MinHash k-mer sketching highlights allopolyploid subgenome sequence differentiation

Gillian Reynolds [1,2], Dr. Veronika Strnadova-Neeley [2], Dr. Jennifer Lachowiec [1]

1. Department of plant science and plant pathology, Montana state University, Bozeman, MT, USA.
2. Gianforte school of computing, Montana state University, Bozeman, MT, USA.

MONTANA STATE UNIVERSITY

Putting the genome back together

Correctly piecing a genome back together following sequencing is, like doing a jigsaw puzzle, critical if you're going to see the big picture

MONTANA STATE UNIVERSITY

Polyploidy Complications

- For polyploids there's an extra layer of complexity

Allopolyploids

↓

2+ genomes, from 2+ distinct taxa

Autopolyploids

↓

2+ genomes, from the same taxa

Segmental allopolyploids

↓

2+ genomes, a little bit of allo, a little bit of auto

Putting the genome back together

- For polyploids there's an extra layer of complexity
  - Chromosomal assignment
  - Subgenomic assignment
- Getting this right is critical
  - Required for downstream comparative genomics analysis
    - Variant identification
  - Required for further study into polyploid genome structure, function and evolution

Computational strategies for labeling

- Traditional bioinformatics methods of assigning sequence ID are based on read and/or genomic alignment to a reference
  - Costly
  - Degrade in performance as sequence divergence grows
  - Prone to reference bias
  - Prone to multi-mapping read loss

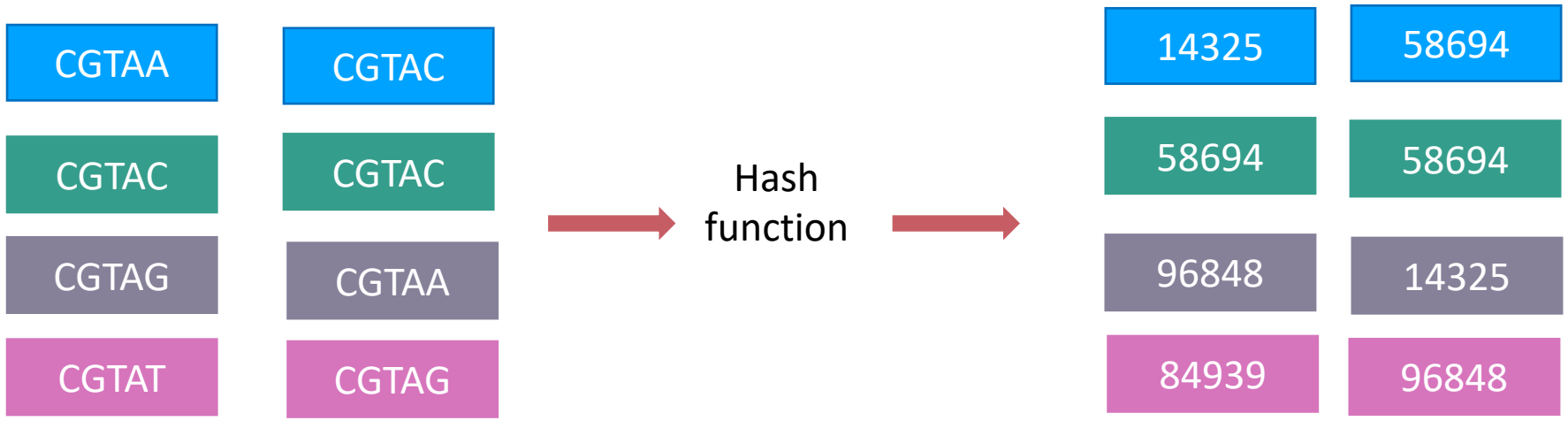
Alignment-free alternatives

- Alignment-free strategies offer a promising new avenue for sequence ID
  - Use statistics to describe features of the sequences
  - Generally linear in time complexity
  - Doesn’t necessarily require reference genomes
    - Less bias
    - Lower space complexity
- One type of alignment-free statistic is MinHash K-mer sketching

