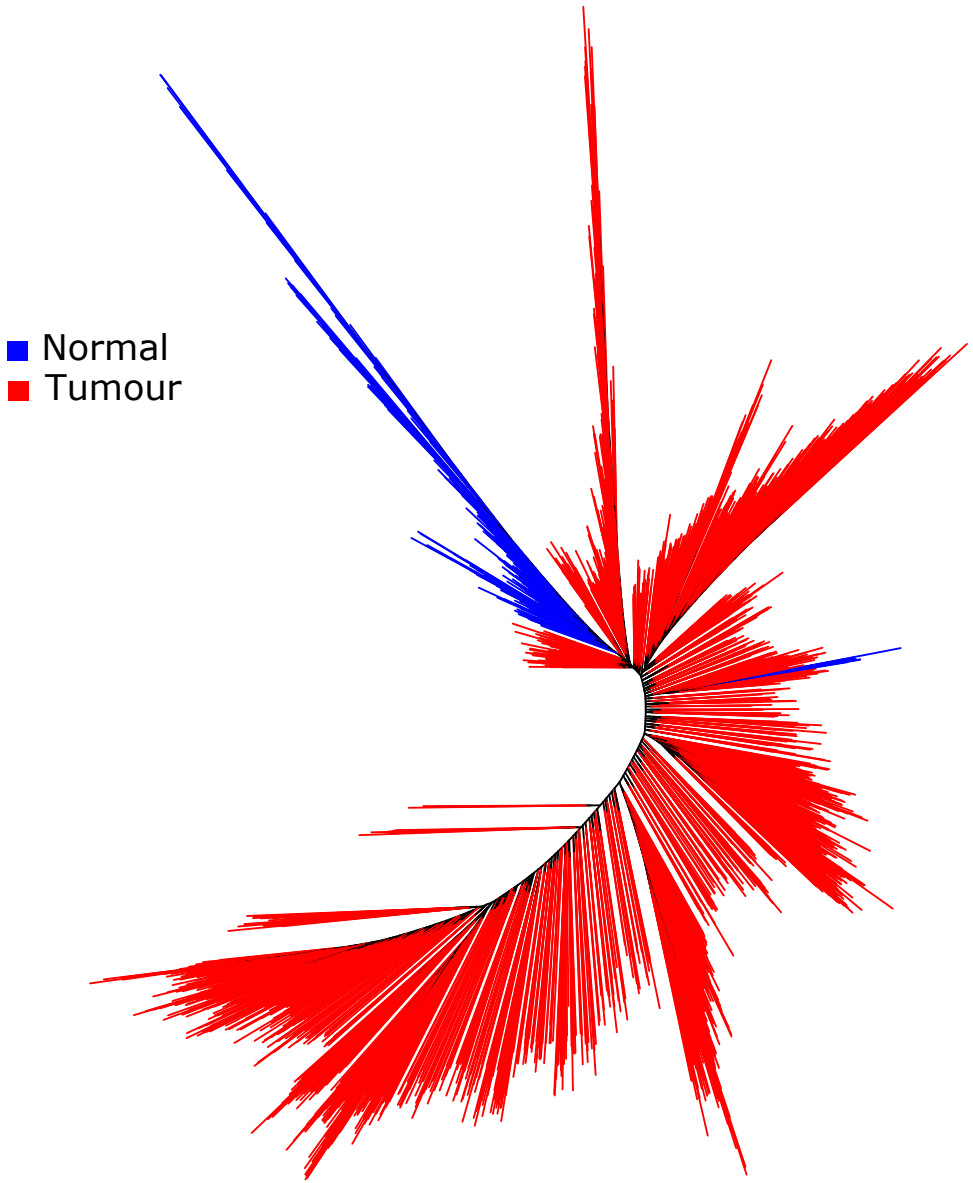


- Ordinal model: neighbouring categories evolve in one another with equal probabilities

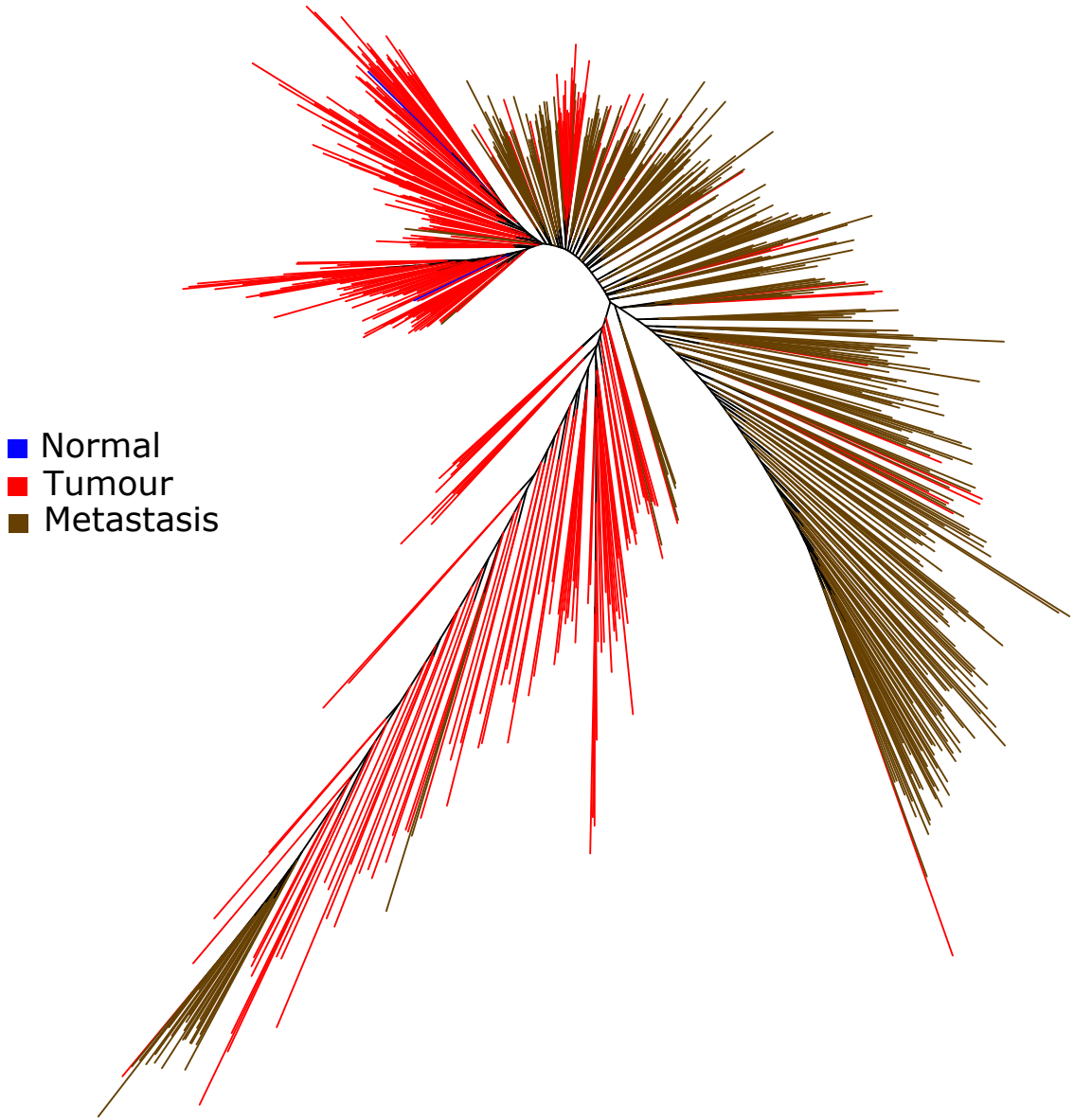
Filtering genes

- Moravec *et al.* 2021: filter cells and genes that show very little expression to obtain 50% data density
- **All** genes expressed at least in one spot
- **Highly variable genes (HVG):** variance higher than a threshold 10 or 2

