

Genetic profiling of 2,683 Vietnamese genomes from non-invasive prenatal testing data

Ngoc Hieu Tran (✉ hieutran1985@gmail.com)

University of Waterloo <https://orcid.org/0000-0001-5186-8852>

Thanh Binh Vo

Gene Solutions

Van Thong Nguyen

Hung Vuong Hospital

Nhat Thang Tran

University Medical Center, Ho Chi Minh city

Thu-Huong Nhat Trinh

Tu Du Hospital

Hong-Anh Thi Pham

Gene Solutions

Hong Thuy Dao

Gene Solutions

Ngoc Mai Nguyen

Gene Solutions

Yen-Linh Thi Van

Gene Solutions

Vu Uyen Tran

Gene Solutions

Hoang Giang Vu

Gene Solutions

Quynh-Tram Nguyen Bui

Gene Solutions

Phuong-Anh Ngoc Vo

Gene Solutions

Huu Nguyen Nguyen

Gene Solutions

Quynh-Tho Thi Nguyen

Medical Genetics Institute

Thanh-Thuy Thi Do

Medical Genetics Institute

Phuong Cao Thi Ngoc

Lund University

Dinh Kiet Truong

Medical Genetics Institute

Hoai-Nghia Nguyen

University of Medicine and Pharmacy at Ho Chi Minh city

Hoa Giang

Medical Genetics Institute

Minh Duy Phan

The University of Queensland

Research article

Keywords: genome sequencing, population genetics, non-invasive prenatal testing

Posted Date: January 3rd, 2020

DOI: <https://doi.org/10.21203/rs.2.20017/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

1 **Genetic profiling of 2,683 Vietnamese genomes from non-invasive**
2 **prenatal testing data**

3 Ngoc Hieu Tran^{1,2,*}, Thanh Binh Vo^{1,3}, Van Thong Nguyen⁴, Nhat Thang Tran⁵, Thu-Huong Nhat
4 Trinh⁶, Hong-Anh Thi Pham^{1,3}, Hong Thuy Dao^{1,3}, Ngoc Mai Nguyen^{1,3}, Yen-Linh Thi Van^{1,3}, Vu
5 Uyen Tran^{1,3}, Hoang Giang Vu^{1,3}, Quynh-Tram Nguyen Bui^{1,3}, Phuong-Anh Ngoc Vo^{1,3}, Huu
6 Nguyen Nguyen^{1,3}, Quynh-Tho Thi Nguyen³, Thanh-Thuy Thi Do^{3,7}, Phuong Cao Thi Ngoc^{3,7},
7 Dinh Kiet Truong³, Hoai-Nghia Nguyen^{8,*}, Hoa Giang^{1,3,*}, Minh-Duy Phan^{1,3,9,*}

8
9 ¹Gene Solutions, Vietnam

10 ²David R. Cheriton School of Computer Science, University of Waterloo, Canada

11 ³Medical Genetics Institute, Vietnam

12 ⁴Hung Vuong hospital, Vietnam

13 ⁵University Medical Center, Ho Chi Minh city, Vietnam

14 ⁶Tu Du hospital, Vietnam

15 ⁷Division of Molecular Hematology, Lund Stem Cell Center, Lund University, Lund, Sweden

16 ⁸University of Medicine and Pharmacy at Ho Chi Minh city, Vietnam

17 ⁹School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane,
18 Australia

19

20 *Corresponding authors.

21 Emails: nh2tran@uwaterloo.ca (Ngoc Hieu Tran); m.phan1@uq.edu.au (Minh-Duy Phan);

22 gianghoa@gmail.com (Hoa Giang); nhnghia81@gmail.com (Hoai-Nghia Nguyen).

23

24 **Abstract**

25 **Background:** The under-representation of Vietnamese ethnic groups in existing genetic
26 databases and studies have undermined our understanding of the genetic variations and
27 associated traits or diseases in the population. Cost and technology limitations remain the
28 challenges in performing large-scale genome sequencing projects in Vietnam and many
29 developing countries. Non-invasive prenatal testing (NIPT) data offers an alternative untapped
30 resource to study genetic variations in the Vietnamese population.

31 **Results:** We analyzed the low-coverage genomes of 2,683 pregnant Vietnamese women using
32 their NIPT data and identified a comprehensive set of 8,054,515 single-nucleotide
33 polymorphisms, among which 8.2% were new to the Vietnamese population. Our study also
34 revealed 24,487 disease-associated genetic variants and their allele frequency distribution,
35 especially five pathogenic variants for prevalent genetic disorders in Vietnam. We also observed
36 major discrepancies in the allele frequency distribution of disease-associated genetic variants
37 between the Vietnamese and other populations, thus highlighting a need for genome-wide
38 association studies dedicated to the Vietnamese population.

39 **Conclusions:** We have demonstrated a successful analysis of NIPT data to reconstruct the
40 Vietnamese genetic profiles. This application provides a powerful yet cost-effective approach for
41 large-scale population genetic studies.

