When a reference genome is not present, reads must be aligned using de novo assembly. A comparative analysis was conducted on two de novo assembly tools, Velvet and Oases, both of which are based on the de Bruijn graph algorithm. This analysis was performed using approximately 2 million short reads obtained from Illumina DNA sequencing data of Escherichia coli.

Fastq files were downloaded from the Sequence Read Archive (SRA) with the accession number SRR21904868. These files underwent preprocessing, which involved adapter trimming and the removal of low-quality reads before assembly. Roughly about 5,000 reads were removed during this process. Both Velvet and Oases were executed with the recommended parameters, with the exception that Velvetg employed the read tracking parameter to offer a more detailed description of the assembly. This parameter is crucial for Oases' algorithm as Oases relies on Velvet for its operations. The two tools were run using seven different kmer lengths: 75, 95, 97, 99, 101, 105, 115.

Velvet assembly process involves two steps: the creation of a hash index using the Fastq files, followed by the actual assembly. Oases is reliant on the preliminary assembly produced by the Velvet assembler.

For the seven different kmer sizes, a comparison was made between the two tools using several assembly metrics, including the N50 score, number of contigs, length of the longest contig, and the total length of assembled contigs. Velvet provided these assembly metrics upon the completion of its run. In contrast, Oases did not directly output these metrics. Instead, it generated a file named ‘transcripts.fa’, which contained the sequences of all contigs in FASTA format. The header of each contig included its length, which was extracted and used to calculate the assembly metrics through a Python script.

Figure 1. Comparison of Velvet and Oases N50 Values by Kmer Size.

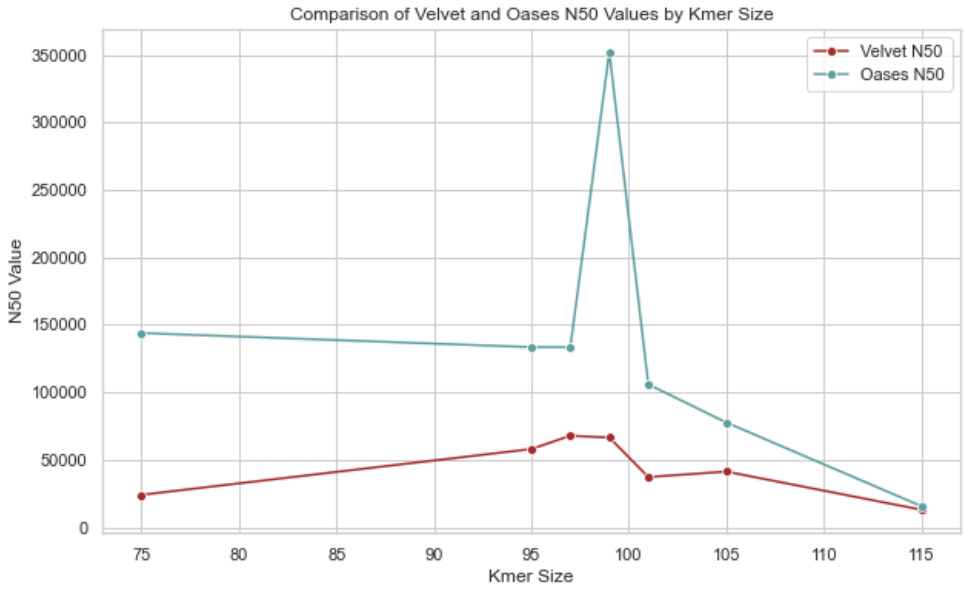

Figure 1 visually depicts the impact of kmer length on the N50 score for both Velvet and Oases. The data indicates that, for both tools, the N50 score reached its peak at a kmer length of 97, which seems to be the optimal choice considering the specific parameters used. Notably, Oases consistently achieved a N50 score higher than Velvet, with values of 130,000 and 70,000, respectively, at a kmer size of 97. Moreover, Oases produced assemblies with fewer, longer contigs, resulting in an overall longer assembled genome compared to Velvet, as detailed in Table 1.

Table 1. Comparison of Velvet and Oases Assembly Metrics

	Velvet				Oases			
kmer size	**N50**	**Num of contigs**	**Size of largest contig**	**Total Length**	**N50**	**Num of contigs**	**Size of largest contig**	**Total Length**
75	24,046	4074	99,347	4,826,068	143,956	630	808,778	11,380,602
95	57,985	957	182,873	4,777,075	133,509	559	753,063	11,575,071
97	67,849	848	232,841	4,774,951	133,513	537	1,075,500	10,338,390
99	66,492	809	182,877	4,774,662	352,270	516	779,645	10,782,566
101	37,239	2564	174,506	4,836,070	105,905	751	734,766	10,608,672
105	41,273	1837	138,140	4,812,862	77,660	796	463,011	9,685,634
115	13,009	1309	48,515	4,777,437	15,672	891	57,390	6,298,773

The justification for the much better performance of Oases can be attributed to the corrective measures it applies after using Velvet's assembly as an input. For instance, Oases uses algorithms similar to TourBus searches to help identify and correct potentially erroneous chimeric contigs. Chimeric contigs can introduce errors in genome assembly, making their correction essential. This distinction could explain why Oases consistently produces superior assembly metrics compared to Velvet, establishing it as the better tools for de novo assembly.