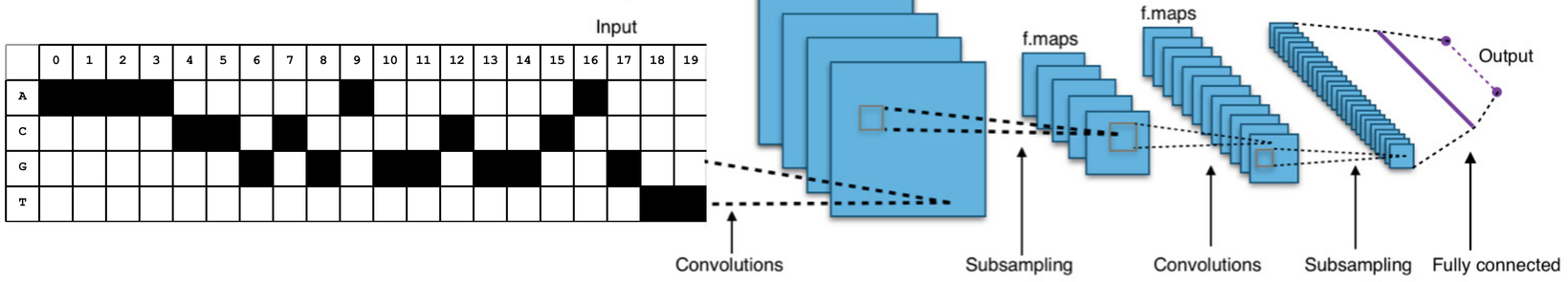
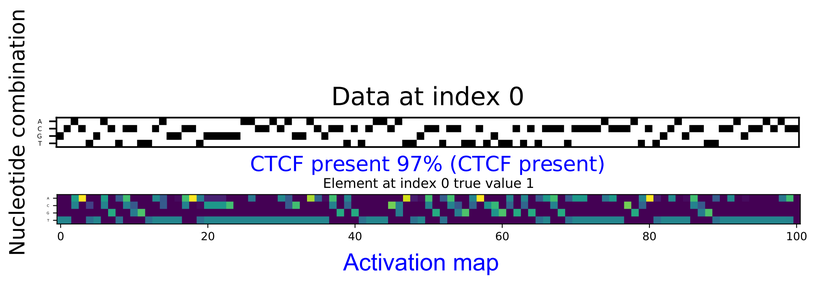
Deep learning has potential for **harmonising** multi-omics data to **discover weak regulatory features**

Tyrone Chen[1], Jason D. Rigby[2], Matthew D. McGee[1], Sonika Tyagi[1]

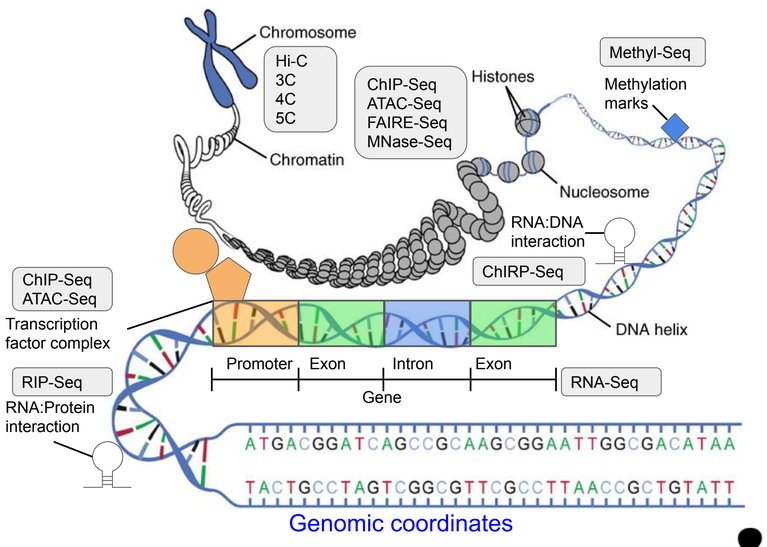
1 School of Biological Sciences, Monash University, Australia; **2** ASPREE National Co-ordinating and Clinical Trial Centre, Monash University, Australia

We have at least 10^8 samples per study for 10^2 studies from databases such as ENCODE and NIH roadmap.

