Pacasus

Removing chimeras in long reads after whole-genome amplification

Sven Warris[1], Elio Schijlen[1], Henri van de Geest[1], Thamara Hesselink[1], Gabino Sanchez Perez[1,3], Dick de Ridder[2]

[1]Applied Bioinformatics, Wageningen UR, Wageningen, The Netherlands
[2]Bioinformatics Group, Wageningen UR, Wageningen, The Netherlands
[3]Genetwister Technologies BV. Wageningen, The Netherlands

Introduction

In many new genomics applications, such as in case of single cells or flow-sorted material, the amount of DNA available for analyses is extremely limiting. Hence there is an urgent demand to reduce the amount of DNA required for sequencing. Whole genome amplification (WGA) is the commonly used option. WGA multiplies small amounts of DNA to generate quantities suitable for sequencing. However, WGA protocols tend to alter the initial template DNA during amplification and introduce chimeric DNA fragments. These chimera severely hamper adequate read mapping and *de novo* assembly, notably with long read technologies such as PacBio.

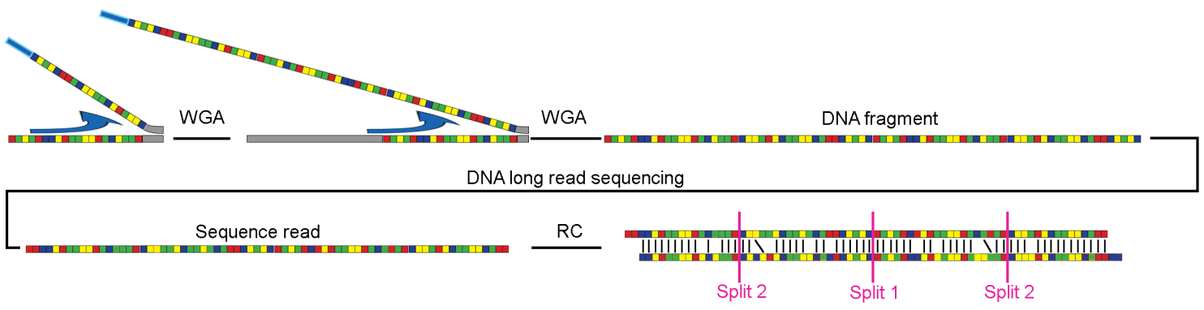

Figure 1: Chimeras are introduced when during WGA the DNA-polymerase continues along an already created WGA product. In this example this incorrect elongation occurs several times, resulting in a DNA fragment containing four copies of the original fragment, which is in turn sequenced. Pacasus detects the palindrome sequence by aligning the read's RC to itself and splits the read in two smaller reads (1). This process is repeated and splits the two resulting reads again (2).

Materials & Methods

To detect chimeras, raw PacBio reads are aligned to their reverse-complement (RC) sequence with Smith-Waterman using pyPaSWAS. The alignment results are inspected for palindromes. If a palindrome sequence is found, the read is split at the start of the palindrome sequence (Figure 1). De novo assemblies are created using Canu, DBG2OLC and Sparse.

Read set	Control	WGA	Pacasus
No. contigs	852	2128	1015
Length (Mbp)	115.6	116.8	123.9
Longest (Mbp)	1.18	0.65	3.40
N50 (Kbp)	292.6	72.6	301.6
L50	117	479	109

Table 1: Assembly statistics for *A. thaliana* using non-amplified DNA ('Control'), amplified DNA ('WGA') and the second data set after processing ('Pacasus').

Results

Pacasus has been applied to PacBio sequence reads of WGA DNA from *Arabidopsis thaliana*. The quality of the *de novo* assembly based on the cleaned set is as high as that of an assembly based on non-amplified DNA (Table 1). The cumulative length plot shows that the PacBio-only assembly performs as good as the hybrid assembly and both are in the same range as the assemblies based on non-amplified DNA (Figure 2).

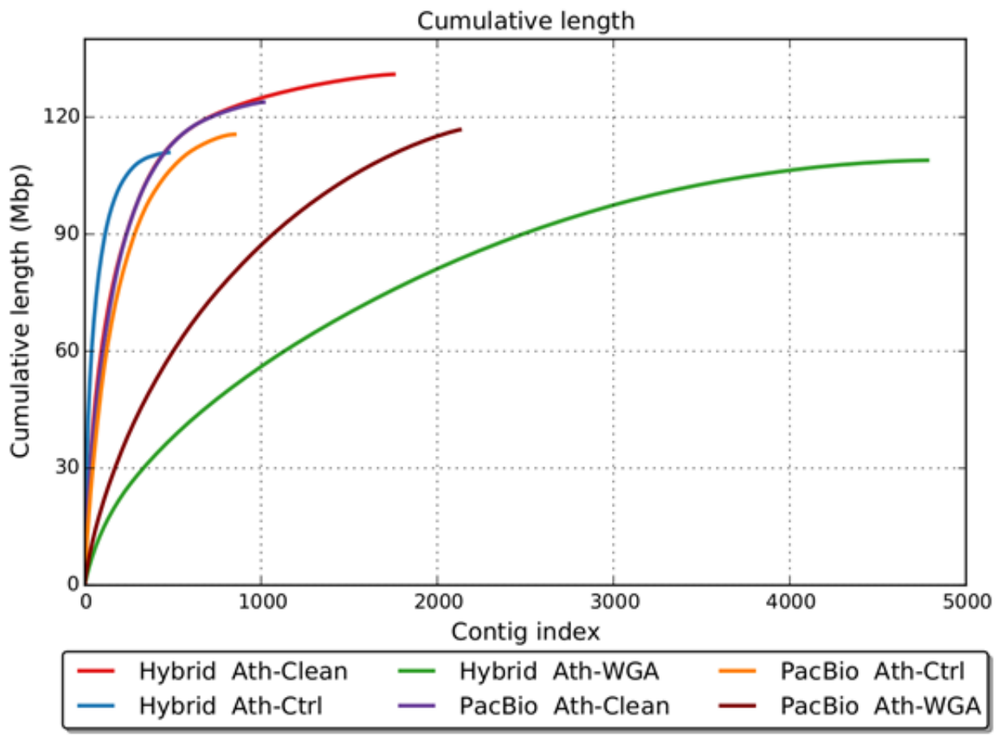

Figure 2: Cumulative lengths of the PacBio-only and hybrid assemblies of Ath-Ctrl, Ath-WGA and Ath-Clean.

Conclusion

Pacasus cleans long, error-rich reads containing chimera introduced after WGA. Pacasus has limited effect on the number of nucleotides in the read set and decreases the average read length by less than 50%. The loss in read length is clearly offset by the removal of incorrect contiguity information. The PacBio-only assembly improved markedly in quality. Pacasus allows to analyse PacBio data obtained from low amounts of DNA from studies of single cells (e.g. in cancer research), of single chromosomes and of environmental samples.

Availability

https://github.com/swarris/Pacasus/

Applied bioinformatics, Bioscience, Wageningen UR, P.O. Box 123, 6700 AB Wageningen
sven.warris@wur.nl
@swarris