Whole genome analysis of a Vietnamese trio

Dang Thanh Hai[1], Nguyen Dai Thanh[1], Pham Thi Minh Trang[1], Le Si Quang[2,*], Phan Thi Thu Hang[2], Dang Cao Cuong[1], Hoang Kim Phuc[1], Nguyen Huu Duc[3], Do Duc Dong[4], Bui Quang Minh[5], Pham Bao Son[1] and Le Sy Vinh[1,4,*]

[1]*University of Engineering and Technology, Vietnam National University Hanoi, Hanoi, Vietnam*

[2]*Wellcome Trust Center for Human Genetics, Oxford University, Oxford, UK*

[3]*High Performance Computing Center, Hanoi University of Science and Technology, Hanoi, Vietnam*

[4]*Information Technology Institute, Vietnam National University Hanoi, Hanoi, Vietnam*

[5]*Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, Vienna, Austria*

**Corresponding authors (Emails, LSQ – quang@well.ox.ac.uk; LSV – vinhls@vnu.edu.vn)*

We here present the first whole genome analysis of an anonymous Kinh Vietnamese (KHV) trio whose genomes were deeply sequenced to 30-fold average coverage. The resulting short reads covered 99.91% of the human reference genome (GRCh37d5). We identified 4,719,412 SNPs and 827,385 short indels that satisfied the Mendelian inheritance law. Among them, 109,914 (2.3%) SNPs and 59,119 (7.1%) short indels were novel. We also detected 30,171 structural variants of which 27,604 (91.5%) were large indels. There were 6,681 large indels in the range 0.1–100 kbp occurring in the child genome that were also confirmed in either the father or mother genome. We compared these large indels against the DGV database and found that 1,499 (22.44%) were KHV specific. *De novo* assembly of high-quality unmapped reads yielded 789 contigs with the length ≥300 bp. There were 235 contigs from the child genome of which 199 (84.7%) were significantly matched with at least one contig from the father or mother genome. Blasting these 199 contigs against other alternative human genomes revealed 4 novel contigs. The novel variants identified from our study demonstrated the necessity of conducting more genome-wide studies not only for Kinh but also for other ethnic groups in Vietnam.

[Hai DT, Thanh ND, Trang PTM, Quang LS, Hang PTT, Cuong DC, Phuc HK, Duc NH, Dong DD, Minh BQ, Son PB and Vinh LS 2015 Whole genome analysis of a Vietnamese trio. *J. Biosci.* **40** 113–124] **DOI** 10.1007/s12038-015-9501-0

1. Introduction

The advent of the next-generation sequencing technology (NGS) has led to an era of personal genomics (Shendure and Ji 2008; von Bubnoff 2008; 1000 Genome Project Consortium 2010; Drmanac 2011). Today a human genome can be sequenced within a week for a cost of around 10,000 USD. This is an astonishing achievement in comparison with the 3 billion USD and 15 years needed to complete the first draft of the human genome (Lander *et al.* 2001; Venter *et al.* 2001; Consortium I.H.G.S. 2004).

A number of large-scale sequencing projects have been conducted, such as the 1000 Genomes Project (Siva 2008; 1000 Genome Project Consortium 2012), the 750

Keywords. Genomic variant analysis; Vietnamese human genome; Whole genome sequencing data analysis

Supplementary materials pertaining to this article are available on the *Journal of Biosciences* Website at *http://www.ias.ac.in/jbiosci/mar2015/supp/Hai.pdf*

Published online: 4 February 2015

Netherlands genomes (Boomsma *et al.* 2014) or the 100 Southeast Asian Malays genomes (Wong *et al.* 2013). Besides, a number of individual human genomes have been sequenced at a high coverage level, such as the Han Chinese genome (Wang *et al.* 2008), Indian genome (Hardy *et al.* 2008), Korean genome (Ahn *et al.* 2009), Japanese genome (Fujimoto *et al.* 2010), Pakistani genome (Azim *et al.* 2013), Turkish genome (Dogan *et al.* 2014) and Russian genome (Skryabin *et al.* 2009).

Being the 14th largest country by population in the world, Vietnam has about 90 million people of 54 different ethnic groups of which more than 80% are Kinh. The 1000 Genomes Project (*http://www.1000genomes.org*) was extended to sequence a number of Vietnamese individual genomes at low coverage. However, such low-coverage sequencing data generated by the 1000 Genomes Project might be biased toward the discovery of high frequency or common variants (Wong *et al.* 2013). A large number of novel variations detected from high-coverage sequencing efforts (Han Chinese, Japanese, Korean, Malaysian, Pakistani, Indian and Turkish) have demonstrated the necessity to deeply sequence more individuals from diverse populations to provide a better and more complete picture of human genome variations.

In this study, for the first time we comprehensively analysed whole genomes of a Kinh Vietnamese (KHV) trio (father, mother and son). The genomes were sequenced to 30-fold average coverage by the Illumina HiSeq 2000 machine. The pedigree information allowed us to verify the detected variants using the Mendelian inheritance law. We used standard methods, software and pipelines to analyse the sequenced genomes. Our study revealed a large number of KHV-specific variants including SNPs, short indels, structural variants and novel contigs. The novel variants and contigs found here suggested that it is necessary to conduct further genome-wide studies not only for the Kinh but also for other ethnic groups to complete the picture of human genome variations for Vietnam.

2. Results

2.1 *Data analysis*

The raw reads were first cleaned by removing the adapter reads, the low-quality reads and the reads with more than 10% of unknown bases. We obtained 578 million (562 million and 493 million) clean paired-end reads of 100 base pair length from the son genome (father genome and mother genome, respectively). Most of the short reads have a high base quality, i.e. ~98% with Phred-score ≥ 20 (supplementary figure 1). Over 99.9% of the short reads were mapped to the NCBI reference genome build 37 (GRCh37d5) with a high mapping quality (~94% with Phred-score ≥ 20). We found that 99.91% of the reference genome (excluding undetermined nucleotides Ns) was covered by at least one read from these genomes. The average coverage of the mapped reads was about 30-fold and uniform across all chromosomes (see supplementary figure 2 for more details). The insert size distributions of paired-end short reads from the child, father and mother genomes are shown in supplementary figure 3. The means (standard deviations) of the insert size distributions in the child, father and mother genomes are 471 (19), 484 (18) and 471 (23), respectively. They are compatible with the expected insert size (500 bps) of the paired-end libraries prepared for deep whole-genome sequencing of the KHV trio.