42 **Keywords:** genome sequencing; population genetics; non-invasive prenatal testing

43

44 Background

45 Following the successful initiative of the 1000 genomes project [1], several large-scale genome
46 and exome sequencing projects have been conducted, either as international collaboration
47 efforts such as ExACT [2], gnomAD [3], or for a specific country or population [4-8]. Those
48 projects have provided comprehensive profiles of human genetic variation in some populations,
49 paving the way for unprecedented advance in treatment of common genetic diseases. However,
50 the lack of diversity and the under-representation of several populations in genome sequencing
51 projects and genome-wide association studies (GWAS) have increasingly become a critical
52 problem [9,10]. For instance, Gurdasani *et al.* found that the representation of ethnic groups in
53 GWAS was significantly biased, with nearly 78% of the participants having European ancestries,
54 whereas the two major populations, Asian and African, only accounted for 11% and 2.4%,
55 respectively [10]. Vietnam has a population of 96.5 million, the 15th highest in the world and the
56 9th highest in Asia. Yet there was merely one dataset of 99 Vietnamese individuals that had
57 been studied as part of the 1000 genomes project (population code KHV, the Kinh ethnic group
58 in Ho Chi Minh City, Vietnam). A recent study has sequenced genomes and exomes of another
59 305 individuals to further expand the Vietnamese genetic database [11]. However, costs and
60 technologies to perform large-scale genome sequencing projects still remain a challenge for
61 most developing countries, including Vietnam.

62 An alternative approach has been proposed recently to re-use the low-coverage genome
63 sequencing data from non-invasive prenatal testing (NIPT) for large-scale population genetics
64 studies [12,13]. NIPT is a method that sequences cell-free DNA from maternal plasma at an
65 ultra-low depth of 0.1-0.2x to detect fetal aneuploidy [14]. By combining a sufficiently large
66 number of NIPT samples, one could obtain a good representation of the population genetic
67 variation. The benefits of re-using NIPT data for population genetics are manifold. First, the data
68 can be re-used at no extra cost given the approval and consent of the participants. As one of

69 the most rapidly adopted genetic tests, NIPT has been successfully established and become a
70 standard screening procedure with thousands to millions of tests performed world-wide,
71 including many developing countries such as Vietnam [14]. Using NIPT data for population
72 genetic studies may also reduce privacy concerns since the genetic variants can only be
73 analyzed by aggregating a large number of samples and the results can only be interpreted at
74 the population level. A single sample tells little about the genetic information of an individual due
75 to low sequencing depth. Last but not least, previous studies have suggested that sequencing a
76 large number of individuals at a low depth might provide more accurate inferences of the
77 population genetic structure than the traditional approach of sequencing a limited number of
78 individuals at a higher depth, especially when the budget is limited [15,16].

79 In this paper, we presented the first study of Vietnamese genetic variations from non-invasive
80 prenatal testing data, and to the best of our knowledge, the third of such kinds in the world
81 [12,13]. We analyzed the NIPT data of 2,683 pregnant Vietnamese women to identify genetic
82 variants and their allele frequency distribution in the Vietnamese population. We also studied
83 the relationships between the Vietnamese genetic profile and common genetic disorders, and
84 discovered pathogenic variants related to prevalent diseases in Vietnam. Finally, we highlighted
85 the differences in the distribution of disease-associated genetic variants between the
86 Vietnamese and other populations, thus highlighting a need for genome-wide association
87 studies dedicated to the Vietnamese population.

88 Results

89 Data collection

90 A total of 2,683 pregnant Vietnamese women who performed non-invasive prenatal testing
91 during the period from 2018-2019 at the Medical Genetics Institute, Vietnam, were recruited to
92 the study. The participants have approved and given written consent to the anonymous re-use
93 of their genomic data for the study. All information of the participants is confidential and not

94 available to the authors, except the records that their NIPT and pregnancy results are normal.
95 The study was approved by the institutional ethics committee of the University of Medicine and
96 Pharmacy, Ho Chi Minh city, Vietnam. The whole genome of each participant was sequenced to
97 an average of 3.6 million paired-end reads of 2x75 bp, which corresponds to a sequencing
98 depth of 0.17x per sample.

99 **Genome coverage and sequencing depth of the NIPT dataset**

100 Data pre-processing was first performed on each of 2,683 samples and the results were stored
101 in binary alignment map (BAM) format, one BAM file per sample. The data pre-processing steps
102 include: quality control of raw data using FastQC [17], trimmomatic [18]; alignment of paired-end
103 reads to the human reference genome (build GRCh38) using bwa [19], samtools [20],
104 MarkDuplicates [21]; and summary of alignment results using Qualimap [22], bedtools [23], IGV
105 [24]. The quality of raw data and alignment results are presented in Supplementary Figures S1
106 and S2. Raw data showed high sequencing quality, no error or bias was observed. The
107 mapping quality and insert size distributions followed closely what expected across the
108 reference genome. The overall sequencing error rate was about 0.3%. More details of the data
109 pre-processing steps can be found in the Methods section.