MinHash K-mer Sketching

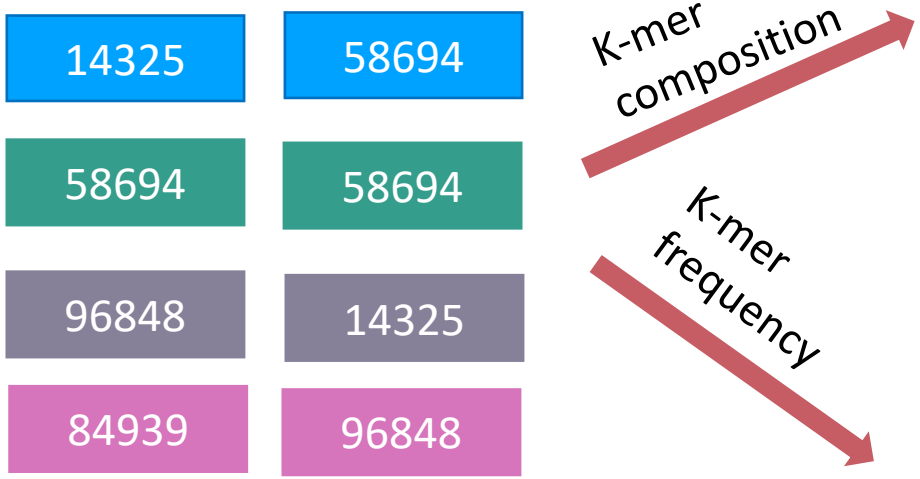
- MinHash K-mer sketching employs the following steps
  1. Break sequences down into their constituent k-mers

2. Convert the K-mers to MinHash sketches via a hash function



MinHash K-mer Sketching

3. Build K-mer composition or k-mer frequency matrix



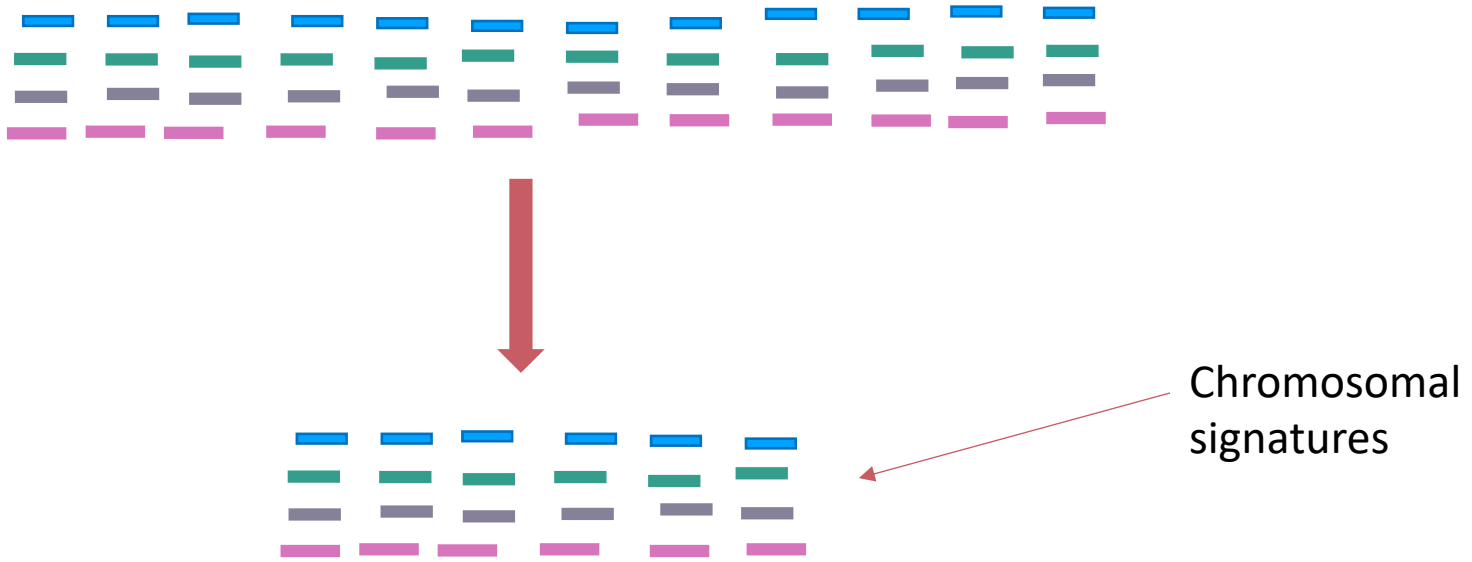
	14325	58694	96848	84939
Chromosome1	1	1	0	0
Chromosome2	0	1	0	0
Chromosome3	1	0	1	0
Chromosome4	0	0	1	1

	14325	58694	96848	84939
Chromosome1	1	1	0	0
Chromosome2	0	2	0	0
Chromosome3	1	0	1	0
Chromosome4	0	0	1	1

MONTANA STATE UNIVERSITY

MinHash K-mer Sketching

4. Downsample the set of MinHash sketches
    - SourMash [1] uses a scaling approach

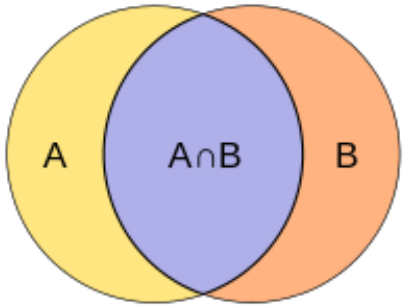


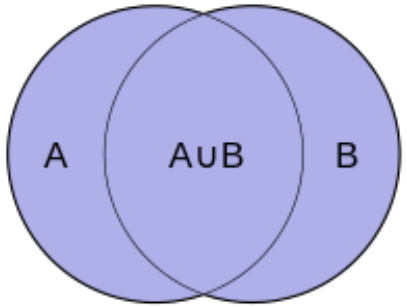
[1] - Pierce, N.T., Irber, L., Reiter, T., Brooks, P. and Brown, C.T., 2019. Large-scale sequence comparisons with sourmash. *F1000Research*, *8*.

