Building the Vietnamese Reference Genome

Ha Anh Tuan Nguyen
hatng@kth.se

Instituto Superior Técnico, Lisboa, Portugal
October 2014

Abstract

With the advent of Next Generation Sequencing technology, the cost of sequencing a full human genome has been reduced dramatically. Several individual genome projects and large-scale sequencing projects such as 1000 genomes project, 750 Dutch genomes, 100 Southeast Asian Malays, YanHuang project, ... have been established to identify the genetic variations in human genomes. The identified genetic variations could become useful information for analyzing genetic diseases and for discovering the genetic diversity between populations. Recent results coming from whole genome sequencing projects have suggested the existence of genetic differences between peoples from diverse populations. Such phenomenon has revealed the limits of the standard NCBI reference genome in population-specific genome-wide studies. To tackle this problem, a population-specific reference genome is needed. Using the Vietnamese Kinh data produced by 1000 genomes project we constructed a reference genome for Vietnamese. Experiments on some chromosomes showed that the Vietnamese-specific reference genome helped improving the mapping quality of short reads and the quality of variants when dealing with Vietnamese genomes. Further population studies revealed the close genetic relationships between Vietnamese Kinh and some Thai and Chinese ethnic groups. The genetic distances between Vietnamese and other Southeast Asian populations were also implied in the results.

Keywords: Bioinformatics, Vietnamese Kinh, Reference genome, Genetic diversity

1. Introduction

In 2000, after spending $3 billion in more than 10 years, the first draft human reference genome was released [7]. It soon became one of the most important research results in the 21st century because the reference genome can act as the guiding sequence for every genome-wide study projects. The first human reference genome also opened a new era of research in the related fields such as molecular medicine and human evolution.

Using the benefits of the reference genome, a number of large-scale sequencing projects were established including 1000 genomes project (1KG) [1], The genome of the Netherlands (or 750 Dutch genomes) [3], and 100 Southeast Asian Malays [15]. The ultimate goal of those large-scale sequencing projects is to discover all genetic variations. Based on those variations, several analyses will be conducted in order to identify the disease-related variations as well as the genetic diversity between populations. While the disease-related variations may help us improving our understanding of disease mechanisms, the genetic diversity could give us the information about ancient human migrations.

Recently, several individual reference genomes have been built such as: Yobura Nigerian [2], Chinese individual genome (or YanHuang Project) [14], Korean individual genome [8], Japanese individual genome [5], Indian individual genome [6], etc. Those genomes reveal millions of population-specific variations. In fact, a country-specific reference genome holds the unique information of the population. Therefore, it works more precisely than the original reference genome in discovering the genetic variations of the individual coming from that population.

Vietnam is one of the most populous countries in the world; it is ranked 14th in the world in population in recent statistical studies. As a part of 1KG, Vietnamese samples have been sequenced and the raw reads have been released at the end of 2013. However, 1KG did not cover all ethnic groups in Vietnam. The project only extended to sequence Vietnamese Kinh population - the most populous ethnic group in Vietnam. Because the data has just been released recently, only a few analyses have been made to study the sequenced data. Further studies on Vietnamese human genomes require a good and precise reference genome for Vietnamese. Such requirement leads to the need of building the Vietnamese Reference Genome (VNRG).

It is believed that data from 1KG also provide an
opportunity to study the human genetic diversity [10]. Although several new populations including Vietnamese have been sequenced, 1KG still cannot cover the genetic diversity in Southeast Asia [4]. Four years before 1KG released Vietnamese data, HUGO Pan-Asian SNP (PASNP) consortium publicized the map of Human Genetic Diversity in Asia [13]. The map includes 75 different populations; most of them come from 10 Asian countries (5 Southeast Asian countries). However, PASNP project did not take any sample from Vietnam leaving the lack of sufficient data for analyzing the relationship between Vietnamese and other Asian populations.

This thesis focuses on two goals. First of all, we would like to construct the VNDRG, which is closer to Vietnamese human genomes than the other references. Besides the first goal, we also want to perform an analysis regarding the genetic diversity on the data produced by 1KG and PASNP project. This study will only concentrate on Asian countries, especially the genetic diversity between Vietnamese Kinh and other ethnic groups in Southeast Asia.

