De Novo Assembly with SPAdes assembler

Overview

We published the whitepaper, "Dell EMC PowerEdge R940 makes *De Novo* Aseembly easier", last year to study the behavior of SOAPdenovo2 [1]. However, the whitepaper is limited to one *De Novo* assembly application. Hence, we want to expand our application coverage little further. We decided to test SPAdes (2012) since it is a relatively new application and reported for some improvement on the Euler-Velvet-SC assembler (2011) and SOAPdenovo[i]. SPAdes is also based on de Bruijn graph algorithm like most of the assemblers targeting Next Generation Sequencing (NGS) data. De Bruijin graph-based assemblers would be more appropriate for larger datasets having more than a hundred-millions of short reads.

As shown in **Figure 1**, Greedy-Extension and overlap-layout-consensus (OLC) approaches were used in the very early next gen assemblers [2]. Greedy-Extension's heuristic is that the highest scoring alignment takes on another read with the highest score. However, this approach is vulnerable to imperfect overlaps and multiple matches among the reads and leads to an incomplete assembly or an arrested assembly. OLC approach works better for long reads such as Sanger or other technology generating more than 100bp due to minimum overlap threshold (454, Ion Torrent, PacBio, and so on). De Bruijin graph-based assemblers are more suitable for short read sequencing technologies such as Illumina. The approach breaks the sequencing reads into successive k-mers, and the graph maps the k-mers. Each k-mer forms a node, and edges are drawn between each k-mer in a read.

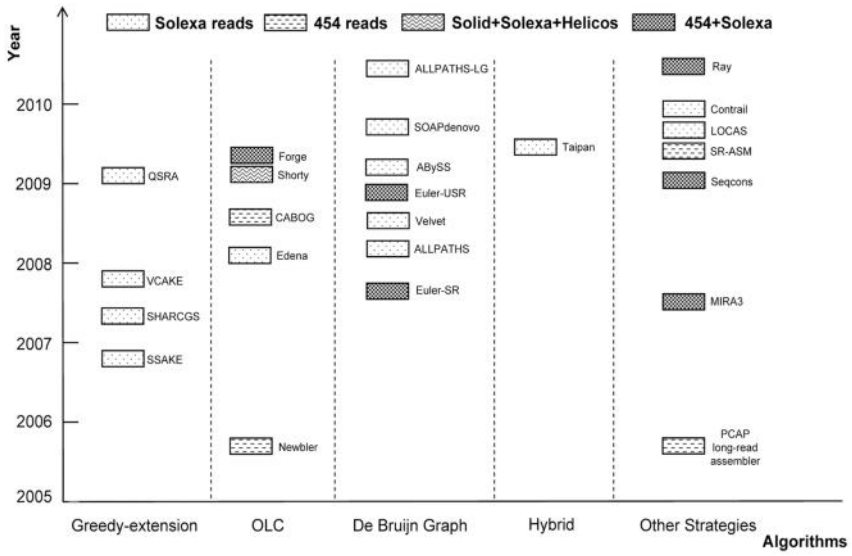

Figure 1 Overview of de novo short reads assemblers. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3056720/

SPAdes is a relatively recent application based on de Bruijn graph for both single-cell and multicell data. It improves on the recently released Euler Velvet Single Cell (E +V- SC) assembler (specialized for single-cell data) and on popular assemblers Velvet and SoapDeNovo (for multicell data).

All tests were performed on Dell EMC PowerEdge R940 configured as shown in **Table 1**. The total number of cores available in the system is 96, and the total amount of memory is 1.5TB.

