Finite-state transducers for inferring tumour evolution from copy number variation profiles

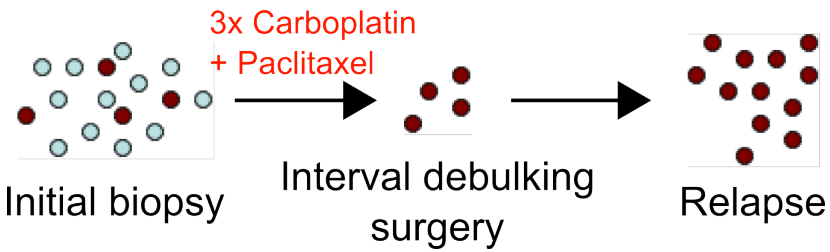
Roland Schwarz

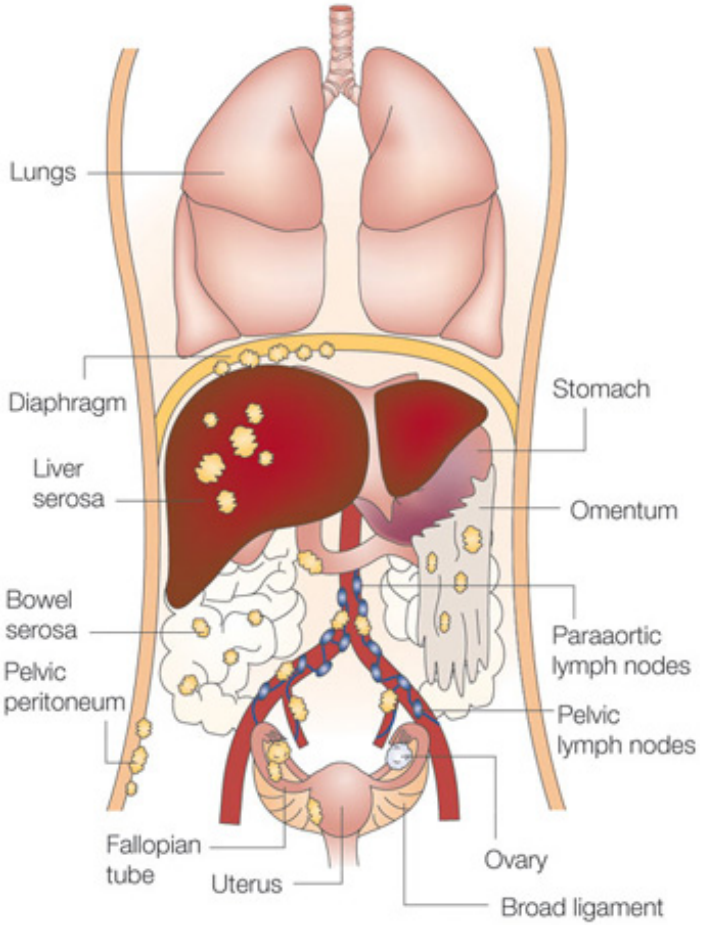
CRUK Cambridge Research Institute
University of Cambridge

September 3, 2010

Intra-tumour heterogeneity

- Ovarian cancer spreads locally
- After first set of chemotherapy often resistant relapse
- Being able to explain tumour evolution and tumour heterogeneity will help to fight resistant relapse