Patient 2: Maximum likelihood phylogeny on filtered spots

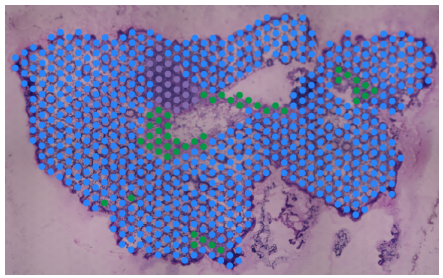


Patient 1: Maximum likelihood phylogeny on filtered spots

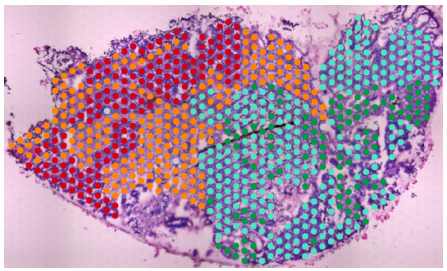


Pathologist guided annotation

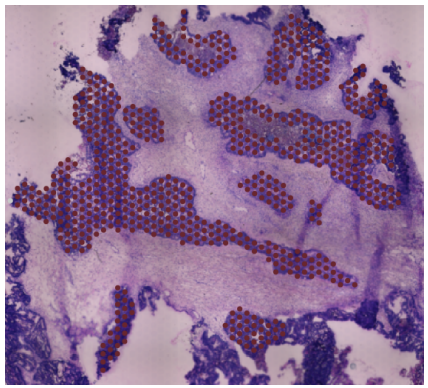
Patient 1 Normal



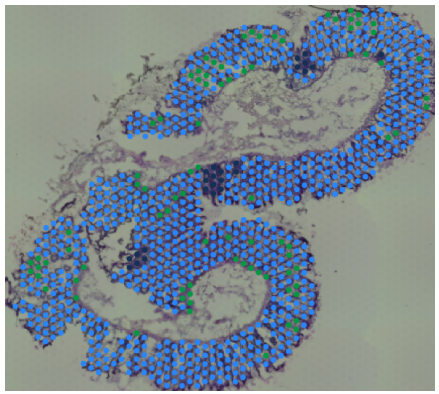
Patient 1 Tumour 2



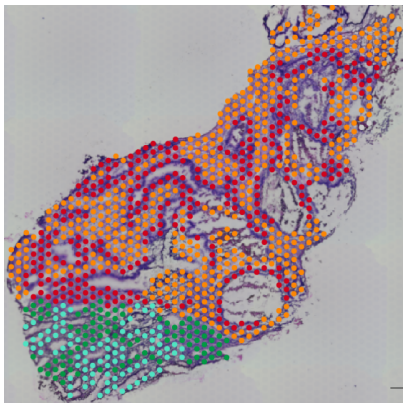
Patient 1 Ovary Metastasis



Patient 2 Normal 1

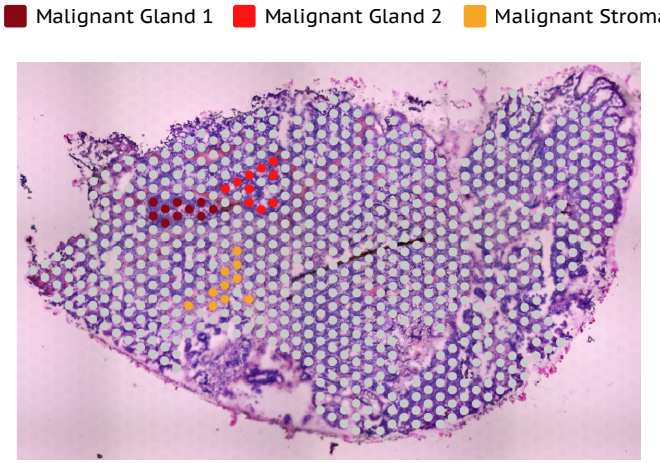


Patient 2 Tumour 2

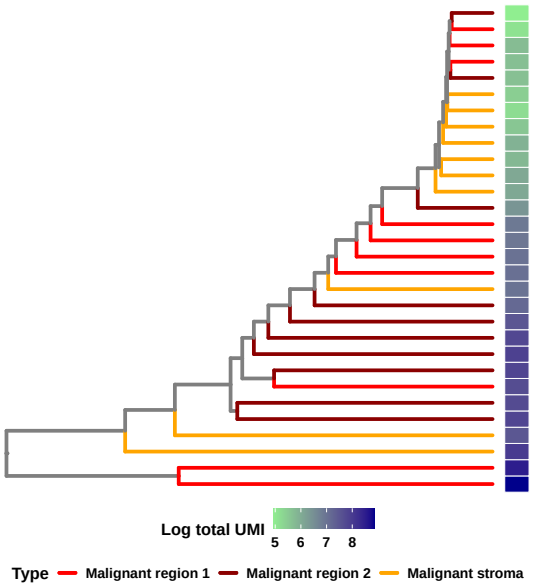


- Benign Colon Epithelium (normal)
- Benign Colon Epithelium (tumour)
- Benign Colon Stroma
- Malignant Glands
- Malignant Glands (metastasis)
- Malignant Stroma

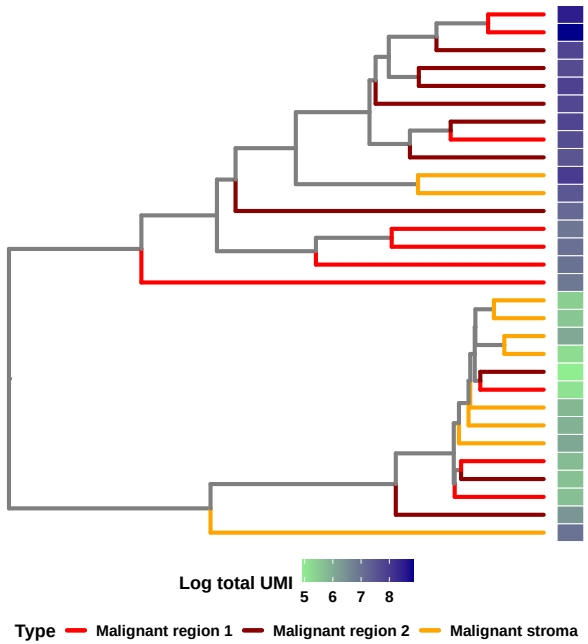
Patient 1, test set (Tumour 2), Bayesian inference, MCC trees



