New Genome Assembly Method

Katy Merritt and Hannah Squier, Dr. Alan Lemmon

Relevant Vocabulary

- Base pairs (bps): nucleotides; ATCG
- Genome: the full collection of genetic material in the nucleus of an organism
- Reads/sequences: string of nucleotides representing a segment of the genome
- DNA sequencing: method to determine the order of nucleotides in a section of DNA
- Genome assembly: the process of reconstructing a genome's sequence using reads
- Kmer: a short sequence of nucleotides of length K (e.g. 5mer, 6mer)
  - SCK: Single-Copy Kmer; a sequence that appears once in the genome
- Repetitive regions: portions of the genome that appear more than once
- Coverage: the number of times a portion of DNA is sequenced
- *Pseudacris feriarum*: upland chorus frog, found in the SE United States
  - Size of genome: 4.5 billion bases (gigabases, Gb)

Background

Current genome assembly methods[1]:

1. De Bruijn graph (**Fig 1**):
   - Stores all sequences and their connections in one large graph.
   - Heuristic methods required to extract genome sequence from graph.
2. Overlap layout consensus method (**Fig 2**):
   - Use similarities between DNA sequences to create longer consensus sequences.
   - Repetition in large genomes makes it easy for this method to incorrectly overlay two sequences that share a common region but are in different parts of the genome.

Problems:

- Both methods are effective at assembling small genomes, but they do not scale well as genome size and data set size increase.
- Poor scaling occurs because repeats are not initially avoided.
- The *Pseudacris feriarum* genome is large (4.5 Gb) and has numerous repeats, making it difficult to assemble with current methods.

Solution:

- Avoid the issues caused by repeats by
  1) identifying SCKs (Single-Copy Kmers),
  2) identifying their relative positions in the genome,
  3) estimating the intervening sequences at the end of the process.

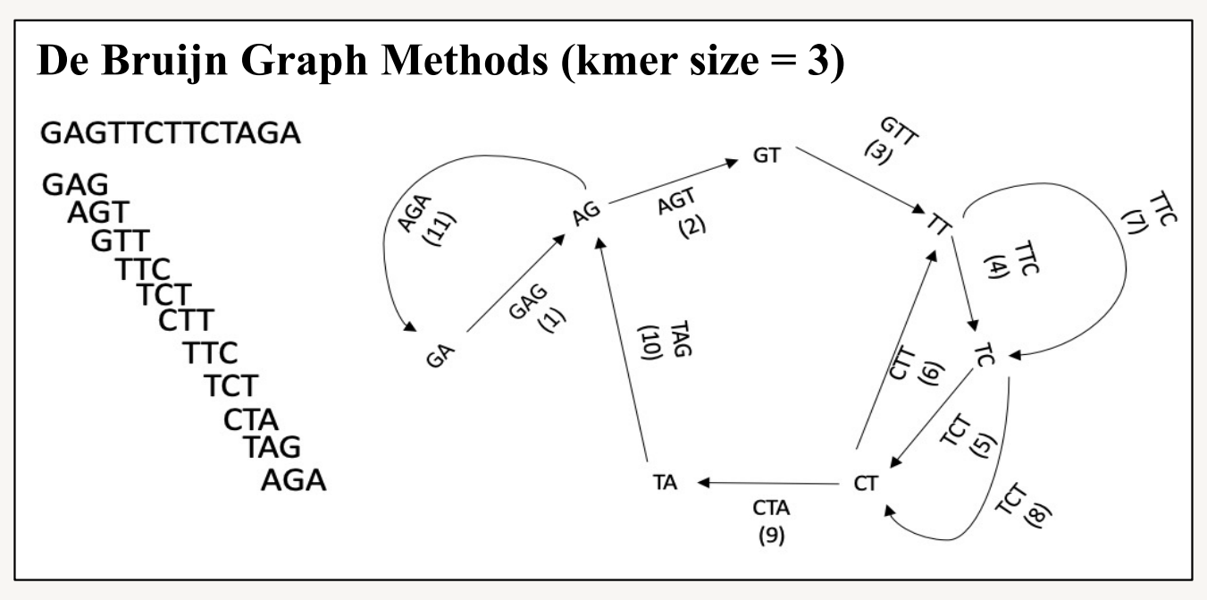


Figure 1 shows an example of a de Bruijn graph. The method breaks sequences into smaller kmers and tracks their connections. The graph contains the entire genome.