MONTANA STATE UNIVERSITY

MinHash K-mer Sketching

- Once a signature has been obtained for every sequence, we can use simple set comparison metrics to compare genomic sequences

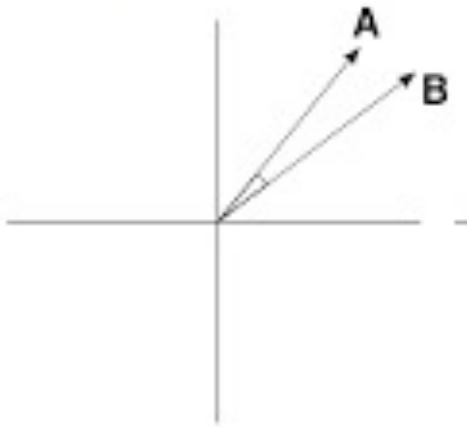


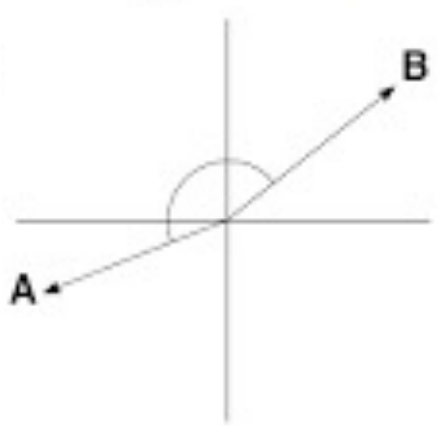


$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

MinHash K-mer Sketching

- Once a signature has been obtained for every sequence, we can use simple set comparison metrics to compare genomic sequences

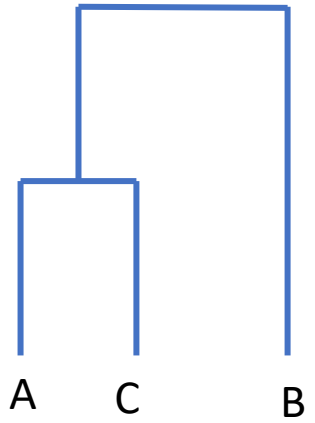




MinHash K-mer Sketching

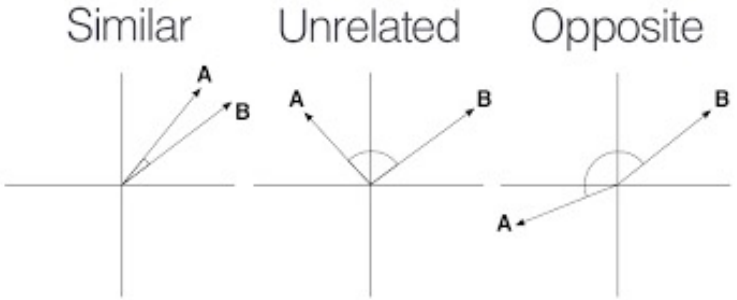
- The result of such a comparison is a pairwise similarity matrix
- This matrix can be used for downstream clustering and visualization of sequence relationships

	Sequence A	Sequence B	Sequence C
Sequence A	1	0.5	0.8
Sequence B	0.5	1	0.2
Sequence C	0.8	0.2	1



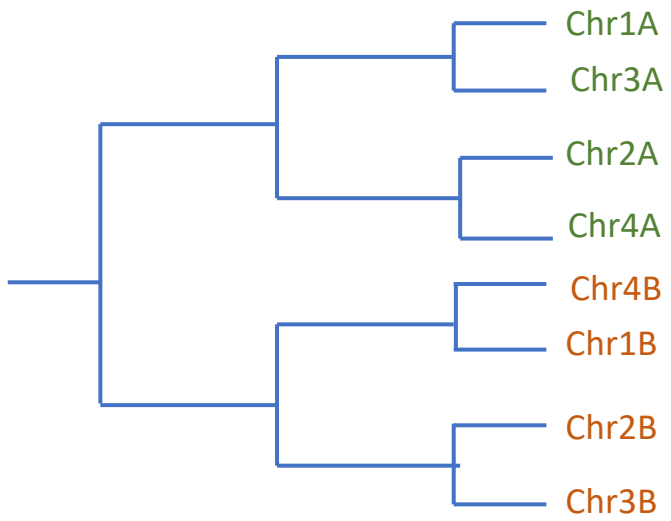
MinHash K-mer Sketching

- This approach has been used, with great success, for metagenomics
  - Problem – One sample, many genomes
  - Solution – Obtain each sequences genomic signature, bin sequences by signature similarities
  - Assumption – Taxonomically related individuals - similar genomic signatures