All genes

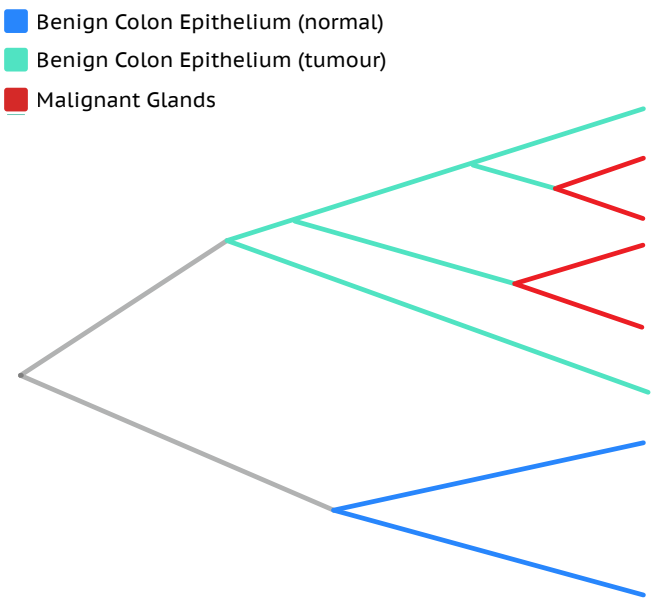


HVG genes

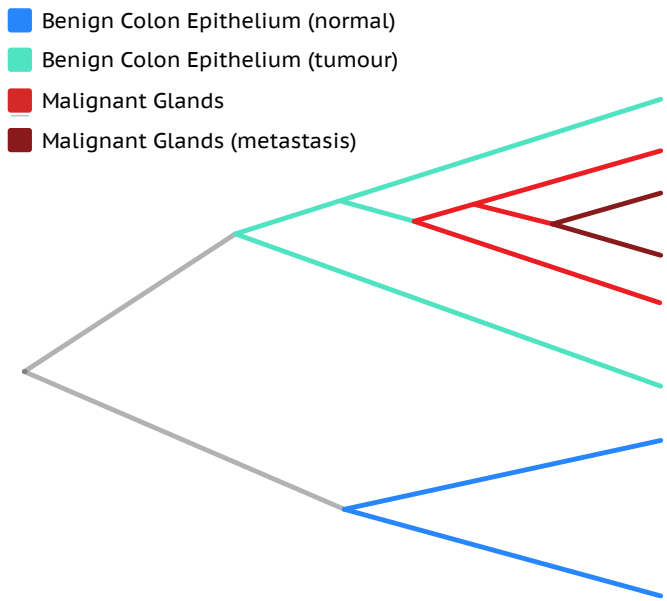


Expected topological relationship among tissue types

- Similar tissue types form clades
- Cancerous clades diverge from normal
- Metastatic clades diverge from cancerous
- Only a small number of cancerous or metastatic clades