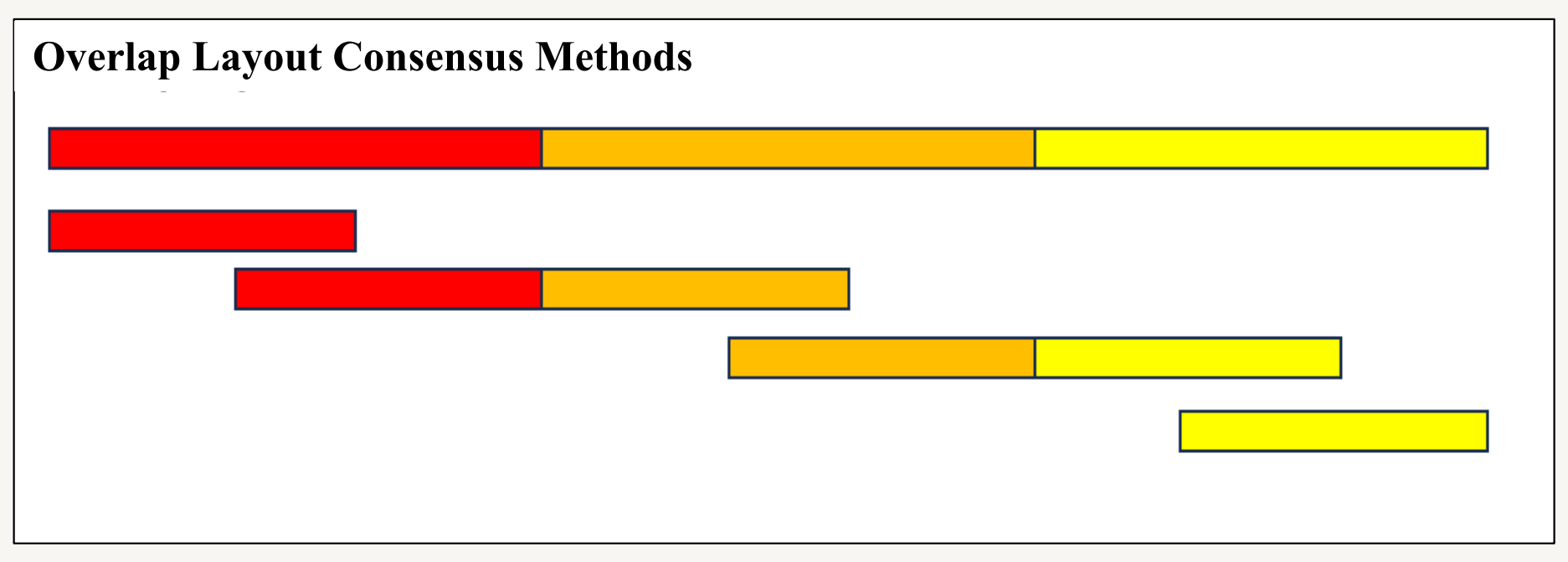