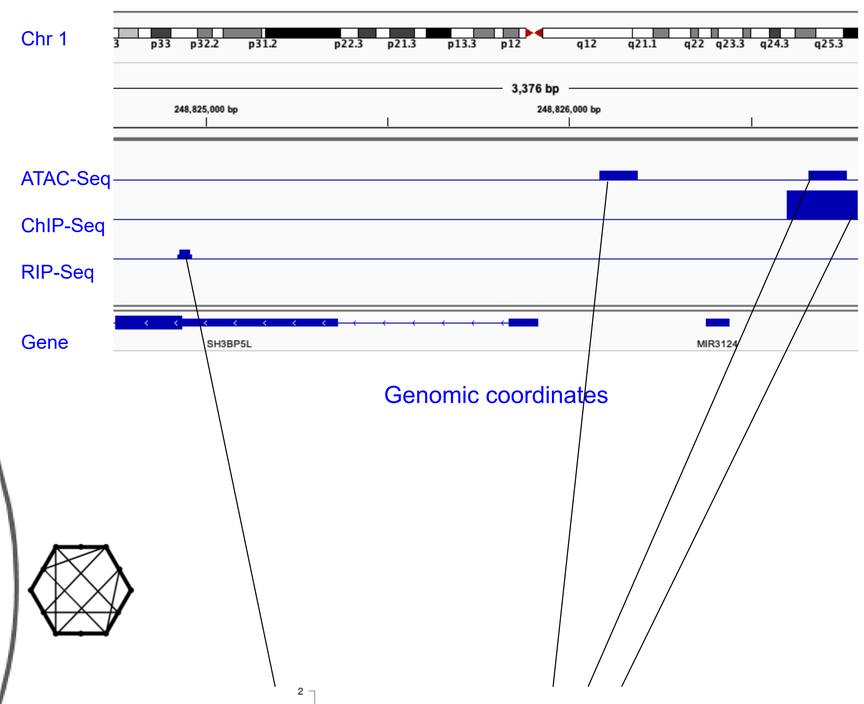


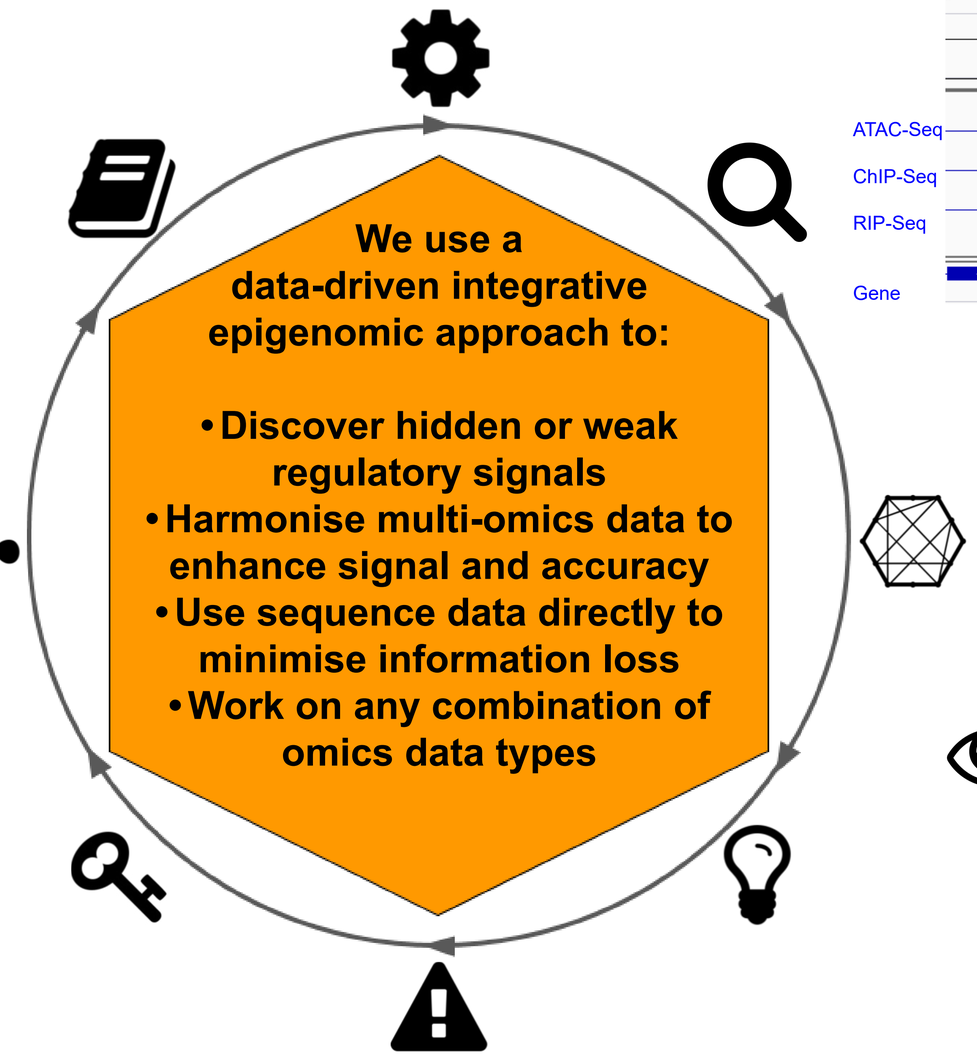


Different assays used to infer the state of biological systems

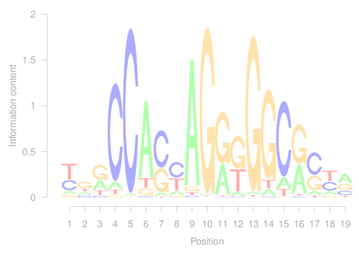


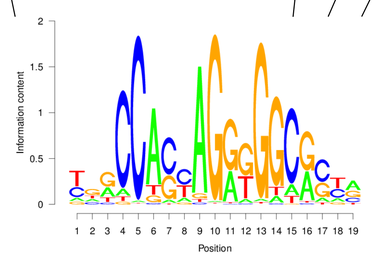
Reads containing the CTCF motif





The CTCF motif is an example of a signal present within data





Expected problems during processing:

- Sequence length distribution differ in datasets
- Signals can be sparse and discontiguous
- Multiple signals may be at the same location
- Category imbalance skews model training

Strategies to resolve expected problems:

- Sequence vectorisation & pyramid pooling can resolve uneven sequence lengths
- Segmentation shows multiple signals in data
- Interleaving and shuffling data redistributes data more evenly

We detect CTCF with an accuracy of 86% in:

- ATAC-Seq
- Chip-Seq
- RIP-Seq data
- 3D genome data integration is in progress

School of Biological Sciences

Genetics, Genomics and Health

0000-0002-9207-0385

tyrone.chen@monash.edu

tyronechen

We acknowledge high performance computing support provided by the Monash eResearch Centre.