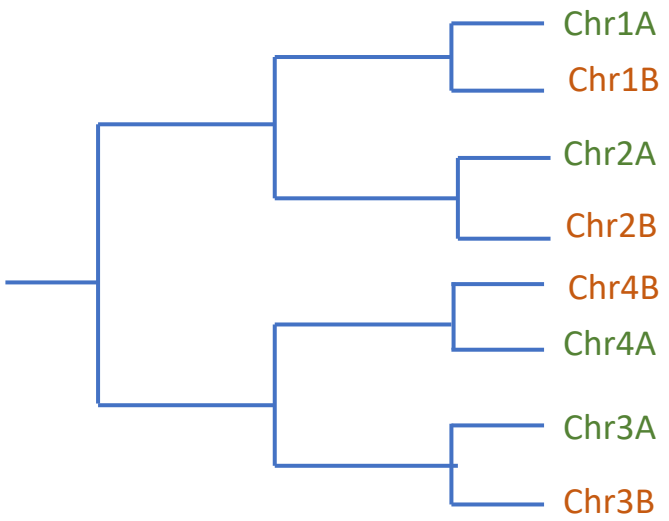
Polyploid Subgenome ID

- Polyploid fragment ID has a similar problem
  - One genome, several subgenomes
- For subgenomes we may see:
  - Chromosomes clustering by subgenome
    - Chromosomes have high intra-subgenomic similarity



Polyploid Subgenome ID

- Polyploid fragment ID has a similar problem
  - One genome, several subgenomes
- For subgenomes we may see:
  - Chromosomes clustering across subgenomes
    - Chromosomes have high inter-subgenomic similarity



Unknown territory

- This strategy has never been applied to the polyploid subgenome differentiation problem before (to the best of my knowledge)
- Aimed to test:
  - How subgenomic sequences cluster
  - How progenitor sequences cluster
  - If different polyploid types exhibited different subgenomic clustering structures

Methods

MONTANA STATE UNIVERSITY

Method Overview

1. Obtain reference chromosomes from NCBI genomes
2. Generate genomic signatures using SourMash4.0
   - Using both composition and frequency
     - Focus today - frequency
   - Tested for a range of k-mers – (4-11, 21, 31, 41, 51, 61)
3. Use SourMash's inbuilt "compare" function to obtain the pairwise similarity matrix
4. Use SourMash's inbuilt "plot" function to cluster the sequences and visualize chromosomal relationships

Genomes for subgenome clustering

- *Brassica carinata* (Ethiopian mustard)
- *Brassica juncea*
- *Brassica napus* (rapeseed)
- *Coffea arabica* (coffee)
- *Gossypium hirsutum* (upland cotton)
- *Gossypium tomentosum* (Hawaiian cotton)
- *Triticum aestivum* (bread wheat)
- *Triticum dicoccoides* (emmer wheat)
- *Triticum Turgidum* (emmer wheat)

Allopolyploids – (genomes obtained from different species)

- *Arachis hypogaea* (peanut)

Segmental allopolyploid – Some allopolyploidy, some autopolyploidy

- *Saccharum spontaneum*
- *Panicum virgatum* (switch grass)

Autopolyploids – genomes obtained from the same species

Our Methods

- The genomes used for progenitor sequence clustering were
  - *Triticum aestivum* (AA,BB,DD)
    - *Triticum Urartu* (AA)
    - *Aegilops tauschii* (DD)
  - *Brassica carinata* (BB,CC)
    - *Brassica nigra* (BB)
    - *Brassica olercea* (CC)
  - *Brassica Juncea* (AA,BB)
    - *Brassica nigra* (BB)
    - *Brassica rapa* (AA)
  - *Brassica napus* (AA,CC)
    - *Brassica olercea* (CC)
    - *Brassica rapa* (AA)
  - *Arachis hypogaea* (AA,BB)
    - *Arachis duranensis* (AA)
    - *Arachis ipaensis* (BB)