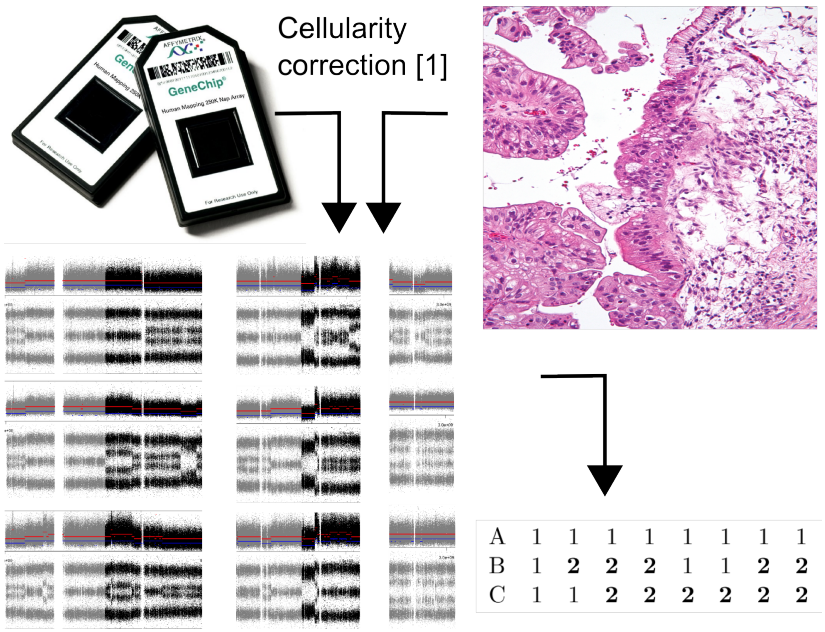


Nature Reviews | Cancer

Data

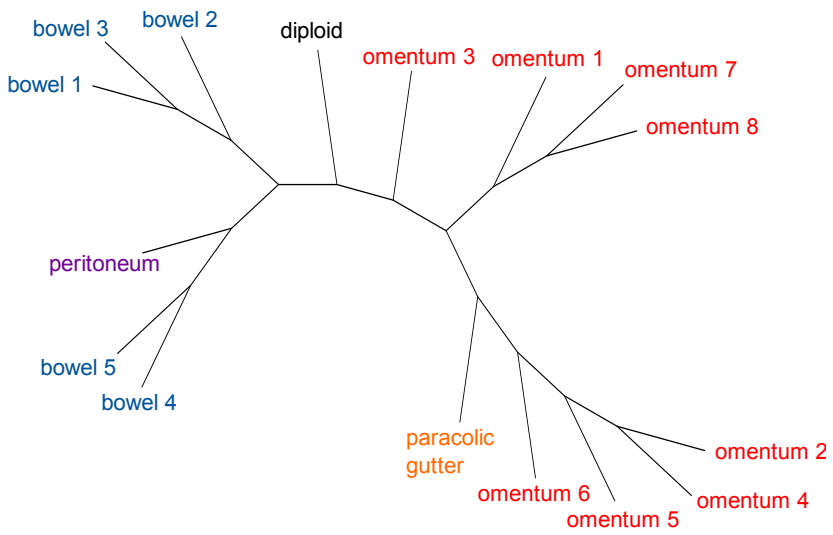
CNV profiling

- We gathered multiple (9 – 30) samples per patient from primary tumours and different metastatic sites
- Samples were H&E stained and hybridised on Affymetrix SNP 6.0 arrays
- After correction for differing cellularity scores CNV profiles were determined



Goals

- Determine evolutionary distances between profiles
- Reconstruct evolutionary histories within a patient
- Work on full profiles, not breakpoints only or collapsed regions



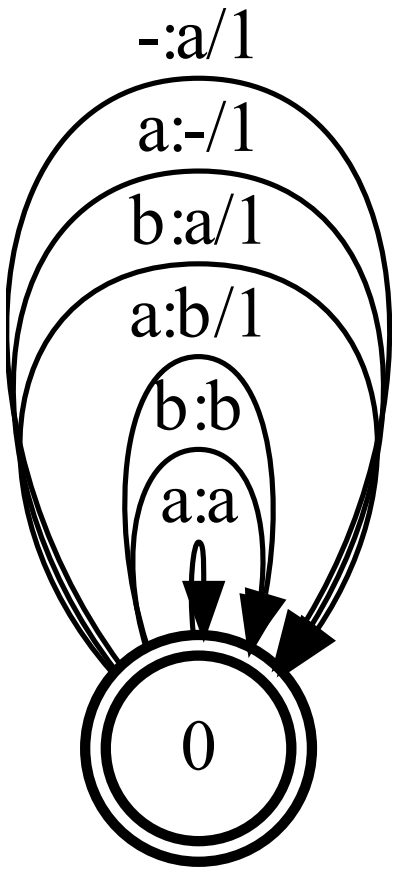
Challenges

- Modelling of sequential data with horizontal dependencies (amplification and deletion events of differing lengths)
- sequences $>$ 1.8 million probes
- divergent data of overlapping events ($<$ 50% profile identity)
- noisy data

