ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A Survey of Error Correction for Next Generation Sequencing

Ramgopal Ahirwal
M.Tech Student
Dept. of Computer science and Engineering,
UIT, RGPV, Bhopal (M.P) (India)

Prof. Anjana Deen
Assistant Professor
Dept. of Computer science and Engineering
UIT, RGPV, Bhopal (M.P) (India)

Dr. Piyush Kumar Shukla
Assistant Professor
Dept. of Computer science and Engineering,
UIT, RGPV, Bhopal (M.P) (India)

Abstract: Next-generation sequences (NGS) is a very important process drawback to review the population sizes of deoxyribonucleic acid molecules and to scale back the redundancies in NGS information. The arrival of next-generation sequencing technologies has enhanced the accuracy and amount of sequence information, gap the door to bigger opportunities in genomic analysis. Error Correction is vital for many next-generation sequencing applications as a result of extremely correct sequenced reads can seemingly cause higher quality results. Several techniques for error correction of sequencing knowledge from next-gen platforms are developed within the recent years. However, compared with the quick development of sequencing technologies, there's a scarcity of standardized analysis procedure for various error-correction ways, creating it troublesome to assess their relative deserves and demerits.

Keywords: Next-generation sequencing (NGS), Error Correction, DNA, RNA.

I.

II. Introduction

There are variety totally different of various NGS platforms exploitation different sequencing technologies, an in depth discussion of that is on the far side the scope of this text. However, all NGS platforms perform sequencing of many little fragments of deoxyribonucleic acid in parallel. Bioinformatics analyses are wont to piece along these fragments by mapping the individual reads to the human reference ordering. Every of the 3 billion bases within the human ordering are sequenced multiple times, providing high depth to deliver correct information and an insight into surprising deoxyribonucleic acid variation. [1]

Next-generation sequencing (NGS), also called high-throughput sequencing, is that the catch-all term wont to describe variety of various trendy sequencing technologies including:

A. Illumina (Solexa) sequencing
B. Roche 454 sequencing
C. Ion torrent: nucleon / PGM sequencing

These recent technologies enable us to sequence deoxyribonucleic acid and polymer far more quickly and cheaply than the antecedently used Sanger sequencing, and in and of itself have revolutionized the study of genetics and biological science.

A. Illumina sequencing

In NGS, huge numbers of short reads are sequenced in a very single stroke.

To do this, first of all the input sample should be cleaved into short sections. The length of those sections can rely on the actual sequencing machinery used. In Illumina sequencing, 100-150bp reads are used. Somewhat longer fragments are ligated to generic adaptors and toughened to a slide exploitation the adaptors. PCR is dole out to amplify every read, making a spot with several copies of constant read. They’re then separated into single strands to be sequenced.

B. Roche 454 sequencing

Roche 454 sequencing will sequence for much longer reads than Illumina. Like Illumina, it will this by sequencing multiple reads quickly by reading optical signals as bases are added. As in Illumina, the deoxyribonucleic acid or ribonucleic acid is fragmented into shorter reads, during this case up to 1kb. Generic adaptors are added to the ends and these are annealed to beads, one deoxyribonucleic acid fragment per bead. The fragments are then amplified by PCR exploitation adaptor- specific primers. Each bead is then placed in a very single well of a slide. Thus every well can contain one bead, lined in several PCR copies of one sequence. The wells additionally contain deoxyribonucleic acid enzyme and sequencing buffers.

C. Ion Torrent: nucleon / PGM sequencing

Unlike Illumina and 454, ion torrent and ion nucleon sequencing don't create use of optical signals. Instead, they exploit the very fact that addition of a dNTP to a deoxyribonucleic acid chemical compound releases an H+ ion. As in other forms of NGS, the input deoxyribonucleic acid or ribonucleic acid is fragmented, this point ~200bp. Adaptors are added and one molecule is placed onto a bead. The molecules are amplified on the bead by emulsion PCR every bead is placed into one well of a slide.

III. OVERVIEW OF ERROR CORRECTION

Error-correction ways designed thus far have in the main targeted haplotype ordering sequencing. During this setting, error correction with reference to a particular genomic position will be achieved by birthing out all the reads covering the position, and examining the bottom in this specific position from of these reads. As errors are rare and random, reads that contain miscalculation in a very specific position will be corrected exploitation the bulk of the reads that have this base properly. This general plan has been enforced altogether error correction algorithms, albeit indirectly. Because the supply ordering is unknown, the reads from constant genomic location are inferred counting on the idea that they generally share sub reads of a set length, like k-mers. Some ways [2, 6] more derive multiple sequence alignment (MSA) of reads that share common k-mers and ask for corrections counting on the MSA, whereas others [7–10, 3–5, 11–14] correct errors at the amount of k-mers or variable length sub reads. In each cases, genomic repeats and non-uniform sampling of ordering could cause multiple equally seemingly correction selections that cause ambiguity in correction we tend to classify error-correction ways into 3 types—k-spectrum based mostly, suffix tree/array-based and MSA-based ways.