2.2 *SNPs analysis*

We identified 4,823,475 single nucleotide polymorphism (SNPs) in KHV trio genomes, of which 3,667,344 (3,689,555 and 3,677,721) SNPs are in the child genome (father genome and mother genome, respectively). Over 2.3 millions SNPs are shared among three genomes (figure 1). The Ti/Tv ratios are 2.063, 2.064 and 2.063 in the child, father and mother genomes, respectively. The number of detected SNPs in each genome is comparable to those reported in other individual genome-wide studies such as 3,132,608 SNPs in the first Japanese individual genome (Fujimoto *et al.* 2010) and 3,439,107 SNPs in the first Korean genome (Ahn *et al.* 2009). Of the KHV SNPs, 4,728,141 (98.02%) were eligible for Mendelian validation (see Materials and methods section). We found that 4,719,412 (99.82%) SNPs fulfill the Mendelian law while 8,729 (0.18%) SNPs violated the law. This hints that the false positive rate of SNP calls is approximately 0.18%. These Mendelian-compatible SNPs are used for downstream analyses. Table 1 shows the genotype distribution of Mendelian-compatible and Mendelian-violated SNPs.

Functional region annotation revealed that there were 1,112,189 (23.6%) SNPs in introns, 789 (0.02%) SNPs in 5′-UTRs, 4,481 (0.09%) SNPs in 3′-UTRs and 29,647 SNPs in coding regions (22,209, 22,405 and 22,246 in the father, mother and child genomes, respectively). These numbers are similar to those reported by the 1000 Genomes Project. Among 29,647 SNPs in the coding regions, 15,039 SNPs are synonymous and the remaining 14,608 SNPs are non-synonymous (see figure 2 for further details). SNPeff classified 19,878 SNPs in the KHV trio as functional SNPs (i.e. non-synonymous, 5′-UTR, 3′-UTR SNPs), of which 14,980, 14,956 and 14,924 are in the father, mother and child genomes, respectively (see figure 2 for further details). The number of functional SNPs in each KHV individuals is compatible with those reported in a large-scale exome study (Tennessen *et al.* 2012).

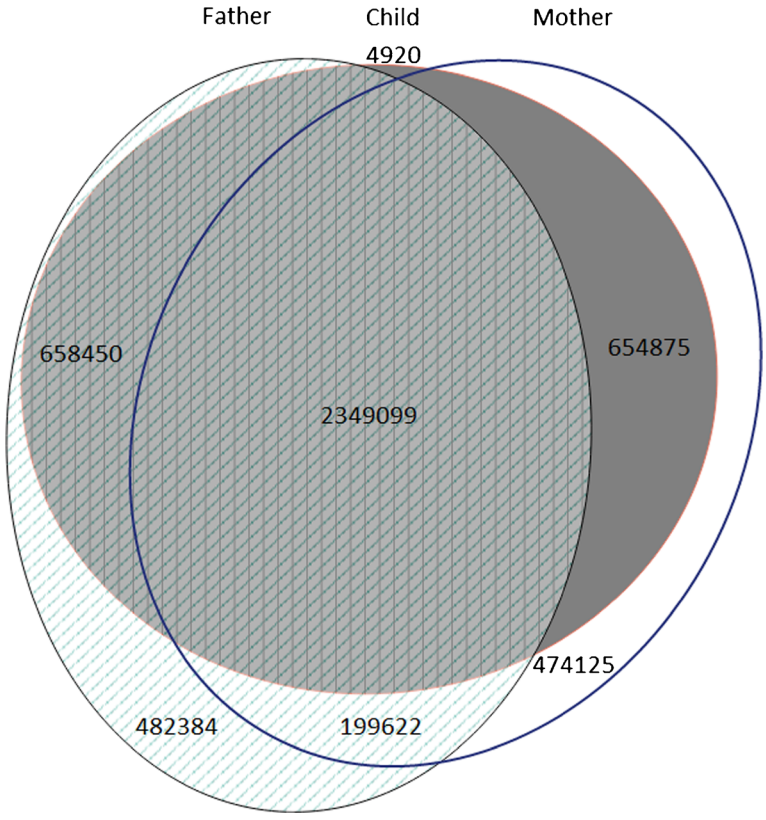

Figure 1. SNP distribution in child, father and mother genomes.

We compared Mendelian-compatible SNPs with the dbSNP (Build 138; Sherry *et al.* 2001) and the 1000 Genomes Project database (2012 release). Note that variant calls of Vietnamese individuals in the 1000 Genomes Project have not been available in this release. There are 109,914 (2.3%) novel or KHV-specific SNPs, i.e. those that were not present in either dbSNP or the 1000 Genomes Project database. These SNPs were categorized into 5′-UTR (25 SNPs), 3′-UTR (112 SNPs), introns (25,749 SNPs) and coding regions (273 synonymous substitutions and 535 non-synonymous substitutions) (see figure 3 for further details). Further analysis for these novel SNPs might reveal specific characteristics of the Kinh trio.

2.3 *SNPs shared between KHV trio genomes and other populations*

We compared SNPs in the KHV trio with SNPs in other populations. To this end, we downloaded all SNPs in 1,092

Table 1. Mendelian analysis of KHV trio-variants

		Mother								
		HOM ref	HET ref	HOM mut	HOM ref	HET ref	HOM mut	HOM ref	HET ref	HOM mut
Father	HOM ref	0%	40.91%	0.03%	0.09%	22.23%	8.27%	0%	0.01%	0.02%
	HET ref	42.08%	16.91%	0.01%	22.91%	18.78%	9.38%	0.02%	11.84%	13.17%
	HOM mut	0.04%	0.02%	0%	8.83%	9.51%	0.01%	0.02%	13.02%	61.90%
		Child: HOM ref			Child: HET ref			Child: HOM mut		

'HOM ref' means homozygotes where both alleles are identical to the reference; 'HOM mut' means homozygotes where both alleles differ from the reference; 'HET ref means heterozygotes where only one allele is identical to the reference. The cells in gray indicate the Mendelian-compatible SNPs.