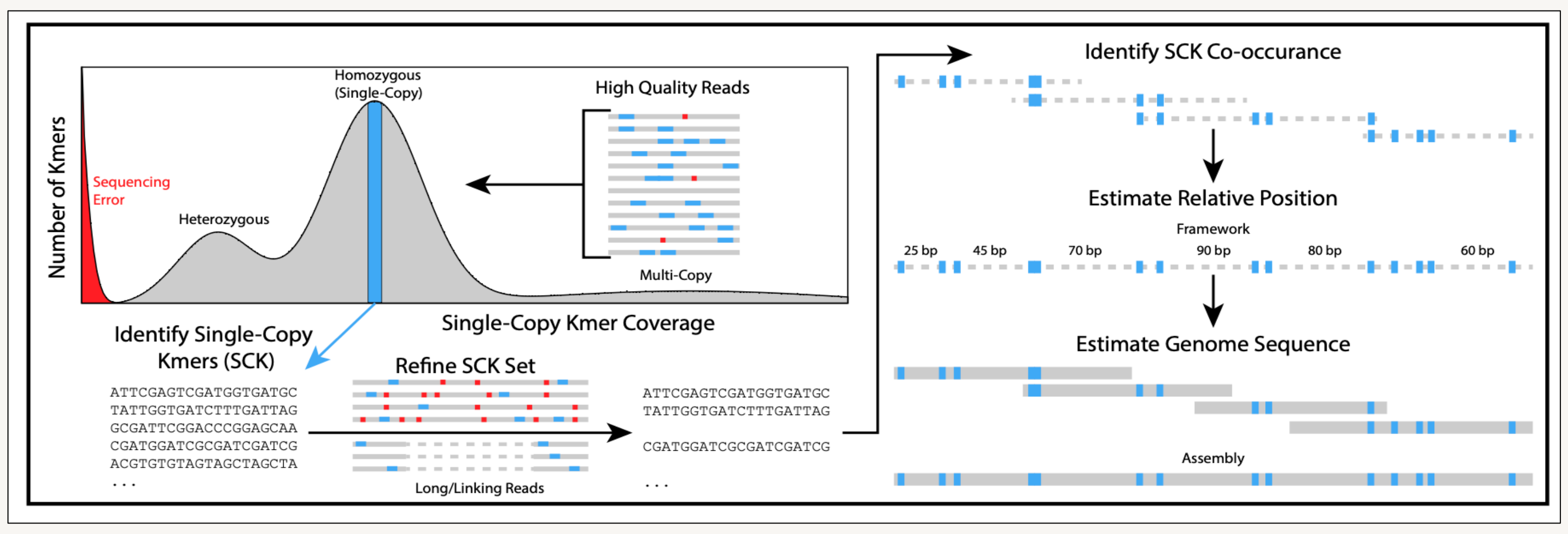