The remaining of this article is organized as follows. The next section describes the details of all materials used in this thesis. Then in Section III, we present our proposed method to achieve the goals of this thesis. The significance of this thesis is shown in Section IV. Finally, our conclusion throughout this thesis is stated in Section V.

2. Materials

There are a total of five different data sources that had been used in this thesis: The raw read data, Omni data, the variant call set, the NCBI human reference genome, and PASNPdb.

First of all, in order to construct the VNDRG, we collected the raw reads of Vietnamese Kinh (encoded as KHV) from 1KG project. There is a total of 100 low-coverage Vietnamese individuals in the selected dataset. The raw reads are paired-end. Thus, each sample has two FASTQ files. Each individual in the dataset has a unique ID, which is generated by 1KG. The raw data coming from those 100 individuals was considered as the major contributor for constructing and validating the VNDRG.

Apart from sequencing human genome at large scale, 1KG also uses several microarrays to generate the highly reliable SNPs data. Since genotyping by using microarray is expensive, it only covers a subset of human genetic variations. 1KG resource provides data coming from Illumina’s Omni2.5 beadchip, one of the most powerful and advanced genotyping microarrays available. The Board Omni data consists of 121 Vietnamese individuals including 100 samples we retrieved from the raw data.

1KG has released a new variant call set of 2504 individuals in August of 2014. The new dataset consists of 13635194 variants of all 22 autosomal chromosomes. 99 Vietnamese Kinh individuals were included in the newly released data. Note that, the variants of sex chromosomes have not yet been called. The variant data were used not only for constructing the reference genome but also for studying the Vietnamese genetic diversity.

The NCBI human reference genome is the most commonly used human reference genome. It is also considered as a standard reference genome for many whole-genome sequencing projects. We retrieved the GRCh37 reference genome from NCBI and used it as a reference in constructing and evaluating the VNDRG.

Lastly, PASNPdb is the most detailed SNP database of Asia. It consists of 73 Asian populations with 1928 individuals, 54794 SNPs on autosomal chromosomes and 1216 SNPs on chromosome X. We used this database for measuring the relationship between Vietnamese Kinh and neighbor populations.

3. Methods

Before getting into the details of our method, let us denote:

- \( G_s \): The standard NCBI GRCh37 reference genome.
- \( G_c \): The VNDRG.
- \( G = \{G^1, \ldots , G^{100}\} \): The set of 100 Vietnamese genomes obtained from the 1000 Genomes projects where \( G^i \) is the genome for the \( i^{th} \) Vietnamese sample.
- \( G_1 = \{G^1, \ldots , G^{50}\} \): The set of 50 genomes used as the training data set to build the Vietnamese reference genome.
- \( G_2 = \{G^{51}, \ldots , G^{100}\} \): The set of 50 genomes used as the testing data set to assess the quality of the Vietnamese reference genome.

Our method for constructing the Vietnamese reference genome can be divided into two phases: building \( G_c \) and evaluating \( G_c \). 50 samples were selected to build the reference genome (\( G_1 \)) and 50 remaining individuals were selected to validate the reference genome (\( G_2 \)). The pipeline for building the reference genome and evaluating it will be separately discussed as follows.

3.1. Constructing the reference genome

Figure 1 shows the workflow that we used for building \( G_c \). The reference genome was built based on a hypothesis: If \( G_c \) is closer to Vietnamese genomes than the standard reference genome \( G_s \), it will have more population-specific alleles than the standard
Reference genome. Using the proposed hypothesis, at each allele position, we calculated the allele frequency of that allele. By calculating the allele frequency, we were able to identify the positions where the reference alleles are different from the majority alleles on the Vietnamese genomes. Finally, we altered the NCBI GRCh37 Reference Genome by replacing the all the identified positions with the alternative alleles.

Figure 2 illustrates the pipeline proposed by Board Institute for calling variants. We applied this pipeline for 50 samples that were selected for building the reference genome. In general, this pipeline can be divided into three sub-processes: processing the collected reads, calling variants, and evaluating the discovered variants. While the second and the third sub-process were done using GATK, the first sub-process required different tools for mapping the reads and evaluating the quality of the mapped reads.

