# SubKluster: Novel method to bin scaffolds from cereal  genomes into subgenomes using substring frequency analysis

Victor Kalbskopf[1*], Nick Sirijovski[123], Dag Ahren[1]

[1]Department of Biology, Lund University, Lund, Sweden
[2]SCANOATS
[3]LTH, Lund University, Lund, Sweden
[*]Corresponding Author
E-mail: vi4227ka-s@student.lu.se

## Abstract

The genome of the Belinda variety of the hexaploid oat (*Avena sativa*) has recently been sequenced and assembled. This project aims to improve the assembly by clustering the thousands of scaffolds into their three ancestral subgenomes using Principle Component Analysis (PCA) of kmer and repeat-element frequencies. The method was developed using a chromosome level assembly of hexaploid Wheat (*Tritium aestivum*), which formed highly distinguishable subgenome true clusters in their PCA graph, which indicates that the method has merit. The longest scaffolds of oats that formed 90% of the genome (N90) were processed in the same manner, and which resulted in 2 clusters, one with about one third of the 3-copy BUSCOs (Benchmarking Universal Single-Copy Orthologs), and another with two thirds. The latter cluster could

then be subdivided into two clusters, with about half of the 2-copy BUSCOs in each cluster. A one:one:one ratio of BUSCOs in each cluster would indicate that the subgenomes are dividing into their respective clusters. The clustering is not neat or as clear as in the wheat example, but the length of the scaffolds or the state of the assembly may have a very large effect on the efficacy of the method. It is hoped that this method, with additional improvements, could be used to assess the assemblies of other large polyploid genomes and be part of a larger pipeline for understanding crop genome evolution.

# Introduction

Like most crops, *A. sativa* has a large, complex genome which has resisted thorough sequencing and assembly. The difficulties are 3-fold: (I). It is filled with repeats, from many sources, including coding regions, like rRNA, and noncoding, like long tandem repeat (LTR) retrotransposons or mini/micro-satellites; (II). Gene duplication and deletion; and (III). Allopolyploidy (1). All together, this results in very large genomes with very low gene density and large regions of heterochromatin. The consequence of which is the C-value paradox. The C-value is the amount of DNA in a haploid genome measured in picograms. It is a paradox because one expects the C-value to scale linearly with gene content, as seen in prokaryotes and 'simple' eukaryotes (2). But when repeat elements make up the majority (90-95%) of most plant genomes (3), we can see why this expectation breaks down.

## Repeat elements complicate assembly

The most commonly used, cheapest sequencing technology used for sequencing whole genomes today is Illumina. It produces reads of 150-300 base-pairs (bp) long. A study on *Triticum aestivum* (bread wheat) found the average length of the longest retroelements, which make up 50% of the chromosomes in question, to be 571bp (4). This number can be much longer or shorter. Because even the longest illumina reads can not span that repeat, the placement of that read to form a scaffold would be not much more than a guess. This is because the repeat, by it's very nature, will occur multiple places in the genome, so the assembler will not know if this is a duplicate read, or a duplicate repeat. To get around this problem, mate-pair libraries are used, whereby the sequencing primers are separated by long fragments (3 kbs, as a typical example) that aren't sequenced, but the regions downstream of the primers are, at 150 bps long. These mate-pairs can span repeat-rich regions and allow the assembler to allocate reads more accurately (5). However, repeats can repeat on themselves, and far exceed the 3kbp of the mate-pair. Another solution is to use linked-read technology. One such technology, 10X Chromium, is a library preparation system that uses unique barcodes added to short reads that originated from one long DNA fragment. These can then be linked in *silico* post sequencing, constructed into their original fragments, and used to span the long repetitive regions (6). This is analogous to using the much more expensive BAC cloning and genetic mapping methods, which was used to sequence the wheat genome that has the same challenges as oats (7). However, 10X Chromium requires very high quality, high molecular weight DNA during the barcoding process, and it is still sensitive to all the weaknesses of Illumina sequencing, as that is how the barcoded fragments are sequenced.

## Polyploidy

*A. sativa* has 3 subgenomes designated A, C, and D. Each subgenome has 14 diploid chromosomes, which means a total count of 42 chromosomes. The allohexaploid we have today was formed by 2 distinct steps. An ancient diploid progenitor genome designated A’, underwent hybridisation with with another diploid C-genome, to form a tetraploid CA’. This is now known as CD, because the A’ progenitor is unrecognisable relative to all known accessions, or the A’-genome progenitor is extinct. CD experienced a hexaploidy event with a more contemporary A genome, to form the ACD (AACCDD) genome we have today (8) (See example of wheat genome evolution in Discussion, Fig. 17). Not only did this process triple the size of a “conventional” diploid genome, which increases the cost of sequencing it, but it also complicates the assembly process. Assemblers require uniqueness in their reads, but 6 similar chromosomes will provide 6 similar reads, assuming coverage of 1x. This complication can result in the construction of chimeric chromosomes, with a mixture of different subgenomes in one scaffold. Scaffolds are assembled by connecting contigs - short spans created by overlapping reads. Contigs are stitched together using mate-pair libraries or/and long-read technologies like 10X Chromium to form scaffolds. Due to increased complexity, scaffolds become short, to avoid chimeracy in low confidence predictions (9). Misassembly can also make downstream analysis difficult. This is exemplified in the allohexaploid *A. sativa*, which was sequenced recently. The quality of the assembly was good, given the low costs involved, as all sequencing used standard Illumina short-read sequencing in addition to 10X Chromium library preparation. The N90 (represents the length of the smallest scaffold which is part of the largest 90% by length) is 2.8 Mbp, and includes 693 scaffolds. The longest is 113.8 Mbp. While this is a great step forward, there is much potential for improvement. But this process could be simplified if the scaffolds could be assigned (binned) to their respective subgenomes.

## Kmer analysis

Polyploidy is not the only source of complexity for assemblers. Metagenomic sequences may contain DNA from hundreds of taxa in a single sample. Assemblers designed for this data use a combination of GC content and kmer frequency analysis, among other things, to bin the reads and contigs into their respective species (10). It was suggested that perhaps a similar approach could be applied to the 523,398 oats scaffolds. The subgenomes may be different enough so that the kmer profile would be unique for each subgenome, and  provide 3 bins - one for each subgenome. Alternatively, it is possible that the similarity is along homeologous chromosomes - 7 bins (one for each chromosome number, 1-7) to which we can assign each scaffold. Kmers are any sequence of length *k*. So the arbitrary ACCTTGA is a kmer of length 7 - a 7mer, and ACGGTACCATA designated Ɲ is a 11mer.  Ɲ has ACGG, CGGT, GGTA, GTAC, TACC, ACCA, CCAT, and, CATA as 4mers (known as tetramers), for example. Since DNA is double stranded, one can also search for kmers in the reverse strand, but only the forward strand was used in this project.  If certain kmers are more populous on one subgenome, thanks to retrotransposon activity or contributions from the parent genome, that may make that subgenome distinct enough to differentiate those scaffolds from the mass of others. This pattern could be revealed by statistical analysis. One typical method is to use Principal Component Analysis (PCA) which ‘summarises’ the effects of multiple variables and reveals them on a coordinate plane, in at least 2 dimensions.

## BUSCO Analysis

The Benchmarking Universal Single-Copy Orthologs (BUSCO) (11) is a method to evaluate the completeness of a genome. A

BUSCO is a sequence, usually a gene, that is expected to be present once in a haploid genome. One can then define a set (database) of BUSCOs for a species or taxa. If one were to do a denovo sequencing of a species, then the BUSCO analysis of the new assembly can be analysed using a BUSCO database from a closely related species. Then one can compare the commonality if the BUSCO sets between the reference taxa and the new genome. If they are close to identical, then one can assume that the denove assembly was reasonably successful.

## Methods

Illustration 1 succinctly describes SubKluster, the pipeline designed to place (bin) scaffolds into subgenome clusters using kmer frequency analysis. The process involves counting kmers, tabulating the counts, and performing a PCA on that table. The results from a BUSCO (Ver 3.0.1) (11) analysis show which complete BUSCOs from the *Zea mayz* (Maize) database can be found in each scaffold. This list is imported into the R script, and is used to assess whether the clusters represent single subgenomes, since a large set of single copy BUSCOs in each cluster would indicate that the scaffolds were binned correctly. Bash scripts turned the various  programs into a pipeline.

At the time of writing, neither the wheat nor oats genomes used in this paper have been published, but both were used with permission. The wheat reference is the IWGSC RefSeq v1.0

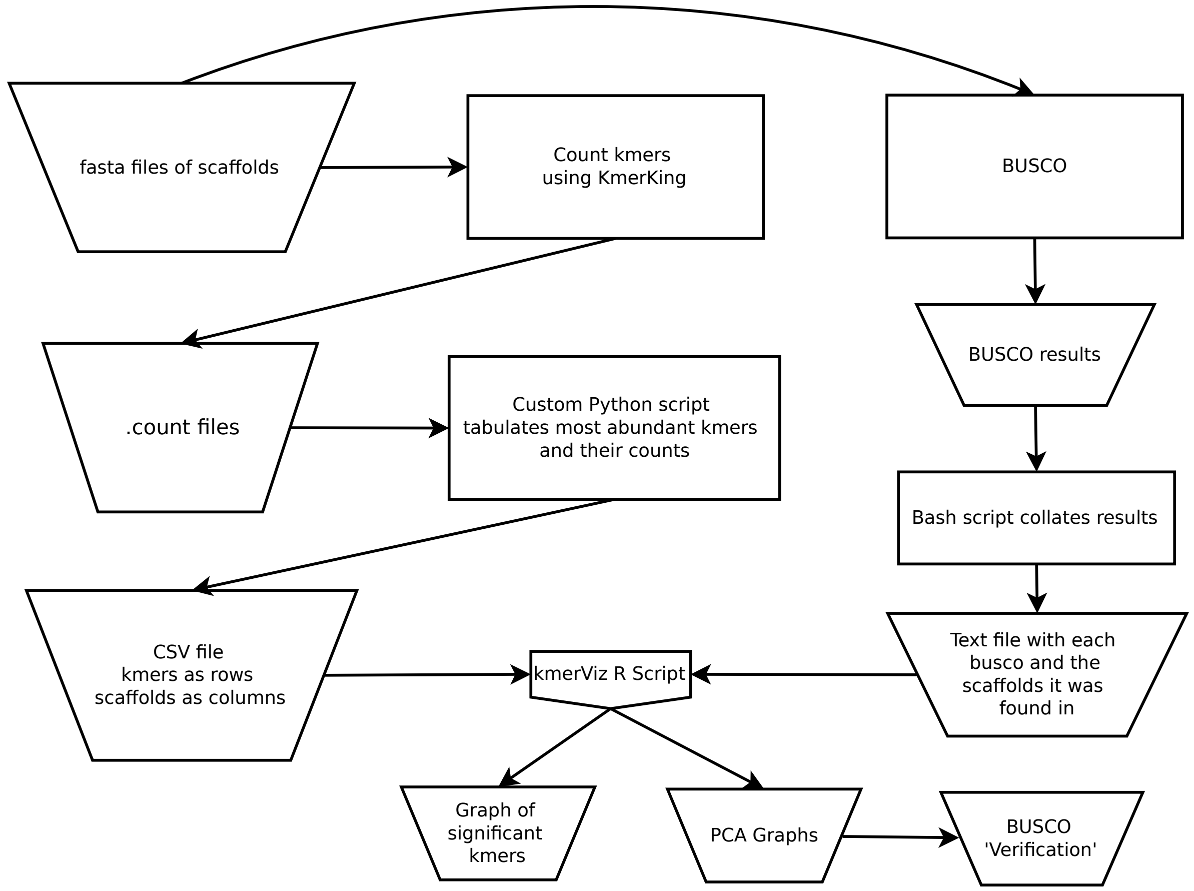

fasta files of scaffolds
Count kmers using KmerKing
BUSCO
.count files
Custom Python script tabulates most abundant kmers and their counts
BUSCO results
Bash script collates results
CSV file kmers as rows scaffolds as columns
kmerViz R Script
Text file with each busco and the scaffolds it was found in
Graph of significant kmers
PCA Graphs
BUSCO 'Verification'

**Illustration 1 : Bioinformatic pipeline of SubKluster.** Fasta files that contain one scaffold each are processed by KmerKing (Canbäck, unpublished), which produces one count file for each scaffold. The count files are tab delimited: first column has the kmer, and the second has that kmer's count in that scaffold. These count files are imported into a custom python script that collects the most abundant kmers given a threshold, and produces a table in text format of the kmers, and their counts for each scaffold. In parallel, BUSCO analysis is performed on the same fasta files. These results are collated, selecting only complete BUSCOs. The text file contains a list of BUSCOs and the scaffolds in which they were found.

