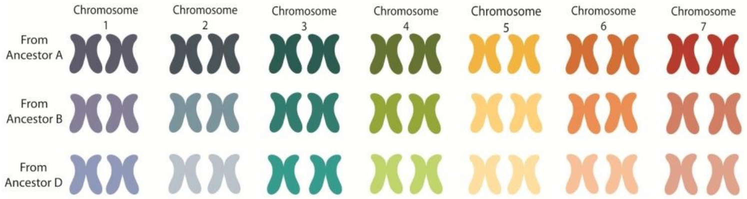
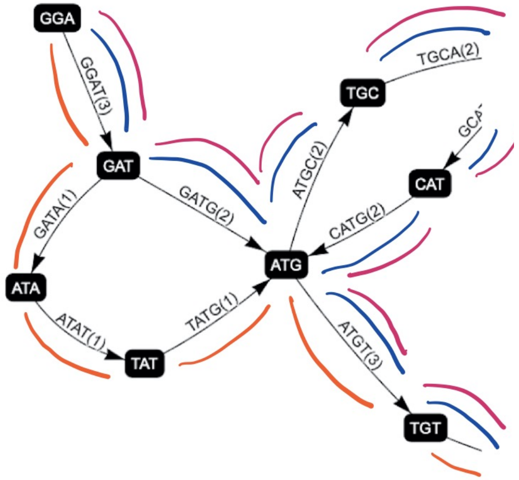
Identifying features for subgenomic sequence identification in a De Bruijn Graph (DBG)

Gillian Reynolds, Jennifer Lachowiec & Veronika Strnadova-Neeley



- When genomes are sequenced, all genomic material is fragmented
- The fragments need to be pieced back together again to be analysed
- A **DBG** is the primary data structure for piecing the genome back together
- Some organisms, like wheat, have multiple genomes – **polyploidy**
- Polyploidy makes traversing a DBG challenging as there are **multiple valid paths in the graph**
- Incorrect path traversal can lead to incorrect genome reconstruction



- We aim to use a **coloured DBG** to assist in graph traversal
- However, the subgenomes still require **pre-graph labeling**
- As such, we have investigated **features** which would allow us to perform fragment labeling prior to graph construction

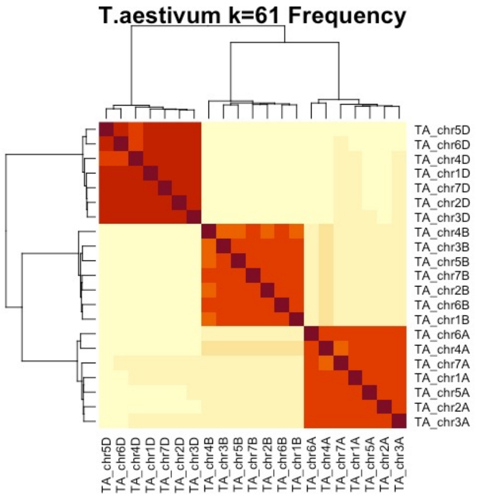
References

1. Pierce, N.T., Irber, L., Reiter, T., Brooks, P. and Brown, C.T., 2019. Large-scale sequence comparisons with sourmash. *F1000Research*, *8*.
2. Bushnell,B. (2017). BBSketch. https://www.biostars.org/p/234837/
3. Wheat genome image obtained from: https://coloradowheat.org/2013/11/why-is-the-wheat-genome-so-complicated

Funding

Research generously funded via the Montana wheat and barley Association

- We have found **k-mer composition** and **frequency** obtained via **MinHash sketches** [1,2] to be excellent features for subgenomic differentiation



Three major challenges remain

1. Determining if these features are abundant in short-read sequence data
2. Evaluating the effectiveness of existing clustering approaches
3. Designing a clustering algorithm for the analysis of large-scale short read data

MONTANA STATE UNIVERSITY