In the first sub-process, the raw reads coming from 50 Vietnamese individuals were mapped against the standard reference genome using BWA-MEM. After that, for each sample, the duplicate reads were marked and removed by applying Picard tools. Note that the mapping result may contain errors. To tackle this problem, we used GATK for realigning and then recalibrating the mapped reads. At the end of this step, we retrieved a set of 50 BAM files, which correspond to 50 Vietnamese individuals. Each BAM file contained the analysis-ready reads for the next sub-process.

The raw variants were discovered by GATK. We used two different strategies for two sets of chromosomes: the autosomal chromosomes and the sex chromosomes. While variants on autosomal chromosomes were discovered by applying the HaplotypeCaller package of GATK, UnifiedGenotyper was used for discovering the variants on the sex chromosomes.

After the variants of all 50 Vietnamese individuals were discovered, they were then evaluated and filtered using VariantRecalibrator and ApplyRecalibration options of GATK. This step marked all the low-quality variants by applying several quality thresholds such as read depth, mapping quality, haplotype quality, etc.

After finishing the pipeline proposed by Board Institute, we computed the allele frequency of every high-quality SNP. The alternate alleles that have higher frequencies than the reference alleles were stored in the “majority allele set $S$”. Then this set was used for constructing the VNRG.

### 3.2. Evaluating the reference genome

In order to measure the effectiveness of the newly built reference genome, we used $G_2$ - the second set of Vietnamese individuals. Two different criteria were selected to compare the performance of $G_v$ and $G_s$: the short reads mapping and the genotype calling. The workflow for calculating those criteria were illustrated in Figure 3.

**Figure 3:** The workflow for evaluating the Vietnamese reference genome

The first criterion, short reads mapping quality was retrieved by applying BWA-MEM algorithm. The new genome set $G_2$ were mapped against both the newly constructed $G_v$ and the standard $G_s$. The process resulted in two different sets of alignments: $A_{1I}$ and $A_{2I}$ respectively. After mapping all the reads coming from 50 Vietnamese individuals, the quality of short reads mapping were calculated. Let $P$ be the set of positions of allele in $S$. A read is called “effective read” if it covers at least one position in $P$. Let $M_i$ be the set of effective reads that were mapped to $G_i$ and $M_e$ be the set of effective reads that were mapped to $G_e$. The mapping quality of $G_v$ is better than $G_s$ if the quality of reads in $M_v$ is better than that in $M_e$. To compare the mapping quality of those sets, three different thresholds
of Phred-scale quality were considered: 20, 40 and 60. For each threshold, we measured the percentages of effective reads that have higher quality than the selected threshold on both sets $M_v$ and $M_s$.

Using two generated alignment sets $A_1$ and $A_2$ we conducted genotype calling for all genomes in the testing dataset $G_2$. For each genome $G_i$ in $G_2$, we measured the genotype accuracy based on its Omni data. Let $O_i = \{O_{i1}, \cdots, O_{i100}\}$ represents the set of Omni SNP data where $O_{ij}$ is the Omni SNPs of $G_i$. Similarly, let $O_s = \{O_{s1}, \cdots, O_{s100}\}$ and $O_v = \{O_{v1}, \cdots, O_{v100}\}$ be the set of SNPs called from $G_2$ using the standard reference genome $G_s$ and Vietnamese reference genome $G_v$ respectively.

For every genotype set $O_s \in O_s$, we denoted:

- $c(i)_s$, the number of SNPs that were found in both genotype sets: $O_s$ and $O^i$.
- $w(i)_s$, the number of SNPs that were not the same in $O_s$ and $O^i$.
- $P_s = \frac{c(i)_s}{c(i)_s + w(i)_s}$, the precision of $O_s$ when comparing it with $O^i$.
- $r(i)_s = \frac{c(i)_s}{|O_s|}$, the recall of $O_s$.
- $f(i)_s = 2 \times \frac{P_s \times r(i)_s}{P_s + r(i)_s}$, the “f-score” of $O_s$.

To generalize the evaluation system, we computed:

- $P_s = \frac{\sum c(i)_s}{\sum (c(i)_s + w(i)_s)}$ where $i = 51..100$: The precision of the set $O_s$.
- $R_s = \frac{\sum c(i)_s}{\sum O_s}$ ($i = 51..100$): The recall of the set $O_s$.
- $F_s = 2 \times \frac{P_s \times R_s}{P_s + R_s}$: The F-score of $O_s$.