Table 1 Dell EMC PowerEdge R940 Configuration

	Dell EMC PowerEdge R940
CPU	4x Intel Xeon Platinum 8168 CPU, 24c @ 2.70GHz (Skylake)
RAM	48x 32GB @2666 MHz
OS	RHEL 7.4
Kernel	3.10.0-693.el7.x86_64
Local Storage	12x 1.2TB 10K RPM SAS 12Gbps 512n 2.5in Hot-plug Hard Drive in RAID 0
Interconnect	Intel® Omni-Path
BIOS System Profile	Performance Optimized
Logical Processor	Disabled
Virtualization Technology	Disabled
SPAdes version	3.10.1
Python version	2.7.13

The data used for the tests is a paired-end read, ERR318658 which can be downloaded from European Nucleotide Archive (ENA). The read generated from blood sample as a control to identify somatic alterations in the primary and metastatic colorectal tumors. This data contains 3.2 Billion Reads (BR) with the read length of 101 nucleotides.

Performance Evaluation

SPAdes runs three sets of de Bruijn graphs with 21-mer, 33-mer, and 55-mer consecutively. This is the main difference with regards to SOAPdenovo2 which run a single k-mer, either 63-mer or 127-mer.

In **Figure 2**, the runtimes, wall-clock times, are plotted in days (blue bars) with various number of cores, 28, 46, and 92 cores. Since we do not want to use the entire cores of each socket, 92 cores were picked as the maximum number of cores for the system. One core per socket was reserved for OS and other maintenance processes. Subsequent tests were done by reducing the number of cores in half. Peak memory consumptions for each case is plotted as a line graph. SPAdes runs significantly longer than SOAPdenovo2 due to the multiple iterations on three different k-mers.

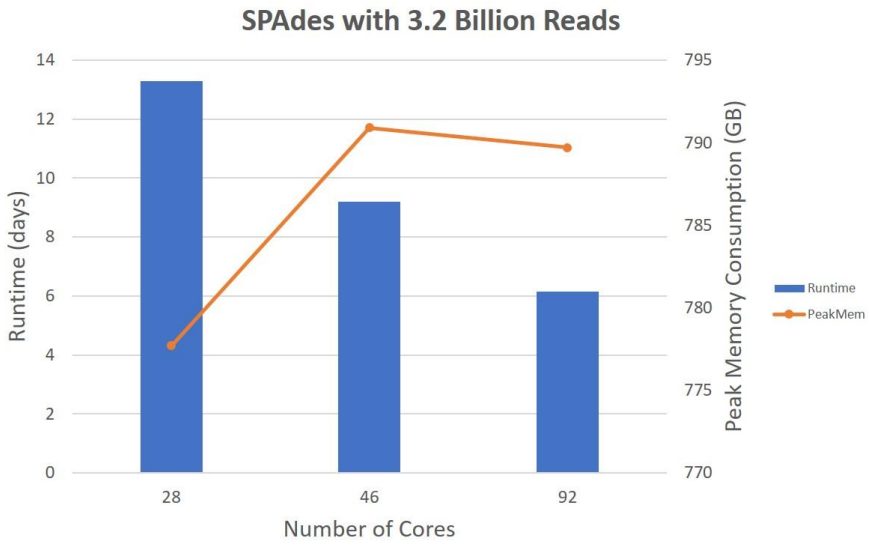

Figure 2 SPAdes tests with various number of cores

The peak memory consumption is very similar to SOAPdenovo2. Both applications require slightly less than 800GB memory to process 3.2 BR.

Conclusion

Utilizing more cores helps to reduce the runtime of SPAdes significantly as shown in **Figure 2**. For SPAdes, it is recommendable to use the highest core count CPUs like Intel Xeon Plantinum 8180 processor with 28 cores and 3.80GHz to bring down the runtime further.

Resources

Internal web page

1. http://en.community.dell.com/techcenter/blueprints/blueprint_for_hpc/m/mediagallery/20444301

External web page

2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2874646/

Contacts

Americas

Kihoon Yoon

Sr. Principal Systems Dev Eng

Kihoon.Yoon@dell.com

+1 512 728 4191

[i] It refers an earlier version of SOAPdenovo, not SOAPdenovo2.