VNU Journal of Science: Comp. Science & Com. Eng., Vol. 30, No. 3 (2014) 31-35

Preliminary Results on the Whole Genome Analysis of a Vietnamese Individual

Dang Thanh Hai[1], Nguyen Dai Thanh[1], Pham Thi Minh Trang[1], Dang Cao Cuong[1], Hoang Kim Phuc[1], Son Bao Pham[1], Le Sy Vinh[1,*], Le Si Quang[2], Phan Thi Thu Hang[2], Do Duc Dong[3], Nguyen Huu Duc[4]

[1]*University of Engineering and Technology, Vietnam National University Hanoi*
[2]*Wellcome Trust Center for Human Genetics, Oxford University, UK*
[3]*Institute of Information Technology, Vietnam National University Hanoi*
[4]*High Performance Computing Center, Hanoi University of Science and Technology*

Abstract

We present preliminary results on the whole genome analysis of an anonymous Vietnamese individual of the Kinh ethnic group (KHV) that was deeply sequenced to 30-fold using the Illumina sequencing machines. The sequenced genome covered 99.8% of the human reference genome (GRCh37). We discovered (1) 3.4 million single polymorphism nucleotides (SNPs) of which 41,396 (1.2%) were novel, (2) 654 thousand short indels of which 35,263 (5.4%) were novel (i.e., not present in the dbSNP and the 1000 genomes project databases). We also detected 10,611 large structural variants (length ≥ 100 bp). This study is our initial step toward large-scale genome projects on Vietnamese population.

Manuscript communication: Received 18 February 2014, revised 25 March 2014, accepted 27 March 2014
Corresponding author: Le Sy Vinh, vinhls@vnu.edu.vn

Keywords: High coverage whole genome sequencing, Variant analysis, Vietnamese human genome.

1. Introduction

The emerging advances of the next generation sequencing (NGS) technologies today have allowed the conduction of a variety of large-scale sequencing projects, such as the 1000 genomes project [1, 2, 3], the 750 Netherlands genomes [4] or the 100 southeast Asian Malays genomes [5]. In addition, due to the low sequencing cost, a number of studies were provoked to sequence individuals at high coverage levels from diverge populations such as Han Chinese [6], Indian [7], Korean [8], Japanese [9], Pakistani [10], Turkish [11] and Russian [12].

Those sequencing efforts for Han Chinese, Japanese, Korean, Malaysian, Pakistani and Indian detected millions of genetic variants, of which an appreciable fraction was population specific i.e., not present in the dpSNP [13] or the 1000 genomes project (1KGP) database. Vietnam with approximate 90 million people of 54 different ethnic groups is the 14th largest country by population in the world. Vietnam plays as an important place in human-being migration routines over thousands of years of history. The 1KGP was extended to sequence genomes of 100

Kinh Vietnamese at a low-coverage (4x). However, such low-coverage sequencing data generated by the 1000 genomes project might be biased toward the discovery of high frequency or common variants. These facts created the impetus for our comprehensive genome-wide study of a Kinh Vietnamese (KHV) individual whose genome was sequenced at a high coverage level (~30x) by the Illumina HiSeq 2000 machine.

We detected an appreciably large number of KHV specific genetic variations (including SNPs, short indels, and structural variations). It indicated the necessity to conduct further large-scale genome-wide studies on not only Kinh group but also other Vietnamese ethic groups to provide a better and more complete picture of Vietnamese human genome variations.

2. Materials and methods

2.1. Data production

The genome of an anonymous male Kinh Vietnamese individual without any obvious genetic disorders was deeply sequenced at 30-fold average coverage by Illumina HiSeq 2000 machine (Illumina Inc.,) at the BGI-Hongkong using two paired-end libraries with the insert size of 500 base pairs and the read length of 100 base pairs. The donor is of the Kinh Vietnamese ethnic group for at least 3 generations. The donor gave written consent for public release of the genomic data for scientific research use.