Applying the identical procedure, we retrieved the precision $P_v$, the recall $R_v$, and the F-score $F_v$ of $O_v$.

A SNP $t$ at location $l_t$ on the genome is called “k-distance” from the majority allele set $S$ if and only if there exists at least one SNP $t' \in S$ at location $l_{t'}$ such that $|l_t - l_{t'}| \leq k$. A subset $O(k)$ of Omni genotype data $O$ is considered “k-distance” from $S$ if every SNP in $O(k)$ is “k-distance” from $S$. We compared the genotype quality of $O_v$ with $O_s$ by taking 12 subsets of $O$ for calculating the precision and recall values. Each subset is “k-distance” from $S$ where $k \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 500, 1000\}$.

3.3. Discovering the genetic diversity of Vietnamese Kinh

After integrating variant data coming from 1KG into PASNPdb, we used Treemix program [11] to reconstruct the phylogenetic tree of PAN-Asian populations. YRI - the African population was picked as the root of the built tree. Several tests have been conducted to measure the stability of the tree topology. For the final produced tree, we assumed that 5 SNPs are grouped together (treemix option: $-k$ 5). We also used the bootstrap option to validate the tree (treemix option: $-\text{bootstrap}$).

EIGENSOFT [9, 12] was used to study the ge-
genic diversity between Vietnamese Kinh and other Asian populations. We conducted a number of PCA tests using different subsets of the full genotype data to identify the closely related populations to Vietnamese Kinh. The first subset contains every available Asian populations. The second subset is the combination of Southeast Asian individuals. And the final subset is the combination of some populations that are geographically close to Vietnam.

4. Experimental results
4.1. Evaluation of the proposed method for constructing the reference genome
The comparison of mapping quality and genotype quality between VNRC and NCBI GRCh37 on chromosome 2 (one of the biggest chromosomes in the human genome) will be discussed in this section.

Using BWA for mapping the raw reads of 50 Vietnamese individuals against two references, we received the differences on the average number of matched bases per read (see Table 1). Note that, the “Improvement” columns were retrieved by subtracting the average number of matches/read on GRCh37 by that on VNRC. According to Table 1, there were improvements on every testing sample. For example, the number of matched bases per read of HG02084 increased from 99.19 bases to 99.76 bases. When millions of reads are taken into account, this difference could lead to a noteworthy enhancement to the alignment quality. The results also implied that VNRC helped increasing the accuracy of BWA-MEM. As a consequence, the new reference genome could be more precise than GRCh37 when calling variants for Vietnamese genomes.

Table 2: The mapping quality of 50 Vietnamese genomes at the altered positions on chromosome 2

<table>
<thead>
<tr>
<th># reads/sample</th>
<th>( G_s )</th>
<th>( G_v )</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.12/06/30</td>
<td>98.56 %</td>
<td>99.24 %</td>
<td>99.76/69</td>
</tr>
<tr>
<td>Phred-scale ≥ 20</td>
<td>97.90 %</td>
<td>95.48 %</td>
<td>99.95/60</td>
</tr>
<tr>
<td>Phred-scale ≥ 40</td>
<td>97.90 %</td>
<td>95.48 %</td>
<td>99.95/60</td>
</tr>
<tr>
<td>Phred-scale ≥ 60</td>
<td>92.18 %</td>
<td>95.22 %</td>
<td>99.95/60</td>
</tr>
</tbody>
</table>

The qualities of the reads mapped at the altered position also showed that VNRC is better than NCBI GRCh37 in term of mapping Vietnamese short reads. This were illustrated in Table 2. As indicated by this table, the altered positions on VNRC were covered by approximately 1333723 reads in average that was 6157 reads more than the standard reference genome performance. Moreover, VNRC had better results than GRCh37 on all Phred-scale thresholds (20, 40 and 60). This pattern indicated the improvement on the quality of 50 alignments that correspond to 50 testing Vietnamese individuals.

Using the genotype results on chromosome 2 of 50 testing genomes, we made a comparison between VNRC and GRCh37 on genotype calling quality criterion. Table 3 showed the number of genomes in which VNRC had better performances than the standard reference genome on three different criteria: precision, recall, and f-score. As reported by Table 3, VNRC outperformed GRCh37 in almost all of the cases. In fact, the newly built reference was worse than the standard reference in only one Vietnamese sample: HG02070.

