XXXXXXXX
doi: 10.1093/XXXXX/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

Sequence Analysis

gencore: an efficient tool to generate consensus reads for error suppressing and duplicate removing of NGS data

Shifu Chen[1,2,†,*], Yanqing Zhou[1,†], Yaru Chen[1], Tanxiao Huang[1], Wenting Liao[1], Yun Xu[1], Zhihua Liu[2], and Jia Gu[2]

[1]HaploX Biotechnology. [2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences.

*To whom correspondence should be addressed.

[†]The first two authors should be regarded as Joint First Authors.

Abstract

Summary: this paper presents an efficient tool *gencore*, to eliminate errors and duplicates of next-generation sequencing (NGS) data. This tool clusters the mapped sequencing reads and merges each cluster to generate one consensus read. If the data has unique molecular identifier (UMI), *gencore* uses it for identifying the reads derived from same original DNA fragment. Comparing to the conventional tool Picard, *gencore* greatly reduces the output data's mapping mismatches, which are mostly caused by errors. This error-suppressing feature makes *gencore* very suitable for the application of detecting ultra-low frequency mutations from deep sequencing data. Comparing to the performance of Picard, *gencore* is about 3X faster and uses much less memory.

Availability and Implementation: *gencore* is an open source tool written in C++. It's hosted in github: https://github.com/OpenGene/gencore

Contact: chen@haplox.com

1 Introduction

High-depth next-generation sequencing (NGS) has been widely used for precision cancer diagnosis and treatment. From such deep sequencing data, somatic mutations can be detected to guide personalized targeted therapy or immunotherapy. Recently, circulating tumor DNA (ctDNA) sequencing has been recognized as a promising biomarker for cancer treatment and monitoring. Since the tumor-derived DNA is usually a small part of the total blood cell-free DNA, the mutant allele frequency (MAF) of the variants detected from ctDNA sequencing data can be very low (as low as 0.1%). To detect such low-frequency variants, we usually increase the sequencing depth (can be higher than 10,000x). However, the processes of making NGS library and sequencing are not error-free. Particularly, the library amplification using PCR technology can produce a lot of errors, and consequently cause some false positive mutations in the result of NGS data analysis.

As a result of library amplification, NGS data can have duplicates. The higher the sequencing depth is, the more duplication the data can have. Traditionally, we just mark the duplicated reads and remove them before downstream analysis. For low-depth paired-end NGS data, the read pairs of same start and end mapping positions can be treated as duplicated reads derived from a same original DNA fragment. Then, the reads clustered together can be merged to be a single read. Due to the nature that errors usually happen randomly, the inconsistent mismatches in the clustered read group can be removed to generate a consensus read.

However, for ultra-deep sequencing, it's possible that two read pairs with same positions are derived from different original DNA fragments. This possibility can be higher when the DNA fragments are shorter. For example, cell-free DNA usually has a peak length of ~167 bp, which is much shorter than the peak length of normally fragmented genomic DNA. To better identify sequencing reads derived from different DNA fragments, a technology called unique molecular identifier (UMI) has been developed. With UMI technology, each DNA fragment is ligated with unique random barcodes before any DNA amplification process. The UMIs can be then used for accurate clustering of sequencing reads.

Currently the conventional de-duplication tool like Picard MarkDuplicates cannot perform consensus read generating well, and is not able to handle the UMI-integrated data. Furthermore, Picard MarkDuplicates is too slow and uses too much memory, which makes it not suitable for cloud-based deployment. These unmet requirements drove us to develop a new tool called *gencore*, which eliminates errors and removes duplicates by generating consensus reads.

2 Implementation

gencore requires an input of sorted BAM file and a reference genome FASTA file. If the FASTQ data has UMIs, it can be preprocessed using

fastp (Chen, Zhou, Chen, & Gu, 2018) to move the UMIs from read sequences to read identifiers.

gencore clusters read pairs by their mapping positions and UMIs (if applicable), and then generates a consensus read for each cluster. The main implementation of *gencore* can be briefly introduced as following steps:

(1) Position clustering: all mapped read pairs are grouped together by their mapping chromosome, start position and end position.

(2) UMI clustering: for each group of same mapping positions, read pairs are then grouped by their UMIs with one base difference tolerance. If the data has no UMIs, this step is skipped.

(3) Pair scoring: for each pair in a cluster, the overlapped region of the paired reads is computed. Each base is initialized with a score. And for each base in the overlapped region, its score is adjusted according to its consistence with its paired base, with the consideration of their quality scores.