Overview

Finite-state transducers

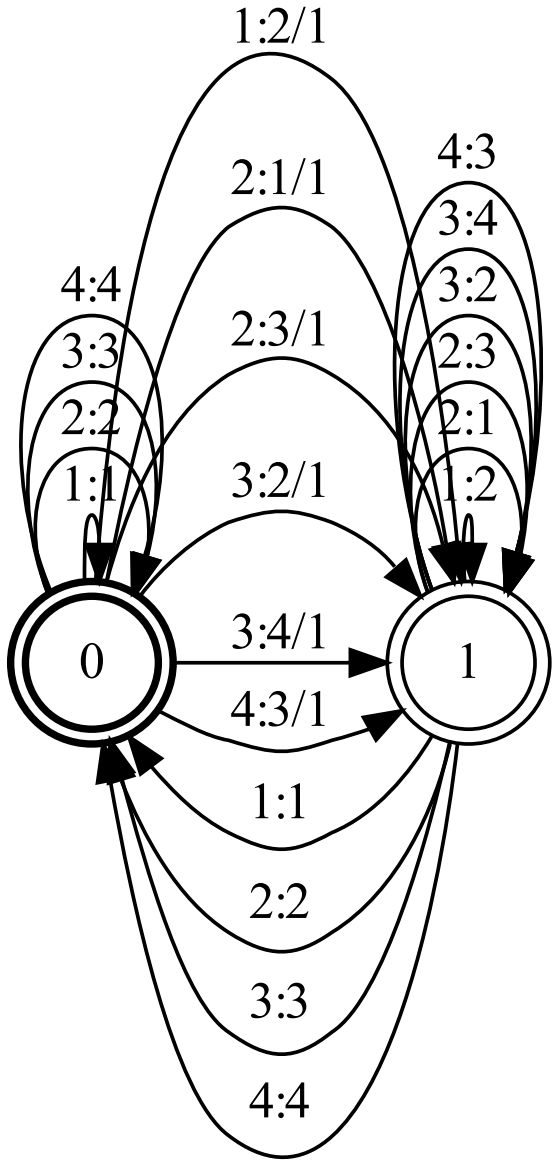
- Finite-state automata with input and output tape
- Transitions are weighted with elements of different semirings
- FSTs assign a value to each pair of sequences (e.g. similarity score or edit-distance)
- can also work on distributions of sequences (noise)