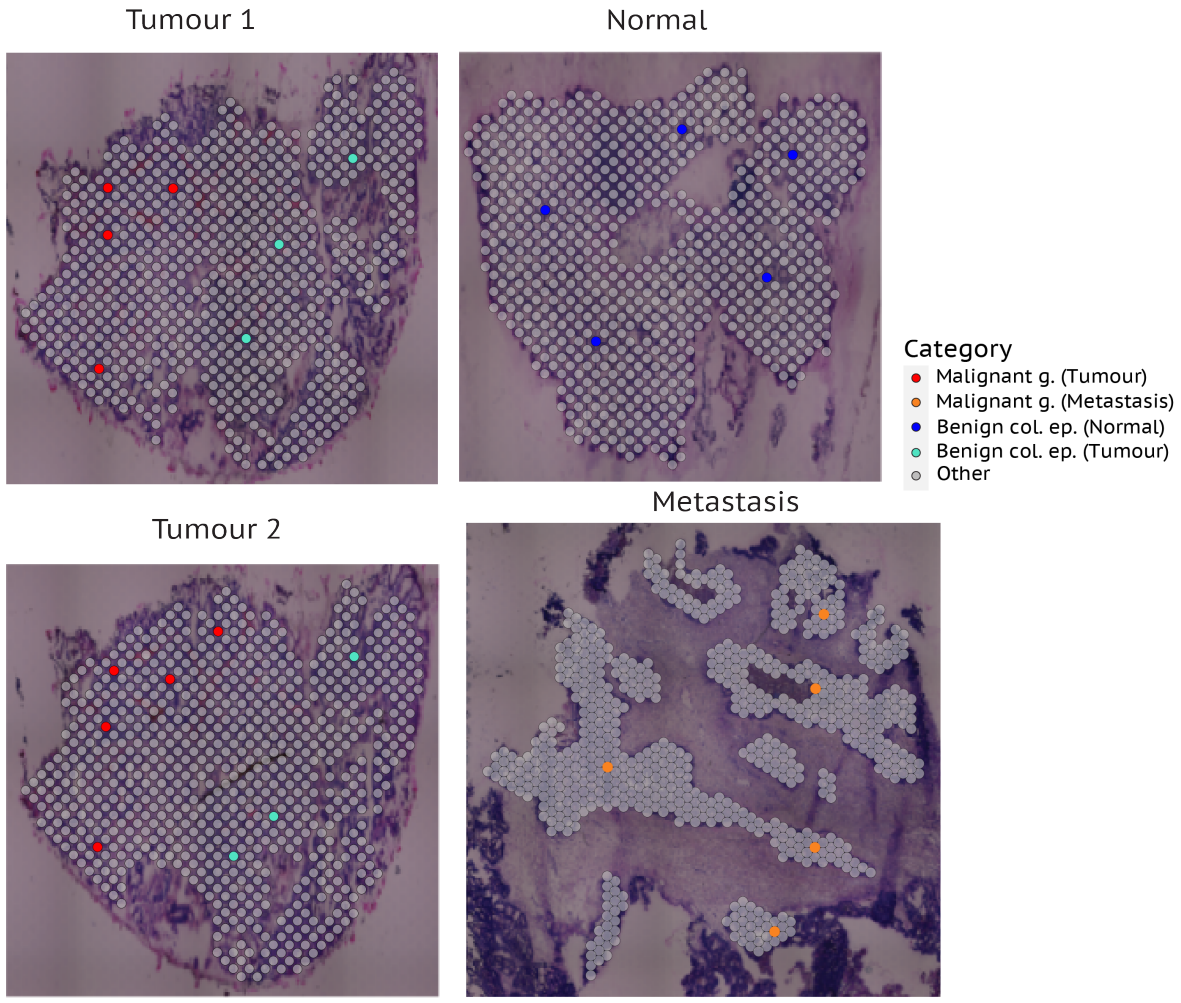
2 malignant clades



1 malignant clade

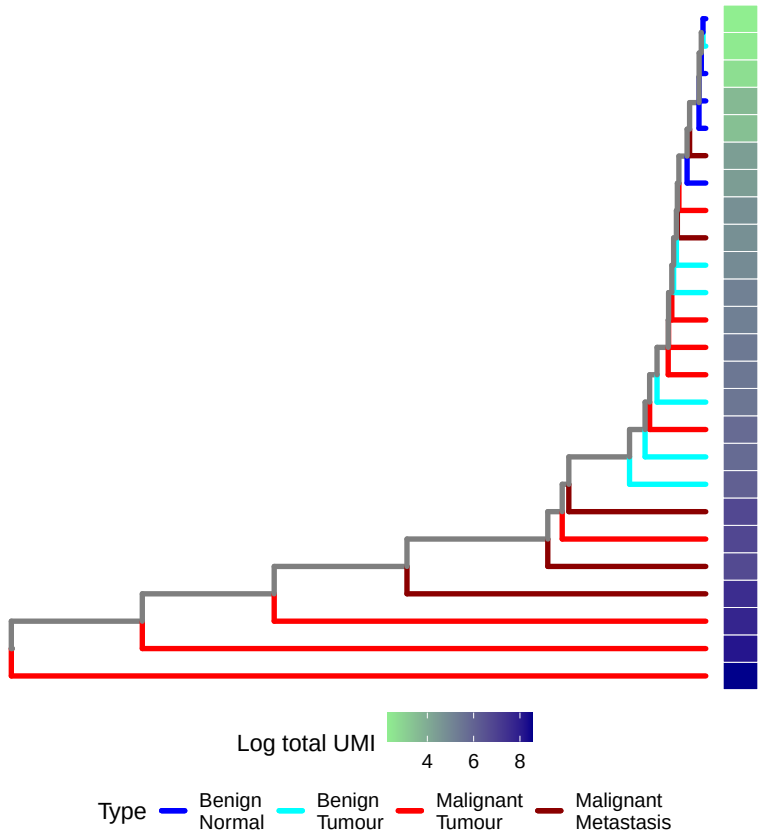
Patient 1, manually selected

Manually selected spots from distant regions



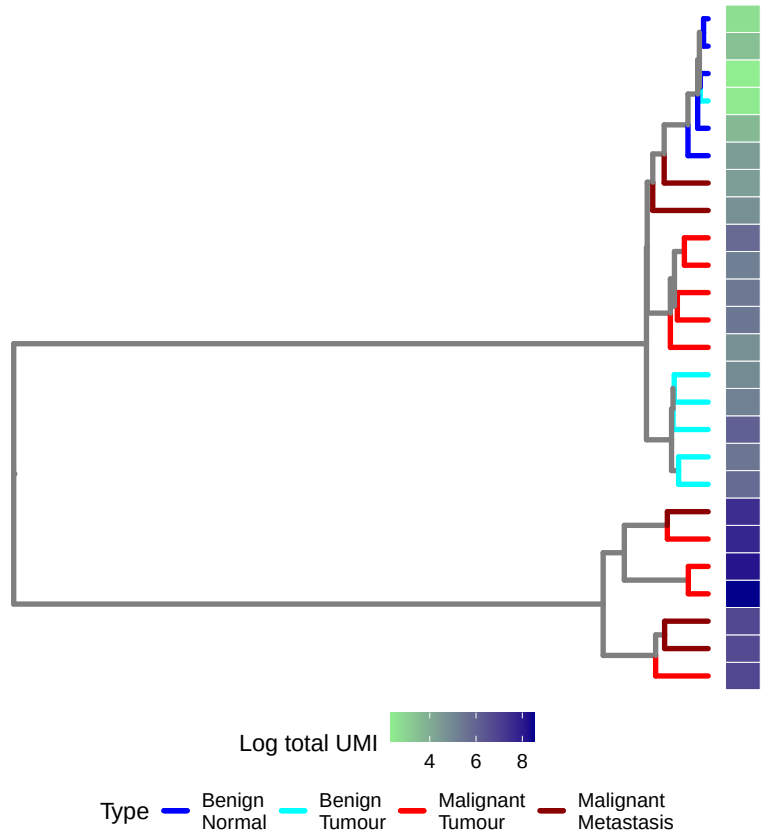
Patient 1, manually selected (Bayesian inference, MCC trees)

All



14 malignant clades

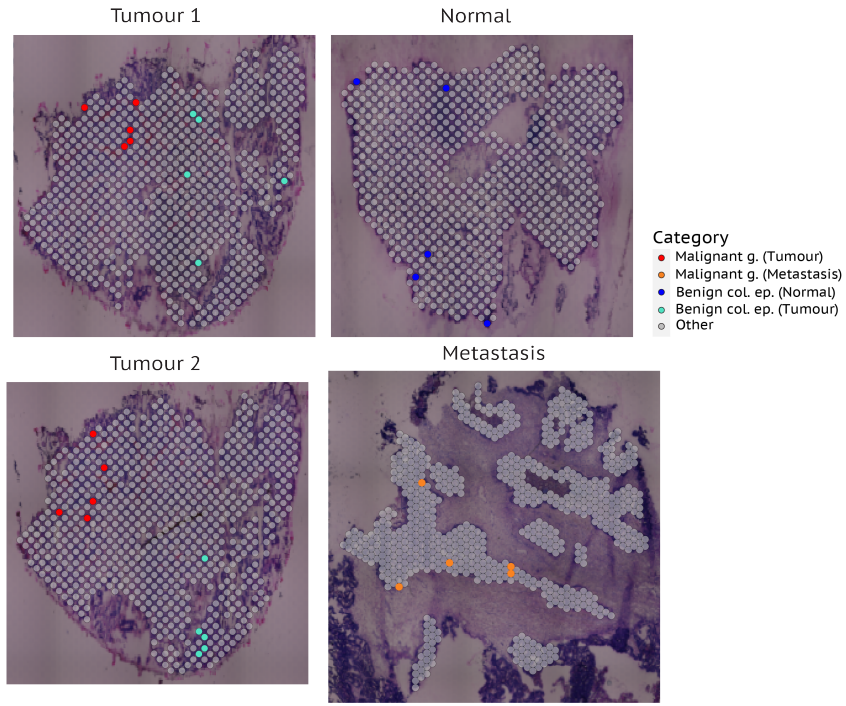
HVG



4 malignant clades

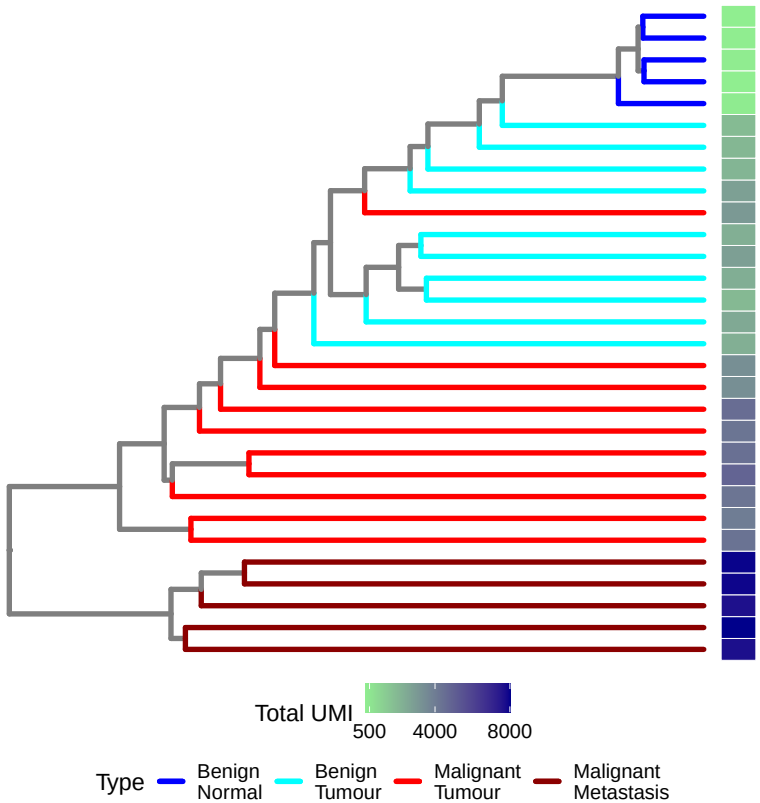
Patient 1, highest total counts among the types

- Split all malignant gland and benign colon epithelium spots in six categories:
  - Benign colon epithelium Normal
  - Benign colon epithelium Tumour 1
  - Benign colon epithelium Tumour 2
  - Malignant gland Tumour 1
  - Malignant gland Tumour 2
  - Malignant gland Ovary metastasis
- Choose five spots with highest total counts from each of the categories



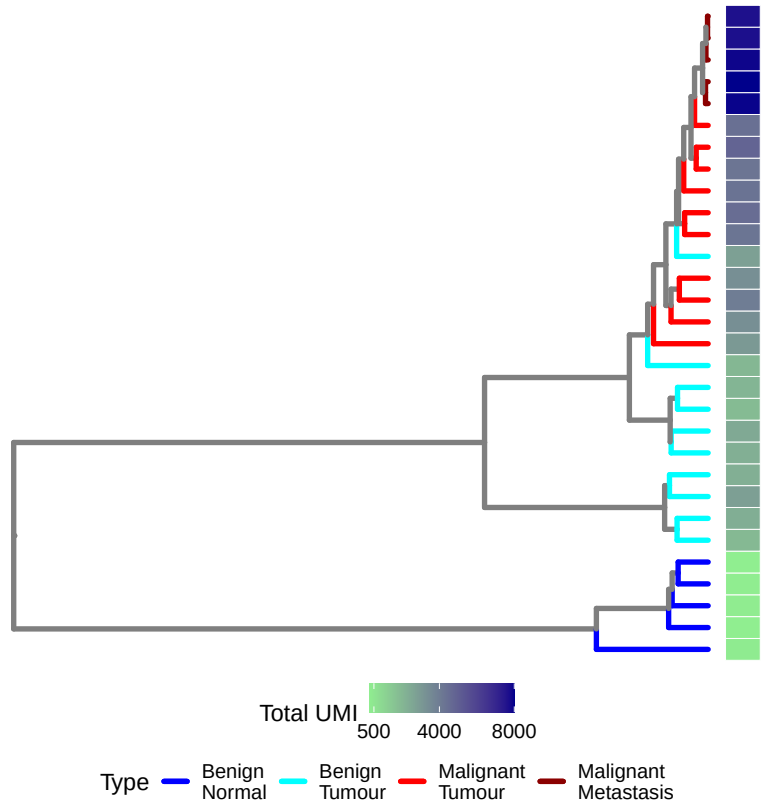
Patient 1, highest total counts (Bayesian inference, MCC trees)