Allopolyploids and progenitors

Segmental allopolyploid and progenitors

MONTANA STATE UNIVERSITY

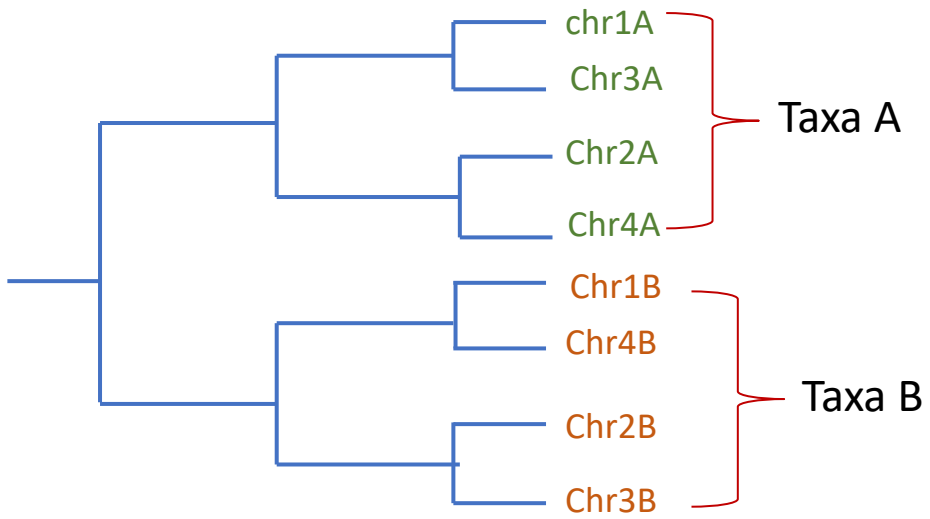
Results

MONTANA STATE UNIVERSITY

K-mer frequency clusters subgenomes

- All allopolyploid show some level of subgenomic clustering for a given k-mer range

- Indicates that allopolyploid subgenomes have higher intra-subgenomic similarity than inter-subgenomic similarity.

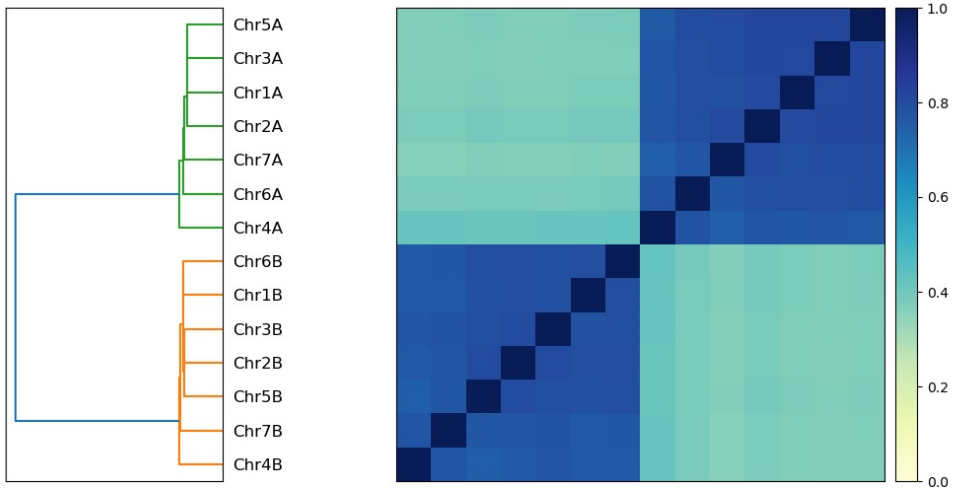


Stable subgenomic clustering

- 6/10 allopolyploid species showed subgenomic clustering for a wide range of k-mer frequency values

T. dicoccoides k=7 frequency

T. dicoccoides k=61 frequency



MONTANA STATE UNIVERSITY

Sporadic subgenomic clustering

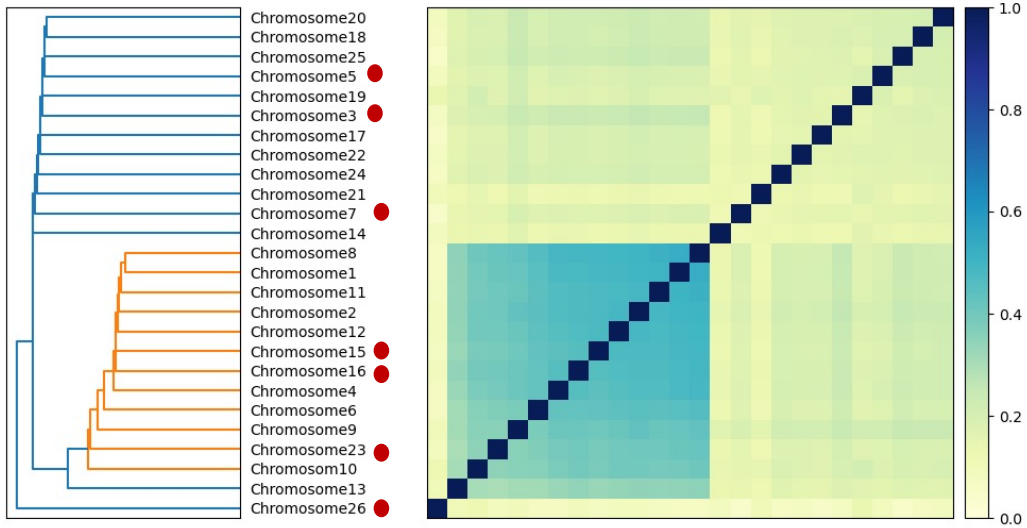
- 3/10 allopolyploids showed shorter, more sporadic ranges
  - K=7, 21-61 (*B.Juncea*)
  - K=8,11-31, 51-61 (*B.napus*)
  - K=21-41, 61 (*C.arabica*)