2.2. Methods

BWA [14] was used to map short reads into the reference genome (GRCh37). To identify SNPs and short indels, we used GATK toolkit from the Boad Institute [15, 16], and followed the recommended best practice workflow.

We compared the detected variants with the dbSNP (Build 138, [13]) and the 1000 genomes project database. The Breakdancer tool (version 1.4.4, [17]) was used with default parameters for calling structural variants from high quality (Phred-score mapping quality $\geq$ 20) mapped paired-end reads. The DGV database of human genomic structural variations (version released on 2013-07-23, [18]) was used to assess the novelty of these predicted structural variants.

3. Results

We obtained 578 million paired-end reads of 100 base pair length of which 98% reads had the quality greater than or equal to 20 (see Figure 1). Most of the reads (99.99%) were mapped to the reference genome and 99.8% of the reference genome (excluding undetermined nucleotides Ns) was covered by at least one read. The mapping quality was high, i.e., 93.8% of reads had the mapping quality score greater than or equal to 20. In total, the average coverage of short reads sequenced from the KHV genome against the reference genome is about 30x and similar for all chromosomes.

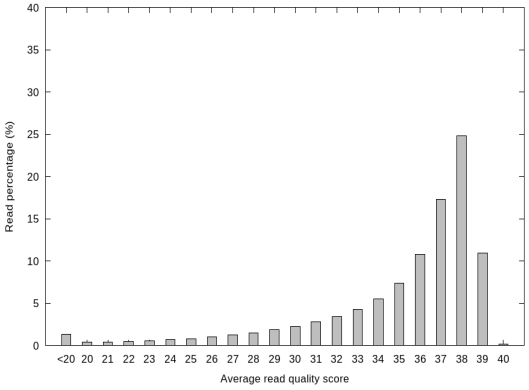

Figure 1. The quality of short reads.

3.1. SNP calling

We identified 3.4 million SNPs (quality score >= 20; depth coverage >= 4, filter = PASS). This number is similar to those reported in other previous genome-wide studies such as 3.1 million SNPs in the first Japanese individual genome [9] and 3.4 millions SNPs in the first Korean genome [8]. There were 41,396 (1.2%) SNPs that were not present in the dbSNP database version 138

(the most comprehensive catalogue of known SNPs from other large-scale genome-wide studies [13]). These were considered as KHV specific SNPs. The number of KHV novel SNPs is smaller than those detected in Ahn et al. (2009) [8] and Fujimoto et al. (2010) [9] because we compared against the latest version (138) of the dbSNP database. 295 of such novel SNPS were located in the coding exon regions of which 98 SNPs are synonymous and 197 are non-synonymous substitutions.

3.2. Indel calling

We identified 654,024 short indels of which 316,802 were insertions while 337,222 were deletions. These numbers are comparable with those detected in the Turkish individual genome [11] and in the Shigemizu genome-wide study [19]. The lengths of these discovered indels were mainly from 1 to 6 (Fig. 2). The number of indels with the length of 1 base pair was 322,544 (54%). The longest insertion and deletion were of 160 bps and 255 bps, respectively. 291,822 (44.6%) of the detected indels were located within gene regions of which 287,678 (98.58%) were found in introns and 3,062 (1.05%) were in coding exons.

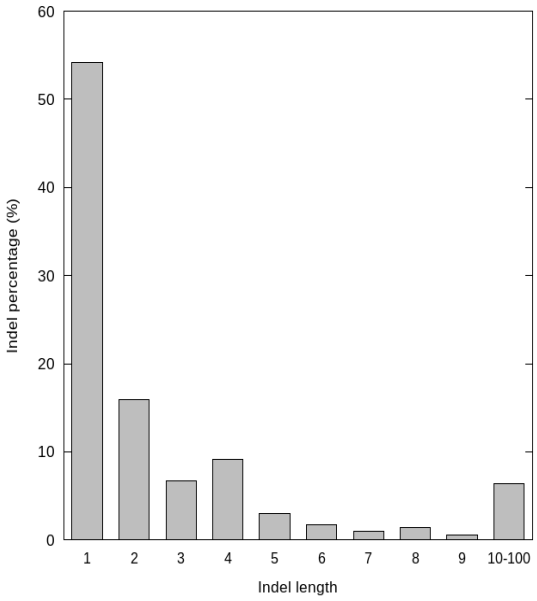

Figure 2. The length of indels detected in the KHV genome.