Model

111111111111111111111

111222222222111111111



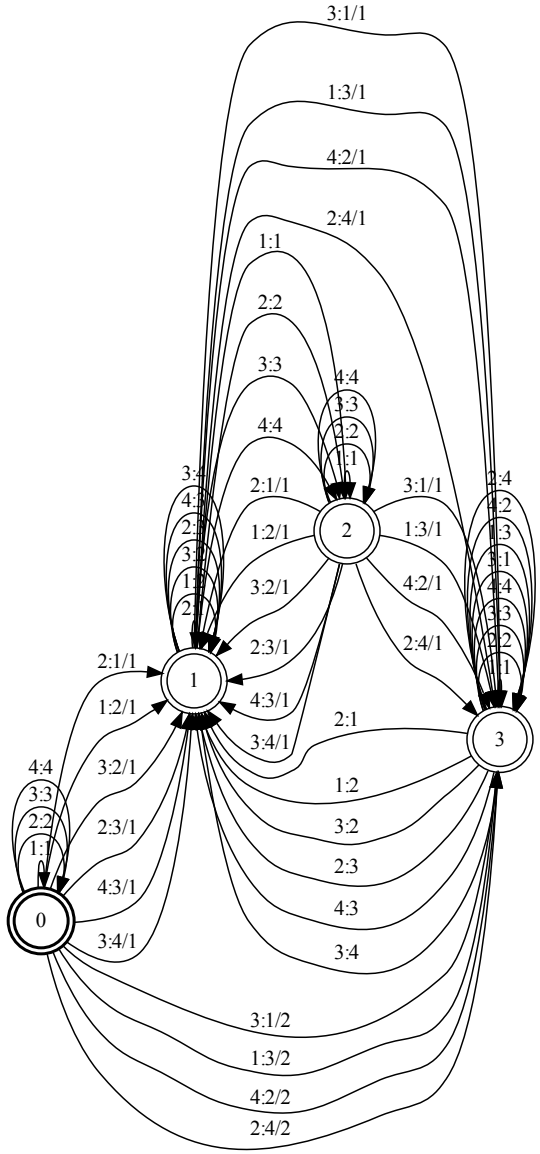
Model

111111111111111111111

111222222222111111

11111111111111111111

111222223333322111111



Model

111111111111111111111