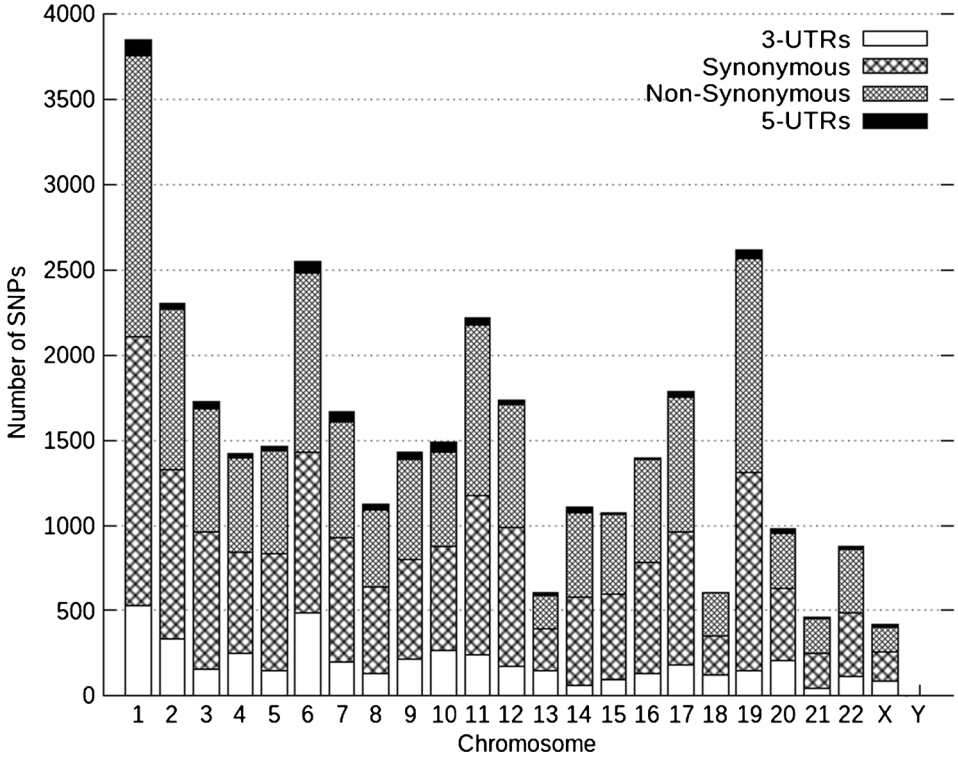

Figure 2. Functional regions of all Mendelian-supported SNPs in the KHV trio across chromosomes.

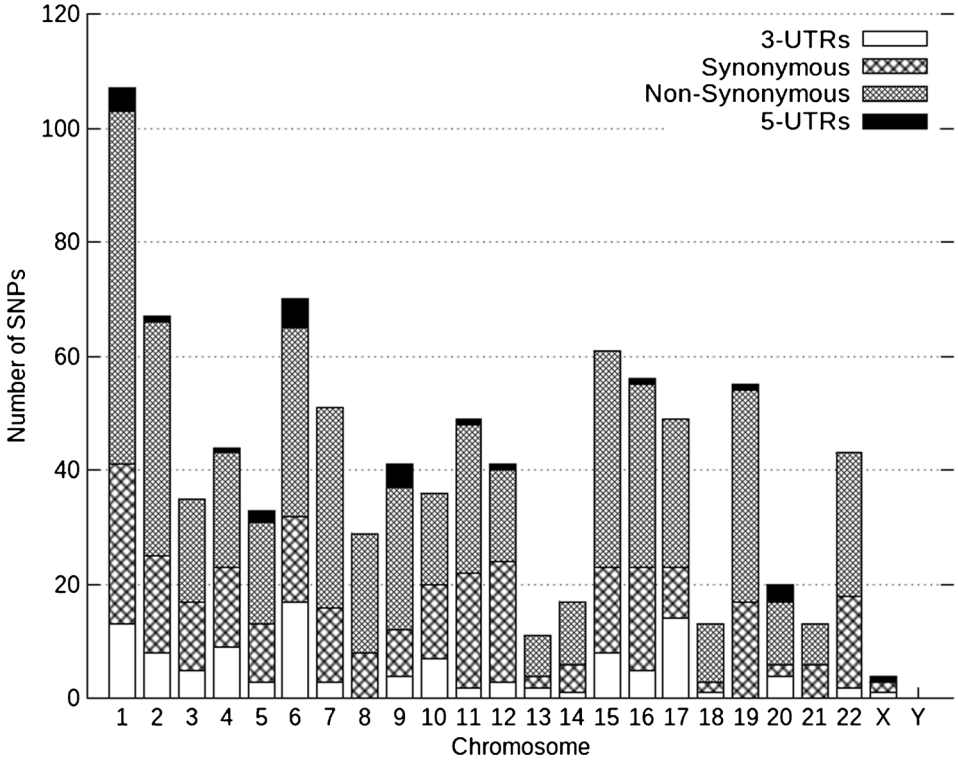

Figure 3. Functional regions of KHV-specific (novel) SNPs in the KHV trio across chromosomes.

human genomes from the 1000 Genomes Project database (the variants of Vietnamese individuals in the 1000 Genomes Project have not been released). From this, we extracted four population-specific SNP subsets corresponding to the Chinese (2,346,268), Japanese (1,128,438), African (3,206,983) and European (9,057,610) populations, respectively. A specific subset of a population contains SNPs that are unique to that population, i.e. not present in other populations. We compared KHV SNPs with these specific subsets and found that 1% of the Chinese subset (0.15% of the Japanese subset, 0.02% of the European subset, and 0.02% of the African subset) shared similarities with 4,719,412 detected KHV SNPs.

2.4 *Short indel calling*

We identified 974,100 short indels (length ≤100bp) in the KHV trio genomes consisting of 465,609 insertions and 508,491 deletions. There are 774,499 indels (375,561 insertions and 398,938 deletions) in the child genome; 763,403 (371,308 insertions and 392,095 deletions) in the father genome and 767,361 (372,316 insertions and 395,045 deletions) in the mother genome (supplementary figure 4). These numbers are similar with those reported in recent individual human genome-wide studies (Dogan *et al.* 2014; Shigemizu *et al.* 2013). Among detected short indels, 834,623 (85.68%) are eligible for Mendelian validation (see Materials and methods section). We found that only 7,238 (0.87%) short indels violate the Mendelian law. The remaining 99.13% of Mendelian-compatible short indels were then used for further analyses. Over 90% of short indels have the length from 1 to 9 bp (figure 4).

Functional region annotation of short indels indicated that there are 203,212 (24.5%) in introns, 4,637 (0.6%) in coding regions, 90 (0.01%) in 5′-UTRs and 927 (0.1%) in 3′-UTRs. Figure 5 and supplementary figure 5 show the functional effect distribution for short indels across chromosomes. We compared the Mendelian-compatible indels with the 1000 Genomes Project database and found that 59,119 (7.15%) indels are novel or KHV specific.