All



8 malignant clades

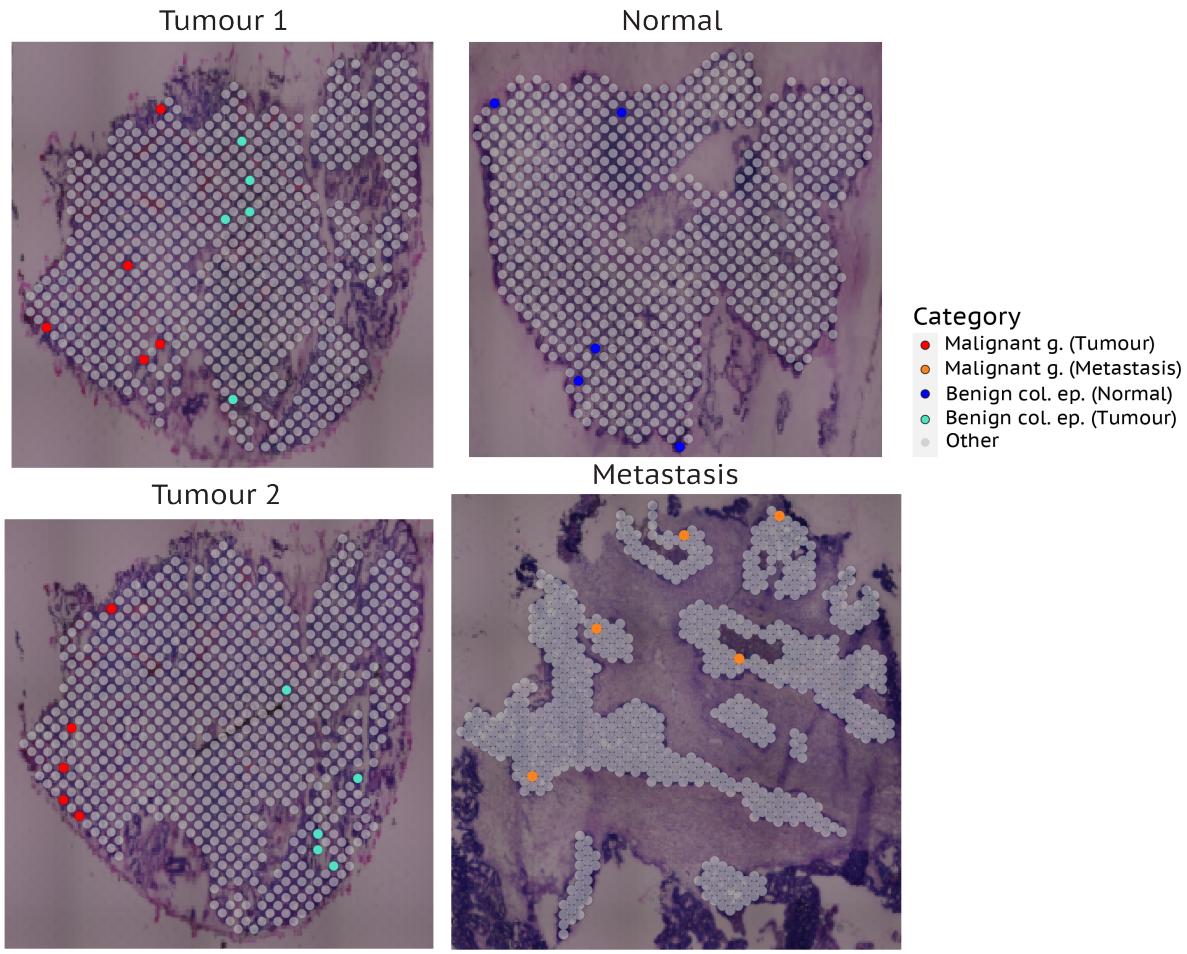
HVG



3 malignant clades

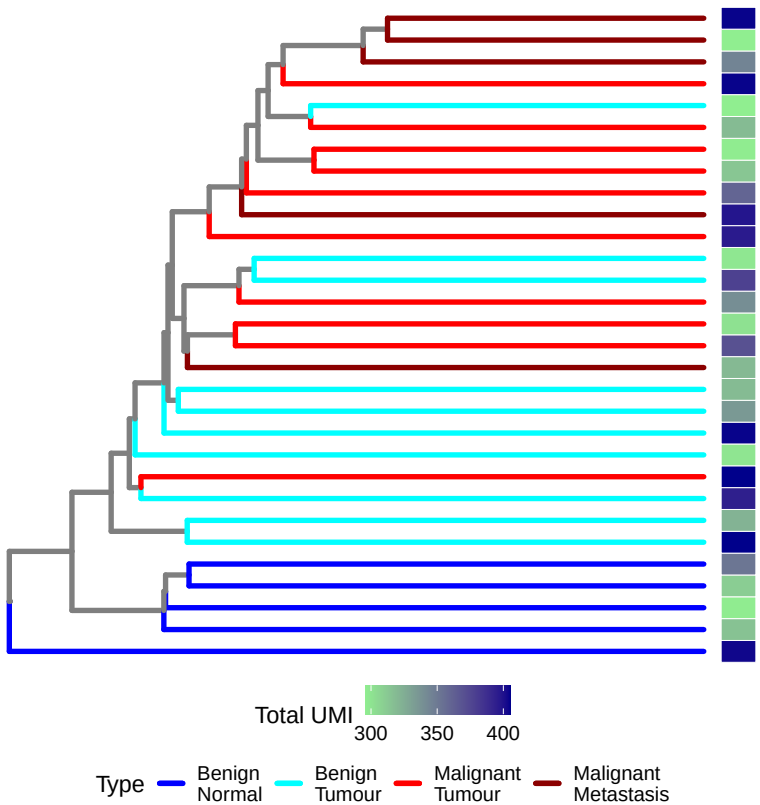
Patient 1, similar total counts

Counts within the highest total counts of normal (between 296 and 405 UMIs)