110 The average genome coverage and depth were 14.59% and 0.17x per sample, respectively,
111 and aggregated to 95.09% and 462x across 2,683 samples (Figure 1a). Although the
112 sequencing depth per sample was low, there might be more than one read from the same
113 sample overlapping at a genome position. This problem may affect the estimation of allele
114 frequency because the estimation is based on the assumption that a sample may contribute
115 only 0 or 1 allele (read) at any given genome position [12,13]. For instance, we found that the
116 average percentage of genome positions with depth 2x (i.e. covered by two overlapping reads)
117 in a sample was 1.75% (Figure 1a). These overlapping reads occurred randomly across the
118 reference genome and the samples. At any genome position, there were on average 47 out of

119 2,683 samples that each contributed two reads (Supplementary Figure S3). To address this
120 problem, we followed a filtering strategy from previous studies [12,13] to keep only one read if
121 there were overlapping reads in a sample. Thus, for every genome position, each sample could
122 only contribute up to one read, and when the samples were aggregated, all reads at any
123 position were obtained from different samples. In addition, we also removed alignments with low
124 mapping quality scores (MAPQ < 30).

125 After the filtering step, the sequencing depth was reduced from 0.17x to 0.12x per sample. The
126 aggregated sequencing depth of 2,683 samples was 364x and the genome coverage was
127 91.56% (Figure 1a). The distributions of genome coverage and sequencing depth are presented
128 in Figures 1b-d. The sequencing depth was approximately uniform across the reference
129 genome, except for low-mappability regions and chromosome Y. The distributions of
130 sequencing depth and MAPQ score also closely followed the mappability of the human
131 reference genome obtained from Umap [25]. The average sequencing depth of chromosome Y
132 was about 2.1% that of the whole genome, consistent with the proportion of fetal DNA in NIPT
133 samples (8-10%) [14].

134 **Variant calling and validation**

135 We aggregated 2,683 samples into one and used Mutect2 from GATK [26,27] for variant calling
136 and allele frequency estimation. In addition to its main function of somatic calling, Mutect2 can
137 also be used on data that represents a pool of individuals, such as our NIPT dataset, to call
138 multiple variants at a genome site [12,13,28]. The called variants were further checked against
139 strand bias, weak evidence, or contamination using FilterMutectCalls. The allele frequencies
140 were estimated based on the numbers of reads aligned to the reference and the alternate
141 alleles. For validation, we compared our NIPT call set to the KHV (Kinh in Ho Chi Minh City,
142 Vietnam) and EAS (East Asian) populations from the 1000 genomes project [1], as well as the
143 dbSNP database (version 151, [29]).

144 We identified a total of 8,054,515 SNPs from the NIPT dataset. The transition to transversion
145 ratio was 2.0 over the whole genome and 2.8 over protein coding regions, which was similar to
146 the observed ratios from previous genome or exome sequencing projects. As expected, a
147 majority of these SNPs, 7,390,020 or 91.8%, had been reported earlier in the KHV call set
148 (Figure 2a). Since the KHV population only had 99 individuals, we further looked into its
149 common SNPs that were shared by at least two individuals. We found that the NIPT call set
150 recovered 90.5% of the KHV common SNPs (Supplementary Figure S4; 6,889,016 / 7,609,526
151 = 90.5%). This sensitivity is in line with the genome coverage reported earlier in Figure 1a. An
152 important advantage of NIPT data is the ability of sampling a large number of individuals to
153 better represent a population and to accurately estimate the allele frequency. We found a strong
154 Pearson correlation of 98.8% between the allele frequency of the NIPT call set and that of the
155 KHV call set (Figure 2b). Furthermore, thanks to its larger sample size, the NIPT allele
156 frequency indeed showed better resolution than the KHV one, as evidenced by vertical trails in
157 Figure 2b or a zoomed-in view in Supplementary Figure S5.

158 Our NIPT call set included 664,495 (8.2%) SNPs that had not been reported in the KHV call set.
159 Among them, 67,153 (0.8%) were found in the EAS call set, another 517,020 (6.4%) were found
160 in the dbSNP database, and the remaining 80,322 (1.0%) were novel SNPs (Figure 2a).
161 Majority of those SNPs had allele frequencies less than 10%. The overall allele frequency
162 distribution of our NIPT call set is presented in Figure 2c.

163 We used VEP (Variant Effect Predictor [30]) to analyze the effects of 8,054,515 variants in the
164 NIPT call set (Figure 2d). About 1.5% of the SNPs were located in the coding and UTR regions,
165 12% in the upstream and downstream regions, and 4% in the TF regulatory regions. More than
166 80% of the SNPs were located in the intron and intergenic regions. We also noted that the new
167 SNPs and those in KHV had similar proportions of coding, UTR, upstream and downstream,
168 and TF regulatory regions (Figure 2d).