2.5 *Structural variant calling*

All mapped reads with quality greater than or equal to 20 were used to identify large structural variants (length ≥100 bp). We identified 10,611 structural variants SVs in the child genome, 9,055 SVs in the father genome, and 10,505 SVs in the mother genome (table 2). Almost all of the SVs (>90%) are large indels (supplementary figure 6). A large indel was considered as a 'Mendelian-supported' indel if it occured in the child genome and in either the father or the mother genome. In this study, we focused on analysing Mendelian-supported large indels. There were 6,681 Mendelian-supported large indels in the range of 0.1–100 kbp consisting of 2,855 insertions and 3,826 deletions. Most of these large indels have length ranging between 100 to 500 bp and there are no insertions longer than 500 bp (figure 6).

Functional region annotation of Mendelian-supported indels using the refGene database (*http://www.ncbi.nlm.nih.gov/refseq/*) indicated that 990 (14.8%) large indels overlap at least 1% with 1004 genes; and 227 (3.4%) large indels overlap at least 1% with 306 coding exons of 219 genes.

We compared the 6,681 Mendelian-supported indels with the curated structural variation DGV database version 2013-07-23 (*http://projects.tcag.ca/variation/*) and found that 5,182 are present in the DGV database. Thus, the 1,499 remaining Mendelian-supported large indels (1387 insertions and 112 deletions) are considered as KHV large novel indels.

2.6 *De novo assembly of unmapped reads*

Unmapped high-quality reads (Phred-score read quality ≥20) were used for *de novo* assembly of contigs using Velvet *de novo* assembler tool (version 1.2.10; Zerbino and Birney 2008). We obtained 235, 279, 275 contigs with the length ≥300 bp from the child, father and mother genome, respectively (table 3 and supplementary figure 7). We used the Blast program to align the contigs from the child genome against those from the mother and father genomes. A contig from the child genome was considered as a 'Mendelian-supported' contig if it could be aligned with at least one contig from either the mother or the father genome with significance. In this study, we focused on analysing these Mendelian-supported contigs.

There were 199 Mendelian-supported contigs with the average length of 583 bp. Most contigs had length from 300 bp to 500 bp (figure 7). We conducted Blast searches of these contigs against alternative human genome assemblies (HuRef, YH, WGSA, GRCh37) and the chimpanzee genome. A large number of those 199 contigs are aligned with significance with these examined genomes, e.g. 140 contigs were aligned with the HuRef genome (see table 4 for further details). Four out of 199 Mendelian-supported contigs did not yield significant alignment with any examined alternative human genomes or the chimpanzee genome. Their lengths are 322, 405, 488 and 1161 bp. As these 4 contigs were assembled from high-quality reads and supported by the Mendelian inheritance law, they are therefore considered as KHV novel contigs.

2.7 *Functional analysis of SNPs*

We conducted functional analysis of 14,608 non-synonymous KHV SNPs. The SIFT program (Kumar *et al.*

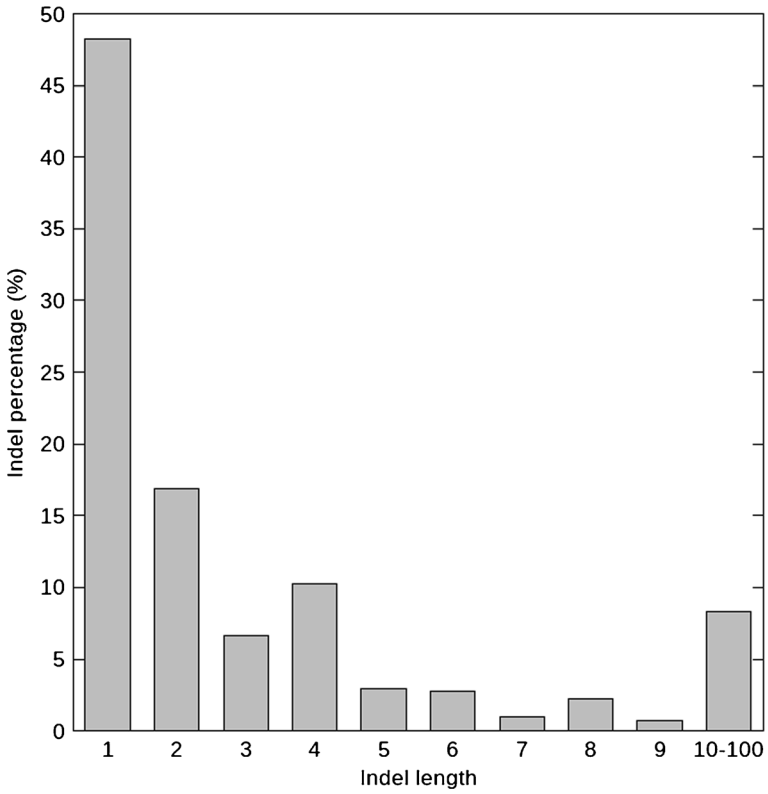

Figure 4. The length (the number of nucleotides) distribution of Medenlian-supported indels in the KHV trio.

2009) predicted that 2,943 (20.15%) SNPs are potentially damaging missense on 2,131 genes. Of these genes, 1,955 are associated with GO terms. The Gorilla tool (Eden et al. 2009) identified 20 enriched GO terms with the corrected *P*-value in range of 10e−4 to 10e−5 (figure 8) of which 'transcription, DNA-templated', 'RNA metabolic process', 'RNA biosynthetic process' and 'cellular nitrogen compound biosynthetic process' were the strongest enrichments. There are 12 genes (*ZNF19, ZNF708, ZNF705G, ZNF224, ZNF93, ZNF780A, ZNF28, ZNF124, ZNF530, ZNF443, ZKSCAN4* and *XRN1*) involved with all these 20 enriched GO terms. The first 11 genes are zinc finger protein family and involved in 12 out of all 20 enriched terms. The last gene *XRN1* involved the other 8 remaining terms. These genes together with related non-synonymous SNPs in the KHV trio are listed in supplementary table 1.

2.8 *Novel allelic genes in the KHV trio*

We followed the workflow described in the Cortex paper (Iqbal *et al.* 2012) to find novel allelic genes in the KHV trio. We assembled all three (trio) genomes independently using the Cortex de novo assembler. Cortex reported 45,186 (43,921 and 44,503) novel contigs (i.e. contigs with the length ≥100 bp and <90% homology to the reference genome GRCh37d5) in the child genome (mother and father, respectively) among which 37,070 (82%) contigs were supported by the Mendelian inheritance law. To find novel allelic genes, these Mendelian-supported contigs were blasted against the RefSeq gene database, and alternative human genome assemblies (HuRef, YH, WGSA). We found 9 contigs that were aligned to 19 RefSeq genes but did not match (≥90% homology) to any alternative human assemblies (supplementary table 2). These 9 contigs are considered as novel allelic genes in the KHV trio. Note that these 9 contigs do not overlap with any novel contigs assembled from unmapped reads.