The table and list are imported into the kmerViz R script which produces PCA graphs that show clusters of scaffolds, and graphs indicating the most influential kmers. A cluster of scaffolds that contains a large set of complete BUSCOs was interpreted as a subgenome.

assembly, kindly provided by the International Wheat Genome Sequencing Consortium, and the oats genome identifier is NRQ-11003, kindly provided by SCANOATS (Industrial Research Centre).

An alternate source of data to kmer frequency are the biological repeats themselves. The Poaceae repeat database was downloaded from http://pgsb.helmholtz-muenchen.de (12), and used as the source for the Blast Like Alignment Tool (BLAT) (13), which searched for the repeats in the fasta files. Each hit was counted using a bash script that produced the same .count files that were imported into the Python script. This proceeded exactly the same way as data obtained from kmer counts.

## Software Details

KmerKing (unpublished) was used to count the kmers in each fasta file. For k>12, it only reported kmers that occurred at least 4 times in that file. This reduced the size of the count files for the next step. The Python script requires version 3.6 and up, but only uses standard modules. It includes optimisations to shorten run-time if it needs to run again on the same data with a different threshold, by storing a compressed version of an intermediate step. After multiple iterations of development, it only reads and writes to the hard drive twice and once respectively, but it can be sped up by using faster storage, like a Solid State Drive (SSD). The script provides information to the user about current progress, and also makes some very loose estimations for how long the current step in the process will take. When generating the PCA, a 1.9 GB table used up to 46.6 GB of RAM, but the generation of the table used 8.5 GB. These values are completely dependant on k and the number of scaffolds being analysed. The R (ver 3.4.4) Script used ggplot2 (ver 2.2.1), ggthemes (3.5.0), ggrepel (ver 0.8.0), plot3D (ver 1.1.1),

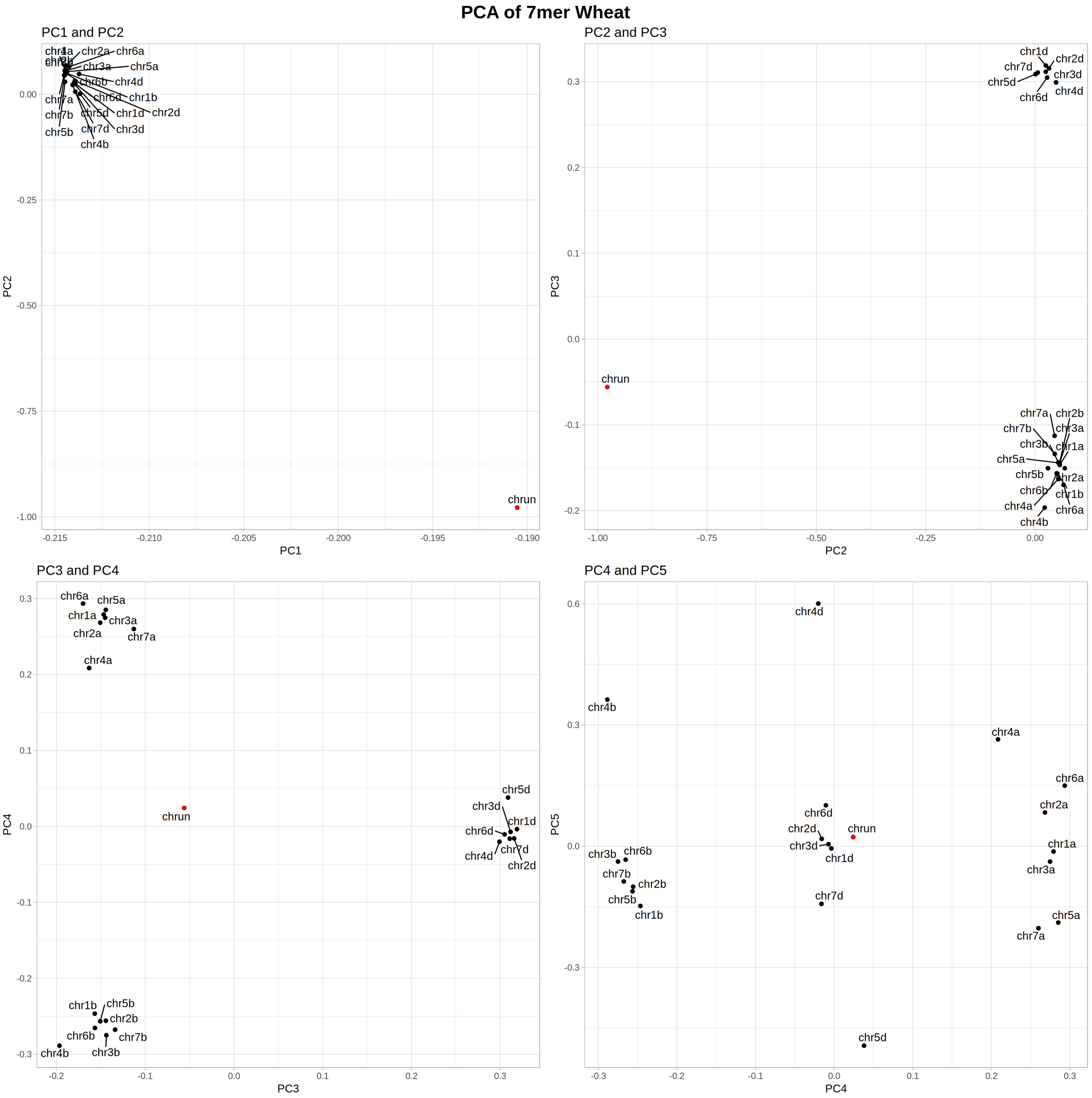


**Figure 1 The first 5 principals of a PCA of 7mer counts of  the wheat pseudomolecule assembly.** Each new component reveals new clusters of chromosomes. The labels indicate what the chromosome is, and the last letter (a, b or d) indicates the subgenome. The ‘chrun’ scaffold is the collection of unknown sequences found during the assembly. It is  responsible for most of the variation for the PC1, PC2, and PC3. However, in PC2 and PC3, we see the chromosomes dividing into 2 groups (not including the unknown ‘chromosome’), with subgenomes A and B in one group, and subgenome D in the other. PC3 and PC4 is the most illustrative, separating out all the subgenomes clearly. PC4 and PC5 also reveals 3 looser clusters, but it’s pointing to similarities in chr4 between the subgenomes.

plot3Drgl (ver 1.0.1), gridExtra ( ver 2.3), and grid (ver 3.4.4) packages , all for plotting.

## Results

The effort began by using 7mers from wheat (Fig. 1). Later, when analysing much greater numbers of scaffolds, the default PCA plotting packages did not work, so the '% of variability' one expects to see in PCA plots was not generated. For the sake of consistency, it is omitted in the wheat genome plots. Figure 1 plots the first 4 components of the 7mer counts. The scaffold (chrun) containing sequences unplaceable by the assembler is responsible for most of the variation for the first 3 components. When this 'chromosome' is removed prior to PCA generation, the clusters form perfectly in the first two components (data not shown). There are hints of homeology in the relative positioning of a few of the other chromosomes seen in PC4 and PC5. The chromosome 4 (chr4) scaffolds are the highest on the PC5 scale for their representative subgenomes, and though less obvious, the same is true for chromosome 6, after chromosome 4. Using a larger k also improves the resolution of clusters (Fig. 3, 11mers are used), but it also increases the number of kmers searched by a very large factor. Assuming a search space of at least 2 x k, the number of kmers (n) is given by n = $z^k$, where z is the size of the alphabet, therefore n scales exponentially with k. So for  standard DNA,

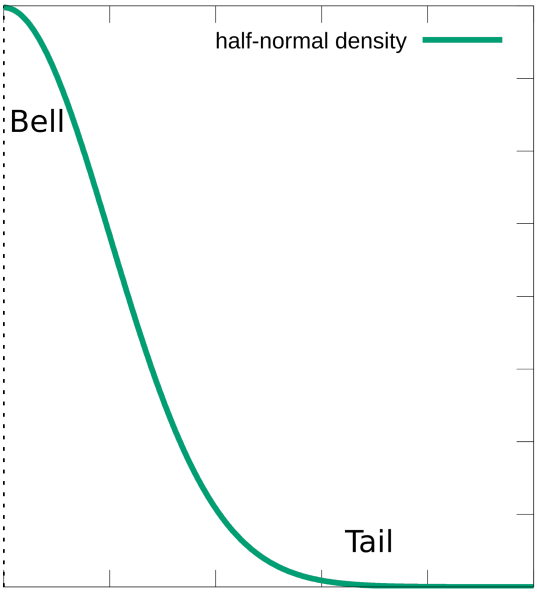


**Figure 2: A generic example of the right side of a normal density distribution graph.** Y-axis is the denisty of any particular value found on the x-axis. The top left part is known as the bell, and the lower right part is the tail.

$n = 4^k$. For k=7, we get 16348 kmers. In Figure 2, k=11, resulting in 4.194304 x$10^6$ 11mers. This can still be analysed on a desktop workstation, given 22 chromosomes, but we will see later, each kmer count has to be represented for 693 scaffolds. This requires RAM available only to large servers with RAM in the hundreds of gigabytes, and analyses taking over 36 hours. Therefore we attempted to reduce the number of kmers required to generate distinct clusters.