Generalizing the results from 50 individuals, we computed the precisions, recalls, and f-scores for the whole testing genome set. It was clearly shown in Table 4 that the VNRC outran NCBI GRCh37 in all twelve different K-distance subsets of Omni data. It is reasonable to see that the differences of f-score decreased when the value of \( K \) increased. This pattern happened because increasing \( K \) widens the radius for searching K-distance SNPs. In another way, it increases the number of SNPs in K-distance subset of Omni data. Therefore, it will narrow the distance between \( F_f \) and \( F_v \).

Table 3: Overall comparison of genotype calling quality between VNRC and GRCh37 on chromosome 2 of 50 Vietnamese individuals

<table>
<thead>
<tr>
<th>( K ) (K-distance)</th>
<th>#(( v_s &gt; p_s ))</th>
<th>#(( v_v &gt; p_v ))</th>
<th>#(( f_s &gt; f_v ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>K=10</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>K=20</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>K=30</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>K=40</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>K=50</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>K=60</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>K=70</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>K=80</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>K=90</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>K=100</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>K=150</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
</tbody>
</table>

4.2. The Vietnamese Reference Genome
The histogram of the frequencies of reference alleles was shown in Figure 4. As can be seen from the Figure, most of the reference alleles had their allele frequencies higher than 0.5. For instance, more than 4.5 million SNPs with reference allele frequency higher than 0.995 were found suggesting that those detected SNPs were extremely rare, only a small fraction of Vietnamese has those variants. However, the reference alleles with frequencies lower than 0.5 still existed. Notably, we found more than 300 thousand of SNPs in which the reference alleles did not appear (the reference allele frequency = 0). It meant that none of the Vietnamese individuals had those reference alleles. As a consequence, they were removed from the VNRC.

Using the data coming from 1KG, over 2 million of alleles were detected in the majority allele set. Figure 4 illustrates the number of alleles found on 24 chromosomes. It is very easy to find that chromosome Y had the least number of alleles; only 1483 alleles were found. This pattern happened because of two reasons. First of all, chromosome Y is one
of the smallest chromosomes in the human genome. Because of that, it is highly possible that the number of SNPs found on chromosome Y is lower than that on other chromosomes. Secondly, not all of the samples collected by 1KG are male. Consequently, it is harder to identify the high-quality SNPs on chromosome Y. As a result, there were fewer alleles on chromosome Y in S than on other chromosomes.

Moving to the autosomal chromosomes and chromosome X, the number of alleles in S seemed to relatively follow the size of the chromosome. On the one hand, the two biggest chromosomes (chromosome 1 and 2) had the highest number of alleles (191406 and 198585 respectively). On the other hand, chromosome 21 and chromosome 22 were found to have the fewer number of alleles than the other chromosomes. Further analyses also suggested that chromosomes with the same length tended to have approximately the same number of alleles in S.