3. Materials and methods

We used standard and high-quality methods and software/pipelines that had been used in the 1000 Genomes Project and other human genome projects to analyse our KHV trio genomic data.

3.1 *Data production*

The genomic DNA used in this study was from an anonymous Kinh Vietnamese (KHV) trio (father, mother and son)

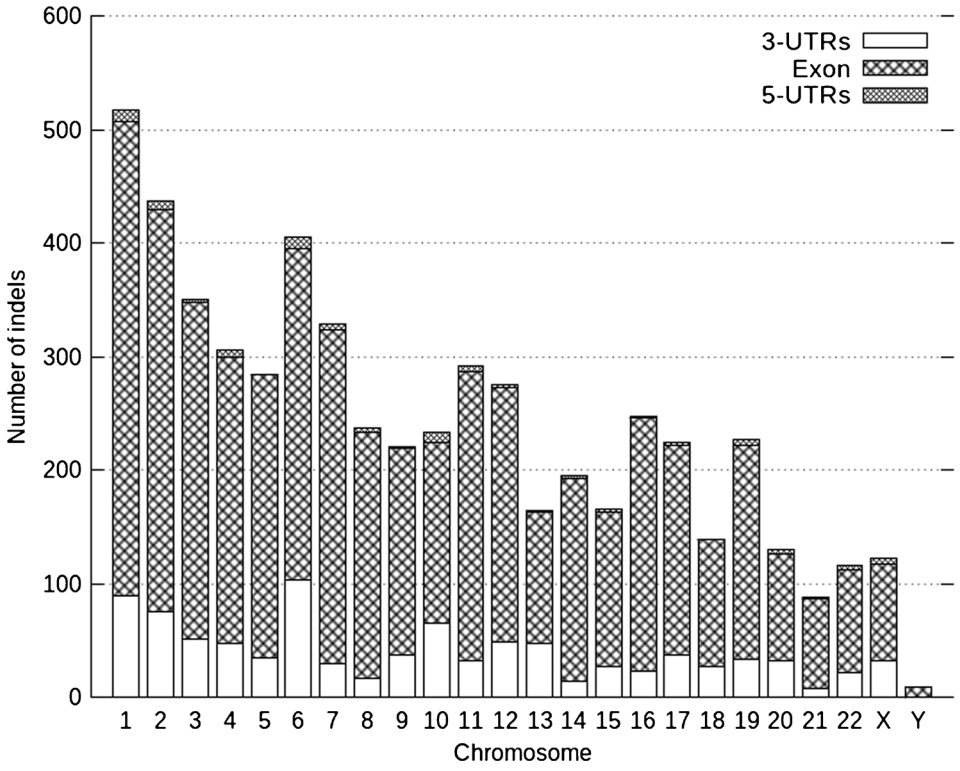

Figure 5. Functional regions of Mendelian-supported short indels in the KHV trio across chromosomes.

without any obviously known genetic disorders. The parents come from Kinh Vietnamese ethnicity for at least five generations (self-reported). The donors gave written consent for public release of the genomic data for the use in scientific researches.This study was approved by the Committee on Ethics in Research on Humans of School of Medicine and Pharmacy, Vietnam National University, Hanoi. The DNA quality, in terms of concentration determination and sample integrity, was tested using Qubit Fluorometer and Agarose GelElectrophoresis. Two paired-end libraries with the insert size of 500 bp were prepared for deep whole-genome sequencing of KHV trio using Illumina HiSeq 2000 machine (Illumina Inc., San Diego, USA) at BGI-Hongkong. The paired-end reads of 100 bp length resulted in about 30-fold average coverage for each genome.

3.2 *Short read mapping*

We used BWA software (Li and Durbin 2009) to map short reads into the reference genome (GRCh37). The BWA software generated SAM files that were consequently converted to BAM files for further analyses. The quality and other statistics of short read mapping were reported using the Samtool (Li *et al.* 2009).

3.3 *SNPs and indel calling*

To identify SNPs and short indels, we used GATK toolkit from the Boad Institute (McKenna *et al.* 2010; DePristo *et al.* 2011), following the best practice workflow: Duplicate mark by Picard,

Table 2. Structural variants detected in the KHV trio genomes

	Indels	CTX	INV	ITX
Child	9617 (90.6%)	331 (3.1%)	357 (3.4%)	306 (2.9%)
Father	8216 (90.7%)	209 (2.3%)	320 (3.5%)	310 (3.4%)
Mother	9771 (93.01%)	168 (1.6%)	295 (2.8%)	271 (2.6%)

CTX is the inter-chromosomal translocation; INV is the inversion; ITX is the intra-chromosomal translocation.

Table 3. Statistics of assembled contigs (length ≥ 300 bp) in the KHV trio genomes

	Child	Father	Mother
The number of contigs	235	279	275
Max length (bp)	1710	2072	1611
Mean length	556.7	566.1	486.5
Total bases	131366	158516	134274
N50	601	613	486
The number of contigs in N50	75	87	96
The number of contigs > 1000 bp	18	21	10

local indel realignment, base quality score recalibration, raw variants (SNPs/short indels) calling, Fisher Exact Test to detect strand bias, and variants recalibration. The HaplotypecCaller (Unified Genotyper) was used to call variants on the autosomes (sex chromosomes). We denoted trio-variant being the variant on the KHV trio. We also denoted a child-variant, father-variant and mother-variant being a variant on the child genome, father genome and mother genome, respectively. A trio-variant was considered a 'good' variant and kept for further analyses if it had a quality score (QUAL) ≥ 30, a depth coverage (DP) ≥ 8, and passed the quality filter from GATK.

3.4 *Mendelian validation*

We used the Mendelian inheritance law for assessing the quality of the detected variants (SNPs/short indels). A variant in a genome was called a 'good' variant if it had an associated genotype quality (GQ) ≥ 30 and the depth coverage (DP) ≥ 4. A trio-variant was considered 'eligible for Mendelian validation' if the child-variant was good, and either the father-variant or the mother-variant was good. All good and 'eligible for Mendelian validation' trio-variants were verified with the Mendenlian law and consequently classified into either Mendelian-compatible or Mendelian-violated variants. Only Mendelian-compatible trio-variants were kept for downstream analyses.