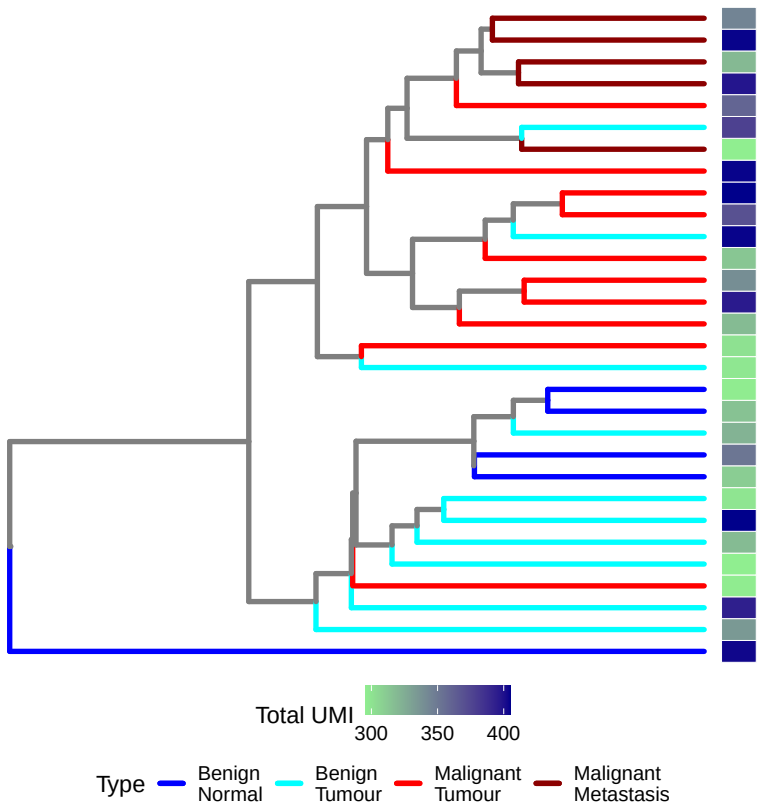
Patient 1, similar total counts (Bayesian inference, MCC trees)

All



9 malignant clades

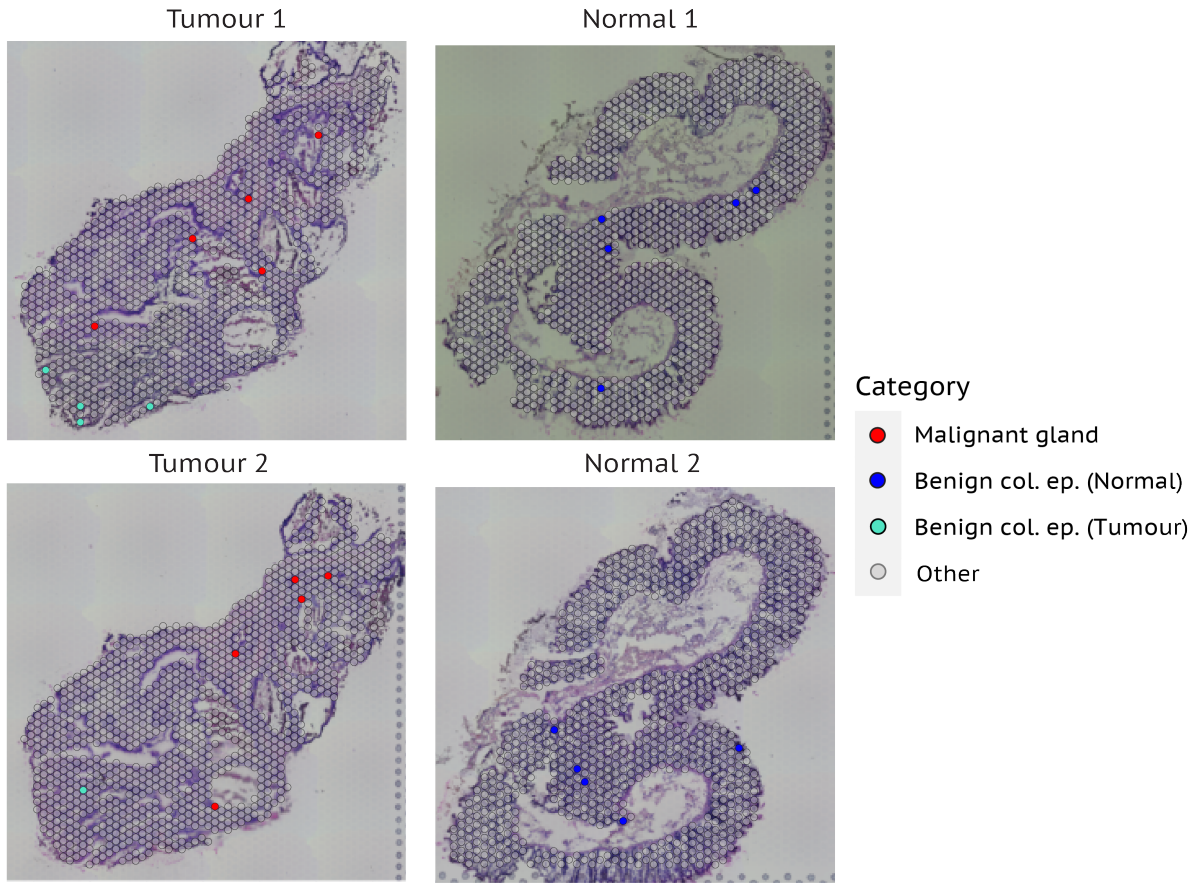
HVG



8 malignant clades

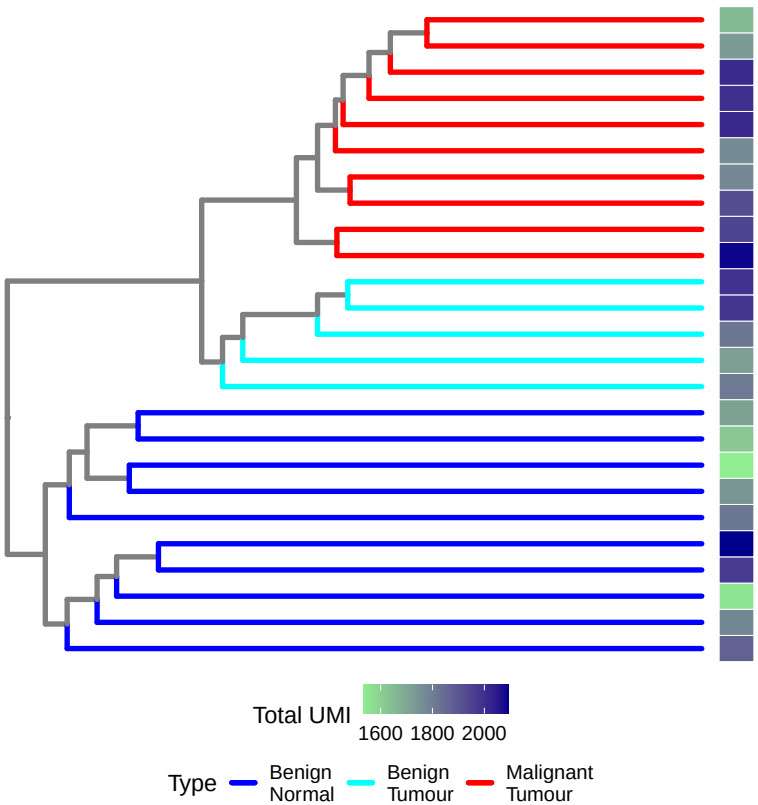
Patient 2, similar total counts (Bayesian inference, MCC trees)

Counts within the highest total counts of normal (between 1535 and 2091 UMIs)