Figure 2 represents the Overlap Layout Consensus method. The method aligns reads based on common regions.



ATTCGAGTCGATGGTGATGC
TATTGGTGATCTTTGATTAG
GCGATTCGGACCCGGAGCAA
CGATGGATCGCGATCGATCG
ACGTGTGTAGTAGCTAGCTA
. . .

ATTCGAGTCGATGGTGATGC
TATTGGTGATCTTTGATTAG

CGATGGATCGCGATCGATCG
. . .

Figure 3 represents the pipeline for our proposed method. Our method relies on identifying SCKs, clustering SCKs, tracking their relative positions to one another, and then finally overlaying reads.

Methods

1. Identify optimal length for SCKs

- Two types of sequencing data
  - Short, high-quality reads & long, low-quality reads
- Issues
  - Kmers need to be long enough to be unique in genome
  - Kmers need to be short enough to minimize sequencing error effects (~13%)
- After intense study, we found 18 base pairs is the optimal kmer length.

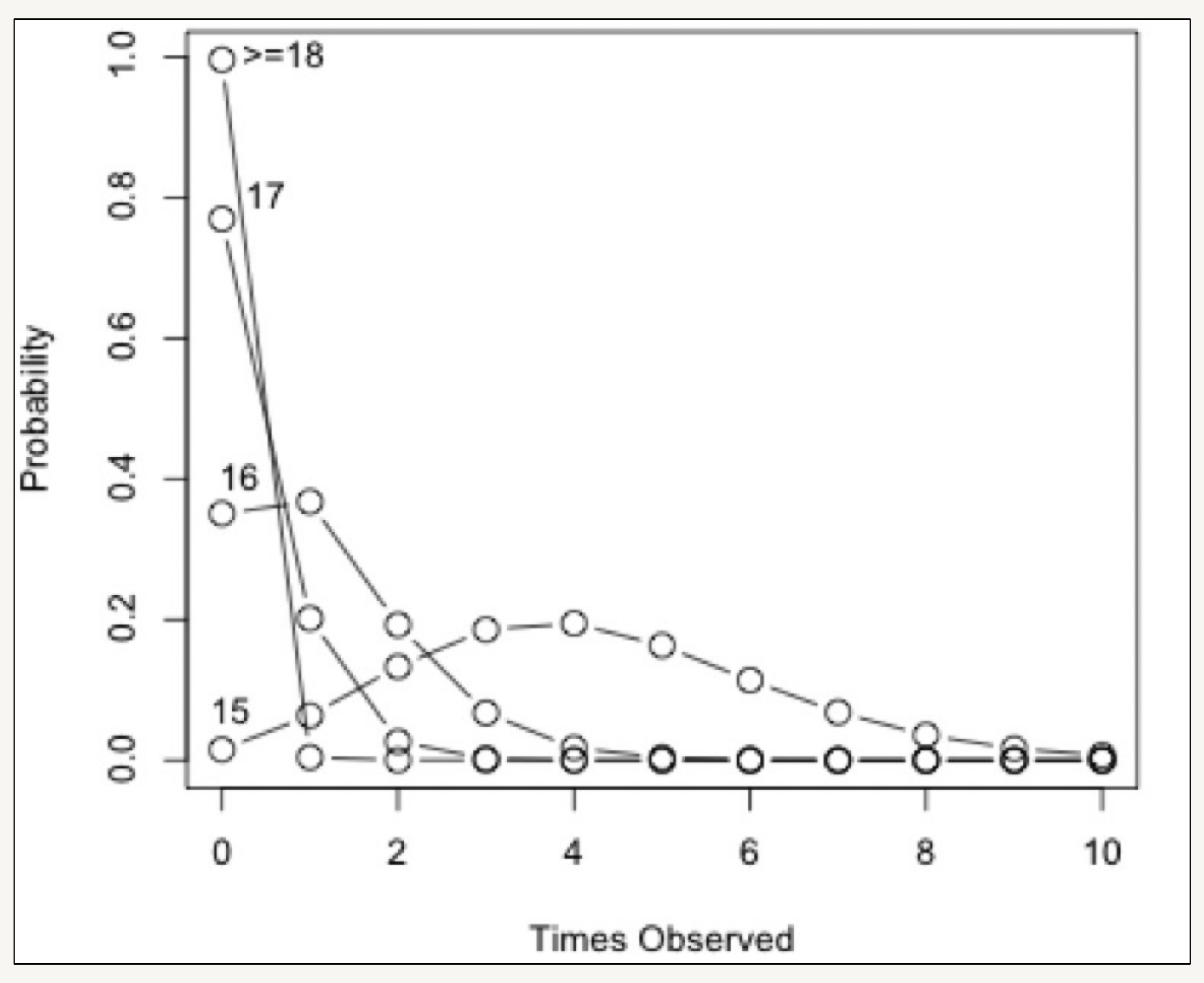


Figure 4 shows the probability of observing a kmer 0-10 times by chance. By 18 base pairs long, it is extremely unlikely to see a given kmer by chance.

