View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

DOCTORAL THESIS ABSTRACT

Genomic variant-detection using long read sequencing and short read sequencing, a comparative analysis

March, 2021 (2021 年 3 月)

Ahmed Nabiel Alkanaq

Department of Human Genetics
Graduate School of Medicine, Yokohama City University

Doctoral Supervisor Adviser: Naomichi Matsumoto, Professor

Comparison of mitochondrial DNA variants detection using short- and long-read sequencing

https://doi.org/10.1038/s10038-019-0654-9

Introduction

The completion of human genome project in 2003 represents a major milestone in understanding the human genome; however, it also discovered that the genome is more complex than we previously thought. To the day of this writing, it has been more than 17 years since the completion of the project, yet the human genome is still not complete.

The later advent of the next-generation sequencing of short-reads and fourth-generation technologies, represented by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), represents revolutionary leaps in genomics.

It is estimated that microsatellites make up to 3% of the human genome (International Human Genome Sequencing Consortium 2001), occurring every 10,000 nucleotides (Edwards et al. 1991, Collins et al. 2003, Subramanian et al. 2003). STRs are very liable for mutations with strand-slippage replication as the predominant mechanism for their polymorphism. It is estimated that the mutation rate for STRs is between 10^{-6} and 10^{-2} nucleotide per generation, higher than SNVs mutations which are at a rate of 10^{-8} nucleotide per generation (Ellegren 2000).

For an integrative and comprehensive understanding of the genome, the main goals of our research are:

A. Identifying the best tools available for accurately detecting single nucleotide variants, using the more accurate short-read sequencing technology, and building an effective pipeline of these tools.

B. Comparing the accuracy and efficiency of both sequencing technologies (short- and long-read sequencing) in determining short nucleotide variants (SNVs).

C. Finding the best methods for identifying short tandem repeat regions in the genome, utilizing the accuracy of short-read sequencing together with the coverage of STR regions using long-read data.

Keywords:

Next generations sequencing, fourth-generation sequencing, short-read sequencing, long-read sequencing, whole genome sequencing, mitochondrial DNA, short tandem repeats

Materials and Methods

A. Seven trio families of children diagnosed with *AiCardi syndrome* were analyzed for possible de novo SNVs. The same basic bioinformatic steps were used for the analysis of both WES and WGS data.

GATK and Bayesian probability calculation for variants were performed followed by a Gradient Boosting model. The machine learning model has been trained with over 50 true positive variants that were confirmed using Sanger sequencing and a cut-off value was selected.

B. Three samples of unrelated individuals that underwent both short- and long-read whole-genome sequencing analyses were selected. For each sample, DNA was extracted from peripheral blood lymphocytes. Short-read sequencing was done with illumina's HiSeqX10. For long-read data, PacBio's single pass subreads were generated through PacBio Sequel.

C. Following an extensive revision of all existing STR-detection software and algorithms, a proposed algorithm was constructed which is comprised of two phases: (1) Statistical analysis of STRs in healthy controls. (2) Profiling of the STRs with their related statistical attributes, and the resulting profiles can be used to compare to affected cases of interest.

Results

A. accurate identification of *de novo* mutations in seven AiCardi cases was performed, and a comparison between the WES and WGS results was compared at the end of the multistep algorithm of the pipeline. While WES identifies more variants, they are determined to be mostly false *de novo* calls through manual checking of bam files. On the other hand, while WGS identifies a smaller number of variants, identified variants were confirmed to be true *de novo* variants.

B. Weighted kappa coefficients indicate that the levels of agreement between long- and short read mtDNA genotyping are “almost perfect” at full coverage. The average coverage of 37 is the required number of long-reads to achieve the short-read’s accuracy for SNV detection.

C. The following conclusion were reached before final algorithmic parameters can be accurately (1) Computational resources for STR can be very expensive. (2) Long-read, unlike short-read, require deeper coverage for elimination of random insertion-deletions before the full extent of an STR region can be accurately identified. (3) Based on the margin of error: to calculate the coefficients of comparing WGS STR variations of a case against normal controls, and to reach a total STR candidate number of 25, which is considered to be cost-effective for manual confirmation the total number of positive controls (having expanded STR) required is 127.

Discussion

With increased affordability and reliability of recent sequencing technologies, it becomes important to replace old WES with WGS; however, a parallel development of efficient bioinformatics algorithms is essential.

Investigating SNV identification in mtDNA provided insight into the different behaviors of long- and short-read sequencing data, and despite the limited number of cases being studied, the downsampling analyses provided a clearer picture on how the same sample can have erroneous variant calls due to systemic or random errors.

The algorithm for STR identification represents the foundation upon which more complex genomic variations can be reliably identified like copy number variations and structural rearrangements.

References

Edwards, A., et al. (1991). DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. American Journal of Human Genetics, 49, 746–756.

Ellegren H (2000). Heterogeneous mutation processes in human microsatellite DNA sequences Nat. Genet., 24, 400-402

International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. Nature, 409, 860–921.

Publication List

1. Alkanaq AN, Hamanaka K, Sekiguchi F, Taguri M, Takata A, Miyake N, Miyatake S, Mizuguchi T, Matsumoto N:

   Comparison of Mitochondrial DNA Variants Detection Using Short- And Long-Read Sequencing,

   Journal of Human Genetics, volume 64, pages1107–1116 (2019)

2. Hamanaka K, Sugawara Y, Shimoji T, Nordtveit TI4, Kato M5, Nakashima M6, Saitsu H, Suzuki T, Yamakawa K, Aukrust I, Houge G, Mitsuhashi S, Takata A, Iwama K, Alkanaq A, Fujita A, Imagawa E, Mizuguchi T, Miyake N, Miyatake S, Matsumoto N:

   De novo truncating variants in PHF21A cause intellectual disability and craniofacial anomalies.

   European Journal of Human Genetics, volume 27, pages378–383(2019)

3. Hamanaka K, Miyatake S, Koshimizu E, Tsurusaki Y, Mitsuhashi S, Iwama K, Alkanaq AN, Fujita A, Imagawa E, Uchiyama Y, Tawara N, Ando Y, Misumi Y, Okubo M, Nakashima M, Mizuguchi T, Takata A, Miyake N, Saitsu H, Iida A, Nishino I, Matsumoto N :

   RNA sequencing solved the most common but unrecognized NEB pathogenic variant in Japanese nemaline myopathy.

   Genetics in Medicine volume 21, pages1629–1638(2019)

4. Hamanaka K, Takata A, Uchiyama Y, Miyatake S, Miyake N, Mitsuhashi S, Iwama K, Fujita A, Imagawa E, Alkanaq AN, Koshimizu E, Azuma Y, Nakashima M, Mizuguchi T, Saitsu H, Wada Y, Minami S, Katoh-Fukui Y, Masunaga Y, Fukami M, Hasegawa T, Ogata T, Matsumoto N:

   MYRF Haploinsufficiency Causes 46,XY and 46,XX Disorders of Sex Development: Bioinformatics Consideration

   Human Molecular Genetics, Volume 28, Issue 14, 15 July 2019, Pages 2319–2329