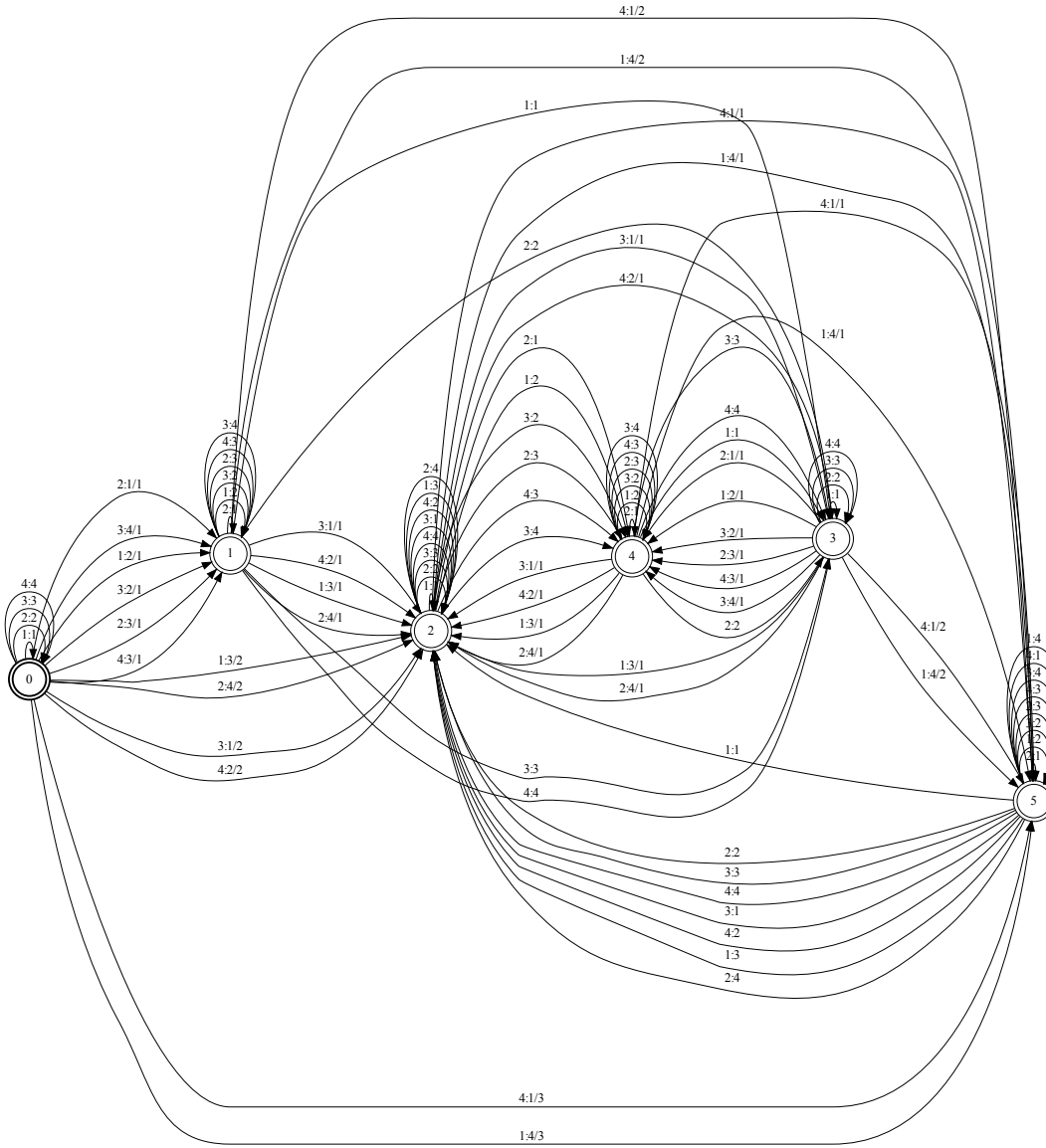
111222222222111111111

111111111111111111111

111222223333322111111

111111111111111111111

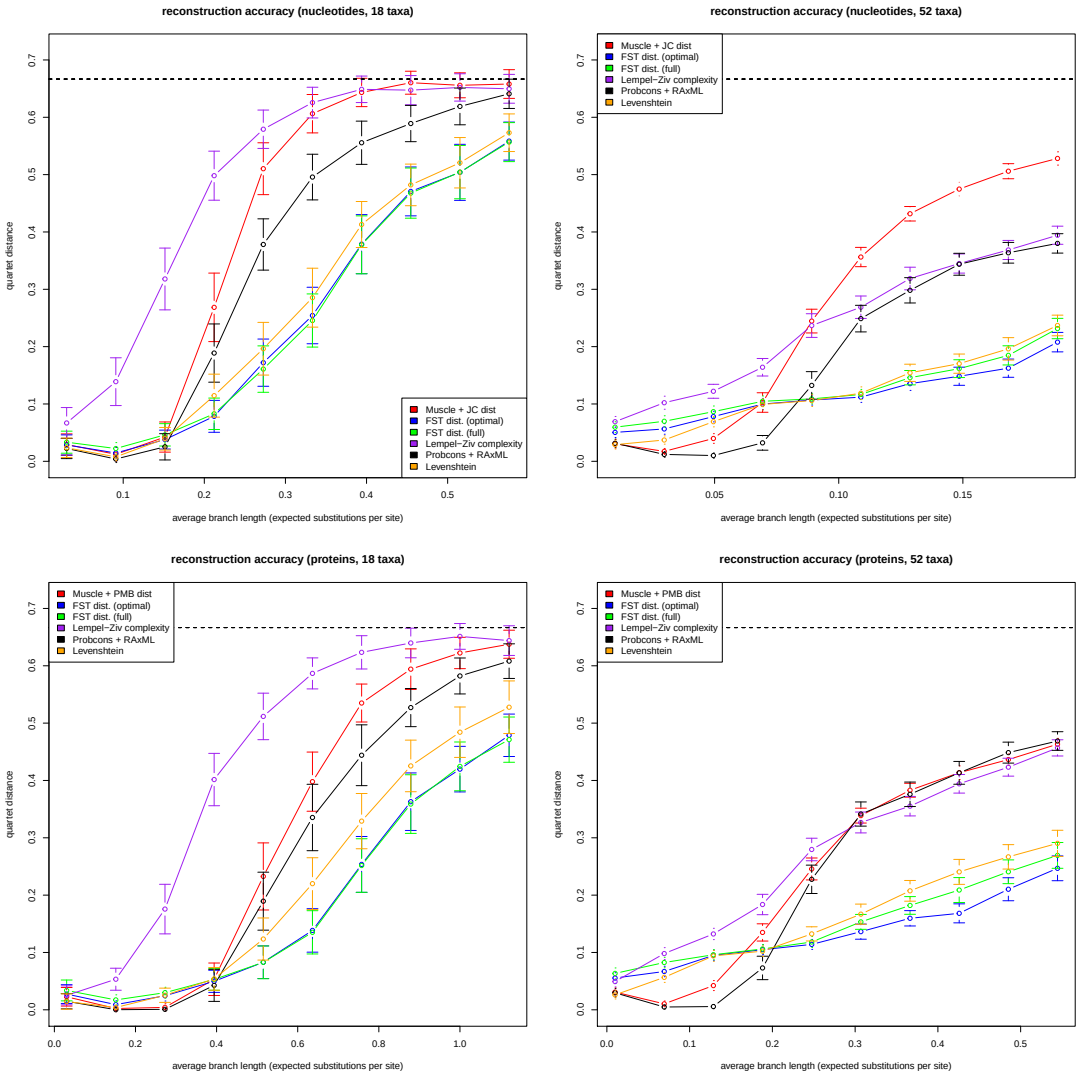
111222223344443322211



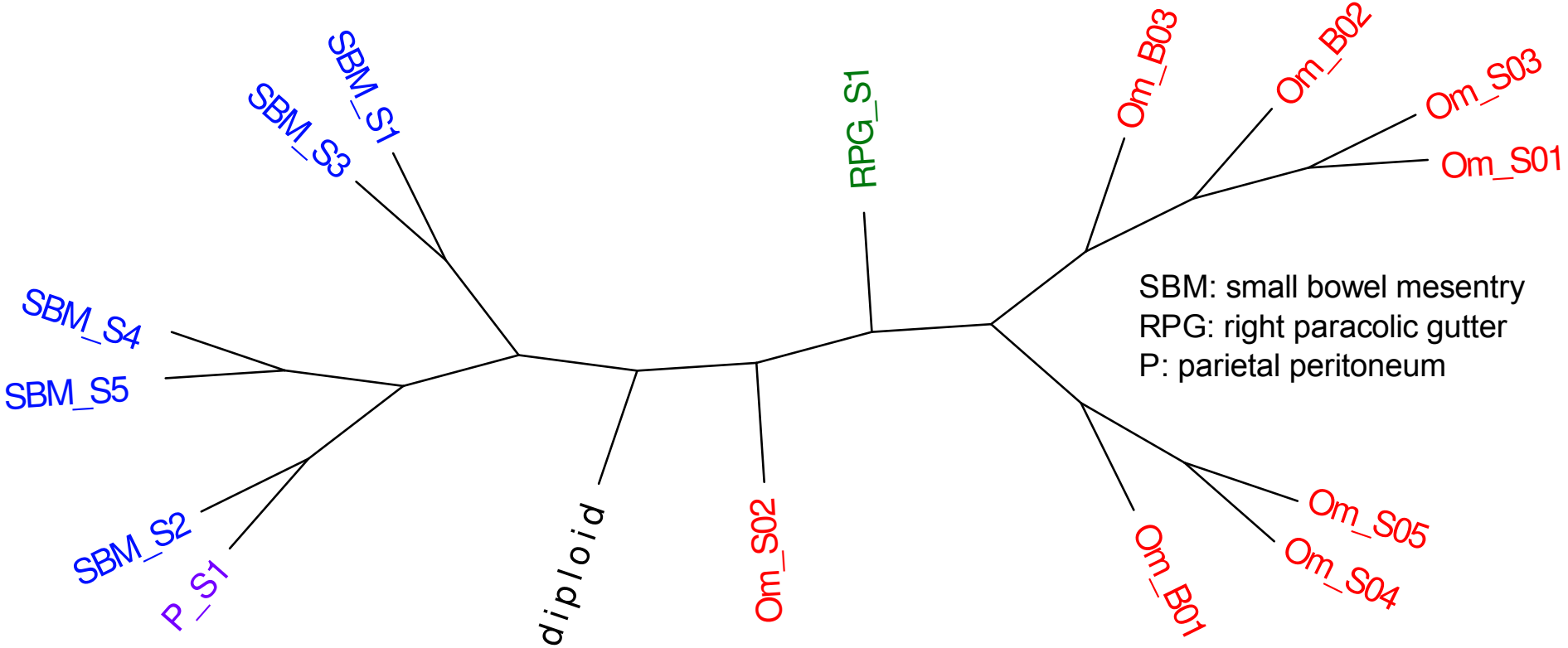
Model

Finite-state transducers

- If the FST is formulated in a specific way ($T = S \circ S^{-1}$), the similarity score is a pd kernel
- This implicitly defines a feature space and an associated metric $d(x,y)^2 = k(x - y, x - y) = k(x,x) - 2k(x,y) + k(y,y)$