3.3. Structural variant calling

Mapped short reads with an average mapping quality greater than or equal to 20 were used for structural variant calling. As a result, 10,611 large SVs (length ≥ 100 bp) were identified. This number was similar to those in other previous individual genome-wide studies [6, 7, 10]. 9,617 (90.6%) out of these large SVs were large indels. The remaining of these large SVs included 331 (3.1%) inter-chromosomal translocations (CTX), 357 (3.4%) inversions (INV) and 306 (2.9%) intra-chromosomal translocations (ITX). Almost all of such large SVs in the KHV genome have the length in between 100 to 500 bps (see Fig. 3).

We compared 9277 large indels (5167 insertions and 4110 deletions) occurring on the same chromosome against the latest version (2013-07-23) of the Database of Genomic Variants (http://projects.tcag.ca/variation/). We found that 1925 insertions and 3978 deletions were present in the DGV database. The remaining 3374 large indels were considered as KHV novel large indels. These novel large indels included 3242 insertions and 132 deletions.

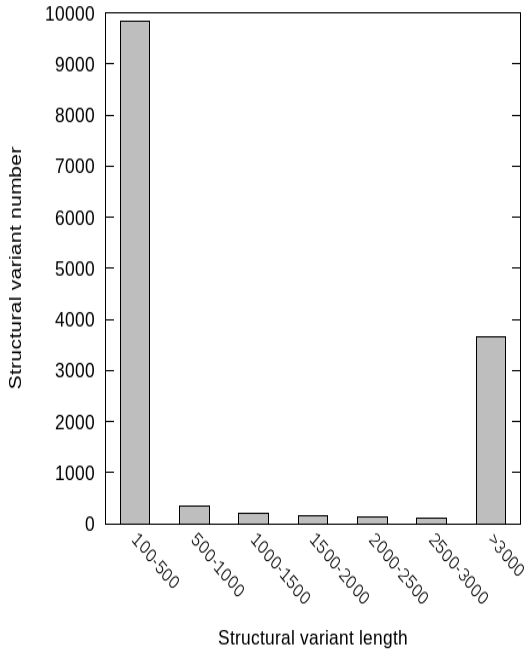

Figure 3. The length of structural variations detected in the KHV genome.

3.4. Conclusion

We have presented the whole genome-wide study of a Vietnamese individual sequenced at a high coverage level (30x). The obtained short reads were of high quality and covered up to 99.8% of the NCBI reference human genome. A substantial number of novel variants including SNPs, indels and large structural variants were detected specific for the Vietnamese individual. These potentially novel findings were demonstrated to associate with known gene functional regions, especially coding-exon regions.

There were 0.01% short reads that were not mapped to the reference genome. These unmapped reads could probably be a valuable genetic source on which we carry out further studies to discover more KHV-specific genetic variants. The study could therefore play an important reference for further large-scale genome-wide studies on Vietnamese population, and hence the development of personalized medicine for Vietnamese people in the near future. It is no doubt that our preliminary results presented here can be refined with the Mendelian law when we have genomes sequenced from a trio (parent, mother and child).

Acknowledgment

We would like to express our special thanks to Prof. Nguyen Huu Duc from Vietnam National University, Hanoi for his continuous encouragements and supports. We thank prof. Jean Daniel Zucker, Dr. Zamin Iqbal and prof. Arndt von Haeseler for their comments on our manuscript. This work was partly financially supported by the Science and Technology Foundation of Vietnam National University, Hanoi.