Table 1: The average number of matched bases per read of 50 Vietnamese individuals on chromosome 2.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>$G_{m}=# of matches/\text{read}$</th>
<th>$G_{v}=# of matches/\text{read}$</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>HG02032</td>
<td>98.01</td>
<td>98.58</td>
<td>0.57</td>
</tr>
<tr>
<td>HG02035</td>
<td>97.81</td>
<td>98.50</td>
<td>0.58</td>
</tr>
<tr>
<td>HG02040</td>
<td>97.87</td>
<td>98.44</td>
<td>0.57</td>
</tr>
<tr>
<td>HG02045</td>
<td>97.86</td>
<td>98.43</td>
<td>0.57</td>
</tr>
<tr>
<td>HG02047</td>
<td>97.90</td>
<td>98.48</td>
<td>0.58</td>
</tr>
<tr>
<td>HG02048</td>
<td>98.08</td>
<td>98.64</td>
<td>0.57</td>
</tr>
<tr>
<td>HG02049</td>
<td>98.14</td>
<td>98.72</td>
<td>0.58</td>
</tr>
<tr>
<td>HG02050</td>
<td>98.06</td>
<td>98.65</td>
<td>0.59</td>
</tr>
<tr>
<td>HG02057</td>
<td>98.06</td>
<td>98.64</td>
<td>0.57</td>
</tr>
<tr>
<td>HG02058</td>
<td>98.32</td>
<td>98.86</td>
<td>0.54</td>
</tr>
<tr>
<td>HG02060</td>
<td>88.16</td>
<td>88.76</td>
<td>0.60</td>
</tr>
<tr>
<td>HG02061</td>
<td>88.14</td>
<td>88.71</td>
<td>0.57</td>
</tr>
<tr>
<td>HG02064</td>
<td>87.27</td>
<td>87.85</td>
<td>0.57</td>
</tr>
<tr>
<td>HG02067</td>
<td>87.45</td>
<td>88.02</td>
<td>0.57</td>
</tr>
<tr>
<td>HG02069</td>
<td>97.92</td>
<td>98.51</td>
<td>0.58</td>
</tr>
<tr>
<td>HG02070</td>
<td>98.09</td>
<td>98.67</td>
<td>0.57</td>
</tr>
<tr>
<td>HG02072</td>
<td>98.14</td>
<td>98.70</td>
<td>0.56</td>
</tr>
<tr>
<td>HG02073</td>
<td>98.03</td>
<td>98.59</td>
<td>0.56</td>
</tr>
<tr>
<td>HG02075</td>
<td>98.99</td>
<td>99.58</td>
<td>0.59</td>
</tr>
<tr>
<td>HG02076</td>
<td>98.92</td>
<td>99.49</td>
<td>0.57</td>
</tr>
<tr>
<td>HG02078</td>
<td>99.16</td>
<td>99.76</td>
<td>0.60</td>
</tr>
<tr>
<td>HG02079</td>
<td>99.03</td>
<td>99.63</td>
<td>0.60</td>
</tr>
<tr>
<td>HG02081</td>
<td>98.06</td>
<td>98.63</td>
<td>0.57</td>
</tr>
<tr>
<td>HG02082</td>
<td>99.01</td>
<td>99.58</td>
<td>0.57</td>
</tr>
<tr>
<td>HG02084</td>
<td>99.19</td>
<td>99.76</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Table 4: Genotype calling quality on chromosome 2. The last column shows the average number of K-distance SNPs from the majority allele set $S$ on Omni chip.

<table>
<thead>
<tr>
<th>$GRCh37 - O_{s}$</th>
<th>$VNRG - O_{s}$</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K=10$</td>
<td>$F_{G} - F_{s}$</td>
<td>$\langle O \rangle$</td>
</tr>
<tr>
<td>$K=20$</td>
<td>$F_{G} - F_{s}$</td>
<td>$\langle O \rangle$</td>
</tr>
<tr>
<td>$K=30$</td>
<td>$F_{G} - F_{s}$</td>
<td>$\langle O \rangle$</td>
</tr>
<tr>
<td>$K=40$</td>
<td>$F_{G} - F_{s}$</td>
<td>$\langle O \rangle$</td>
</tr>
</tbody>
</table>

Figure 4: The histogram of allele frequency of reference allele on 24 chromosomes.

Figure 5: The distribution of majority allele set on 24 chromosomes.
All of the alleles in $S$ were used for altering the standard reference genome GRCh37. By modifying those locations on GRCh37, we successfully constructed the first VNRG.

4.3. Integrating 1KG data into PASNPdb

To study the population structure of the Asian ethnic groups and particularly in Vietnamese Kinh, we used the combination of two different data sources: PASNPdb and 1KG variant data.

Combining two data sources is not straightforward because of the differences in genotyping technologies and references. Therefore, we decided to use only the annotated SNPs that exist in both sources. As the result, the final data compromised 76 different populations with 2027 individuals and 49835 SNPs. 4 populations: YRI, CEU, JPT, and CHB belong to Hapmap project. They were used to verify the result of any analysis that uses the genotype data. For instance, Chinese populations must have characteristics similar to that of CHB.

Figure 6 illustrated the locations of all populations in the final dataset. It is very easy to find that most of the data come from the Southeast Asia. A small fraction of data comes from the East Asia and South Asia. Only two non-Asian populations were used, one African population (YRI) and another one population with ancestry from Northern and Western Europe (CEU).