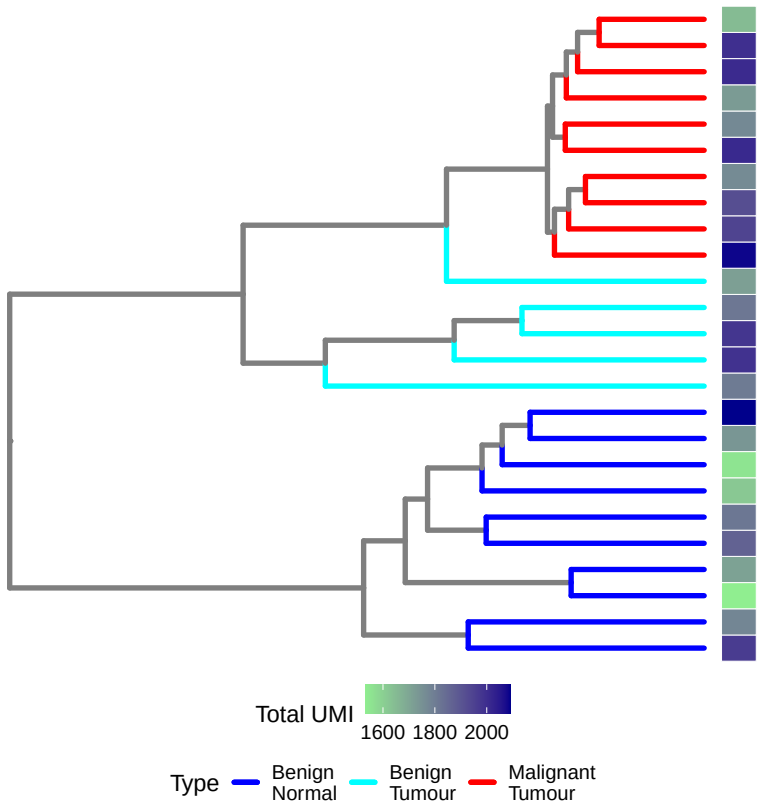
Patient 2, similar total counts (Bayesian inference, MCC trees)