169 **Analysis of pathogenic SNPs and their allele frequencies in Vietnamese population**

170 We searched the NIPT call set against the ClinVar database (version 20191105, [31]) to explore
171 the associations between Vietnamese genomic variants and common genetic diseases. We
172 identified 24,487 SNPs with ClinVar annotations that have been reviewed by at least one
173 research group (Table 1). Among them, five SNPs were classified as “Pathogenic” or “Likely
174 pathogenic”, 117 SNPs were found to affect a “drug response”, and a majority of the remaining
175 were classified as “Benign” or “Likely benign”. We also noted that 391 ClinVar-annotated SNPs
176 (1.6%), including 1 pathogenic SNP, had not been reported in the KHV call set.

177 Table 2 and Supplementary Table S1 present the details of five pathogenic SNPs identified in
178 our NIPT call set. Their associated genetic diseases include: erythropoietic protoporphyria, non-
179 syndromic genetic deafness, Joubert syndrome, hemochromatosis type 1, and 5-alpha
180 reductase deficiency. The SNP rs9332964 C>T in *SRD5A2*, which is associated with 5-alpha
181 reductase deficiency, had not been reported in the KHV call set. 5-alpha reductase deficiency is
182 an autosomal recessive disorder that affects male sexual development. This SNP is rare in the
183 world and East Asia populations, but was found to be more common in the Vietnamese
184 population (allele frequency of 0.05%, 0.67%, and 2.90%, respectively). This SNP had also
185 been reported in a recent study [11] at a very low allele frequency of 1.36%.

186 We noticed that the allele frequencies of the five pathogenic SNPs varied considerably between
187 the Vietnamese, the East Asia, and the world populations (Table 2). For instance, the SNP
188 rs72474224 C>T in *GJB2* is commonly linked to non-syndromic hearing loss and deafness, which
189 is also the most prevalent genetic disorder in the Vietnamese population. We found that its
190 allele frequency in the Vietnamese population was 60% higher than in the East Asia population,
191 which in turn was an order of magnitude higher than in the world population (13.40%, 8.35%,
192 and 0.76%, respectively). The allele frequency was consistent with the estimated carrier
193 frequency of 1 in 5 in the Vietnamese population. Similarly, the allele frequency of rs2272783

194 A>G in *FECH*, which is associated with erythropoietic protoporphyria, was nearly three times
195 higher in the Vietnamese and East Asia populations than in the world population (28.10%,
196 32.57%, and 11.23%, respectively). The prevalence of this pair of SNP and disease in East and
197 Southeast Asia has been reported previously in [32]. On the other hand, the allele frequency of
198 rs1799945 C>G in *HFE*, which is associated with hemochromatosis type 1, was about two and
199 three times lower in the Vietnamese and East Asia populations than in the world population
200 (5.10%, 3.41%, and 10.82%, respectively). Such discrepancies were also observed for “Benign”
201 variants, e.g., those related to autosomal recessive non-syndromic hearing loss (Supplementary
202 Table S2). The variations strongly suggest that population-specific genome-wide association
203 studies are required to provide a more accurate understanding of the clinical significance of
204 genetic variants and the true disease prevalence in the Vietnamese population.

205 Discussion

206 In this study, we analyzed the genomes of 2,683 pregnant Vietnamese women from their non-
207 invasive prenatal testing data. The genomes were originally sequenced at a low depth of
208 approximately 0.17x per sample for the purpose of fetal aneuploidy testing [14]. Here we
209 combined the 2,683 samples to a total sequencing depth of 364x and performed variant calling
210 and analysis for the Vietnamese population. We identified a comprehensive set of 8,054,515
211 SNPs at a high level of sensitivity and accuracy: 90.5% of Vietnamese common SNPs were
212 recovered; 99% of identified SNPs were confirmed in existing databases; and a strong
213 correlation of 98.8% to known allele frequencies. The results were exciting given that the total
214 sequencing depth of our dataset, 364x, was merely equivalent to sequencing 20 individuals at a
215 moderate depth of 20x. It also suggests that there is still plenty of room for improvement by
216 increasing the number of NIPT samples. For instance, Liu *et al.* have demonstrated a large-
217 scale population genetic analysis based on hundreds of thousands of NIPT samples for the
218 Chinese population [12].

219 Another benefit of using NIPT data is cost-effective. In our study, the dataset was re-used at no
220 cost with written consent from the participants. The whole analysis pipeline was done within a
221 week on a cloud computing platform for a few hundred dollars. Thus, the overall cost was
222 negligible compared to that of a typical genome sequencing project. The cost advantage of this
223 approach may play a major role in large-scale genome sequencing projects, especially in
224 developing countries where technologies and resources are still limited.

225 Our study revealed 24,487 disease-associated genetic variants, especially five pathogenic
226 variants for prevalent genetic disorders in the Vietnamese population. We also found major
227 discrepancies in the allele frequency distribution of genetic variants between the Vietnamese,
228 the East Asia, and the world populations. Thus, a comprehensive genetic profile and genome-
229 wide association studies dedicated to the Vietnamese population are highly desired. Knowing
230 the distribution of genetic disorders in the population will be useful for public health policy and
231 planning, preventive medicine, early genetic screening strategies, etc.