The dataset can also be clustered by language family. There were 10 different language families in total; all of them were highlighted in different colors in Figure 6. Some of the families can be clearly clustered using the geographic map, but there were locations where many different language families coexist. On the one hand, Austronesian and Indo-European can be easily detected using the geographic properties. The Austronesian speaking populations can be found at the South of Southeast Asia, and the Indo-European speaking groups stay in India. On the other hand, Mainland Southeast Asia compromises a very complex ethnic pattern. Particularly, in a small region in the North of Thailand, 4 different language families were found.

4.4. The phylogenetic tree of Asian populations

Figure 7 shows the full maximum likelihood tree of the integrated data. Analyzing the tree, we found that the populations that speak the same language family tended to stay in the same cluster. Furthermore, the phylogenetic tree also revealed many migration events. For instance, the Melanesians (AX-ME) were found to have close relationships with the Indonesians. It suggested that those Indonesian populations may share a common ancestor with AX-ME and AX-ME were the result of a migration event coming from Indonesia. Most of Chinese populations were detected as the descendants of Thai populations. The source of this migration started from the North of Thailand, heading East and Northeast. Korean and Japanese can be seen as the results of a migration coming from Beijing, going through Korean Peninsula, and then finally ended up in Japan.

4.5. The genetic diversity of Vietnamese Kinh

By removing YRI and CEU from the full dataset, we created a subset that contained only Asian countries. Figure 8 illustrated all the Asian populations and their connections using the two best principal components generated by EIGENSOFT for this dataset. Analyzing the results from Figure 8, we found that there were a distance between Indians and the other populations. Only one group of Singaporean (SG-ID, an Indian ethnic group in Singapore), and the Uyghur (CN-UG, a Chinese ethnic group that lives close to India) were found to have the overlapping pattern with the Indian populations. Two Malaysian populations (MY-KS, MY-JH) were also found in an isolated cluster. This pattern explained why in the maximum likelihood tree (Figure 7) those populations were not clustered with other Austro-Asiatic populations. We also found the overlapping region between Vietnamese Kinh, Chinese, and Thai populations suggesting the close genetic relationship between them. The Korean and Japanese were found in the top right corner of the scatter plot implying the correlation between those two populations and the diversities between them and the Southeast Asian populations.

Moving to the Southeast Asia in Figure 9, Vietnamese Kinh was found in a very small and dense region along with Thai populations. This phenomenon supports the result we received from the maximum likelihood tree where we found Vietnamese Kinh belonged to the same clade with many Thai and Chinese populations.

Most of the Indonesian populations did not have
the genetic relationship with Vietnamese Kinh according to the PCA result. The separation suggested that the ancestors of Vietnamese Kinh and Indonesian were separated in the early day creating two different sets of genotypes for each population. SG-CH was found to have overlapping with many KHV individuals. This phenomenon could be explained by the genetic relationship between Chinese and Vietnamese Kinh found in Figure 8. Since SG-CH is a group of Chinese in Singapore, they could also inherit some characteristics that are similar to that of KHV.

After analyzing two subsets, there was an agreement between the constructed phylogenetic tree and the PCA results in which Vietnamese Kinh was found to have the genetic connections with some Thai and Southern Chinese populations. To reveal those connections, we created the last subset for the PCA analysis. This data compromised 5 populations in the North of Thailand, 5 populations from China (mostly in South China, except CHB), 4 populations from Taiwan and the last population was KHV. We did not consider other ethnic groups from other countries because according to the previous result, there was no strong genetic connection between KHV and the ethnic group coming from those countries. The result of this subset was shown in Figure 10(a). With the appearance of Thai populations and Taiwanese populations, CHB formed a cluster with TW-HA, TW-HB and CN-SH. This cluster was completely isolated with the cluster that compromised KHV. The similar patterns can be applied for CN-HM and TH-YA. The two native Taiwanese ethnic groups also formed two isolated clusters according to the result.