References

[1] N. Siva, (2008). 1000 Genomes project. Nature biotechnology, 26(3), 256-256.

[2] 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. Nature, 467(7319), 1061-1073.

[3] 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature, 491(7422), 56-65.

[4] D. I. Boomsma, C. Wijmenga, E. P. Slagboom, M. A. Swertz, L. C. Karssen, A. Abdellaoui, & P. I. de Bakker (2013). The Genome of the Netherlands: design, and project goals. European Journal of Human Genetics.

[5] L. P. Wong, R. T. H. Ong, W. T. Poh, X. Liu, P. Chen, R. Li, ... & Y. Y. Teo (2013). Deep Whole-Genome Sequencing of 100 Southeast Asian Malays. The American Journal of Human Genetics.

[6] J. Wang, W. Wang, R. Li, Y. Li, G. Tian, L. Goodman, ... & J. Ye (2008). The diploid genome sequence of an Asian individual. Nature, 456(7218), 60-65.

[7] B. J. Hardy, B. Séguin, P. A. Singer, M. Mukerji, S. K. Brahmachari & A. S. Daar (2008). From diversity to delivery: the case of the Indian Genome Variation initiative. Nature Reviews Genetics, 9, S9-S14.

[8] S.M. Ahn, T.H. Kim, S. Lee, D. Kim, H. Ghang, D.S. Kim, B.C. Kim, S.Y. Kim, W.Y. Kim, C. Kim, et al. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res. 2009;19:1622-1629.

[9] A. Fujimoto, H. Nakagawa, N. Hosono, K. Nakano, T. Abe, K. A. Boroevich, M. Nagasaki, R. Yamaguchi, T. Shibuya, M. Kubo, et al. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nat. Genet. 2010;42:931-936.

[10] M. K. Azim, C. Yang, Z. Yan, M. I. Choudhary, A. Khan, X. Sun, ... & Y. Zhang (2013). Complete genome sequencing and variant analysis of a Pakistani individual. Journal of human genetics, 58(9), 622-626.

[11] H. Dogan, H. Can and H. H. Otu (2014). Whole Genome Sequence of a Turkish Individual. PloS one, 9(1), e85233.

[12] K. G. Skryabin, E. B. Prokhortchouk, A. M. Mazur, E. S. Boulygina, S. V. Tsygankova, A. V. Nedoluzhko, ... & M. V. Kovalchuk (2009). Combining two technologies for full genome sequencing of human. Acta naturae, 1(3), 102.

[13] S. T. Sherry, M. H. Ward, M. K. holodov, J. Baker, L. Phan, E. M. Smigielski, & K. Sirotkin (2001). dbSNP: the NCBI database of genetic variation. Nucleic acids research, 29(1), 308-311.

[14] H. Li and R. Durbin (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.

[15] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297-303.

[16] M. DePristo, E. Banks, R. Poplin, K. Garimella, J. Maguire, C. Hartl, A. Philippakis, G. del Angel, M.A. Rivas, M. Hanna, A. McKenna, T. Fennell, A. Kernytsky, A. Sivachenko, K. Cibulskis, S. Gabriel, D. Altshuler and M. Daly (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics.* 43:491-498.

[17] K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, S. D. McGrath et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature methods 6, no. 9 (2009): 677-681.

[18] J. R. MacDonald, R. Ziman, R. K. Yuen, L. Feuk, S. W. Scherer (2013). The database of genomic variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2013 Oct 29. PubMed PMID: 24174537.

[19] D. Shigemizu, A. Fujimoto, S. Akiyama, T. Abe, K. Nakano, K. A. Boroevich, ... & T. Tsunoda (2013). A practical method to detect SNVs and indels from whole genome and exome sequencing data. Scientific reports, 3.