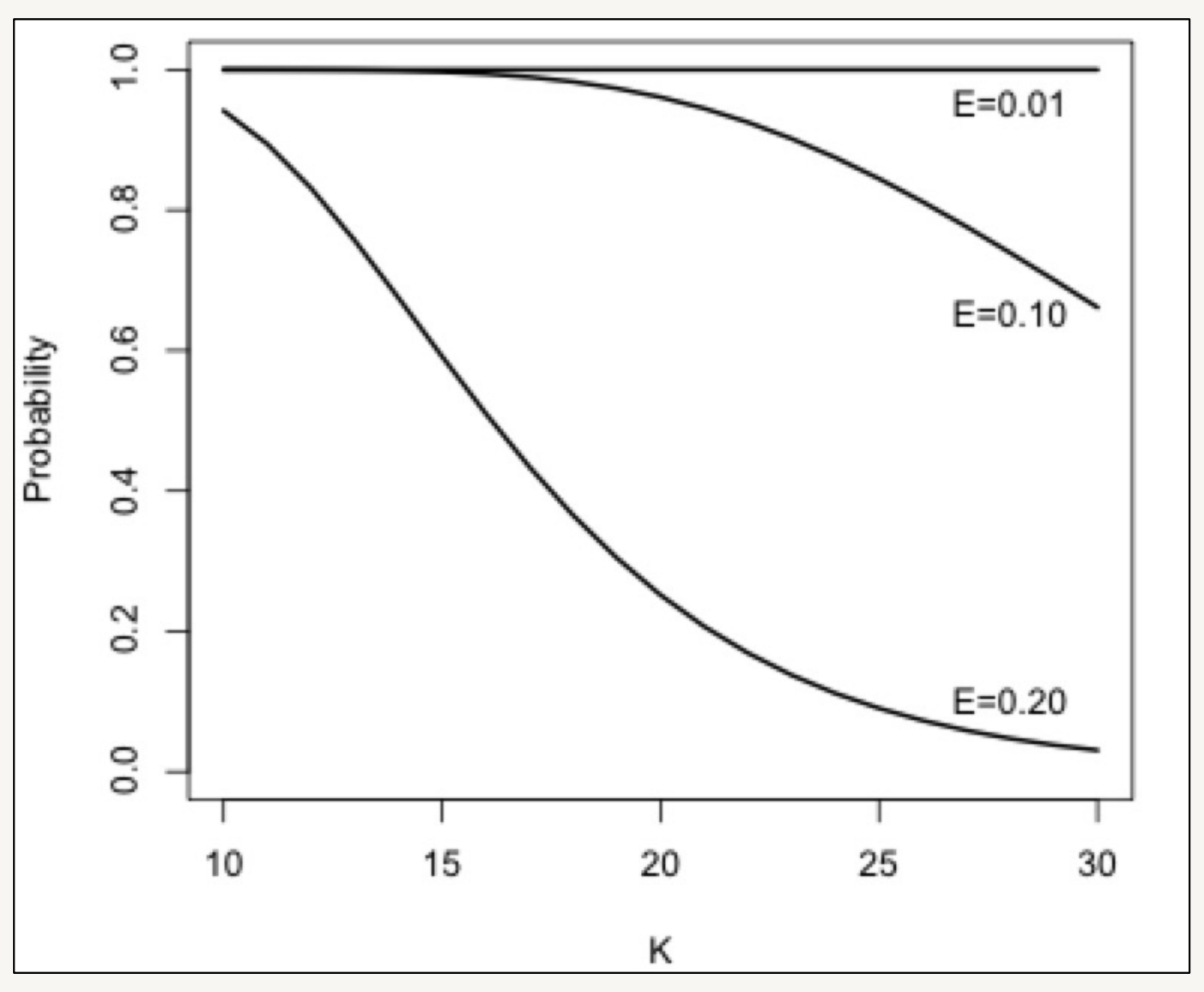


Figure 5 shows the probability of finding at least 1 kmer in 25 reads without an error for the given kmer sizes. With an error rate of 0.13 and a kmer size of 18, we expect to see almost every SCK at least once.

2. SCK Distribution

- Identified homozygous and heterozygous peaks using 18mer coverage values.
- Chose homozygous SCKs to avoid phasing issues.
- Selected 18mers within the coverage range (50x-58x), determined using the observed SCK homozygous peak and avoiding overlap with the heterozygous peak (**Fig 6**).