After analyzing two subsets, there was an agreement between the constructed phylogenetic tree and the PCA results in which Vietnamese Kinh was found to have the genetic connections with some Thai and Southern Chinese populations. To reveal those connections, we created the last subset for the PCA analysis. This data compromised 5 populations in the North of Thailand, 5 populations from China (mostly in South China, except CHB), 4 populations from Taiwan and the last population was KHV. We did not consider other ethnic groups from other countries because according to the previous result, there was no strong genetic connection between KHV and the ethnic group coming from those countries. The result of this subset was shown in Figure 10(a). With the appearance of Thai populations and Taiwanese populations, CHB formed a cluster with TW-HA, TW-HB and CN-SH. This cluster was completely isolated with the cluster that compromised KHV. The similar patterns can be applied for CN-HM and TH-YA. The two native Taiwanese ethnic groups also formed two isolated clusters according to the result.

The cluster that consisted of KHV individuals was shown in Figure 10(b). This cluster contained 7 populations in total: 4 from Thailand, 2 from China, and one from Vietnam. Clearly, all of them located very close to each other in the phylogenetic tree. Among 7 populations, KHV was the only population that uses Austro-Asiatic language. The remaining 6 populations belonged to Tai-Kadai speaking group. Moreover, KHV was almost separable from other groups, only a small part of KHV was overlapped. This phenomenon indicated that the genetic distances between KHV and other populations were still available.

5. Conclusions and future work
In the first part of this thesis, we presented a method for constructing and evaluating a population-specific genome. Taking the advantage of 100 Vietnamese genomes sequenced at low coverage from 1000 genomes project, we demonstrated the significance of the proposed method. The experiment results on two chromosomes: chromosome 2 and chromosome 20 showed that the newly constructed Vietnamese reference genome not only improved the mapping quality of short reads, but also
Figure 7: The Asian phylogenetic tree was constructed by Treemix using 76 different ethnic groups and over 49000 SNPs. YRI was selected as the root of the three. The tree was highlighted according to the language families.

significantly enhanced the genotype calling quality of 50 Vietnamese individuals in the testing dataset. Because the selected genomes that we used in this thesis are unrelated, the proposed method could become a generic method for assembling the population-specific genome.

Applying the method we proposed into the data recently released by 1000 genomes project, we successfully constructed the first Vietnamese reference genome. By substituting over 2 million alleles on the NCBI GRCh37, the assembled genome is expected to be closer to Vietnamese genomes than the original one. Additionally, the Vietnamese reference genome could become a baseline for many other Vietnamese-related genome-wide studies.

In the second part of this thesis, we performed a population study for Vietnamese Kinh - the most populous ethnic group of Vietnam. By integrating the variant database of Vietnamese Kinh with PAN-Asian SNP database, we were able to construct a maximum likelihood phylogenetic tree of 76 different populations (74 of them are Asian ethnic groups). The phylogenetic tree suggested the genetic connections between Vietnamese Kinh and
Figure 10: (a) The PCA results showed the relationship between KHV and some populations that are geographically close to Vietnam. (b) A closer look at the cluster that has KHV in (a) several Thai and South Chinese populations. A similar phenomenon was also indicated in the principal component analyses indicating the genetic similarities between them.

Although the Vietnamese reference genome was successfully built, improvements are still needed on the whole genome, especially on chromosome Y. In 100 Vietnamese genomes released by 1000 genomes project, only 45 of them have chromosome Y. The lack of sufficient data on chromosome Y in the database has massively reduced the quality of the discovered alleles, and the number of alleles in the majority allele set. Increasing the number of Vietnamese samples, particularly male samples could help building a more precise and more detailed reference genome for Vietnamese.

PAN-Asian SNP database is known as the most detailed genotype data for Asian. However, it cannot demonstrate the whole picture of Asian populations. In the Southeast Asia region, even when being integrated with genotypes of KHV in 1000 genome projects, the database still lacks the data coming from Cambodia and Laos. Because of that, we could not determine the origin of Vietnamese Kinh in the second part of our thesis. This problem could be solved by integrating extra genotype data that covers every region that is geographically close to Vietnam. Once the extra data is collected, the same procedure that we used could be re-applied to measure the genetic relationship between Vietnamese and other populations.

Project contributors
The performance of the Vietnamese Reference genome was partly tested by the author including chromosome X, chromosome Y, chromosome 13, chromosome 14, chromosome 15. The rest of the reference genome was evaluated by other members of the research group from the University of Technology under Vietnam National University (VNU).

The analyses on the genetic diversity of Vietnamese Kinh were fully conducted by the author of this article.

References