IV. Review of previous method

A. *"Direct detection of deoxyribonucleic acid methylation throughout single-molecule, time period sequencing" Benjamin A Flusberg, dale R Webster, 2010.*

We describe the direct detection of deoxyribonucleic acid methylation, while not bisulfite conversion, through single-molecule, time period (SMRT) sequencing. In SMRT sequencing, deoxyribonucleic acid polymerases catalyze the incorporation of fluorescently labeled nucleotides into complementary super molecule strands. The arrival times and durations of the ensuing light pulses yield data regarding enzyme dynamics and permit direct detection of changed nucleotides within the deoxyribonucleic acid temple, as well as N6-methyladenine, 5-methylcytosine and 5-hydroxymethylcytosine. Measuring of enzyme dynamics is an intrinsic a part of SMRT sequencing and doesn't adversely have an effect on determination of primary deoxyribonucleic acid sequence. The varied modifications have an effect on enzyme dynamics otherwise, permitting discrimination between them. We tend to use these kinetic signatures to spot purine methylation in genomic samples and located that, together with circular agreement sequencing, they will change single-molecule identification of epigenetic modifications with base-pair resolution. This methodology is amenable to long scan lengths and can doubtless alter mapping of methylation patterns in even extremely repetitive genomic regions. [15]

B. *"Reptile: representative application for brief scan error correction" Xiao yang 2010.*

Error correction is essential to the success of next generation sequencing applications, like re sequencing and de novo genome sequencing. It's particularly vital for prime output short-read sequencing, wherever reads are much shorter and additional abundant and errors additional frequent than in ancient Sanger sequencing. [16].

C. *"Quake: quality-aware detection and correction of sequencing errors" David R Kelley 2010.*

We introduce Quake, a program to find and proper errors in deoxyribonucleic acid sequencing reads. Employing a most probability approach incorporating quality values and ester specific misname rates, Quake achieves the best accuracy on realistically simulated reads. We further demonstrate substantial improvements in de novo assembly and SNP detection after using Quake. Quake can be used for any size project, including more than one billion human reads. [17].

D. *"Correction of sequencing errors in a mixed set of reads" Leena Salmela 2010.*

High-throughput sequencing technologies produce large sets of short reads that may contain errors. These sequencing errors make de novo assembly challenging. Error correction aims to reduce the error rate prior assembly. Many de novo sequencing projects use reads from several sequencing technologies to get the benefits of all used technologies and to alleviate their shortcomings. However, combining such a mixed set of reads is problematic as many tools are specific to one sequencing platform. The SOLiD sequencing platform is especially problematic in this regard because of the two base colors coding of the reads. Therefore, new tools for working with mixed read sets are needed. [18]

E. *"PSAEC: An Improved Algorithm for Short Read Error Correction Using Partial Suffix Arrays" Zhiheng Zhao, Jianping Yin, 2011.*

Sequencing errors in high-throughput sequencing data constitute one of the major problems in analyzing such data. Error correction can reduce the error rate. However, it is a computation and data intensive process for large-scale data. This poses challenges for more efficient and scalable algorithms. In this paper, we propose PSAEC, an improved algorithm for short read error correction using partial suffix arrays in high-throughput sequencing data. [19]

After study of this paper, In this table contribution of many error correction technique and comparison between these technique and some parameter. Technique is Single molecule real time sequencing, Ion semiconductor, pyro sequencing (454), Sequencing synthesis (illumina), Sequencing by ligation (SOLiD) sequencing and Chain termination (Sanger sequencing). And parameter is Read length, Accuracy, Read per run, Time per run, Cost per 1 million bases, Advantage and Disadvantage.

V. Comparison of previous method

In this comparative table I compare the different method of error correction for next generation sequencing data there are different parameters are used compare a previous work parameters are Read length, Accuracy, Read per run, Time per run, Cost per 1 million bases, Advantage and Disadvantage. Also shows the name of different method and its acronyms Single molecule real time sequencing, Ion semiconductor, Pyroseqyencing (454), Sequencing by synthesis (illumina), Sequencing by ligation (SOLiD sequencing), Chain termination (Sanger Sequencing).