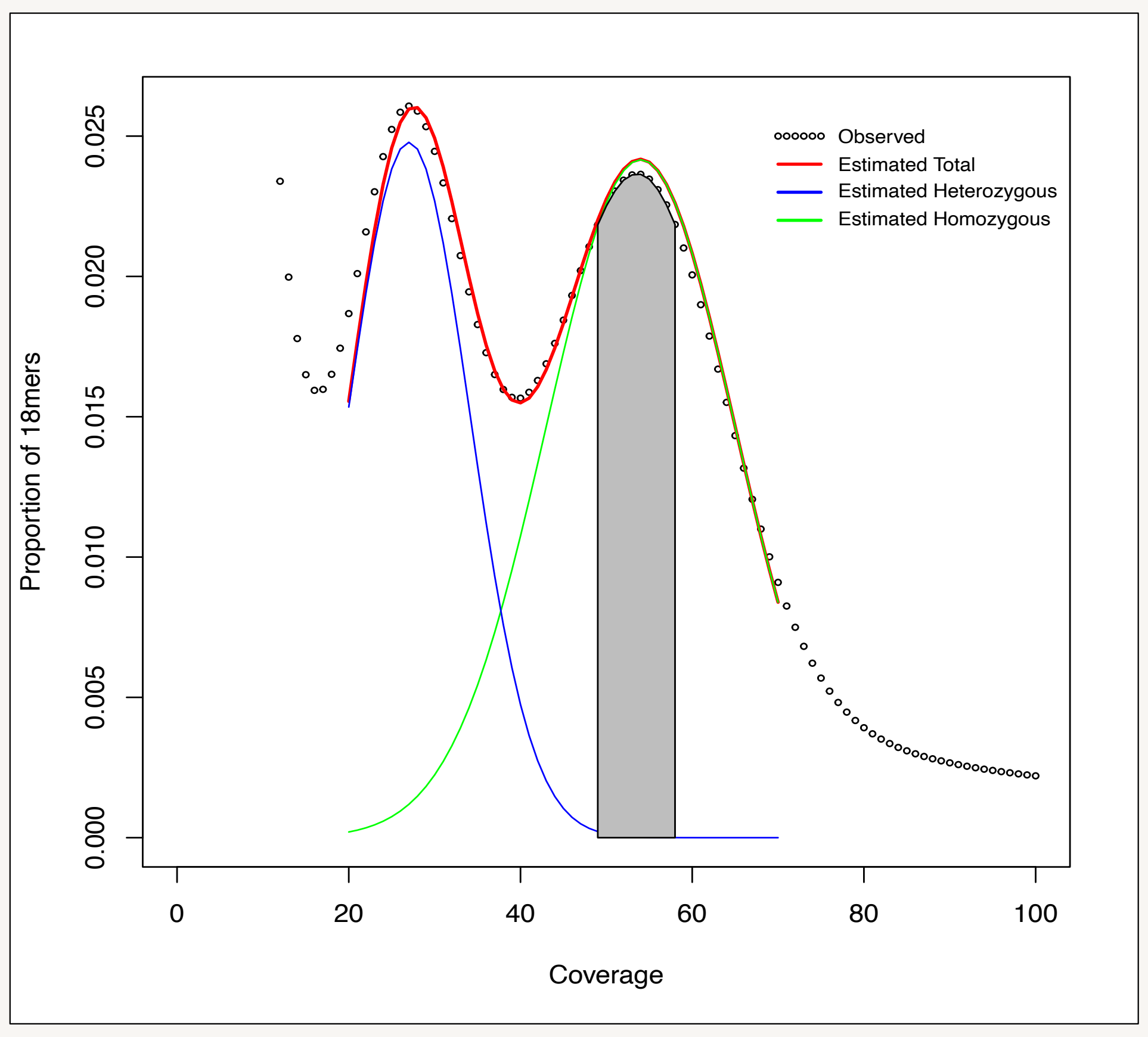


Figure 6 represents the observed distribution of 18mers over different coverage values. We wrote a method to identify two distinct kmer peaks: heterozygous (left) and homozygous (right). The chosen SCK coverage range (50x-58x) is highlighted in gray.

3. Final SCK Selection

- First filtered SCKs by short, high-quality reads. Selected initial SCKs that appeared within the desired coverage and did not contain homo-polymers likely resulting from sequencing error.
- SCKs from previous step were then filtered using long, low-quality reads. With a 13% sequencing error, each SCK is expected to appear in only 2 reads. To make sure that only SCKs are selected, only kmers that appear 2-6 times in the long reads were selected.

4. Chosen SCKs

- After the above steps, we obtained 11,813,926 SCKs to be used for assembly

Future Directions

Next steps:

- Use chosen SCK set to estimate relative positions of SCKs along each chromosome.
- Previous linking using short reads on mouse chromosomes allowed for grouping of non-repetitive regions.
- Use linking data and long reads to span large repeat regions.

Computational Costs:

- Our goal is to reduce the computational cost of genome assembly since current methods require large amounts of disk space and/or RAM.
- Currently our kmer selection process can be performed in 7 hours using 300 gigabytes of RAM. Additional steps are expected to be efficient.

Future considerations and implications:

- Use previously-assembled genomes to determine if proposed method can successfully re-assemble other genomes.
- Only 1% of eukaryotic genomes have currently been assembled[2]. There is need for a new, faster, and computationally-cheaper method.