## Data reduction through filtering

The distribution of kmer counts looks like the right half of a binomial distribution (Fig. 2), but the middle, “bell” is thinner, and the “tail” is much

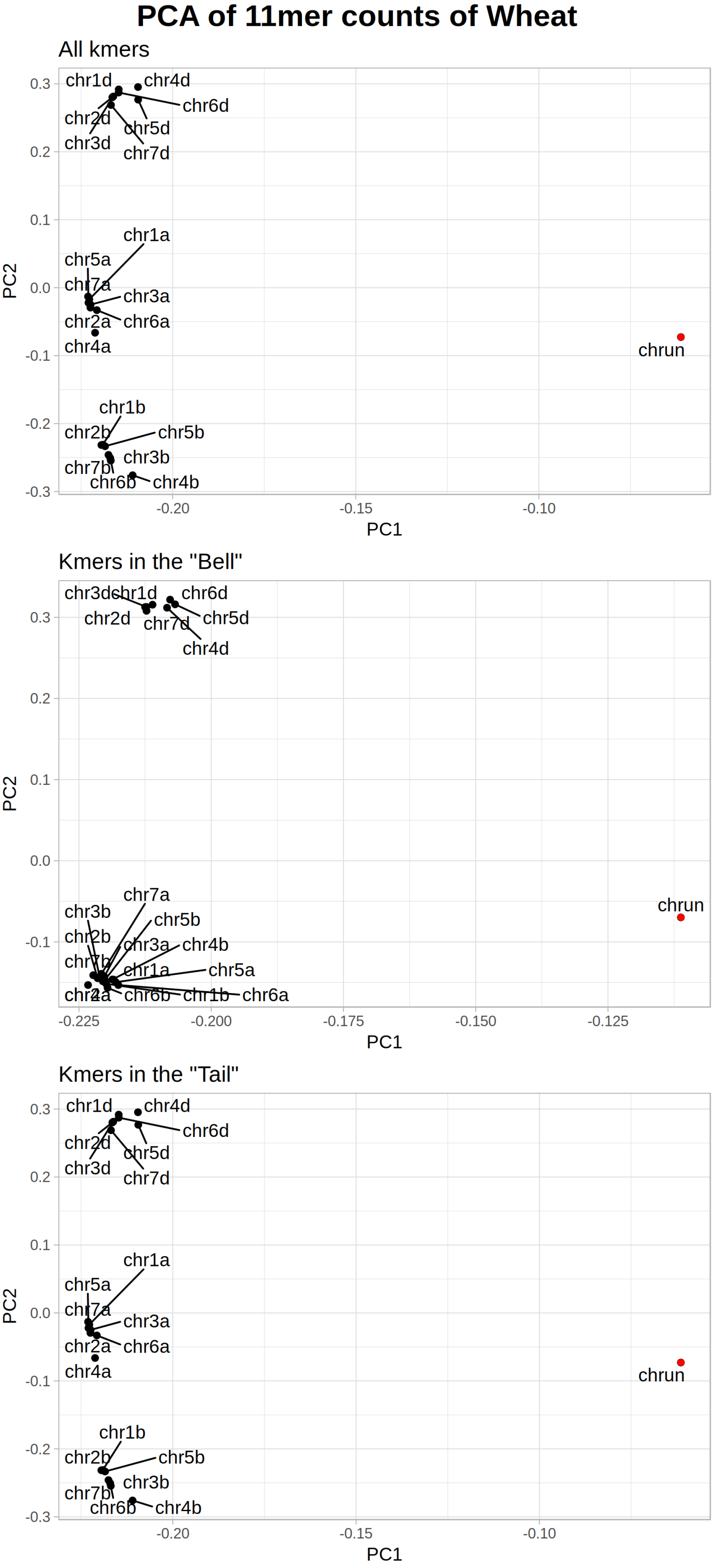
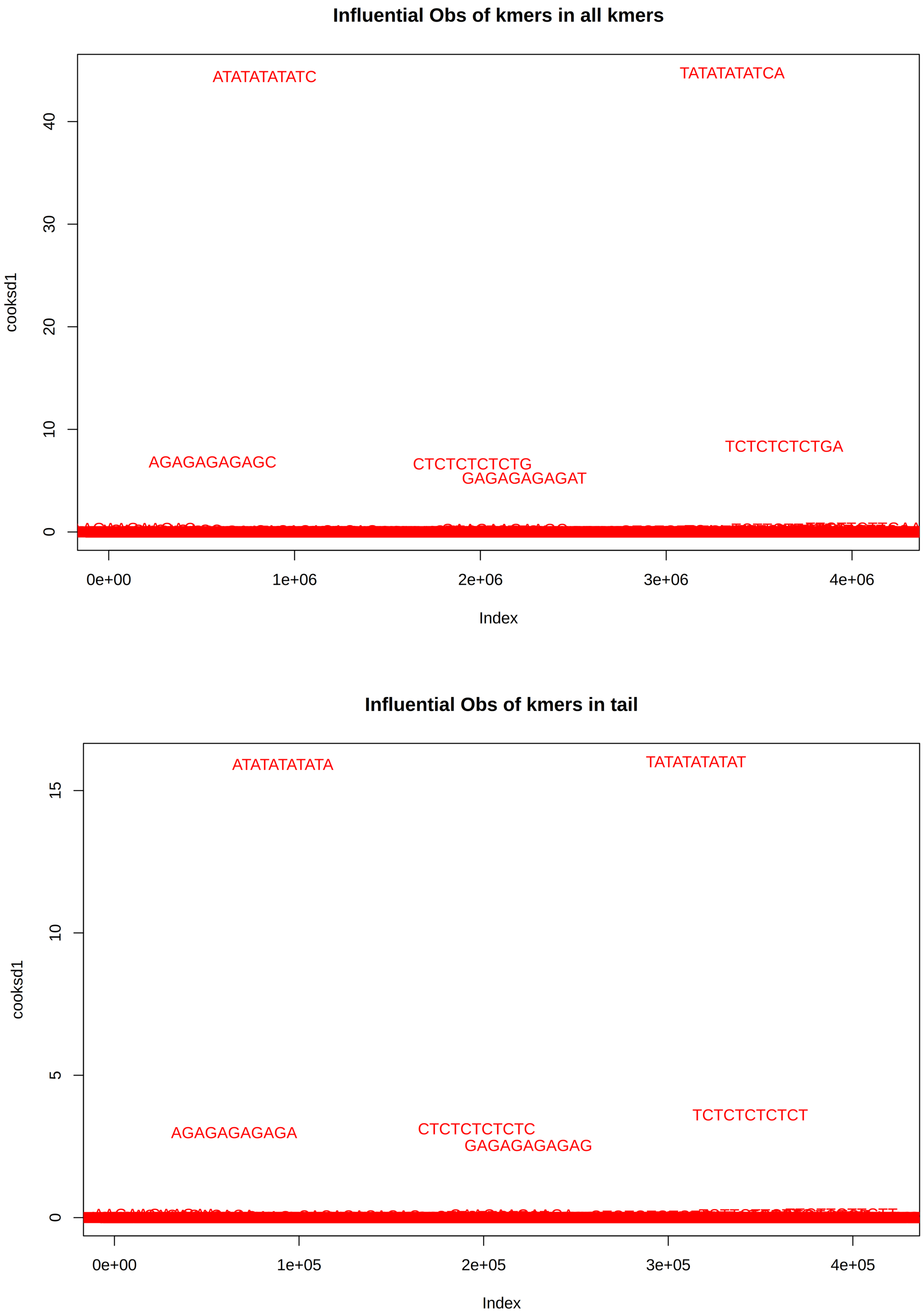


**Figure 3: PCA graphs of 11mer counts.** At the top PCA, all of the kmers were included, in the middle, the kmers with lowest 90% by abundance, found in the bell of the distribution curve, and at the bottom, the 10% most abundant kmers found in the tail of the distribution curve. It looks like one only needs 10% of the data to reproduce a very similar graph.



Influential Obs of kmers in all kmers

Influential Obs of kmers in tail

**Figure 4: The influential observations of 11mers used to construct the all inclusive and tail PCAs measured by Cook's Distance for the wheat genome.** The y-axis describes how much of an outlier a particular kmer is, and the x-axis is their position on a list, in alphabetical order. As you can see, the graphs are nearly identical. The outlying kmers are all repetitive dimers. But not all the possible dimers are represented (GT, GC, AC).

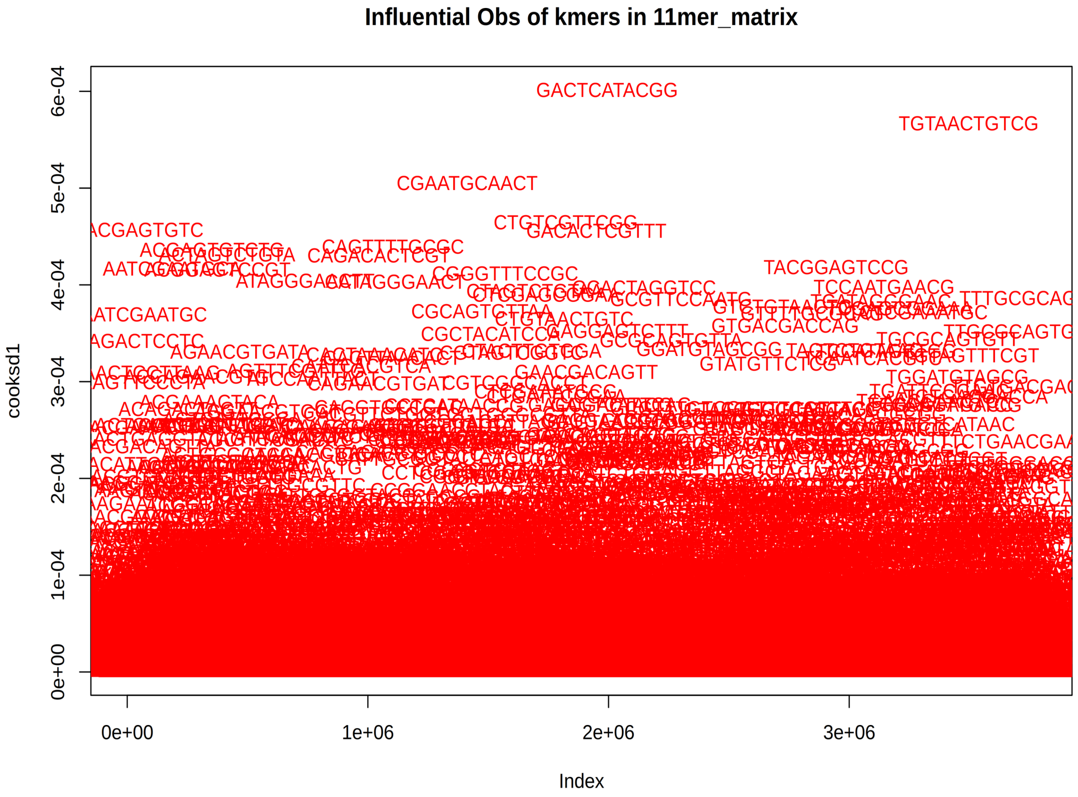


**Figure 5: The influential observations of 11mers used to construct the Bell PCA measured by Cook's Distance for the wheat genome.** The y-axis describes how much of an outlier a particular kmer is, and the x-axis is their position on a list, in alphabetical order. This graph is completely different from those in Fig. 4. These resulted in only two out of three possible clusters, as seen in Fig. 3. And these kmers are all close to zero in their cooksD value.

longer. If the x-axis is the count for a particular kmer, and the y-axis is the number of kmers with that count, then there are many kmers that occur very infrequently, and a few that are abundant. In Figure 3, a 10% cut-off was used to separate the bell from the tail, and the 11mer counts were used to construct PCA graphs. By selecting only the most abundant kmers, one can reproduce almost the same graph, using all the kmers available. This greatly reduces the system requirements and time required to perform the analysis.

To understand why this may be, the loadings of each kmer in the PCA was extracted and compared. The most influential kmers were identified by looking for outliers in a linear regression using the Cook's Distance (cooksD) test (14). This is not a typical use for that test, but the underlying principal seems to work for this application.

In Figures 4 and 5, one can perhaps see why the 'all kmers' and 'tail kmers' PCA graphs are so similar, while the 'bell kmers' PCA is so different. Those kmers in Fig. 4 which have cooksD values above 1 aren't present in the Bell kmer set, and this appears to have a very large effect. However, despite that lack, the Bell PCA still correctly divided subgenome D from A and B.

The nature of the influential kmers indicates that they are both abundant and are of low complexity. This provided another avenue of enquiry.