232 Some technical and design limitations in our study could be addressed in future research to
233 improve the application of NIPT data in population genetics studies. First, currently there is no
234 variant calling tool that is designed specifically for NIPT data. Here we used Mutect2 and
235 previous studies also used similar somatic calling tools with the purpose of identifying all
236 possible variants at a genome site [12,13]. Thus, we took a conservative approach to consider
237 only SNPs but not indels to ensure a reliable call set. Another limitation was the exclusion of
238 chromosome Y due to its low coverage as a result of limited amount of fetal DNA in NIPT
239 samples. This problem could be addressed by increasing the number of samples to obtain
240 enough sequencing coverage for reliable variant calling. NIPT data is also biased by sex, with
241 only ~5% of the data coming from male population (assuming a 10% cell-free fetal DNA fraction
242 with 50% male fetuses). In general, sufficiently large sample size is essential to use NIPT data

243 for population genetics research. Thus, privacy policy, code of ethics, and standards of practice
244 need to be established to protect the confidential information and data of the participants.

245 Conclusions

246 We showed that non-invasive prenatal testing data could be reliably used to reconstruct the
247 genetic profile of the Vietnamese population. Our study identified pathogenic variants for
248 prevalent genetic diseases in the Vietnamese population and called for a need for population-
249 specific genome-wide association studies. The results demonstrated that non-invasive prenatal
250 testing data provides a valuable and cost-effective resource for large-scale population genetic
251 studies.

252 Methods

253 **Sample preparation**

254 Cell-free DNA (cfDNA) in maternal plasma was extracted using MagMAX Cell-Free DNA
255 Isolation Kit from Thermo Fisher Scientific (Waltham, MA, USA). Library preparation was done
256 using NEBNext Ultra II DNA Library Prep Kit from New England BioLabs (Ipswich, MA, USA).
257 The samples were sequenced on the NextSeq 550 platform using paired-end 2x75 bp Reagent
258 Kit from Illumina (San Diego, CA, USA).

259 **Bioinformatics analysis pipeline**

260 Quality check of raw sequencing data was performed using FastQC (version 0.11.8, [17]).
261 Paired-end reads were trimmed to 75 bp, adapters (“TruSeq3-PE-2.fa”) and low-quality bases
262 were removed using trimmomatic (version 0.39, [18]). We only kept pairs with both reads
263 surviving the trimming (about 95.7% of the dataset). The reads were then aligned to the human
264 reference genome, build GRCh38 (hg38), using bwa mem (version 0.7.17-r1188, [19]).
265 Supplementary hits were marked as secondary for Picard compatibility. Alignment results were

266 sorted and indexed using samtools (version 1.9, [20]). Potential PCR duplicates were marked
267 using MarkDuplicates from GATK (version 4.1.1.0, [26]).

268 Alignments with mapping quality scores less than 30 were discarded. In-house Python scripts
269 were developed to mark overlapping alignments and to keep only one of them. Qualimap
270 (version 2.2.1, [22]), bedtools (version 2.25.0 [23]), and IGV (version 2.4.19, [24]) were used to
271 summarize the alignment results and to calculate genome coverage and sequencing depth.

272 Mutect2 from GATK (version 4.1.1.0 [26,27]) was used in tumor-only mode for variant calling. All
273 samples were assigned the same sample name to combine them before variant calling.

274 FilterMutectCalls was used to exclude variants with weak evidence, strand bias, or
275 contamination. bcftools (version 1.9, [33]) was used to filter, summarize, and compare VCF
276 (Variant Call Format) files. VEP (version 98, [30]) was used to predict the effects of variants and
277 to annotate them against dbSNP (version 151, [29]) and ClinVar (version 20191105, [31])
278 databases.

279 The whole analysis pipeline and parameter settings can be found in the attached Python scripts.

280 **Declarations**

281 **Ethics approval and consent to participate**

282 The study was approved by the institutional ethics committee of the University of Medicine and
283 Pharmacy, Ho Chi Minh city, Vietnam. The participants who performed NIPT triSure at Medical
284 Genetics Institute, Vietnam, have approved and given written consent to the anonymous re-use
285 of their genomic data for this study.

286 **Consent for publication**

287 All authors have read and approved the manuscript for publication.

288 **Availability of data and materials**

289 The NIPT variant call set and Python scripts for the bioinformatics analysis pipeline can be
290 found on GitHub: https://github.com/nh2tran/NIPT_WGS

291 **Competing interests**

292 NHT, TBV, HATP, HTD, NMN, YLTV, VUT, HGV, QTNB, PANV, HNN, HG and MDP are current
293 employees of Gene Solutions, Vietnam. The other authors declare no competing interests.