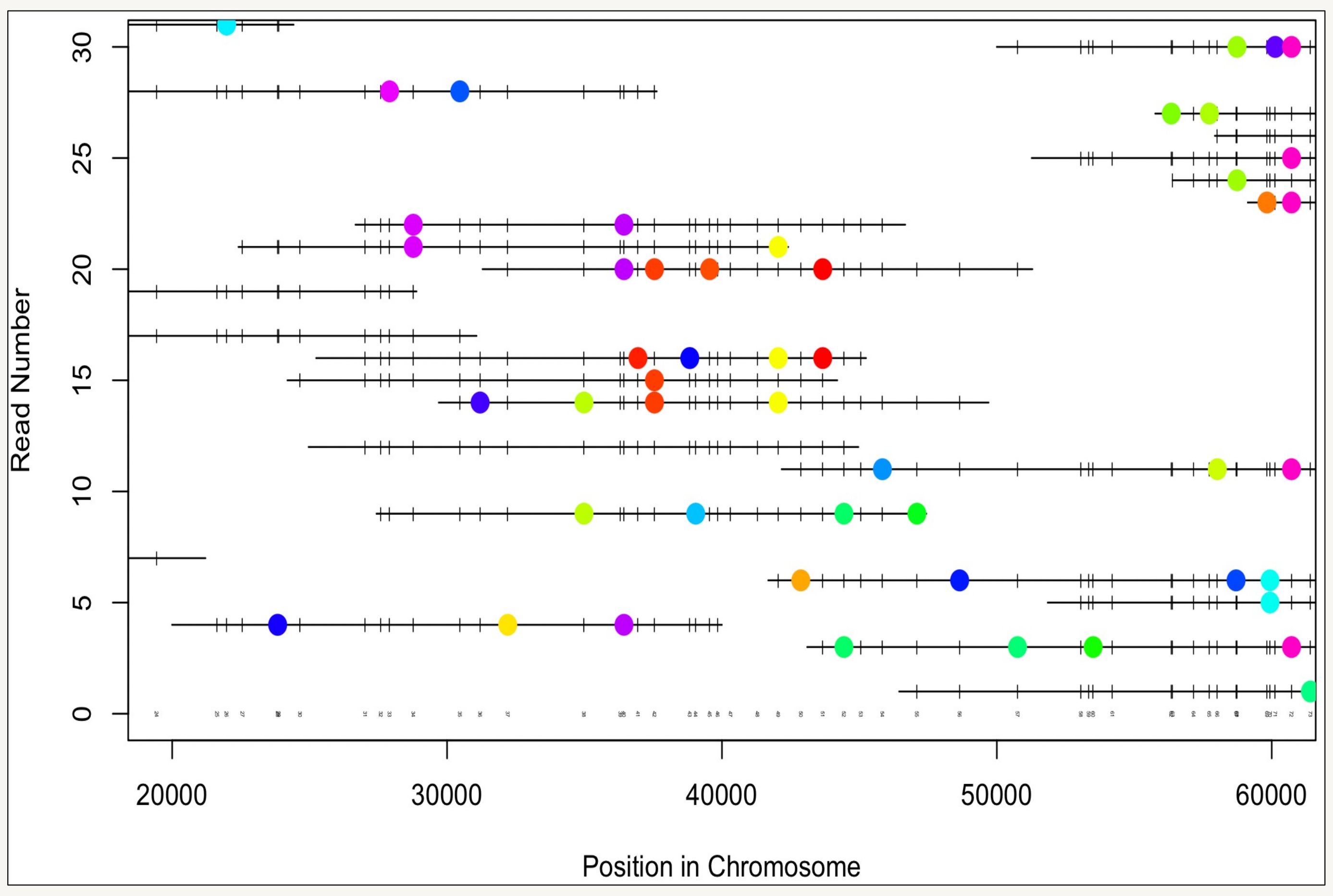


Figure 7 depicts a simulated representation of long reads from a region of a chromosome. The ticks represent sequencing error in the reads while the dots represent error-free SCKs. To maximize use of long reads, reads will have to be aligned based on shared SCKs.

References

1. Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., & Fan, W. (2012). Comparison of the two major classes of assembly algorithms: Overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics*, *11*(1), 25–37. https://doi.org/10.1093/bfgp/elr035
2. Hotaling, S., Kelley, J. L., & Frandsen, P. B. (2021). Toward a genome sequence for every animal: Where are we now? *Proceedings of the National Academy of Sciences*, *118*(52), e2109019118. https://doi.org/10.1073/pnas.2109019118

Acknowledgements

We would like to thank Dr. Alan Lemmon for his guidance, time, and mentorship throughout this project. We would also like to thank Dr. Emily Lemmon and the entire Lemmon Lab for their continuous support and uplifting community.