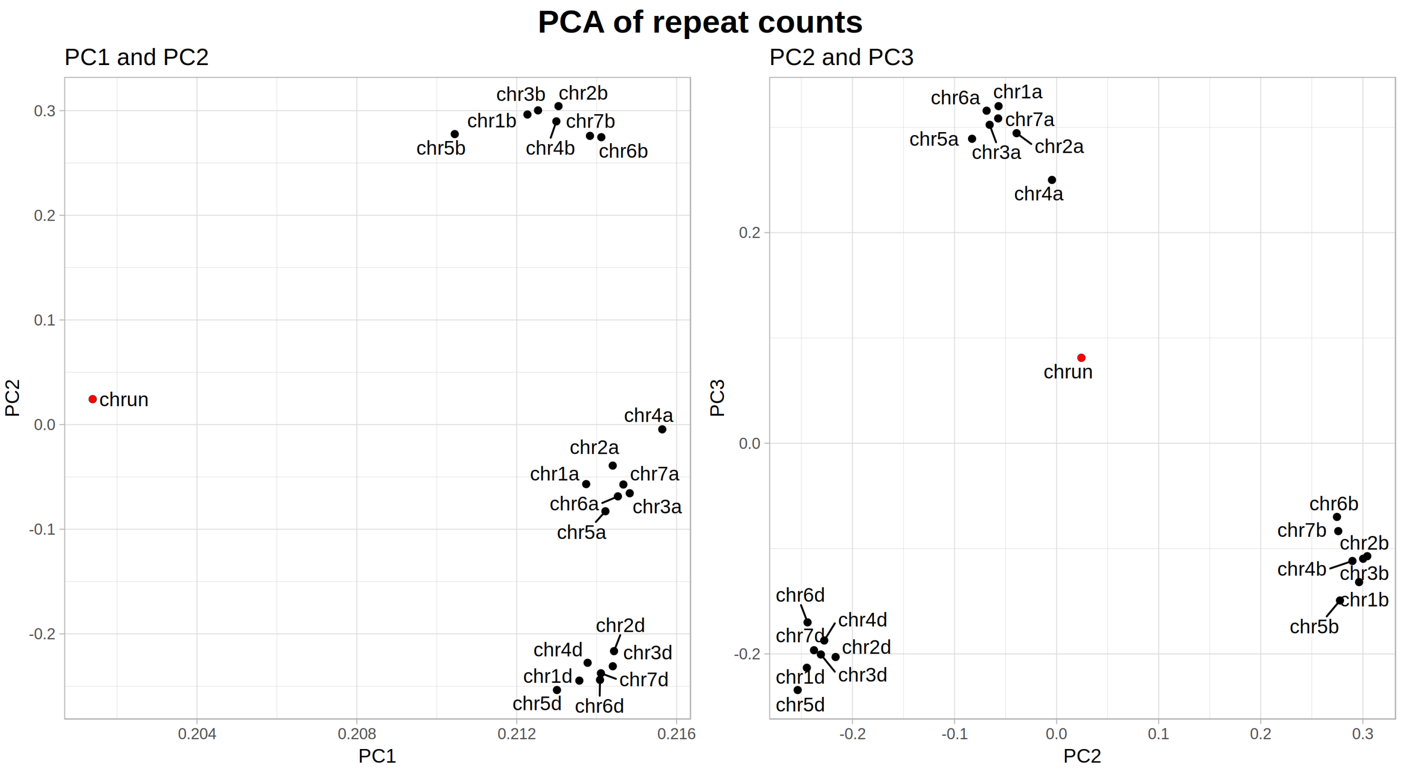


**Figure 6: The first 3 components of a PCA of repeat element counts in the wheat genome.** This shows how well the repeat profile differentiates subgenomes.

## Transposable Elements: The Problem and the Solution?

Araceli *et al* performed fluorescent in situ hybridisation (FISH) using the $(AC)_{10}$ microsatellite on various hexaploid, tetraploid, and diploid oats species. They identified unique physical maps using the $(AC)_{10}$ microsatellite, which was used to identify translocations and preferential distribution patterns unique to each chromosome or subgenome *(15).* We had kmers that resembled this microsatellite, so if the authors were able to use a single 20mer (AC x 10 = 20), a much

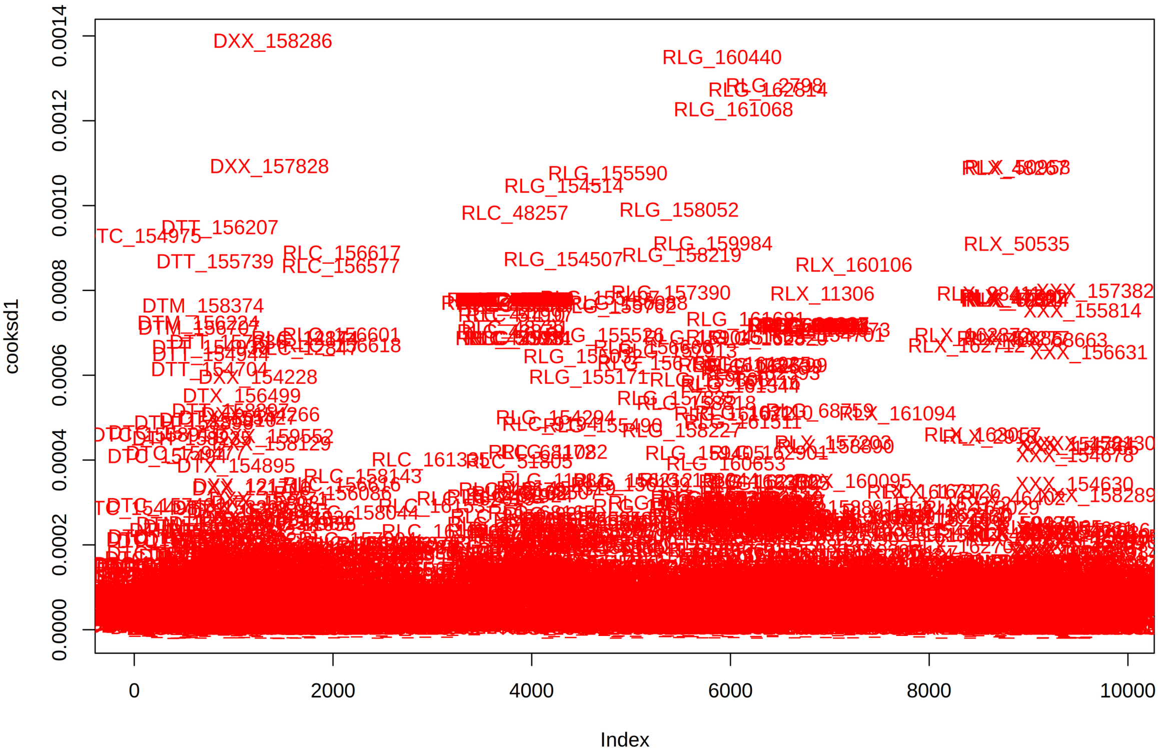


**Figure 7: The influential observations of repeat elements used to construct the wheat genome repeat PCA measured by Cook's Distance.** While most of the repeats weren't as influential, the maximum Cook's Distance is very small, when compared to the that of the kmers. The IDs are defined by the Plant Genome and Systems Biology institute (12).

larger set of transposable elements may reveal new information that the kmers are only just touching on. This way we may use the cause of our difficulties, large repetitive regions of DNA, as a tool for solving the problem.

The very first attempt was successful in binning the scaffolds. Using counts of only 9871 repeats elements, a very clear picture was formed, with 3 distinct clusters (Fig. 6). Components 4 and 5 aren’t shown, as they didn’t have any particular pattern or clustering of note. When analysing which repeats in particular might be influential (Fig. 7), it was found that that there wasn’t much difference. The outliers were not that far from the mean. Though perhaps DXX_158286 and RLG_160440, as well as other repeats elements with the highest Cook’s distance, may be of interest for further work, as they my be important for the evolution of the subgenomes.

## Application in Oats

In parallel to the work on wheat, the analysis on the 693 oat scaffolds (N90 scaffolds) was performed. It started with 7mers (Fig. 8). There is no useful clustering present. The 12mer attempt went better, with PC1 and PC2 revealing two distinct clusters separated by a smear of scaffolds. Certain vague shapes seen in the 7mer PCAs (Fig. 9) resolve themselves into more defined forms using 12mers. The 7mer graph was formed by using all 16 384 kmers, but the 16 777 216 12mers was too large a dataset, creating a 23GB

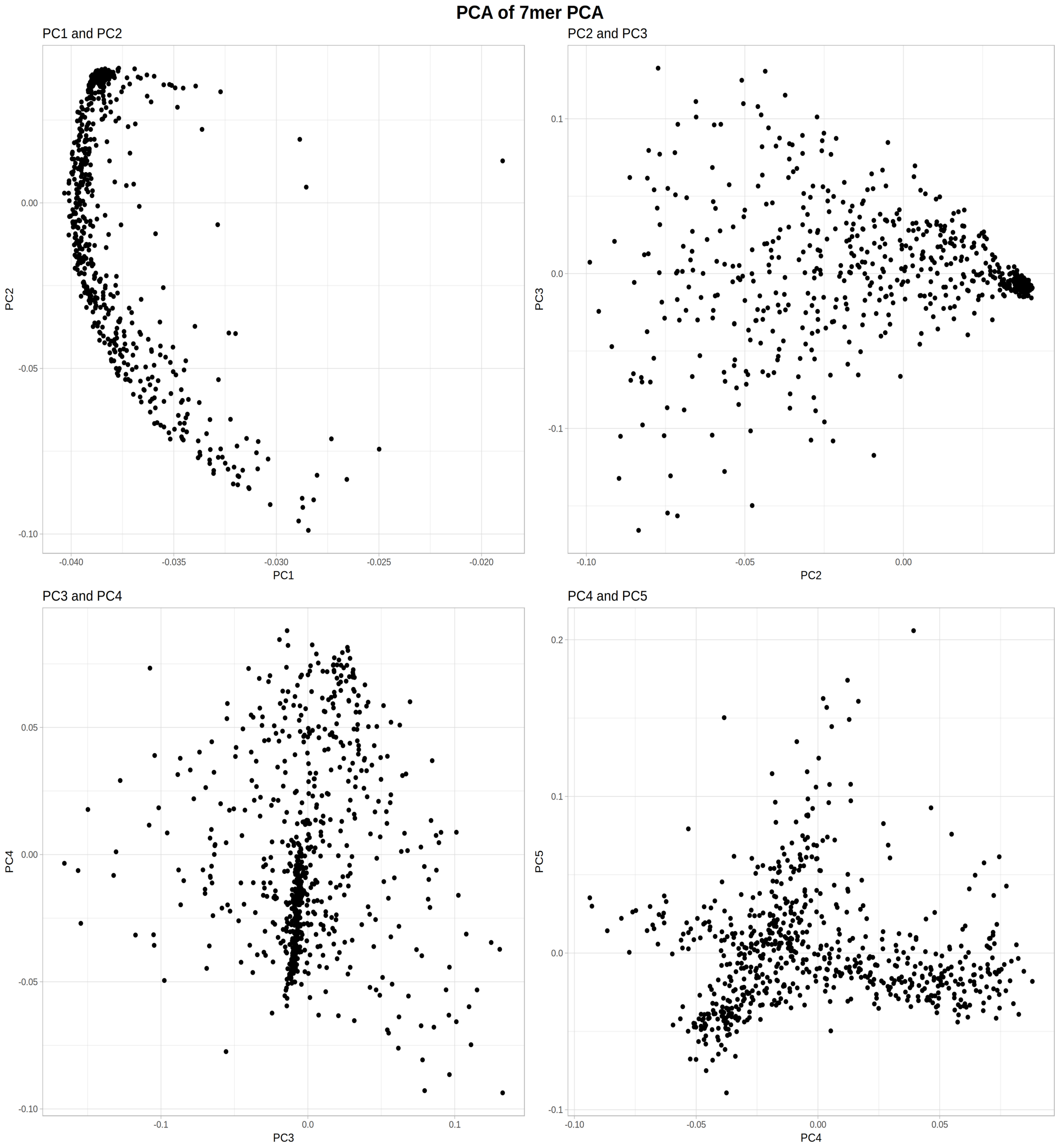


**Figure 8: PCA of 7mer counts in the 693 largest scaffolds from oats.** There is no distinct clustering in any of the first five components.

table. After the filtering method was developed, all kmer based PCAs were limited to the 13 million most abundant kmers. This made it possible to use much longer kmers, with the hope of increasing the resolution of clusters into subgenomes, which was seen in wheat.