294 **Funding**

295 This study was funded by Gene Solutions, Vietnam. The funder did not have any additional role
296 in the study design, data collection and analysis, decision to publish, or preparation of the
297 manuscript.

298 **Authors' contributions**

299 TBV, HATP, HTD, NMN, YLTV, VUT, HGV, QTNB, PANV performed experiments.

300 VTN, NTTM, THNT, TTTD recruited patients and performed clinical analysis.

301 HNN, QTTN, PCTN, DKT designed experiments and analyzed data.

302 HNN supervised the project.

303 NHT, HG, MDP analyzed the data and wrote the manuscript, designed experiments and
304 analyzed sequencing data.

305 **Acknowledgements**

306 The authors thank Dr. Kwok Pui Choi for critical reading of the manuscript. NHT was partially
307 supported by the Canada NSERC grant (OGP0046506) and the Canada Research Chair
308 program.

309 **References**

- 310 1. The 1000 Genomes Project Consortium. A global reference for human genetic variation.
311 Nature 2015;526:68-74.
- 312 2. Lek M et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature
313 2016;536:285-91.
- 314 3. Karczewski KJ et al. Variation across 141,456 human exomes and genomes reveals the
315 spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv
316 2019; doi: 10.1101/531210.
- 317 4. The UK10K Consortium. The UK10K project identifies rare variants in health and disease.
318 Nature 2015;526:82-90.
- 319 5. Gudbjartsson DF et al. Large-scale whole-genome sequencing of the Icelandic population.
320 Nat Genet. 2015;47:435-44.
- 321 6. Maretty L et al. Sequencing and de novo assembly of 150 genomes from Denmark as a
322 population reference. Nature 2017;548:87-91.
- 323 7. The Genome of the Netherlands Consortium. Whole-genome sequence variation,
324 population structure and demographic history of the Dutch population. Nat Genet.
325 2014;46:814-25.
- 326 8. Wu D et al. Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations
327 in Singapore. Cell 2019; 179:736-749.
- 328 9. Editorial. Diversity matters. Nat Rev. Genet. 2019;20:495.
- 329 10. Gurdasani D, Barroso I, Zeggini E, Sandhu MS. Genomics of disease risk in globally
330 diverse populations. Nat Rev Genet. 2019;20:520-535.
- 331 11. Le VS et al. A Vietnamese human genetic variation database. Hum Mutat. 2019;40:1664-
332 1675.

- 333 12. Liu S et al. Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic
334 Associations, Patterns of Viral Infections, and Chinese Population History. *Cell*
335 2018;175:347-359.
- 336 13. Budis J et al. Non-invasive prenatal testing as a valuable source of population specific
337 allelic frequencies. *J Biotechnol.* 2019;299:72-78.
- 338 14. Phan MD et al. Establishing and validating noninvasive prenatal testing procedure for fetal
339 aneuploidies in Vietnam. *J Matern Fetal Neonatal Med.* 2019;32:4009-4015.
- 340 15. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing:
341 implications for design of complex trait association studies. *Genome Res.* 2011;21:940-51.
- 342 16. Fumagalli M. Assessing the effect of sequencing depth and sample size in population
343 genetics inferences. *PLoS One* 2013;8:e79667.
- 344 17. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- 345 18. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence
346 data. *Bioinformatics.* 2014;30:2114-20.
- 347 19. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
348 arXiv:1303.3997v2 [q-bio.GN].
- 349 20. Li H et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.*
350 2009;25:2078-9.
- 351 21. <http://broadinstitute.github.io/picard/>
- 352 22. Okonechnikov K, Conesa A, Garcia-Alcalde F. Qualimap 2: advanced multi-sample quality
353 control for high-throughput sequencing data. *Bioinformatics.* 2016;32:292-4.
- 354 23. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
355 *Bioinformatics.* 2010;26:841-2.
- 356 24. Robinson JT et al. Integrative Genomics Viewer. *Nat Biotechnol.* 2011;29:24-6.
- 357 25. Karimzadeh M, Ernst C, Kundaje A, Hoffman MM. Umap and Bismap: quantifying genome
358 and methylome mappability. *Nucleic Acids Res.* 2018;46:e120.

- 359 26. Van der Auwera GA et al. From FastQ data to high confidence variant calls: the Genome
360 Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43:11.10.1-33.
- 361 27. Cibulskis K et al. Sensitive detection of somatic point mutations in impure and
362 heterogeneous cancer samples. *Nat Biotechnol*. 2013;31:213-9.
- 363 28. <https://software.broadinstitute.org/gatk/documentation/article?id=11136#2.1>
- 364 29. Sherry ST et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*.
365 2001;29:308-11.
- 366 30. McLaren W. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17:122.
- 367 31. Landrum MJ et al. ClinVar: public archive of relationships among sequence variation and
368 human phenotype. *Nucleic Acids Res*. 2014;42:D980-5.
- 369 32. Gouya L et al. Contribution of a common single-nucleotide polymorphism to the genetic
370 predisposition for erythropoietic protoporphyria. *Am J Hum Genet*. 2006;78:2-14.
- 371 33. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and
372 population genetical parameter estimation from sequencing data. *Bioinformatics*.
373 2011;27:2987-2993.