(4) Cluster scoring: for each position in a cluster, the total score of different bases (A/T/C/G) is computed by summarizing the scores that are computed in last step of each base.

(5) Consensus read generating: for each position in a cluster, its base diversity is computed according to the scores of different bases computed in last step. If *gencore* finds one dominant base, this base will also be presented in the consensus read. Otherwise the corresponding base in the reference genome will be used.

After the processing is done, *gencore* will generate a summary of the data before and after processing. Some metrics like mapping rate, duplication rate, passing filter rate and mismatch rate are reported in a JSON format report. Furthermore, *gencore* computes the number of clusters for each different duplication levels, and reports it as a duplication level histogram.

3 Application

Since *gencore* can be used to reduce sequencing errors, it is very useful for the application of detecting low-frequency somatic mutations from cancer sequencing data. Particularly, when the samples are from blood, urine or malignant effusion, the MAF of variants can be even much lower than 1%. The detection of such low-frequency variants can be seriously affected by the errors, which are usually introduced by library preparation and sequencing. *gencore* can significantly reduce the sequencing errors of deep sequencing data, and consequently reduce the false positive calling rate.

To evaluate how *gencore* can help the low-frequency variant detection, we conducted an evaluation experiment using the Horizon Multiplex I cfDNA Reference Standard Set (HD777, HD778). The HD777 is a reference standard set with 8 known mutations at the EGFR, KRAS, NRAS and PIK3CA genes with expected allelic frequency of 5%. The HD778 reference standard set has the same mutations at these four genes, but the expected allelic frequency is 1%.

Sequencing libraries for HD777 and HD778 were made using IDT xGen Dual Index UMI Adapters, and captured with a 451-gene cancer panel. Libraries were sequenced using Illumina NovaSeq 6000 sequencers. The output FASTQ data are 32.6Gb and 32.7Gb respectively.

The FASTQ files were preprocessed by fastp, and then mapped to reference genome hg19 using BWA (Li & Durbin, 2009). After the mapped bam file was sorted using Samtools, the sorted bam files were then be processed by Picard and *gencore* respectively. VarScan2 (Koboldt et al., 2012) was then used to call SNVs from the processed bam files. The missense variants detected in the coding sequences of EGFR/KRAS/NRAS/PIK3CA genes were then filtered with a condition (supporting reads ≥ 4). Then the variant calling results and running performance were evaluated. The comparison result is shown in Table 1.

Table 1. A comparison of Picard and *gencore* results

	Picard Tool	***gencore* UMI mode**	***gencore* non-UMI mode**
HD777 with 8 true positive mutations of %5 MAF			
Depth	637.7X	643.5X	650.6X
Mismatch Rate	0.024%	0.010%	0.012%
Variants (TP+FP)	15 (8+7)	8 (8+0)	8 (8+0)
PPV	53.3%	100%	100%
Running Time	88m	38m	44m
Memory Usage	25.4G	3.08G	3.09G
HD778 with 8 true positive mutations of 1% MAF			
Depth	663.2X	671.9X	679.9X
Mismatch Rate	0.020%	0.010%	0.010%
Variants (TP+FP)	14 (8+6)	8 (8+0)	8 (8+0)
PPV	57.1%	100%	100%
Running Time	77m	22m	22m
Memory Usage	25.5G	3.14G	3.15G

TP = True Positives; FP = False Positives. The values of Depth, Mismatch Rate, Variants and PPV are evaluated with the data processed by Picard Tool and *gencore* respectively. For data with UMI, *gencore* supports UMI mode and non-UMI mode. In the non-UMI mode, *gencore* ignores the UMI and clusters the reads only based on the mapping positions. According to the specification of the reference standard sets, Only the variants at EGFR/KRAS/NRAS/PIK3CA genes are evaluated.

From Table 1, we can find that the *gencore* processed data contained much fewer mismatches than the Picard processed data. With downstream analysis, the *gencore* processed data was detected with only all 8 true positives, while the Picard processed data was detected with many false positives. Besides the improvement of the accuracy, *gencore* was much faster and memory efficient.

Funding

The presented study was funded by Shenzhen Science and Technology Innovation Committee Technical Research Project (grant No. JSGG20180703164202084) and Shenzhen Strategic Emerging Industry Development Special Fund (grant No. 20170922151538732).

References

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34*(17), i884-i890. doi:10.1093/bioinformatics/bty560

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., . . . Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res, 22*(3), 568-576. doi:10.1101/gr.129684.111

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324