Table I. Comparative study in different technique of error correction & NGS

Method	**Single molecule real time sequencing**	**Ion semiconductor**	**Pyroseqyencing (454)**	**Sequencing by synthesis (illumina)**	**Sequencing by ligation(SOLiD sequencing)**	**Chain termination(Sanger Sequencing)**
Read length	2900 bp average	200 bp	700 bp	50 to 250 bp	50+35 or 50+50 bp	400 to 900 bp
Accuracy	87 % read length mode, 99 % accuracy mode	98 %	99.9 %	98 %	99.9 %	99.9 %
Read per run	35-75 thousand	up to 5 million	1 million	up to 3 billion	1.2 to 1.4 billion	N/A
Time per run	30 minutes to 2 hours	2 hours	24 hours	1 to 10 days (depending upon sequencer and specified read length)	1 to 2 week	2. minutes to 3 hours
Cost per 1 million bases	$2	$1	$10	$0.05 to $0.15	$0.13	$2400
Advantage	Longest read length. Fast, Detects 4mC, 5mC, 6mA	Less expensive equipment. Fast	Long read size. Fast	High throughput / cost	Low cost per base	Long individual read, useful for many application
Disadvantage	Low yield at high accuracy. Equipment can be very expensive.	Homopolymer errors.	Runs are expensive. Homopolymer errors.	Equipment can be very expensive.	Slower than other methods.	More expensive and impractical for larger sequencing projects.

VI. CONCLUSION

In this survey paper comparison different method of error correction in Next Generation Sequencing, also discuss different technique for error correction and Next Generation Sequencing. Comparative study is totally different technique shown in table one. There stay many further challenges in next-generation sequencing error correction. One challenge is to tell apart errors from polymorphisms, as an example, single nucleotide polymorphisms (SNPs).

VII. REFERENCES

[1] Sam Behjati, Patrick S Tarpey "What is next generation sequencing?" Behjati S, et al. Arch Dis Child Educ Pract Ed 2013;98:236–238.

[2]. Kao W-C, Chan AH, Song YS. ECHO: a reference-free short-read error correction algorithm. GenomeRes 2011;21: 1181–92.

[3] Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. Genome Biol 2010;11:R116.

[4] Qu W, Hashimoto S-I, Morishita S. Efficient frequency-based de novo short-read clustering for error trimming in next-generation sequencing. Genome Res 2009;19:1309–15.

[5] Salmela L. Correction of sequencing errors in a mixed set of reads. Bioinformatics 2010;26:1284–90.

[6] Salmela L, Schroder J. Correcting errors in short reads by multiple alignments. Bioinformatics 2011;27:1455–61.

[7] Chaisson MJ, Brinza D, Pevzner PA. De novo fragment assembly with short mate-paired reads: does the read length matter? Genome Res 2009;19:336–46.

[8] Chaisson M, Pevzner P, Tang H. Fragment assembly with short reads. Bioinformatics 2004;20:2067–74.

[9] Chin FYL, Leung HCM, Li W-L, et al. Finding optimal threshold for correction error reads in DNA assembling. BMC Bioinformatics 2009;10(Suppl 1):S15.

[10] Ilie L, Fazayeli F, Ilie S. HiTEC: accurate error correction in high-throughput sequencing data. Bioinformatics 2011;27: 295–302.

[11] Schroder J, Schroder H, Puglisi SJ, et al. SHREC: a short-read error correction method. Bioinformatics 2009;25:2157–63.

[12] Wijaya E, Frith MC, Suzuki Y, et al. Recount: expectation maximization based error correction tool for next generation sequencing data. Genome Inform Int Conf Genome Inform 2009;23:189–201.

[13] Yang X, Aluru S, Dorman KS. Repeat-aware modeling and correction of short read errors. BMC Bioinformatics 2011;12(Suppl 1):S52.

[14] Yang X, Dorman KS, Aluru S. Reptile: representative tiling for short read error correction. Bioinformatics 2010;26:2526–33.

[15] Benjamin A Flusberg, Dale R Webster, "Direct detection of DNA methylation during single-molecule, real-time sequencing" 2010.

[16] Xiao Yang, "Reptile: representative tiling for short read error correction" 2010.

[17] David R Kelley "Quake: quality-aware detection and correction of sequencing errors" 2010.

[18] Leena Salmela, "Correction of sequencing errors in a mixed set of reads" 2010.

[19] Zhiheng Zhao, Jianping Yin, "PSAEC: An Improved Algorithm for Short Read Error Correction Using Partial Suffix Arrays" 2011.