374 Figure Legends

375 **Figure 1.** Distributions of genome coverage and sequencing depth of the NIPT dataset. (a)

376 Average genome coverage and sequencing depth per sample and from all samples combined.

377 (b) Summary histogram of sequencing depth over all genome positions. (c) Distribution of

378 sequencing depth per chromosome. (d) IGV tracks of sequencing depth, bwa MAPQ score, and

379 Umap k50 mappability across the whole genome.

380 **Figure 2.** Summary of the NIPT call set. (a) Venn diagram comparison between the NIPT call

381 set, the KHV and EAS call sets from the 1000 genomes project, and the dbSNP database. The

382 percentages were calculated with respect to the NIPT call set. (b) Scatter plot comparison of

383 allele frequency estimated from the NIPT and the KHV call sets. (c) Allele frequency distribution
384 of the NIPT call set. (d) Distribution of locations and effects of variants in the NIPT call set.

Figures

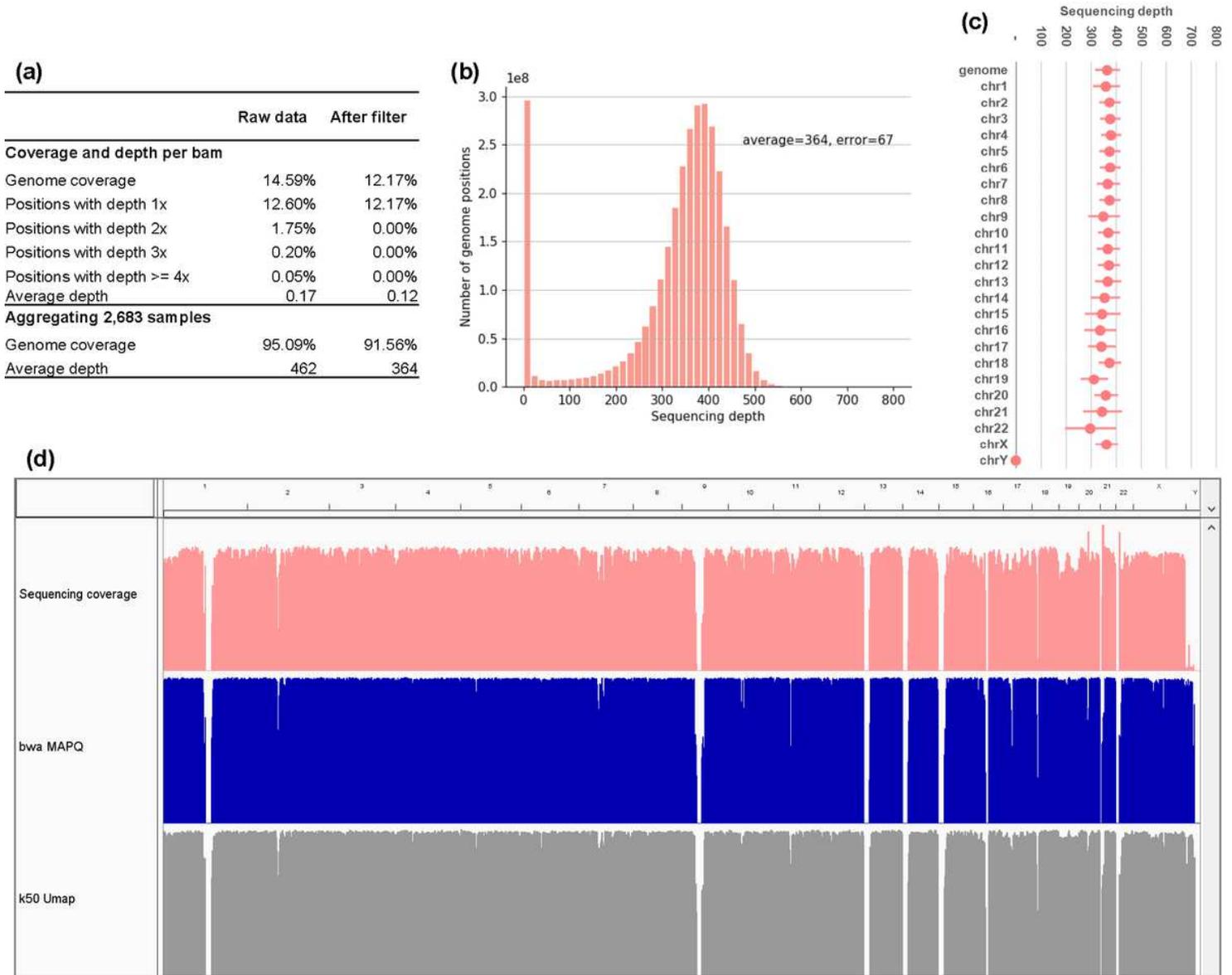


Figure 1

Distributions of genome coverage and sequencing depth of the NIPT dataset. (a) Average genome coverage and sequencing depth per sample and from all samples combined. (b) Summary histogram of sequencing depth over all genome positions. (c) Distribution of sequencing depth per chromosome. (d) IGV tracks of sequencing depth, bwa MAPQ score, and Umap k50 mappability across the whole genome.