3.5 *Functional region annotation and analysis*

Functional effects of Mendelian-compatible variants (SNPs and indels) were annotated with the SNPeff tool (version 3.5; Cingolani *et al.* 2012). Since SNPeff might return different effects for each variant, the strongest effect measured by the variantAnnotator (version 2.8.2, GATK toolkit) was assigned and considered as the effect of each variant.

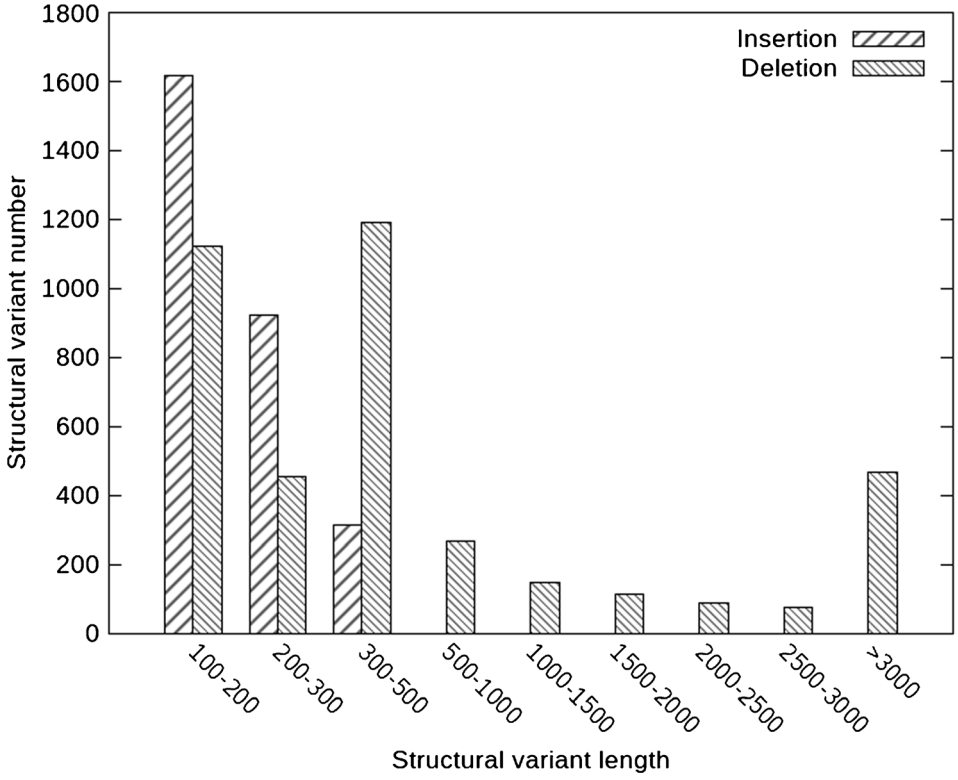

Figure 6. The length (the number of nucleotides) distribution of Mendelian-supported structural variants in the KHV trio.

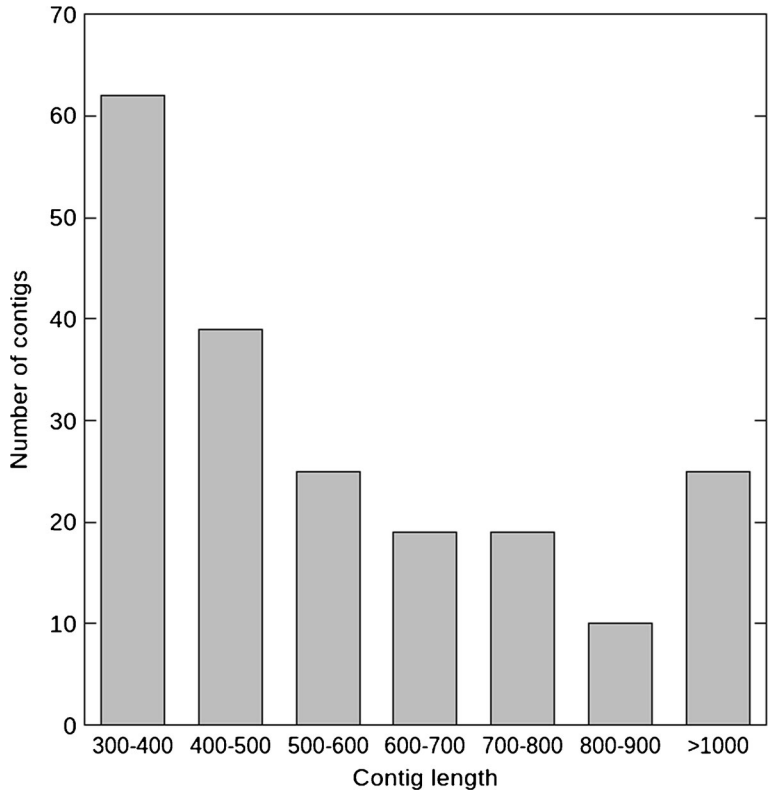

Figure 7. The length (the number of nucleotides) distribution for Mendelian-supported contigs in the KHV trio

The SIFT program (latest update on 04 February 2014; Kumar *et al.* 2009) was used to detect the damaging effects of missense mutations from non-synonymous SNPs. Genes annotated with damaging effects by SIFT were ranked according to the damaging score and then taken as input to the Gorilla program (latest update on 15 February 2014; Eden *et al.* 2009) for functional GO enrichment analysis.

3.6 *Structural variation calling*

The Breakdancer program (version 1.4.4, Chen *et al.* 2009) was used with default parameters for calling structural variants from high quality (Phred-score mapping quality ≥20) mapped paired-end reads. The DGV database of human genomic structural variants (version released on 23 July 2013 for the reference human genome GRCh37; MacDonald *et al.* 2014) was used to assess the novelty of predicted structural variants.

Table 4. Blast searches of Mendelian-supported contigs against alternative human genomes and the chimpanzee genome

Alternative genome	The number of aligned contigs	The number of hits
HuRef (Craig Venter)	140 (70.3%)	297
YH (Han Chinese)	175 (87.9%)	336
WGSA (Celera)	139 (69.8%)	283
GRCh37	61 (30.7%)	239
Chimpanzee genome	179 (89.9%)	351

3.7 *Contig assembly from unmapped reads*

We used the Velvet *de novo* assembler tool (version 1.2.10; Zerbino and Birney 2008) to assemble the unmapped reads into contigs. The Blast program (Altschul *et al.* 1997) was used with default settings (expectation value = 10) to compare the assembled contigs against alternative human genomes (Venter, YH, WGSA, GRCh37) and the chimpanzee genome (release 2.1.4). A contig was considered as a KHV novel contig if it was not aligned with any examined genomes.

