Evaluating Genomic Variation

Katy Merritt and Dr. Alan Lemmon

Abstract

In a recently patented method, polymerase chain reaction is possible using only one primer. The method uses two common adapter sequences and a primer designed to match a repetitive element from a genome. A particular primer can be designed to target a particular number of loci. The focus of this study was to apply a computer program to several example genomes to isolate the best candidate primers. The candidate primers had to meet several requirements including an appropriate melting temperature, length, mapability of flanking regions, and expected number of loci. A maximum of 20 candidate primers were selected for each genome. Some of the primers have been tested, but future work is needed to demonstrate that the theoretical results match the experimental results across species. If the experimental values verify that the primer selection method works as expected, then in the future the method could be applied for less expensive gene sequencing, forensics, as well as paternity tests.

Introduction

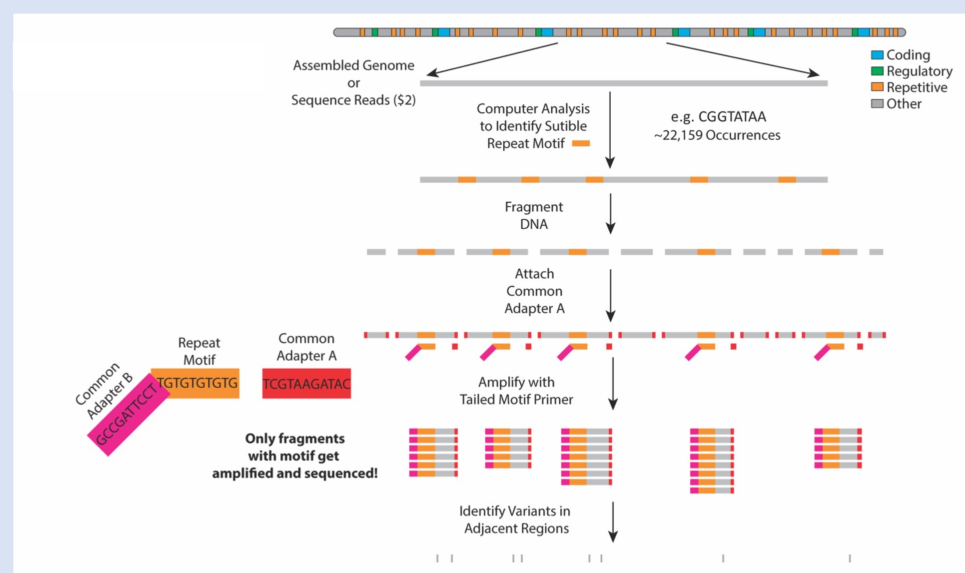

Fig 1. A method recently developed subsamples genomes with a single primer and two common adapters. The primer is a sequence of nucleotide bases that is repeated throughout the genome. The current study is based around finding these repetitive sequences that return areas of interest in the genome.

Species	Common Name	Genome Size (Gb)	Expected number of loci	Melting Temp (*C)	Avg. distance to first single copy kmer
Apis mellifera	Honey Bee	0.23	11,800	28	3.54
Bos taurus	Bovine	2.67	228,400	64.4	39.82
Bos taurus	Bovine	2.67	140,800	49.5	71.34
Danio rerio	Zebrafish	1.68	91,200	59.2	18.33
Gallus gallus	Chicken	1.07	40,800	50.0	0.59
Gallus gallus	Chicken	1.07	21,200	52.9	1.52
Homo sapiens	Human	3.20	746,000	63.0	3.75
Homo sapiens	Human	3.20	66,200	63.1	4.22
Hordeum vulgare	Barley	4.23	258,200	48.5	-
Hordeum vulgare	Barley	4.23	224,600	54.2	-
Mus musculus	Mouse	2.73	108,400	52.9	14.12
Mus musculus	Mouse	2.73	77,200	51.1	-
Oryza Sativa	Rice	0.38	5,800	91.3	-
Psudacris feriarum	Chorus Frog	4.20	447,000	60.4	69.03
Psudacris feriarum	Chorus Frog	4.20	354,000	58.4	-
Python molurus	Python	1.44	40,200	57.9	44.57
Python molurus	Python	1.44	17,000	53.7	8.50
Strongylocentrotus purpuratus	Sea Urchin	0.92	72,200	70.9	95.53
Zea mays	Corn	2.18	162,200	70.3	96.54
Zea mays	Corn	2.18	57,200	55.5	103.24

Table 1. A table summarizing properties of some of the most prevalent kmers in 12 different species. Properties such as melting temperature, average distance to mapable region, and the expected number of loci that will be returned.