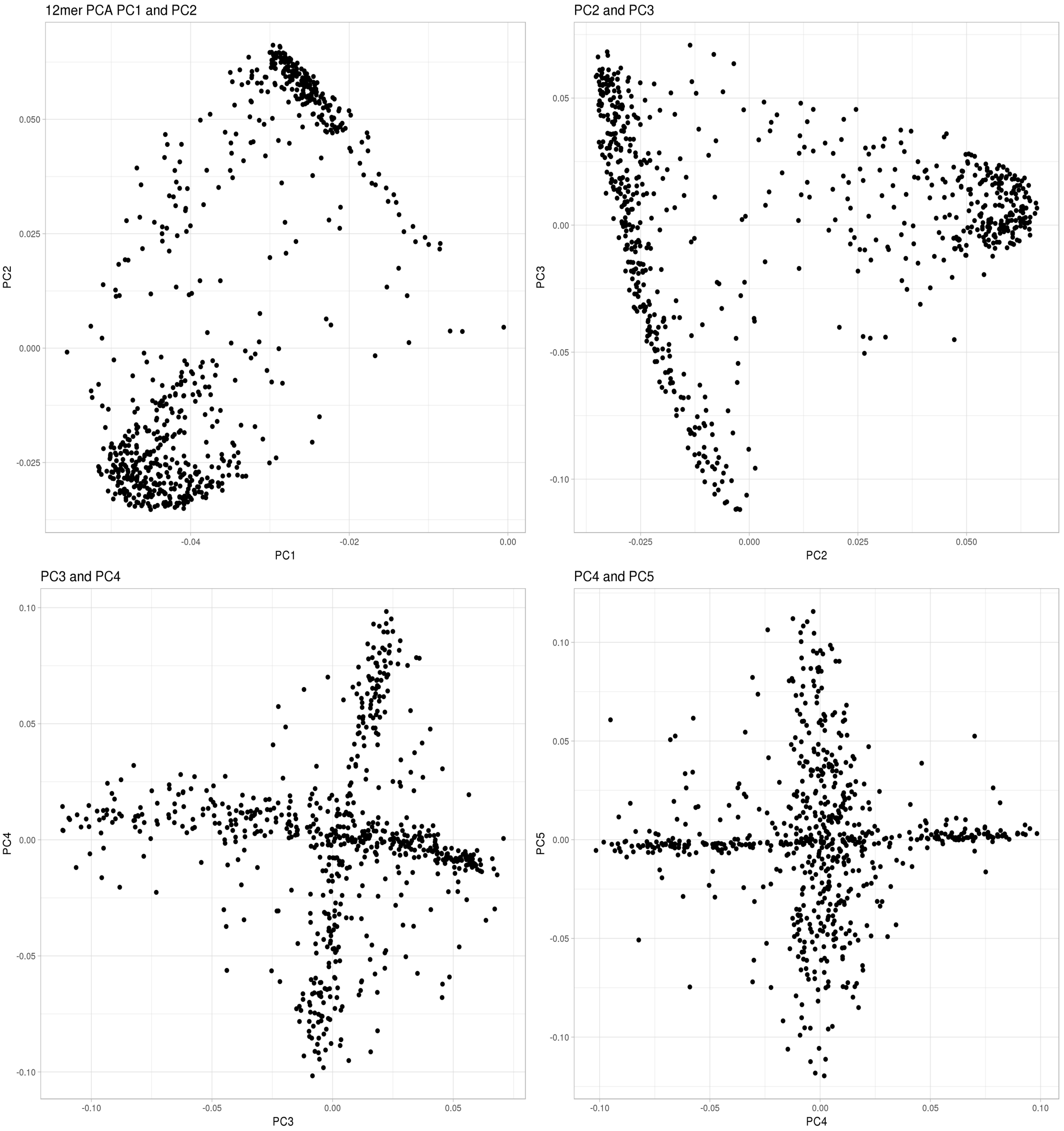


**Figure 9: PCA of 12mer counts in the 693 longest scaffolds from oats.** There appears to be some separation, especially using PC1 and PC2. The cross shape that emerges in the graphs using PC3, 4 and 5 can also be seen in figure 6, in PC4 and PC5.

The 35mer PCA (Fig. 10) did indeed increase the density of the clusters, but the signal to noise ratio is still quite high. Furthermore, the three clusters one would expect are not present. But it was possible that perhaps one of the clusters represented two subgenomes. To identify if this was the case, all the 3-copy BUSCOs (one for each subgenome) were identified. The the clusters were

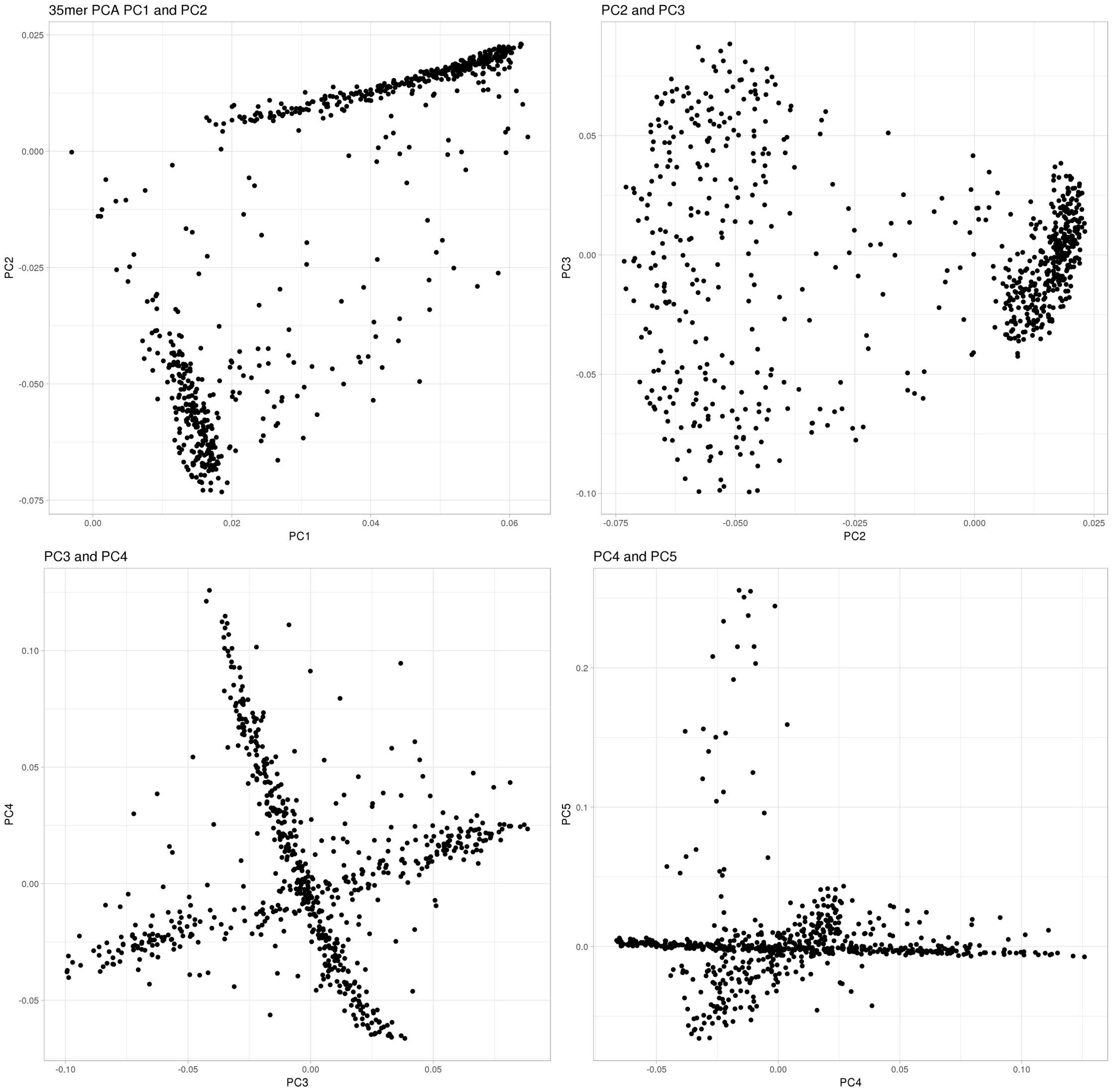


**Figure 10: PCA of 35mer counts in the 693 largest scaffolds from oats.** Once again PC1 and PC2 show 2 dense clusters, though in perpendicular dimensions. The other 3 graphs look similar to the last 3 in the 12mer graph (Fig. 9) and the cross motif is present, though morphed.

divided along the x-axis, which also happened to divide the graph in half. The top cluster is designated cluster 1, and the bottom cluster 2 (Fig 11). As shown in Fig. 12, cluster 2 has two copies of most of the 3-copy BUSCOs. This would imply that cluster 2 has scaffolds originating from two subgenomes, as we are using a complete set of BUSCOs as a representative for a subgenome.

The scaffolds from cluster 2 were used for a new analysis. Their count numbers were subjected to a separate PCA, and new clusters were designated cluster 2A and cluster 2B for the top and bottom

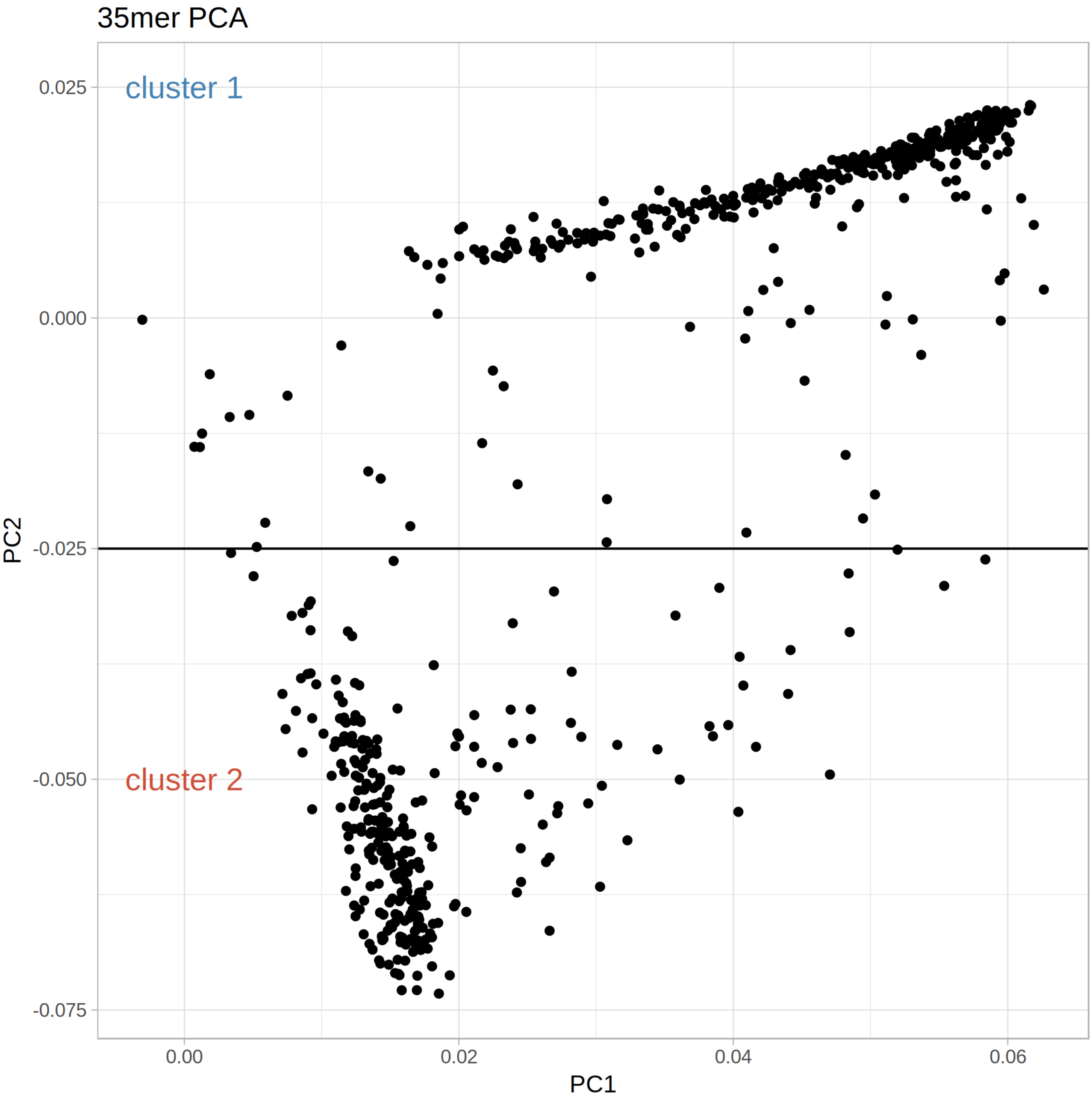


**Figure 11: PCA showing how cluster 1 and cluster 2 are divided in the 35mer counts in the oats genome.** The fact that the line divides the graph in perfect halves is a coincidence.