4. Discussion

The short reads had high quality and covered almost all (~99.91%) positions of the human reference genome. A large number of variants (SNPs, short indels, structural

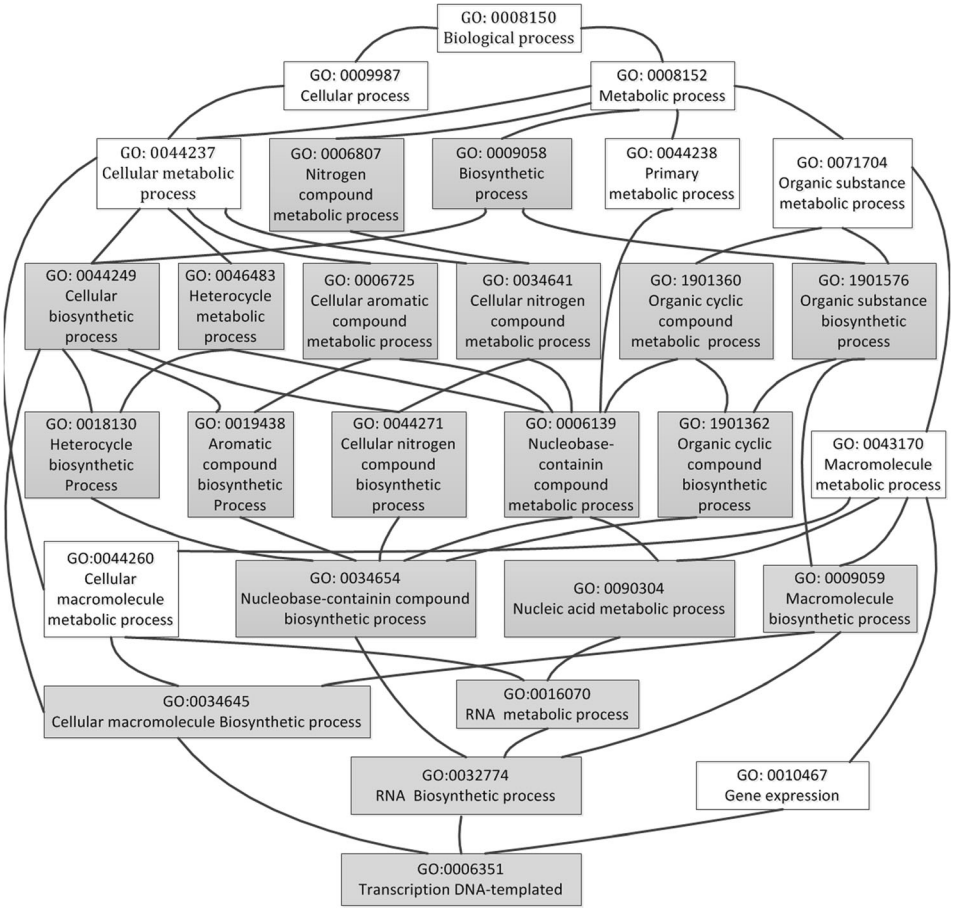

Figure 8. GO graph of significantly enriched GO terms (highlighted) with the corrected *P*-value < 0.001 for missense SNPs in the KHV trio genomes. The corrected *P*-value was calculated by the Gorilla program for multiple testing using the Benjamini and Hochberg method.

variants and assembled contigs) were identified. These findings were similar to those reported in other previous genome-wide studies for individuals from different populations.

We kept all KHV trio-variants that followed the Mendelian inheritance law for further downstream analyses, i.e. 4,719,412 (99.82%) SNPs and 827,385 (99.13%) short indels. This strategy guaranteed the high quality of called variants. The chromosome Y in the father genome was almost identical to that in the son genome. The results demonstrated the paternity and maternity among three KHV individuals in this study.

We compared the variants in the KHV trio with the recently released 1000 genomes genotype calls (2014 release) and found that 73,845 SNPs and 47,070 short indels detected in the KHV trio are novel. Note that 524,165 out of 827,385 Mendelian-supported indels in the KHV trio are in repeat regions. These indels might be the result of mapping artefact and would deserve additional analyses in the future.

Our results revealed that there is an appreciably large number of novel variants including SNPs, short indels and large structural variants. A small number of novel SNPs are non-synonymous substitutions associated with some enriched GO terms.

The comparison between KHV SNPs with those in other populations confirmed a closer relationship between the KHV trio and the Asian populations (including Chinese and Japanese) than the African and European peoples. Within Asian people, the KHV trio showed more genetic variants in common with Chinese people than with Japanese people. Interestingly, we found that the KHV trio were equidistant to African and European peoples.

A number of whole genome studies on trios have been conducted to utilize the pedigree information in genomic trio data. The first group of such studies on trios focus on targeted sequencing of trios associated with specific genetic diseases/risks (Roach *et al.* 2010; He *et al.* 2014). These studies made use of the pedigree information to filter out variants that are inconsistent with the Mendel's laws of inheritance. Roach *et al.* (2010) have shown that the pedigree information helped them to identify a smaller number of potential causal genes associated with autosomal recessive Miller syndrome. Interestingly, Roach

and colleagues incorporated inheritance patterns with the transmission disequilibrium test (TDT) framework (Spielman *et al.* 1993) to identify new and rare autism gene candidates (He *et al.* 2014).

The second group of studies on trios is typically conducted in the pilot phase of large-scale projects where researchers focus on general whole genome analyses, such as variant calling and annotating, novel variants identification, etc. (the 1000 Genomes project; Boomsma *et al.* 2014). These studies utilized the pedigree information to assess the quality of detected variants using the Mendelian inheritance law. For example, DePristo and Mark used the Unified Genotyper in the GATK toolkit to call variants from CEU and YRI trios; then evaluated the quality of called variants using the Mendelian inheritance law (DePristo and Mark 2010). Our study on the KHV trio falls into the second group where general analyses were conducted by recent standard toolkits (e.g. HaplotypeCaller, an improvement of the Unified Genotyper). We used the Mendelian inheritance law not only to evaluate the quality of called variants but also to filter violated variants out of downstream analyses. The number of SNPs and the Mendelian violations in the KHV trio are similar to those in the CEU and YRI trios reported by DePristo and Daly. The Mendelian violation rates in CEU trio and YRI trio are slightly lower than the mutation rate in the KHV trio. This is likely because the depth coverage of CEU and YRI trios (i.e. ≥100x) are much higher than that of the KHV trio (i.e. 30x). The knowledge gained from studies on trios will help us to design large-scale projects with more samples to study the Vietnamese population and diseases.