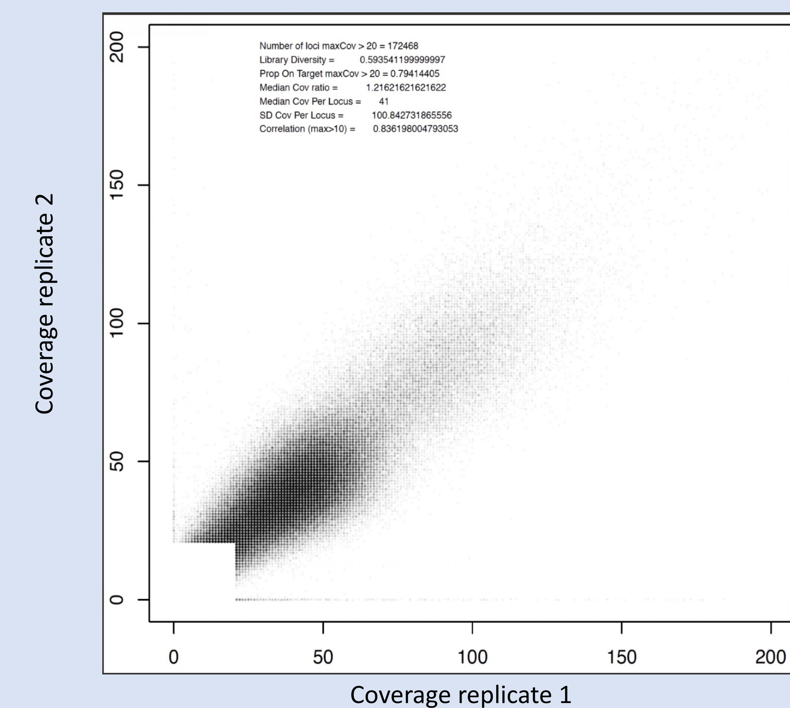

Fig 2. The new method is highly repeatable across technical replicates. Here, we present the coverage from two replicates showing the repeatability of the method. The two clouds of points above result from heterozygous (cov ~ 40) and homozygous (cov ~ 80) loci. The coverage after sequencing. The spread of this coverage would imply that this particular individual is heterozygous, so there are two distinct groups of coverage.

The program used to locate these primers takes a fasta and scans 1% of the file to quickly get an idea of the prevalent kmers. Once the program has identified possible kmers, it goes through and eliminates kmers that appeared too few times or had too many tandem repeats. Once the initial candidates have been determined, they are sorted into clusters to account for kmers that are identical except for one or two base pairs. Finally, the dominant kmer for each cluster is printed.

Acknowledgements

I would like to thank Dr. Alan Lemmon for all of his help and time the past two semesters. I would also like to thank the UROP program for providing this opportunity.

Results

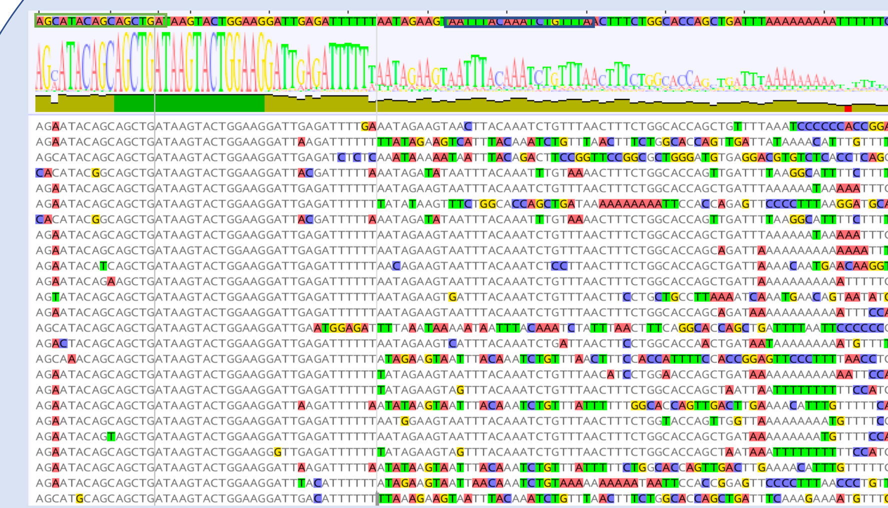

Fig 3a. Original sequence retrieves 166,400 loci.