clusters respectively. In this case, since we know cluster 2 has two copies of the BUSCOs, only the 2-copy BUSCOs were identified. There is nearly a perfect division of BUSCOs between cluster 2A and 2B (Fig. 14) . And thus, with a fair bit of uncertainty and error, we have divided the scaffolds into 3 clusters. However, there are shortcomings. Many BUSCOs aren’t evenly divided between clusters. And an automatic method that did not require manual

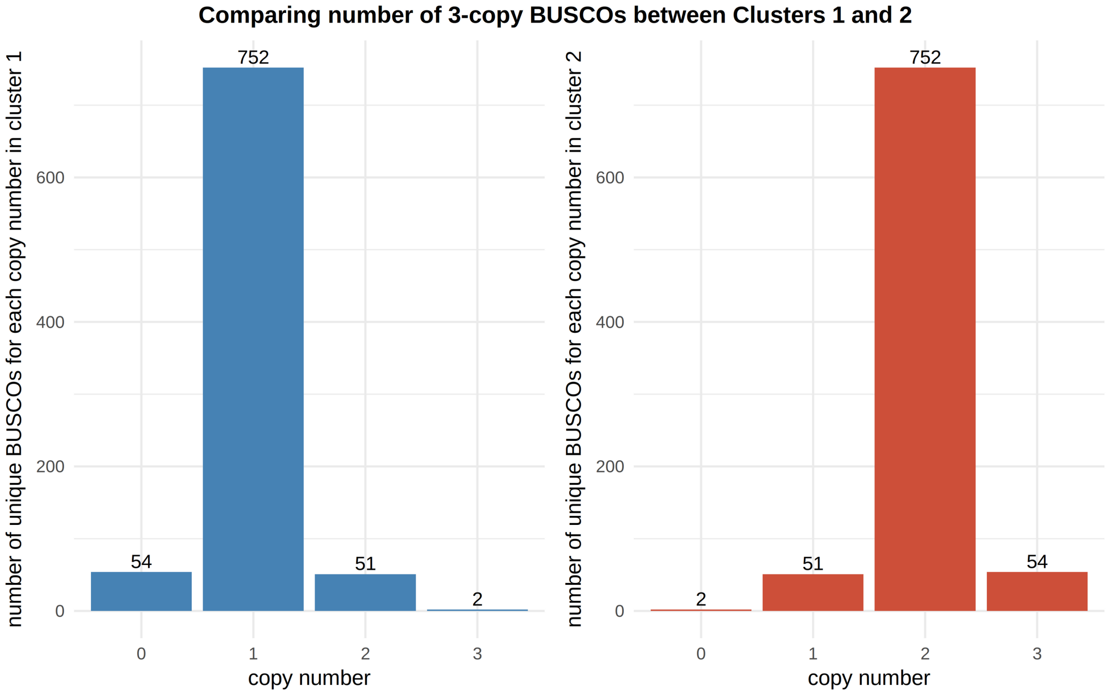


**Figure 12: Comparing number of 3-copy BUSCOs in clusters 1 and 2.** Of all the 3-copy BUSCOs (859), cluster 2 has two copies of nearly all of them (752).

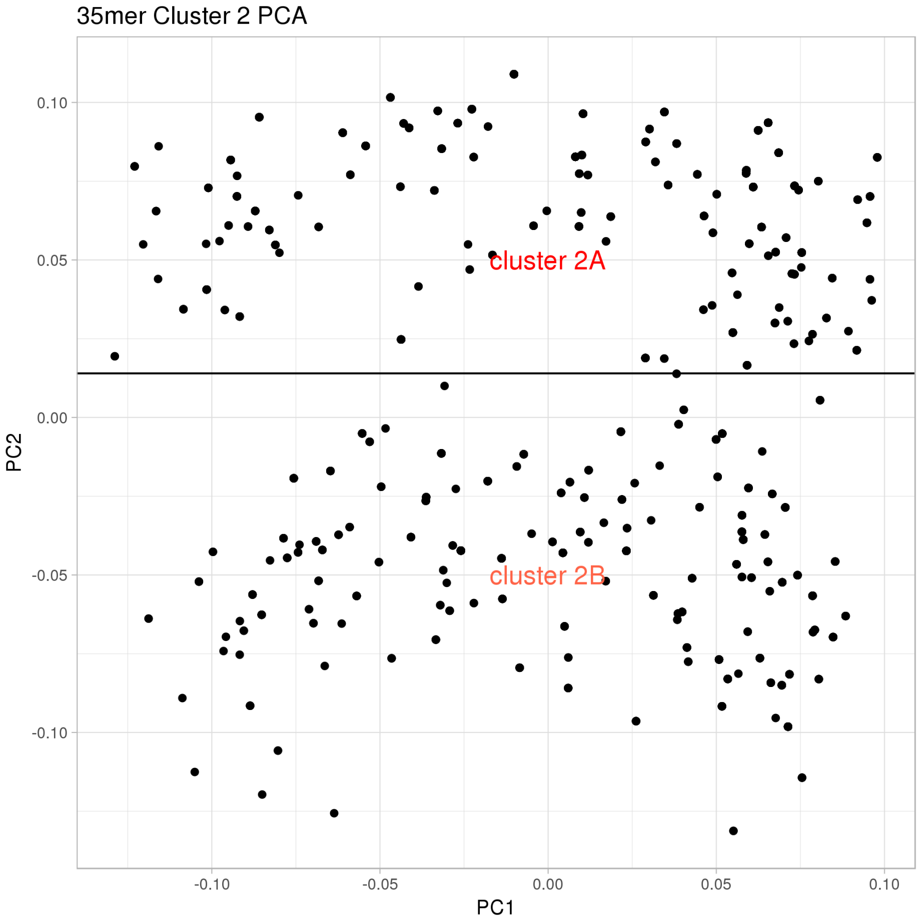


**Figure 13: PCA of 35mer counts only including scaffolds part of cluster 2 in Fig. 11.** The line dividing the clusters was drawn by eye, and creates clusters 2A and 2B, top and bottom, respectively.

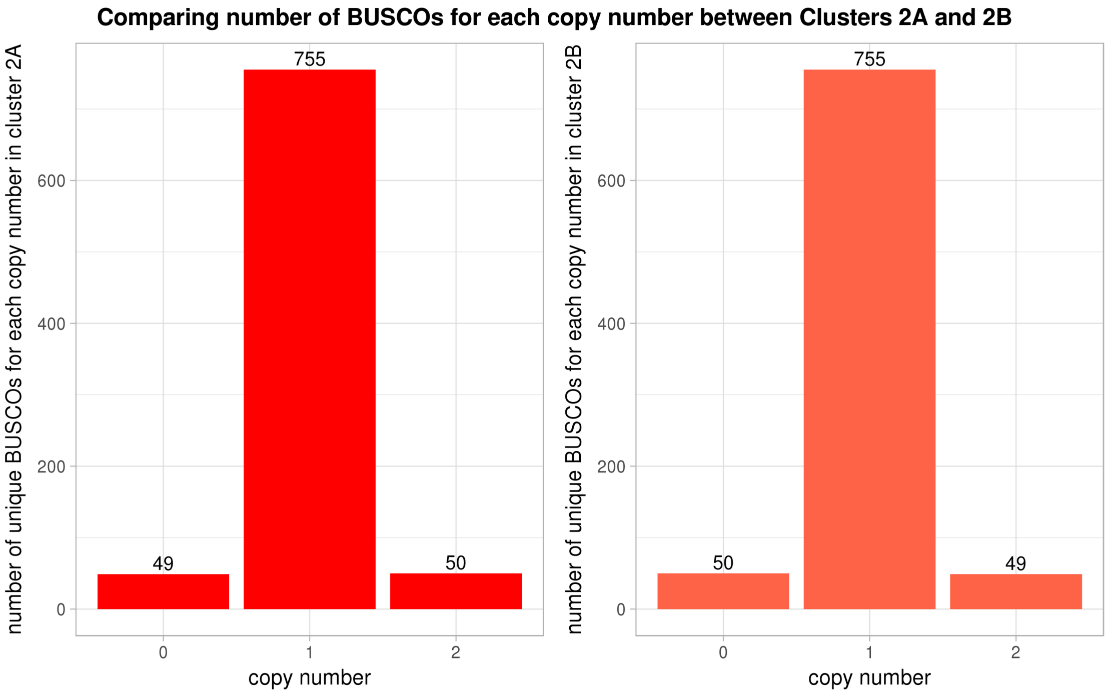


**Figure 14: Comparing number of 2-copy BUSCOs in clusters 2A and 2B.** Each cluster has 1 copy of nearly all of the BUSCOs (854 total)

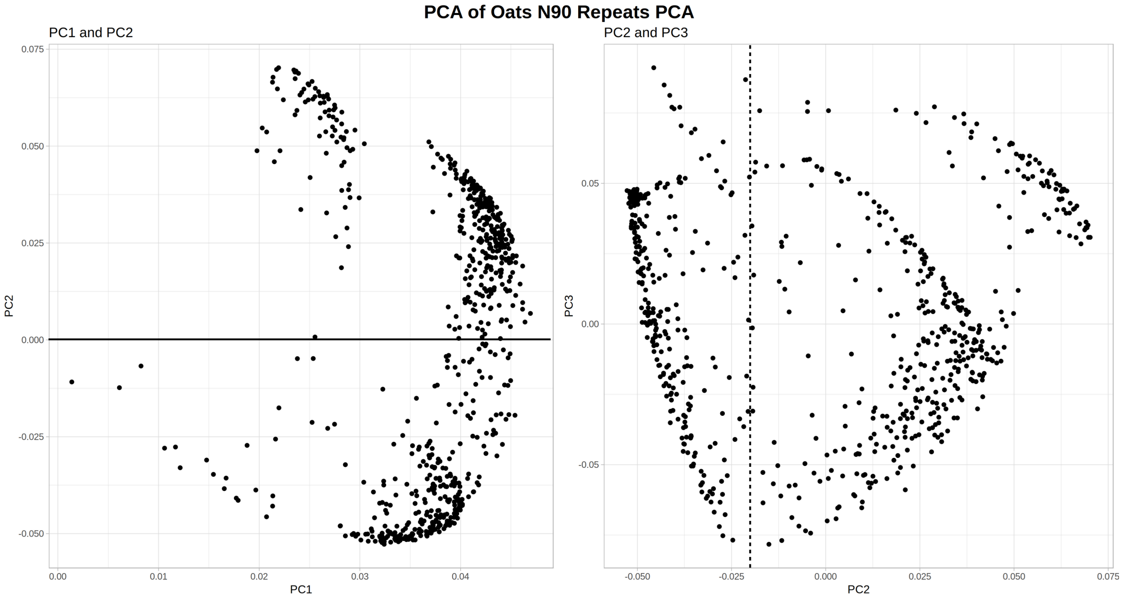


**Figure 15: PCA of repeat element counts in the N90 scaffolds of oats.** There appears to be some separation if one were to draw a solid line along PC2 = 0 for the first 2 components, though the clusters are very loose. For PC2 and PC3, the dotted line may be drawn at PC2 = -0.024. But once again, not ideal clusters.

inspection should create the clusters, perhaps k-means clustering (16). This would reduce user bias, and can be applied at scale.