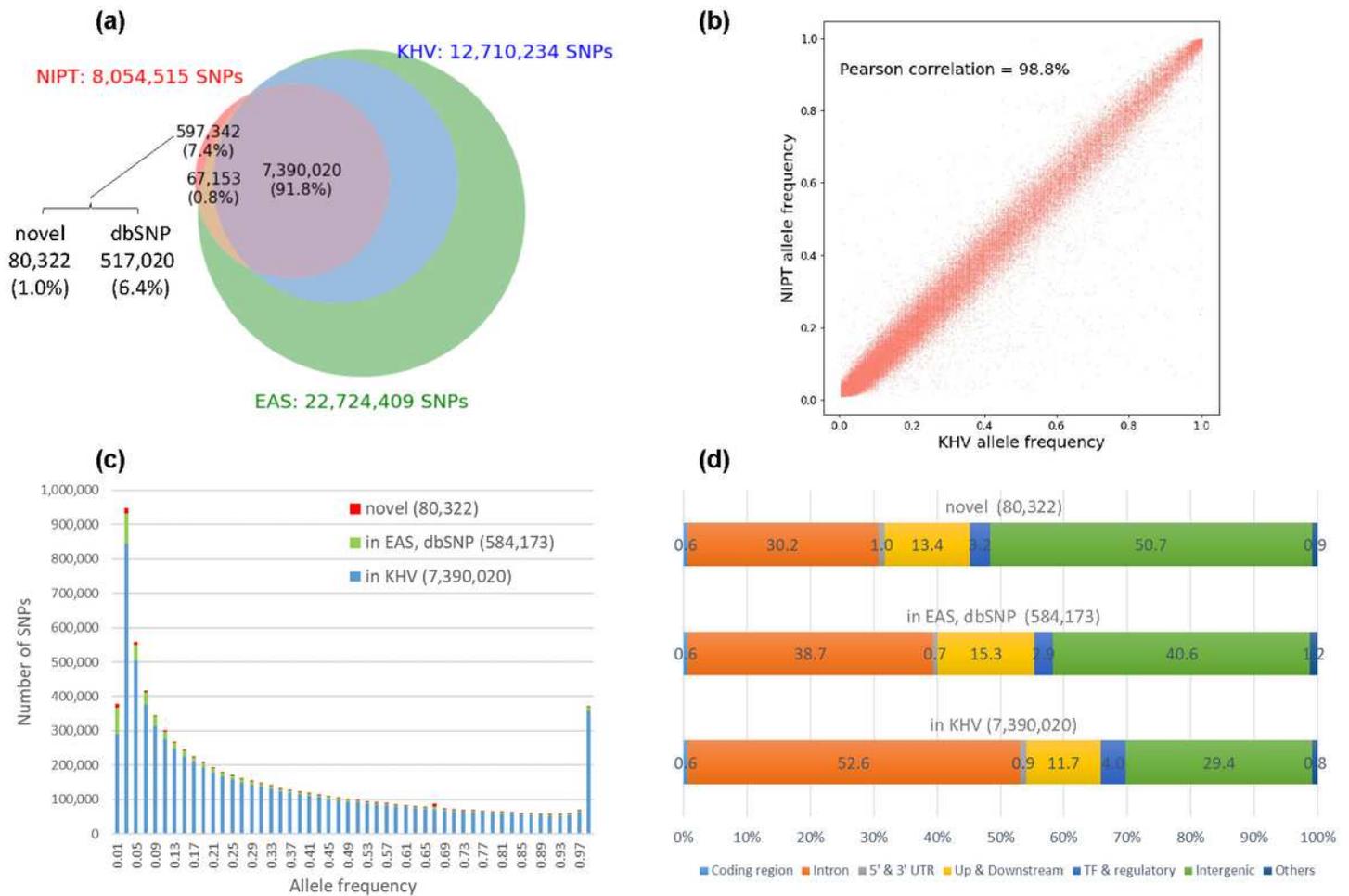


Figure 2

Summary of the NIPT call set. (a) Venn diagram comparison between the NIPT call set, the KHV and EAS call sets from the 1000 genomes project, and the dbSNP database. The percentages were calculated with respect to the NIPT call set. (b) Scatter plot comparison of allele frequency estimated from the NIPT and the KHV call sets. (c) Allele frequency distribution of the NIPT call set. (d) Distribution of locations and effects of variants in the NIPT call set.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Manuscript.supfigurestables.pdf](#)
- [Tables12.pdf](#)