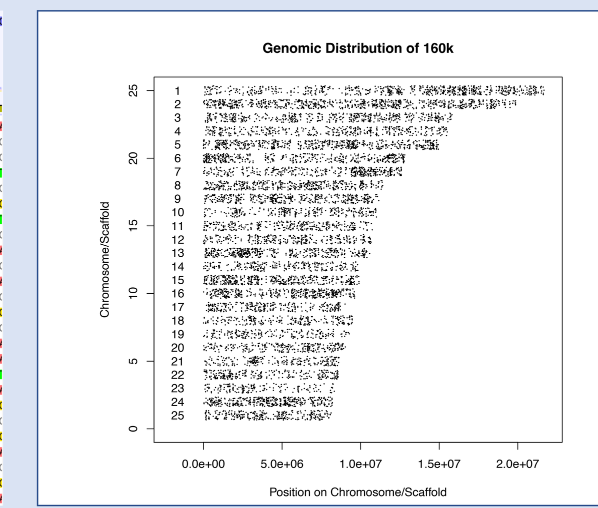

Fig 3b. Expected spread of retrieved loci throughout genome for original sequence

Fig 4a. Slightly shifted primer sequence retrieves 79,480 loci.

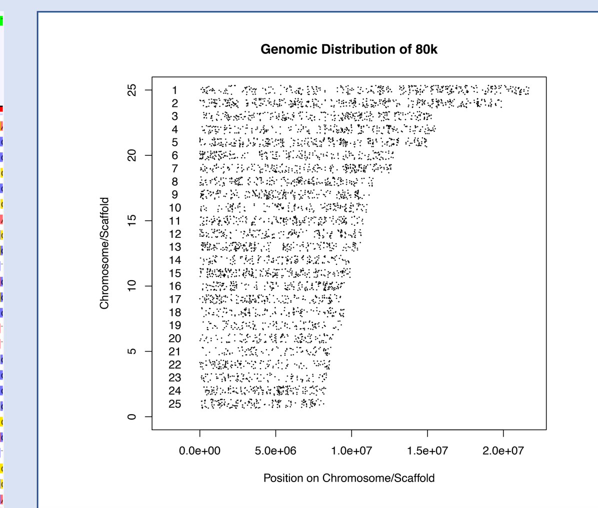

Fig 4b. Expected spread of retrieved loci throughout genome for sequence shown in fig 4a.

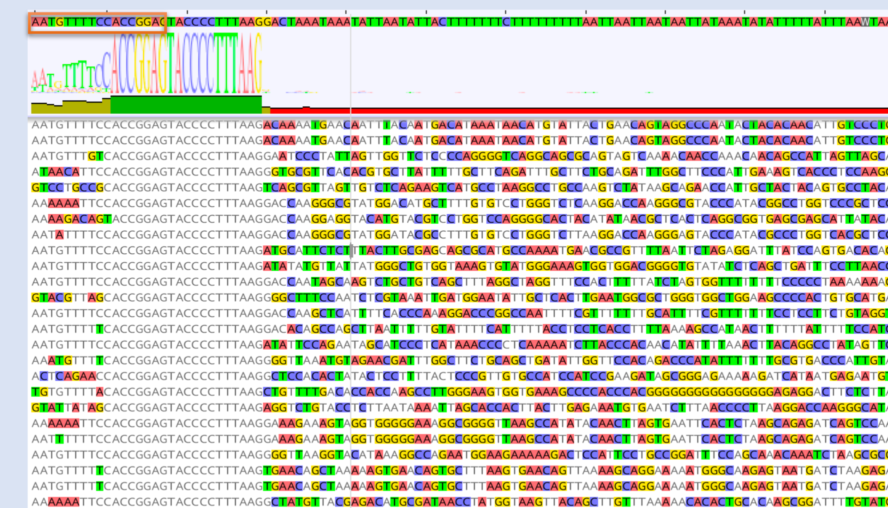

Fig 5a. Further shifted primer sequence retrieves 10,952 loci. Possible variations in nucleotide bases leads to the sequence not looking identical to the sequences above