## A Refrain on Repeats

Just as in wheat, the repeats were used to analyse Oats (Fig. 15). However, the results were not promising. The 4820 repeat counts is half of the 9871 repeats found in wheat. Perhaps this is responsible for performance even worse than the 35mer PCA. Next, a hybrid approach was used. All of the 20mers from from the Poaceae repeat database were extracted, and then only these kmers were counted in the N90 fasta files. The PCA (Fig. 16) is superior to all previous attempts, in terms of density of clusters, and reduced scatter of non-clustered scaffolds. But it still only results in 2 clusters. Only the first 3 components were used here, as components 4 and 5 did not improve cluster separation or reveal any interesting patterns, and looked nothing like those of the 35mer PCA or the full repeat PCA.

But, after the same BUSCO analysis, it was found that the top cluster (cluster 1) contains 2 copies of the 3-copy BUSCOs, and the bottom cluster (cluster 2) only has 1 copy of that same set. This would imply that cluster 1 contains two of the subgenomes, and cluster 2 represents the third.

Cluster 1 was subjected to the same analysis as described above, but this is where the analysis ends, for there was no clustering at all. Cluster 1 did not divide into 2 subgenomes.

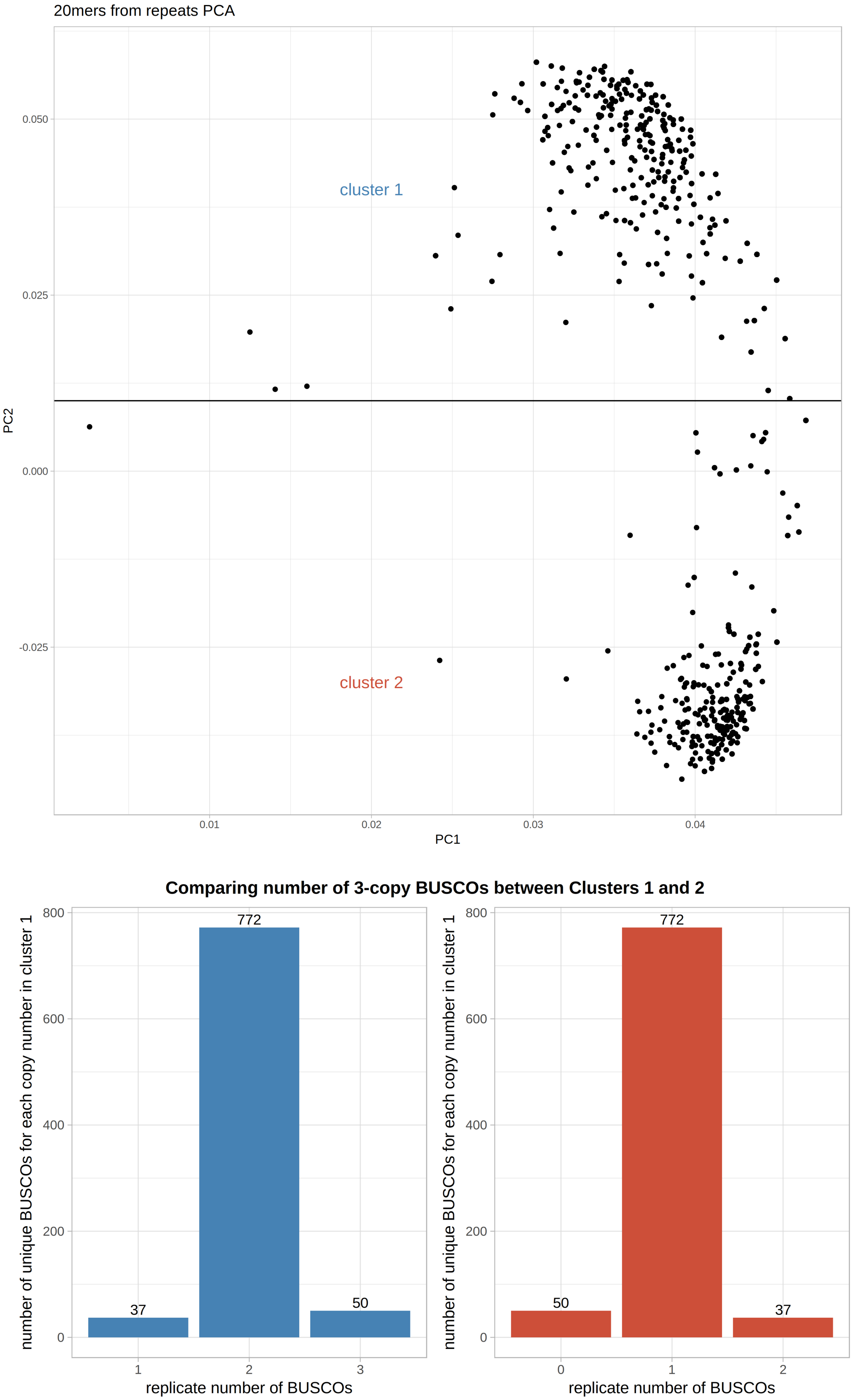


**Figure 16: PCA and cluster analysis of 20mers counts derived from a repeat database using BUSCO copy number to represent completeness of subgenomes.** Of all the 3-copy BUSCOs, cluster 1 has two copies of nearly all of them (772). This implies that cluster 1 represents 2 subgenomes.

# Discussion

When comparing Figures 3 and 6, one can see a slight contradiction. In figure 3, subgenomes A and B are closer together, and even cluster together when looking at the bell PCA. This would indicate greater homology and a closer evolutionary relationship, since they have more similar kmer profiles. However, the repeat element profile in Fig. 6 indicates that subgenomes A and D are are more homologous. Literature (17,18) supports the theory that A and B are more closely related (Fig. 17) . The difference may be due to retrotransposon activity that occurred after the hexaploid was formed, but was only suppressed in the B subgenome, which would bring A and D closer together, from a repeat-element profile perspective.

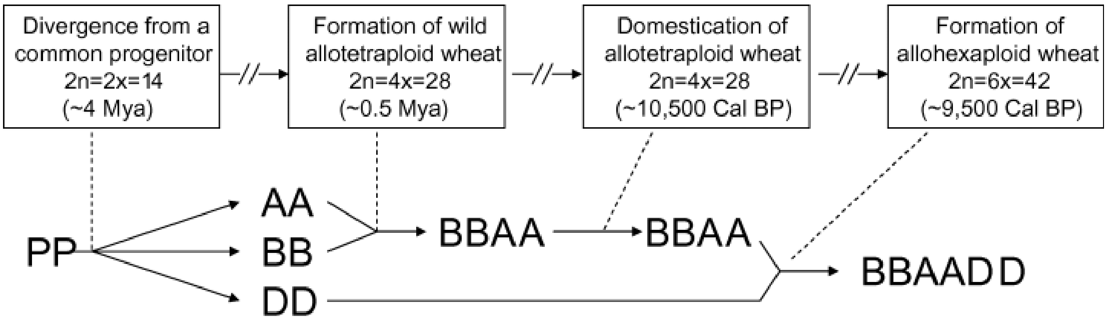

Divergence from a common progenitor 2n=2x=14 (~4 Mya)
Formation of wild allotetraploid wheat 2n=4x=28 (~0.5 Mya)
Domestication of allotetraploid wheat 2n=4x=28 (~10,500 Cal BP)
Formation of allohexaploid wheat 2n=6x=42 (~9,500 Cal BP)
PP
AA
BB
DD
BBAA
BBAA
BBAADD

**Figure 17: A theory on the evolution of the genome architecture of modern bread wheat.** (Adapted from Levy and Feldman, 2004)

In tests where the wheat chromosomes were broken into 60mbp fragments, the clustering got extremely loose. The 14mbs fragments didn’t cluster at all, but formed a smear on the graph. So it would seem that the longer the scaffolds are, the more easily they can be accurately binned. This is problematic when draft assemblies often result in thousands of scaffolds in the kilo-base pair range. But after re-examining the 41mbp fragmented wheat scaffolds, we found that there was still some useful clustering (Fig. 18, 19). But it appears that any less resolution of the clusters would completely obscure them.

However, it may be interesting if a similar kmer based method could detect inter-chromosomal translocations. If fragments associate (cluster) with fragments from a different chromosome, it may indicate just such a translocation. Cross-subgenome

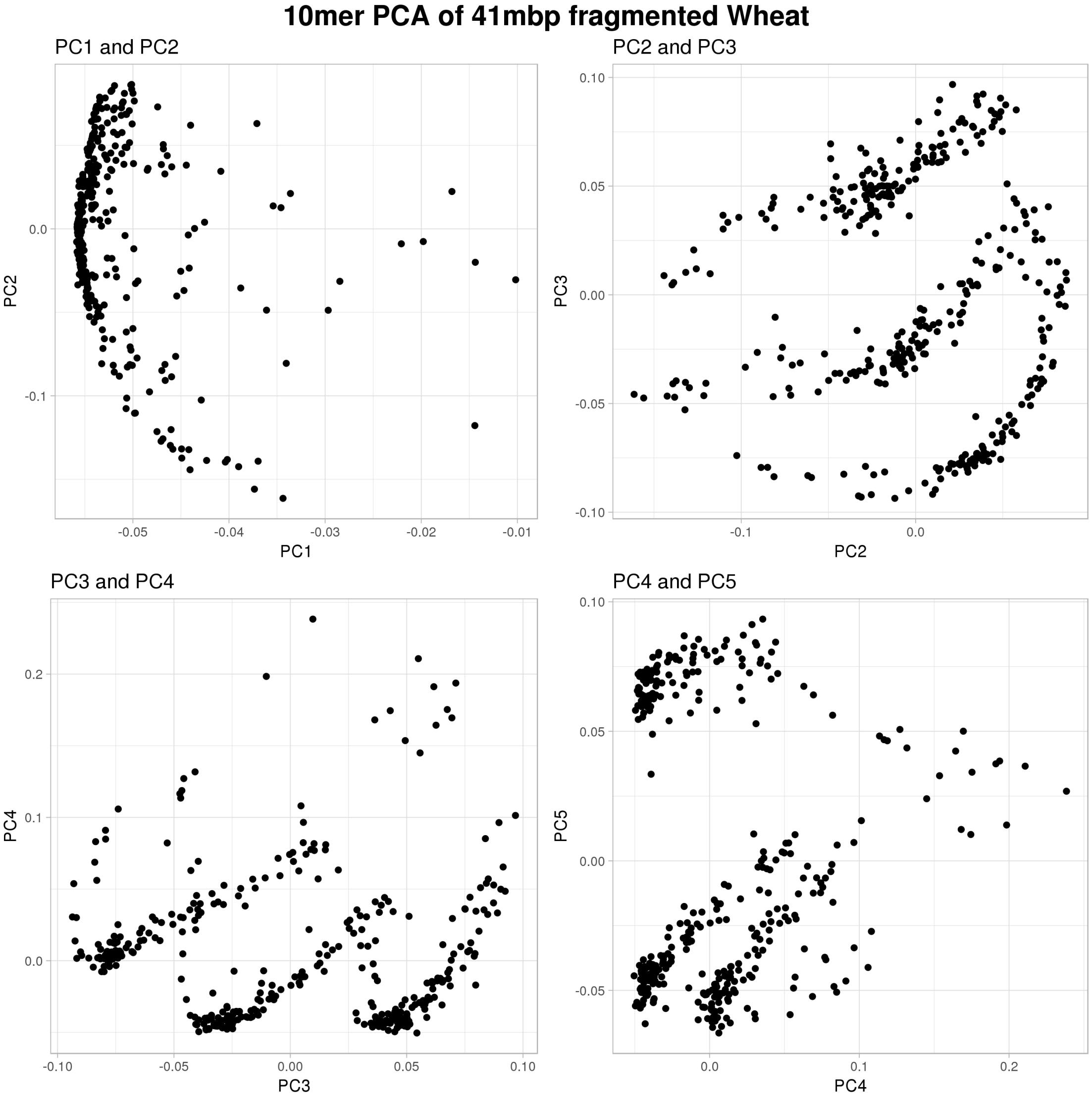


**Figure 18: PCA of 10mer counts from 41mbp fragmented wheat.** The subgenome clustering reveal themselves in PC2 and onwards.