To the best of our knowledge, this is the first Vietnamese whole genome-wide study at a high coverage level. We believe that this study will be an important reference for further genomic studies of Vietnamese and Southeast Asian populations. Finally, the novel variants identified from the KHV trio genomes demonstrated the necessity of conducting more genome-wide studies for Vietnamese and other populations to complete the picture of human genome variations.

All raw sequence data are available at the Sequence Read Archive (RSA) database of NCBI (*http://www.ncbi.nlm.nih.gov/bioproject/259581*). Other data are available upon request to the corresponding authors.

Acknowledgements

We would like to express our special thanks to Prof Nguyen Huu Duc from Vietnam National University, Hanoi, for his constant encouragement and support. We thank Prof Jean Daniel Zucker, Dr Zamin Iqbal and Prof Arndt von Haeseler for providing useful inputs to our manuscript. This project is partly financially supported by the Science and Technology Foundation of Vietnam National University, Hanoi (grant no. QKHCN.13.01). We also would like to thank the Center for Integrative Bioinformatics Vienna for providing computational resources. BQM acknowledges financial support by the Austrian Science Fund - FWF (grant no. I760-B17).

References

1000 Genomes Project Consortium 2010 A map of human genome variation from population-scale sequencing. *Nature* **467** 1061–1073

1000 Genomes Project Consortium 2012 An integrated map of genetic variation from 1,092 human genomes. *Nature* **491** 56–65

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ 1997 Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.* **25** 3389–3402

Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim DS, Kim BC, Kim SY, *et al*. 2009 The first korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* **19** 1622–1629

Azim MK, Yang C, Yan Z, Choudhary MI, Khan A, Sun X, Li R, Asif H, *et al*. 2013 Complete genome sequencing and variant analysis of a pakistani individual. *J. Hum. Genet.* **58** 622–626

Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, Ye K, Guryev V, *et al*. 2014 The genome of the netherlands: design, and project goals. *Eur. J. Hum. Genet.* **22** 221–227

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, *et al*. 2009 Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6** 677–681

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, *et al*. 2012 A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of *Drosophila melanogaster* strain w1118 iso-2 iso-3. *Fly* **6** 80–92

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, *et al*. 2011 A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nat. Genet.* **43** 491–498

DePristo M and Mark D 2010 Mendelian violations in the CEU andYRI Pilot 2 Trios. Technical report at broad Institute of Harvard and MIT

Dogan H, Can H and Otu HH 2014 Whole genome sequence of a turkish individual. *PLoS One* **9** 85233

Drmanac R 2011 The advent of personal genome sequencing. *Genet. Med.* **13** 188–190

Eden E, Navon R, Steinfeld I, Lipson D and Yakhini Z 2009 Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinform.* **10** 48

Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, Nagasaki M, Yamaguchi R, *et al*. 2010 Whole-genome sequencing and comprehensive variant analysis of a japanese individual using massively parallel sequencing. *Nat. Genet.* **42** 931–936

Hardy BJ, Seguin B, Singer PA, Mukerji M, Brahmachari SK and Daar AS 2008 From diversity to delivery: the case of the indian genome variation initiative. *Nat. Rev. Genet.* **9** 9–14

He Z, O'Roak BJ, Smith JD, Wang G, Hooker S, Santos-Cortez RLP, Li B, Kan M, *et al*. 2014 Rare-variant extensions of the transmission disequilibrium test: Application to autism exome sequence data. *Am. J. Hum. Genet.* **94** p33–46

International Human Genome Sequencing Consortium 2004 Finishing the euchromatic sequence of the human genome. *Nature* **431** 931–945

Iqbal Z, Caccamo M, Turner I, Flicek P and McVean G 2012 De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44** 226–232

Kumar P, Henikoff S and Ng PC 2009 Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nat. Protoc.* **4** 1073–1081

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, *et al*. 2001 Initial sequencing and analysis of the human genome. *Nature* **409** 860–921

Li H and Durbin R 2009 Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* **25** 1754–1760

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, *et al*. 2009 The sequence alignment/map format and samtools. *Bioinformatics* **25** 2078–2079

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, *et al*. 2010 The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res.* **20** 1297–1303

MacDonald JR, Ziman R, Yuen RK, Feuk L and Scherer SW 2014 The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42** 986–992

Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, *et al*. 2010 Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328** 636–639

Shendure J and Ji H 2008 Next-generation dna sequencing. *Nat. Biotechnol.* **26** 1135–1145

Sherry ST, Ward MH, Kholodov M, BakerJ PL, Smigielski EM and Sirotkin K 2001 dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res.* **29** 308–311

Shigemizu D, Fujimoto A, Akiyama S, Abe T, Nakano K, Boroevich KA, Yamamoto Y, Furuta M, Kubo M, Nakagawa H, *et al.* 2013 A practical method to detect snvs and indels from whole genome and exome sequencing data. *Sci. Rep.* **3**

Siva N 2008 1000 genomes project. *Nat. Biotechnol.* **26** 256–256

Skryabin K, Prokhortchouk E, Mazur A, Boulygina E, Tsygankova S, Nedoluzhko A, Rastorguev S, Matveev V, *et al*. 2009 Combining two technologies for full genome sequencing of human. *Acta Naturae* **1** 102

Spielman RS, McGinnis RE and Ewens WJ 1993 Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52** 506–516

Tennessen J, Bigham A, O'Connor T, *et al*. 2012 Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337** 64–69

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, *et al*. 2001 The sequence of the human genome. *Science* **291** 1304–1351

von Bubnoff A 2008 Next-generation sequencing: the race is on. *Cell* **132** 721–723

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, *et al*. 2008 The diploid genome sequence of an Asian individual. *Nature* **456** 60–65

Wong LP, Ong RTH, Poh WT, Liu X, Chen P, Li R, Lam KKY, Pillai NE, *et al*. 2013 Deep whole-genome sequencing of 100 Southeast Asian Malays. *Am. J. Hum. Genet.* **92** 52–66

Zerbino DR and Birney E 2008 Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18** 821–829

MS received 11 June 2014; accepted 31 December 2014

Corresponding editor: Partha P Majumder