Consistently incorrect

- 1 allopolyploid (*G. hirusutum*) showed subgenomic clustering with consistent outliers
  - Chromosomes 3, 5, 7 and 15,16 and 23 were consistently clustered in the incorrect subgenome
  - For larger K values (21-61) chromosome 26 was consistently an outlier sequence

G. hirusutum k=7 frequency

G. hirusutum k=61 frequency

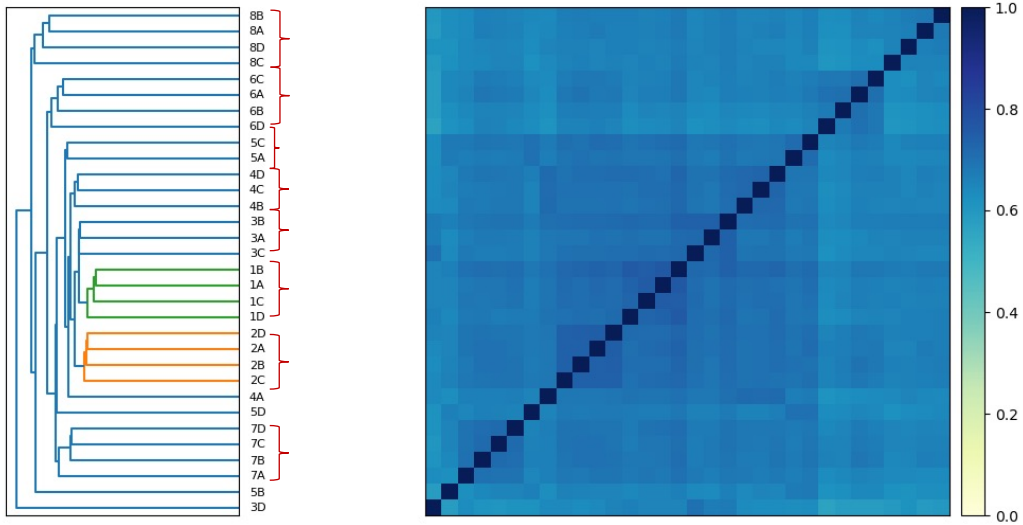


MONTANA STATE UNIVERSITY

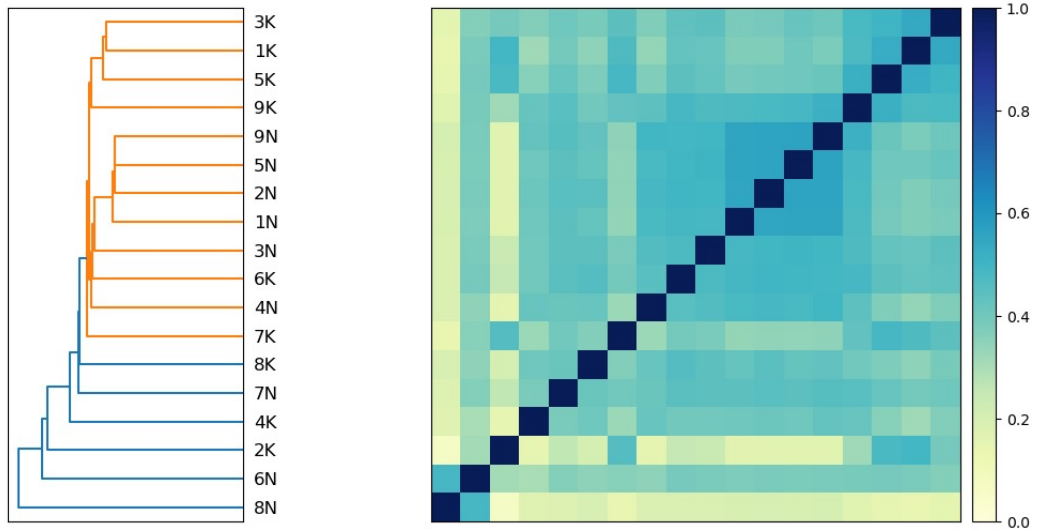
Autopolyploids don’t follow the trend

- Neither autopolyploids clustered by subgenome
  - 1 showed a tendency to cluster by inter-subgenomically (*S.spontaneum*)
  - 1 showed no discernable clustering structure (*P.Virgatum*)