translocations may be responsible for the lack of resolution in the oat clusters. If the assembler created chimeric scaffolds at a large scale, the chances for success using this method would be low.

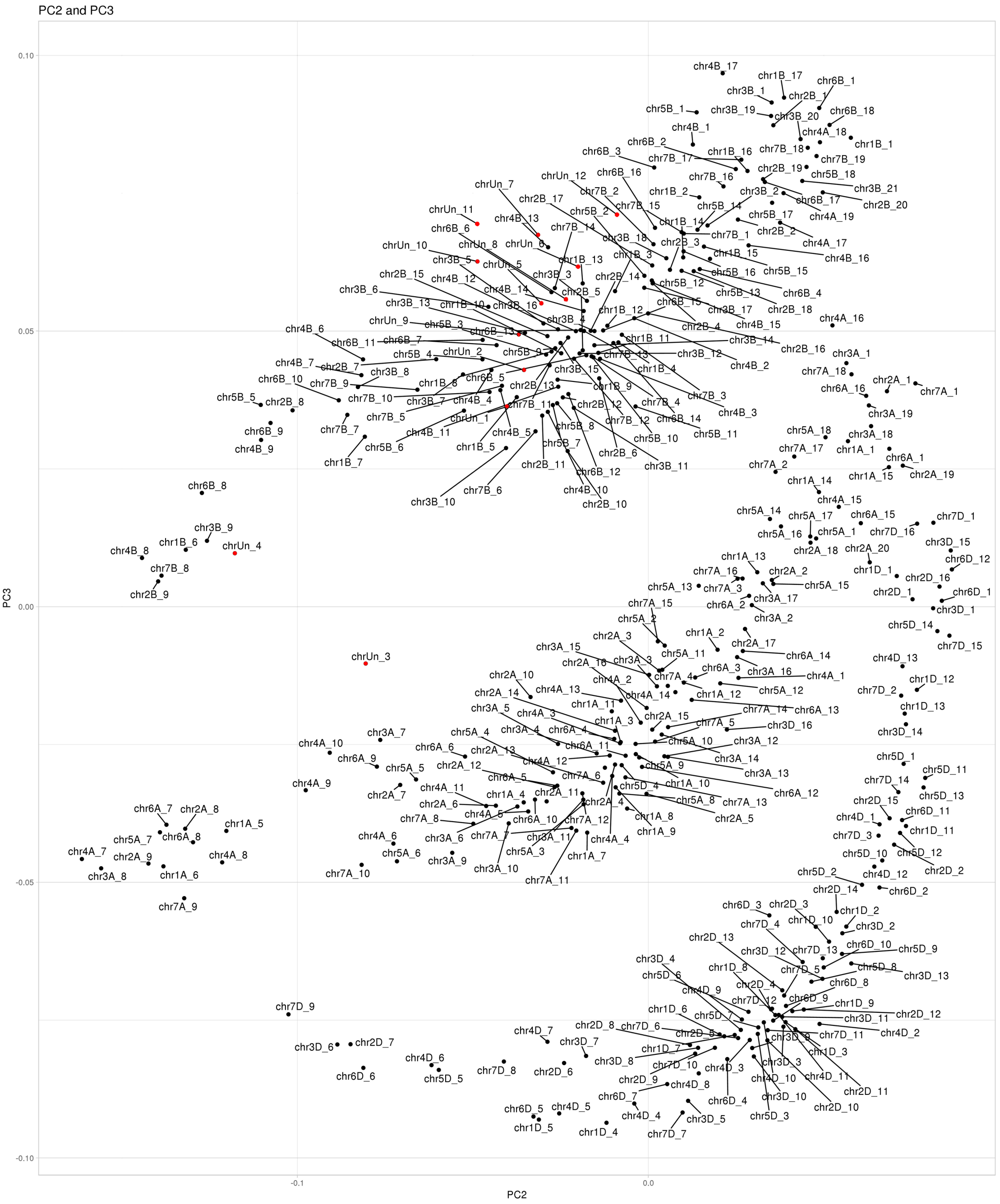


**Figure 19: Components 2 and 3 of PCA 10mer counts from 41mbp fragmented wheat, with labels.** The labels reveal that the clusters formed true subgenomes, though the chrun chromosome fragments got mixed in predominantly with the chrB (subgenome B) cluster.

## Conclusion

Using substrings like kmers and repeat elements to bin scaffolds into subgenomes was validated with wheat. SubKluster works. But it is highly dependant the quality of the draft genome, particularly the length of the scaffolds. We hope to present confirmation that the method will work on other plant draft genomes soon.

## Further Work

As of now, the pipeline requires about 45 GB of RAM to perform the PCA on 1.8 GB of data (about 13 million rows with 693 columns). But if the flat CSV could be placed in a database, then a slower but more memory efficient PCA could be performed. This could also be spread over a computing cluster and calculated in parallel.

With further development of SubKluster, it is hoped that multiple sources of substrings could be used in one PCA. Clustering may improve when mixing the most influential kmers and repeats in the same PCA. It should also be investigated if the length of the scaffolds influenced where they clustered, as the length may not have been accounted for completely as part of the scaling function in the PCA function.

## Acknowledgements

I would like to thank my supervisors Nick and Dag. This was a partnership that accepted crazy ideas, coming and going. They

supported me far beyond I expected, and vitally, provided chocolate during thesis crunch-time. I also want to thank Björn, for letting me keep 2 monitors, and providing the use of his lightning-quick kmer counter. And of course this wouldn't be possible without access to the unpublished Belinda genome, provided by the SCANOATS consortium, who saw fit to trust me with their precious intellectual property. Despite the stimulating work, it would have been a dull office with out my coffee mates and intellectual sparing partners, who provided excellent distractions.

## Appendix

Table comparing wheat and oats assembly stats

| | **Wheat** | **Oats** |
|---|---|---|
| Subgenomes | AABBDD | AACCDD |
| Genome size | ~15 Gbp | ~12 Gbp |
| N50 | 709.8 Mbp | 17.7 Mbp |
| N90 | 509.9 Mbp | 2.8 Mbp |
| Complete BUSCOs | | 1409 |

## Bibliography

1. Shcherban AB. Repetitive DNA sequences in plant genomes. Russ J Genet Appl Res [Internet]. 2015;5(3):159–67. Available from: http://link.springer.com/10.1134/S2079059715030168

2. Greilhuber J, Doležel J, Lysák MA, Bennett MD. The origin, evolution and proposed stabilization of the terms "genome size" and "C-value" to describe nuclear DNA contents. Ann Bot. 2005;95(1):255–60.

3. Mehrotra S, Goyal V. Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. Genomics,

Proteomics Bioinforma [Internet]. Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China; 2014;12(4):164–71. Available from: http://dx.doi.org/10.1016/j.gpb.2014.07.003

4. Garbus I, Romero JR, Valarik M, Vanžurová H, Karafiátová M, Cáccamo M, et al. Characterization of repetitive DNA landscape in wheat homeologous group 4 chromosomes. BMC Genomics. 2015;16(1).

5. Van Nieuwerburgh F, Thompson RC, Ledesma J, Deforce D, Gaasterland T, Ordoukhanian P, et al. Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. Nucleic Acids Res. 2012;40(3).

6. Hulse-Kemp AM, Maheshwari S, Stoffel K, Hill TA, Jaffe D, Williams SR, et al. Reference quality assembly of the 3.5-Gb genome of Capsicum annuum from a single linked-read library. Hortic Res [Internet]. Springer US; 2018;5(1). Available from: http://dx.doi.org/10.1038/s41438-017-0011-0

7. Zimin A V., Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL. The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. Gigascience. 2017;6(11):1–7.

8. Liu Q, Lin L, Zhou X, Peterson PM, Wen J. Unraveling the evolutionary dynamics of ancient and recent polypoidization events in Avena (Poaceae). Sci Rep [Internet]. Nature Publishing Group; 2017 Feb 3;7(December 2016):41944. Available from: http://dx.doi.org/10.1038/srep41944

9. Claros MG, Bautista R, Guerrero-Fernández D, Benzerki H, Seoane P, Fernández-Pozo N. Why Assembling Plant Genome Sequences Is So Challenging. Biology (Basel) [Internet]. 2012;1(3):439–59. Available from: http://www.mdpi.com/2079-7737/1/2/439/

10. Girotto S, Pizzi C, Comin M. MetaProb: Accurate metagenomic reads binning based on probabilistic sequence signatures. Bioinformatics. 2016;32(17):i567–75.

11. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.

12. Spannagl M, Nussbaumer T, Bader KC, Martis MM, Seidel M, Kugler KG, et al. PGSB plantsDB: Updates to the database framework for comparative plant genome research. Nucleic Acids Res. 2016;44(D1):D1141–7.

13. Kent WJ. BLAT — The BLAST -Like Alignment Tool. Genome Res. 2002;12:656–64.

14. Aguinis H, Gottfredson RK, Joo H. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. Organ Res Methods [Internet]. 2013 Apr 14;16(2):270–301. Available from: http://journals.sagepub.com/doi/10.1177/1094428112470848

15. Fominaya A, Loarce Y, Montes A, Ferrer E. Chromosomal distribution patterns of the (AC) 10 microsatellite and other repetitive sequences, and their use in chromosome rearrangement analysis of species of the genus Avena [Internet]. Vol. 60, Genome. 2017. Available from: http://www.nrcresearchpress.com/doi/10.1139/gen-2016-0146

16. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics [Internet]. Berkeley, Calif.: University of California Press; 1967. p. 281–97. (Fifth Berkeley Symposium on Mathematical Statistics and Probability). Available from: https://projecteuclid.org/euclid.bsmsp/1200512992

17. Feldman M, Levy A, Chalhoub B, Kashkush K. Genomic Plasticity in Polyploid Wheat. In: Soltis PS, Soltis DE, editors. Polyploidy and Genome Evolution [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 109–35. Available from: https://doi.org/10.1007/978-3-642-31442-1_7

18. Levy AA, Feldman M. Genetic and epigenetic reprogramming of the wheat genome upon allopolyploidization. Biol J Linn Soc. 2004;82(4):607–13.