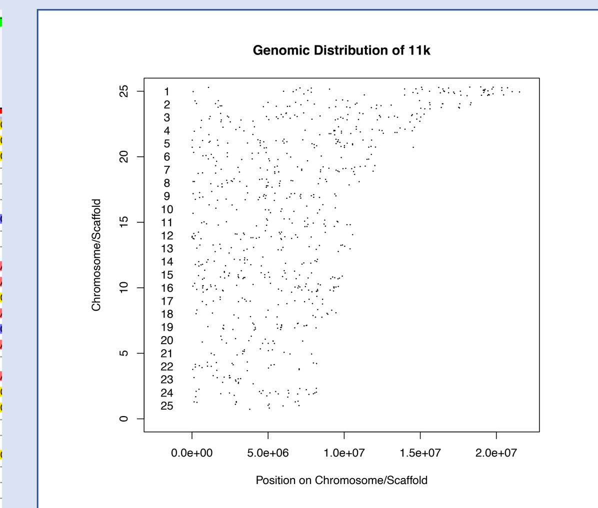

Fig 5b. Expected spread of retrieved loci throughout genome for sequence shown in fig 5.

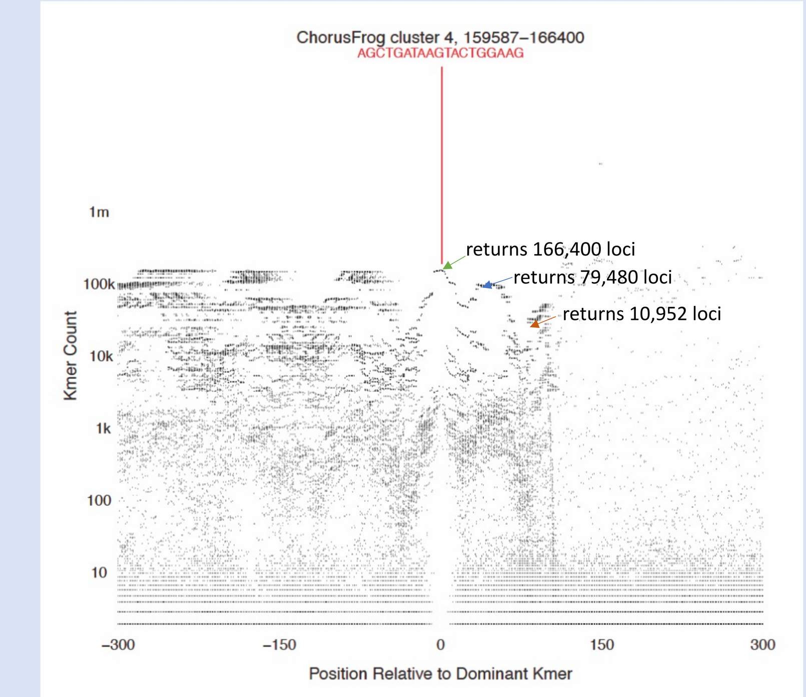

Fig 6. As reading frame of primer is shifted, fewer loci are retrieved until single copy regions are reached

Discussion

Figures 3-5 represent the versatility of this sequencing method when it comes to choosing a primer. Figure 3 represents the original primer sequence that in *P. feriarum* can yield information from over 160,000 loci. Figures 4 and 5 demonstrate that by shifting the starting point of the primer, the number of loci can be reduced to the desirable number. Overall, this method is incredibly versatile due to its ability to be used across species as well as its ability to be modified to retrieve a desired number of loci. Going forward, the next steps following this study will be to test the selected primers for each of the genomes to see if the number of loci returned matches the expected number.

References

Garibyan, L., & Avashia, N. (2013). Polymerase chain reaction. *The Journal of investigative dermatology*, 133(3), 1-4. https://doi.org/10.1038/jid.2013.1