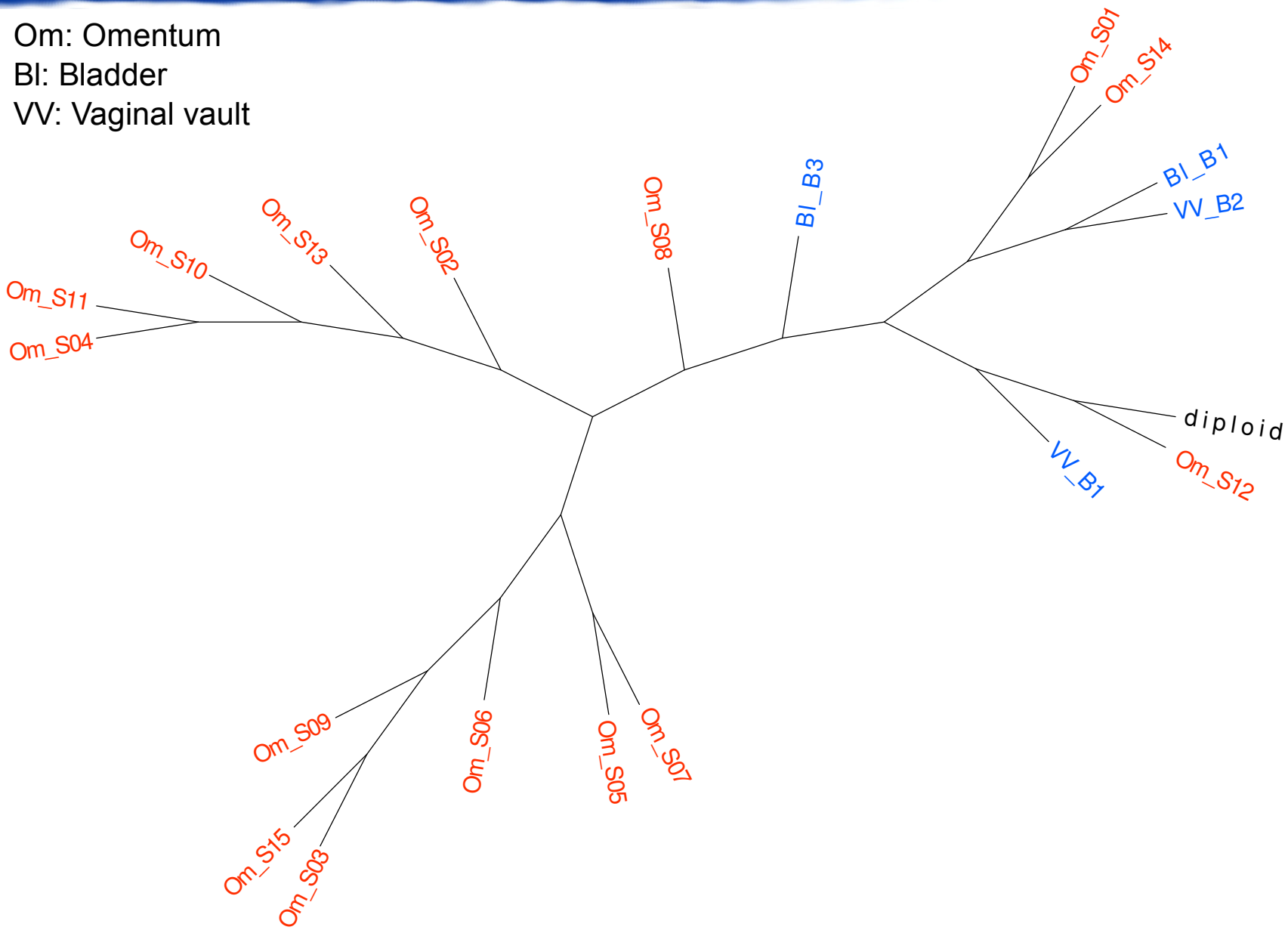


Results


SBM_S1
SBM_S3
SBM_S4
SBM_S5
SBM_S2
P_S1
diploid
Om_S02
RPG_S1
Om_B03
Om_B02
Om_S03
Om_S01
Om_B01
Om_S05
Om_S04
SBM: small bowel mesentry
RPG: right paracolic gutter
P: parietal peritoneum

Results

Om: Omentum
Bl: Bladder
VV: Vaginal vault



Summary

Conclusions

- Finite-state transducers are general tools for sequence comparison
- They allow for both probabilistic or non-probabilistic modelling of events and noise
- Results can be interpreted in a kernel way, i.e. directly applied to e.g. classify the profiles

What's next?

- Estimation of event probabilities and size distributions
- Precursor estimation
- Identification of autapomorphies
- Identification of resistance-driving events

Thanks

Yinyin Yuan, Xin Wang, Mauro Castro and Florian Markowetz

Charlotte Ng, Susie Cooke and James Brenton

CANCER RESEARCH UK

UNIVERSITY OF CAMBRIDGE