S. Spontaneum k =31 frequency



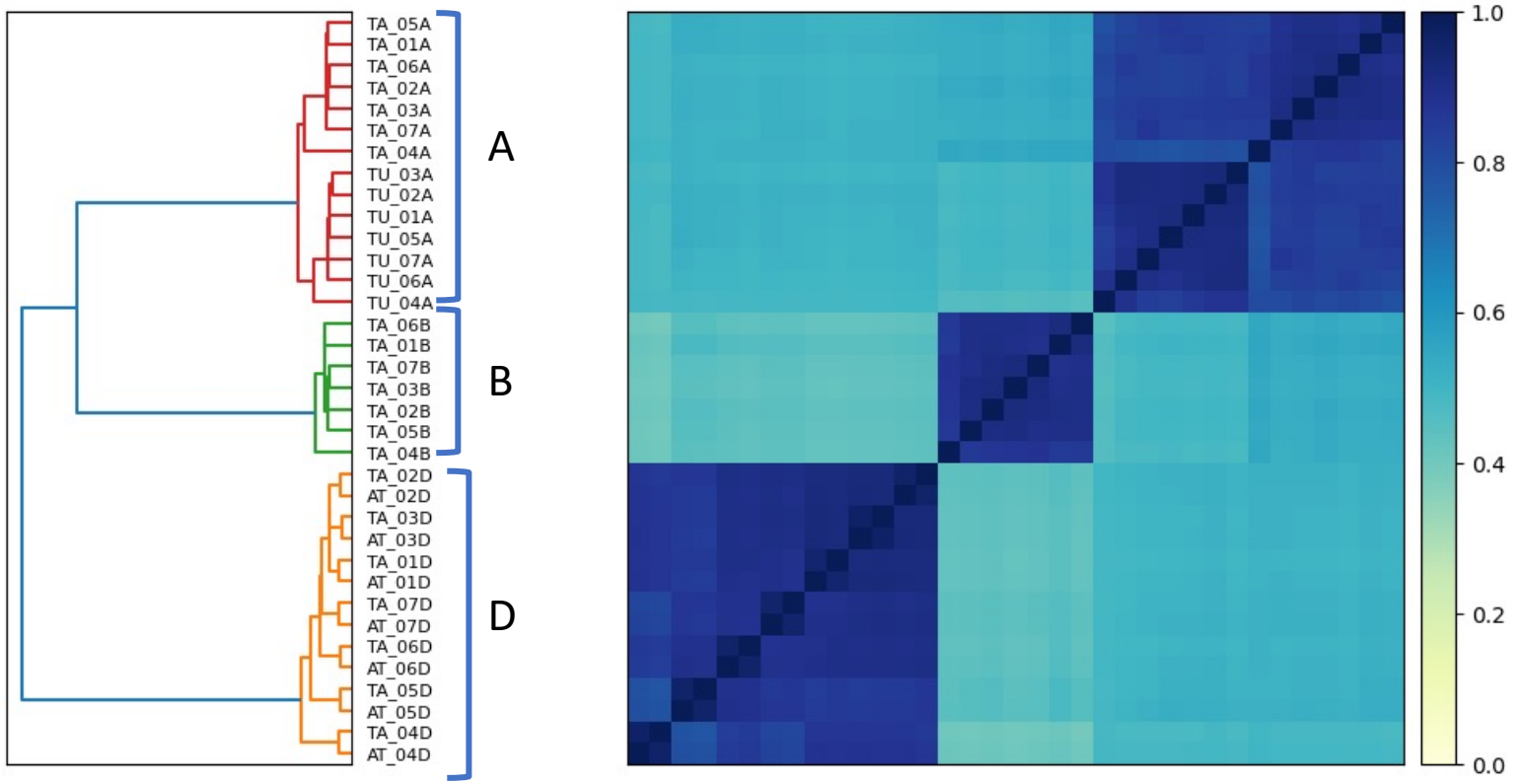
P. Virgatum k=31 frequency



MONTANA STATE UNIVERSITY

K-mer frequency clusters subgenomic progenitors

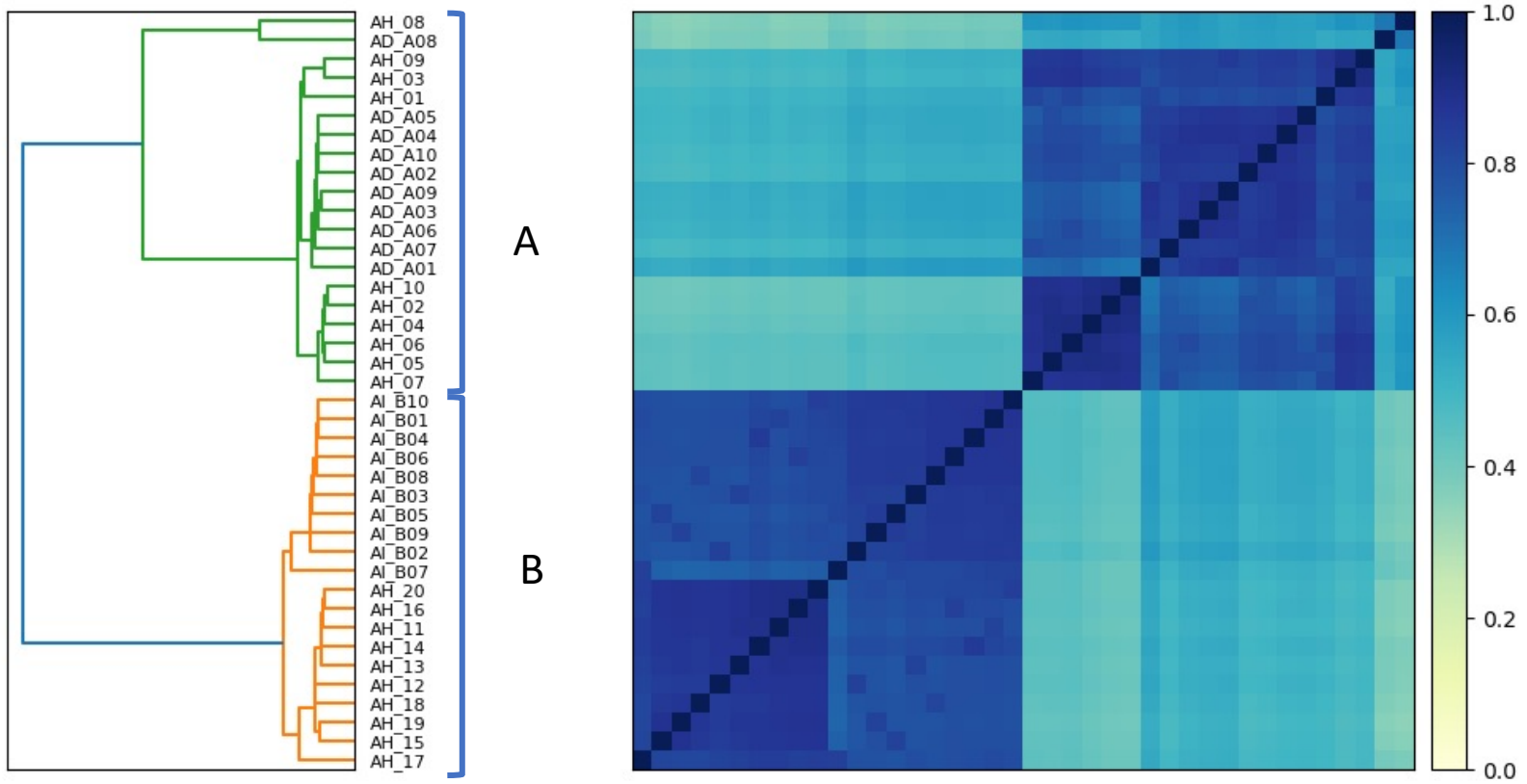
T.aestivum(AABBDD), *T.urartu* (AA), A.tauschii(DD), k=21, k-mer frequency



MONTANA STATE UNIVERSITY

Works for the segmental allopolyploid too

A.duranensis, A.hypogaea, A.ipaensis, k=21, k-mer frequency



MONTANA STATE UNIVERSITY

Conclusions

- K-mer frequency is a reliable subgenomic signature which may be applied to problems such as
  - Subgenomic clustering
  - Progenitor clustering
  - Ploidy typing

Future work

1. Testing a wider variety of taxa
2. Explore the application of this method to draft genome assemblies
3. Explore application of this method for unassigned sequences (’Un chromosome’)
4. Exploring biological significance of k-mer frequency profiles

Acknowledgements

Thank you my two supervisors, Dr. Jennifer Lachowiec and Dr. Vernokia Strnadova-Neeley for their continued help, support and encouragement with this project

Thank you to Dr. Brendan Mumeys group and Dr. Lachowiecs lab for helpful feedback

Thank you to the Montana Wheat and Barley Association for their funding

Thank you Dr. C Titus Brown (co-author of sourmash) for his help extending sourmashs' capabilities

MONTANA STATE UNIVERSITY

Summary

- Investigated the ability of MinHash sketching to identify inter- and intra-subgenomic similarity
- Identified that k-mer frequency MinHash sketching accurately reflects:
  - Subgenomic assignment of chromosomes
  - Progenitor origin of subgenomic sequences
  - Polyploidy type
- We have identified a number of areas for further study

Questions?

MONTANA STATE UNIVERSITY