All



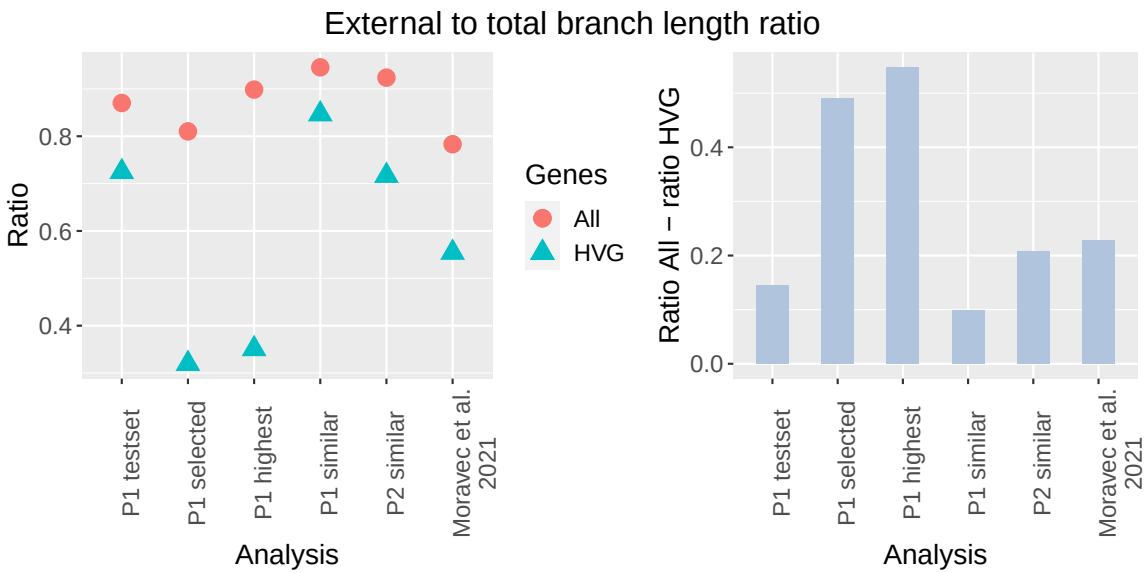
1 malignant clade

HVG

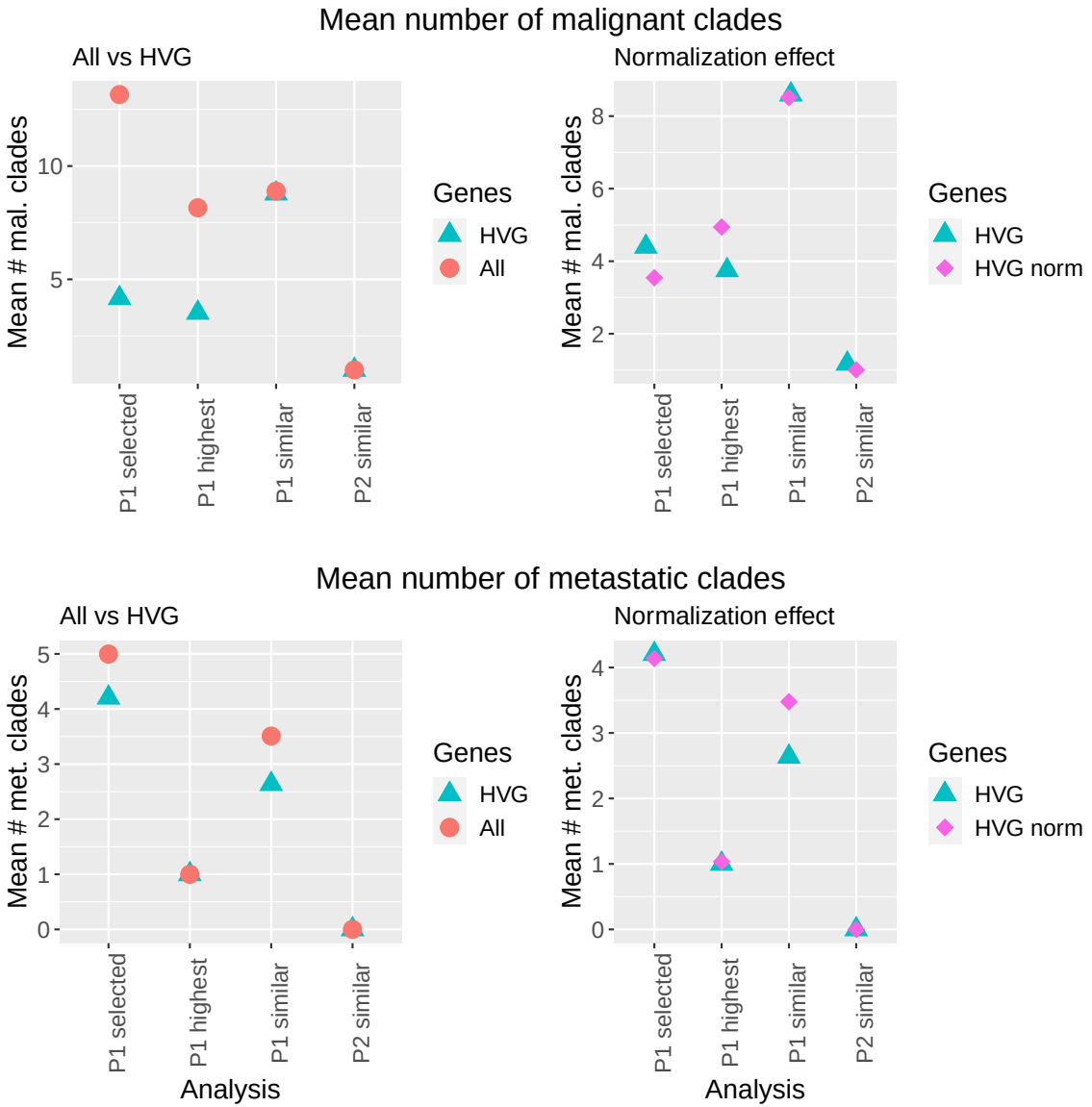


1 malignant clade

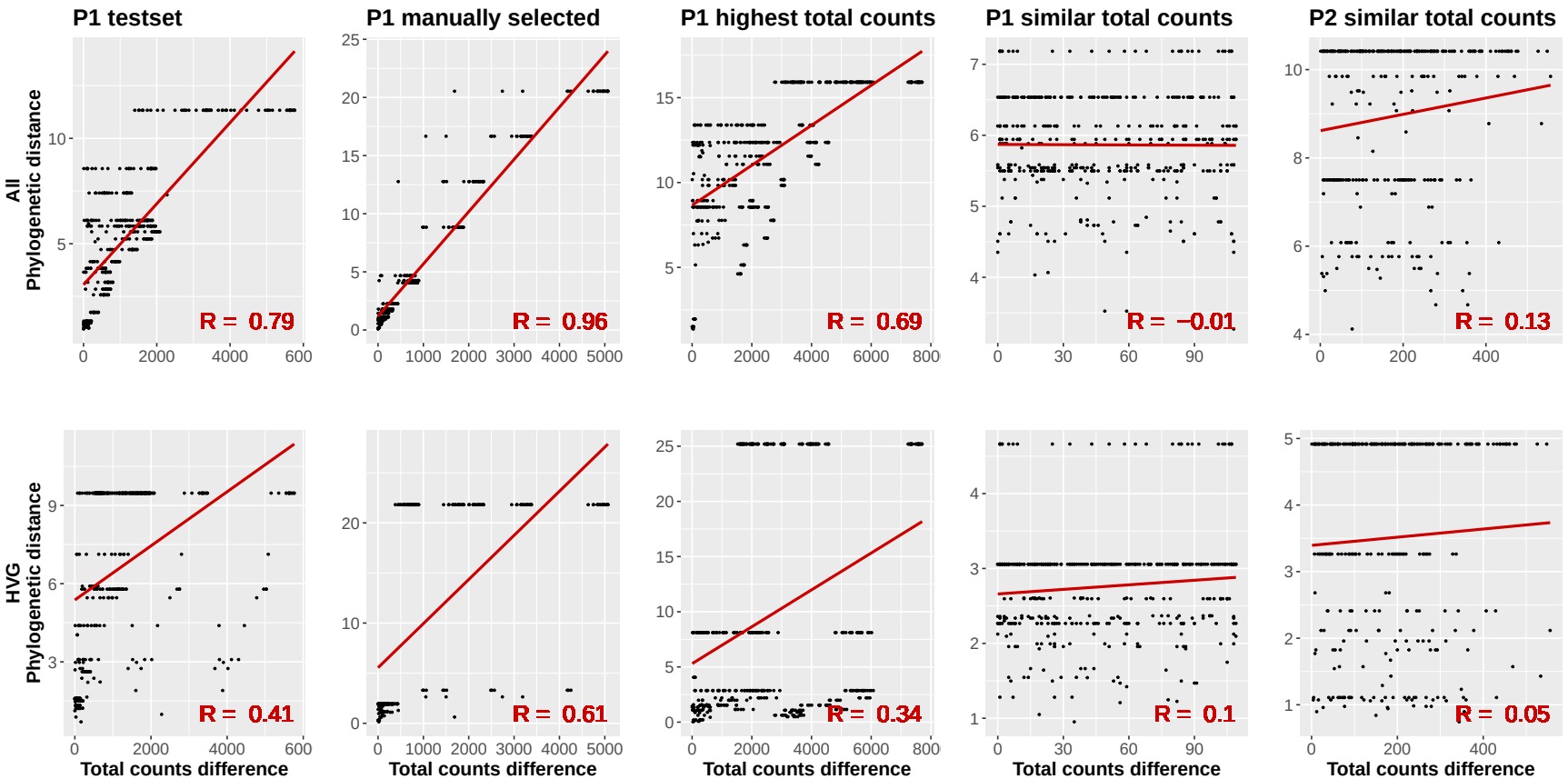
The effect on the external branch lengths (All vs HVG)



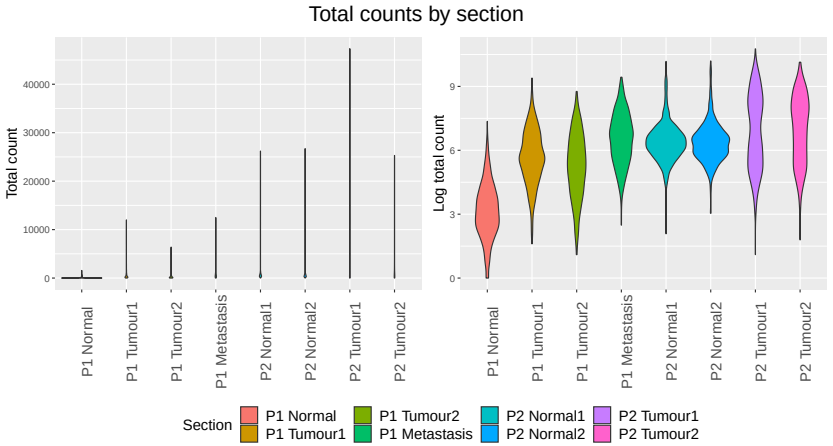
The effect on the number of malignant clades

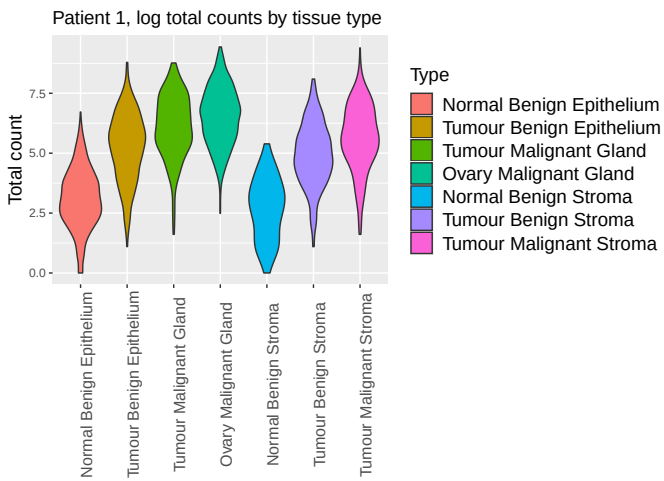


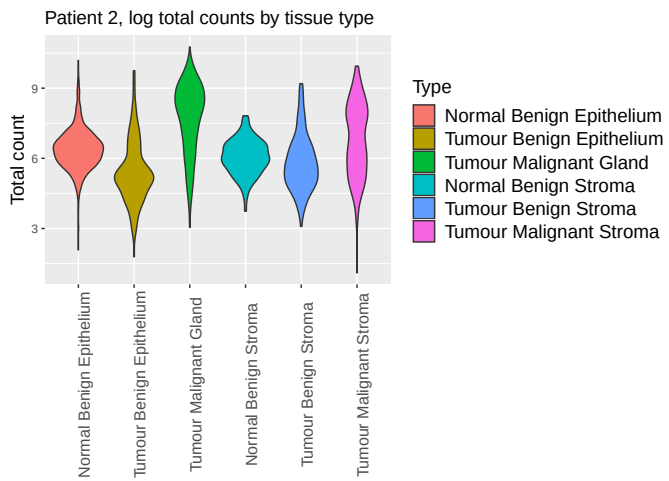
Pairwise phylogenetic distance against total counts difference



Total counts







Limitations and conclusions

- Multicellular resolution is not addressed
- HVG can still contain misleading phylogenetic signal
- Only ordinal model with strict clock was used

To conclude:

- There is a potential in using scRNA-seq expression data for reconstructing phylogenetic trees
- Using only highly variable genes improves phylogenetic clustering of similar tissue types and regions and produces trees with shorter terminal branches
- There are still many problems to address for robust phylogenetic analysis of scRNA-seq expression data

Thank you!

Funding:

- Ministry of Business, Innovation, and Employment of New Zealand Endeavour Smart Ideas grant
- Data Science Programmes grant

Mah and Dunn 2023:

- scRNA-seq expression data of different cell types from different species
- First 20 principal components
- Continuous counts, Brownian motion model