Analyzing Population Genetics Using the Mitochondrial Control Region and Bioinformatics

Takumi Sato, Bonnie Phillips,Sandra M. Latourelle, Nancy L. Elwess [1]

[1]Department of Biological Sciences, State University of New York at Plattsburgh, 101 Broad St, Plattsburgh, NY 12901

Email: nancy.elwess@plattsburgh.edu

Abstract: The 14-base pair hypervariable region in mitochondrial DNA (mtDNA) of Asian populations, specifically Japanese and Chinese students at Plattsburgh State University, was examined. Previous research on this 14-base pair region showed it to be susceptible to mutations and as a result indicated direct correlation with specific ethnic populations. Earlier studies provided the 14-base pair region sequence analysis for Asians in general. This inquiry-based project was generated in a junior/senior college general genetics course. The project produced sequences of the 14-base pair hypervariable region for Japanese and Chinese populations, specifically. The investigation examined 28 Japanese and 27 Chinese international students. A Control group, consisting of 20 random students was also sampled. The 14-base pair hypervariable region for each sample was analyzed through comparison using bioinformatics. Analyzing the samples showed that, overall, 46% of the Japanese samples, 59% of the Chinese samples, and 75% of the Control samples had the same 14-base pair sequence. In addition to this, it was observed that the Japanese, Chinese, and Control group all had their own specific mutations within the 14-base pair hypervariable region.

Keywords: DNA, mitochondria, hypervariable region, bioinformatics, data mining

Introduction

In the summer of 2001, the authors of this journal article were chosen to attend the Vector Bioinformatics Workshop held at the Trudeau Institute in Saranac Lake, N.Y (http://www.dnalc.org/ddnalc/about/annreppdf/annrep2001.pdf, 2001). The workshop was held under the auspices of the Dolan DNA Learning Center (DNALC) located in Cold Spring Harbor, NY and was funded through a Howard Hughes Grant. The five day workshop focused on analyzing patterns in DNA sequences, online algorithms helpful in identifying gene features and using genome browsers to find genes in online databases. Participants were taught to recognize chromosome locations, identify homologs in other organisms, and explore their involvement in normal and disease processes. The workshop also centered on the use of the DNALC's custom *Bioservers*. Participants provided a sample of their own mitochondrial DNA, prepared the sample (http://www.geneticorigins.org/mito/intro.html), using Polymerase Chain Reaction (PCR), and submitted it to the DNALC for sequencing. This professional development opportunity subsequently opened the doors to a variety of learning experiences that could be incorporated into an undergraduate genetics' curriculum. One such opportunity is the focus of this article.

The Nature of Science is such that as one slowly collects information, formulates hypotheses, and explores testable possibilities, the end result is always another fork in the road. Answering one question only creates many more questions. Two General Genetics students did not stop at the knowledge of their own mitochondrial control region sequence of DNA returned to them by the DNALC *Bioserver*. Once the Vector Bioinformatics Workshop was completed, participants were allowed to submit their students' DNA samples for the sequencing of the mtDNA control region. As a result, undergraduate genetics students were able to use their *own* DNA sequences to investigate some population genetics questions but the answers still did not satisfy the curiosity factor. The student team found that reading some related journal articles (Anderson et al., 1981; Horai and Hayakawa, 1990; Horai et al., 1993; Lewis et al., 2007) presented information about a 14 base pair hypervariable section (found between mitochondria bases 16180-16193) within the control region of the mtDNA

.

The team's subsequent research endeavor was designed to compare Japanese students, Chinese students, and a control group (no Asians were included), using the garnered14-base pair hypervariable region within mitochondrial DNA (mtDNA). It was hypothesized that Chinese and Japanese students would have a 14-base pair region that was more similar than those of the control group. Additionally, the team wanted to know if testing specific Asian populations (Japanese and Chinese) would support the analyzed data published in the Horai et al. (1993) literature that investigated Asians as a whole entity. It should be noted that one of the student team members was from Japan.

Research was initiated after reading the previously mentioned journal articles that analyzed the 14-base pair region of different populations of people, including Asians, Africans, Native Americans and Europeans. For each of these groups, a common 14-base pair sequence was identified (Horai and Hayasaka, 1990; Horai et al., 1993; Lewis et al., 2007). In the Horai et al., (1993) literature, Asians were tested (along with Europeans, Africans and Native Americans). However, Asian groups were not specified (Japanese, Chinese, or Korean, etc). Noting this, the team chose to test DNA samples taken from two specific Asian groups, Japanese and Chinese, to compare the 14-base pair region. There is a high population of International students on the State University of New York at Plattsburgh campus (7-8% of total student population). A vested interest in the investigation outcome (Japanese team member) and availability of test subjects allowed for the study to be done.

Using the campus student population and under the auspices of the Committee on the Protection of Human Subjects (COPHS), DNA samples were obtained from 28 Japanese international students, 27 Chinese international students, and 20 students from within the genetics class, the latter being used as the Control group. The Chinese, Japanese, and Control group samples were randomly collected (again with proper, signed consent forms). Both the Chinese and Japanese students represented many regions across China and Japan (Figures 1 and 2).

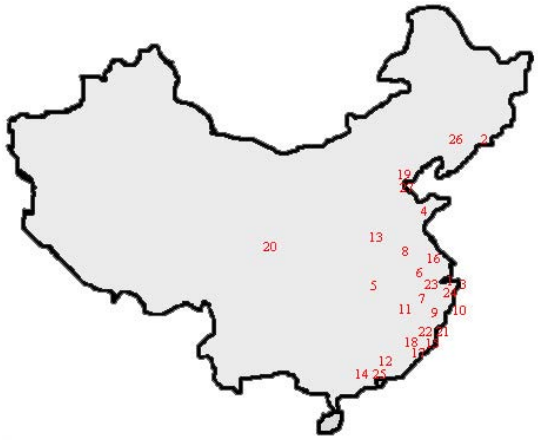

Figure 1. Map of China; the numbers indicate original r
esidence of the International Chinese students from whom DNA was collected.

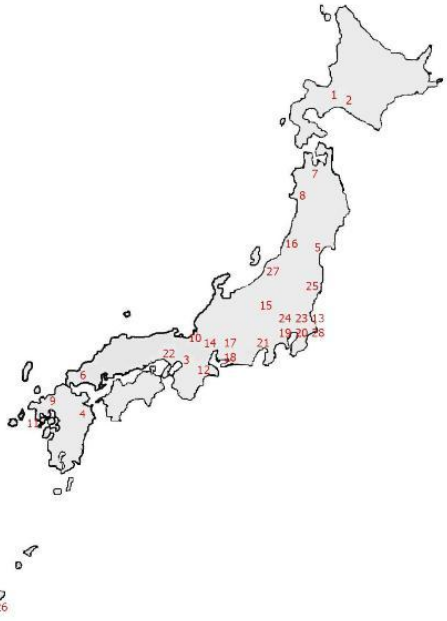

Figure 2. Map of Japan; the numbers indicate the original residence of the International Japanese students
from whom the DNA samples were collected.

The focal point for the research was the 14-base pair hypervariable region found within the mitochondrial DNA control region (Figure 3). This 14-base pair sequence is supported as being

hypervariable, since it is more susceptible to mutations, including insertions, deletions, transversions, and transitions (Walker et al. 2003). These point mutations or single nucleotide polymorphisms (SNPs) accumulate at a rate 10 times that of nuclear DNA (Walker et al. 2003).

Employing protocols from both molecular genetics and bioinformatics, DNA sequence comparisons of the 14-base pair region were done.

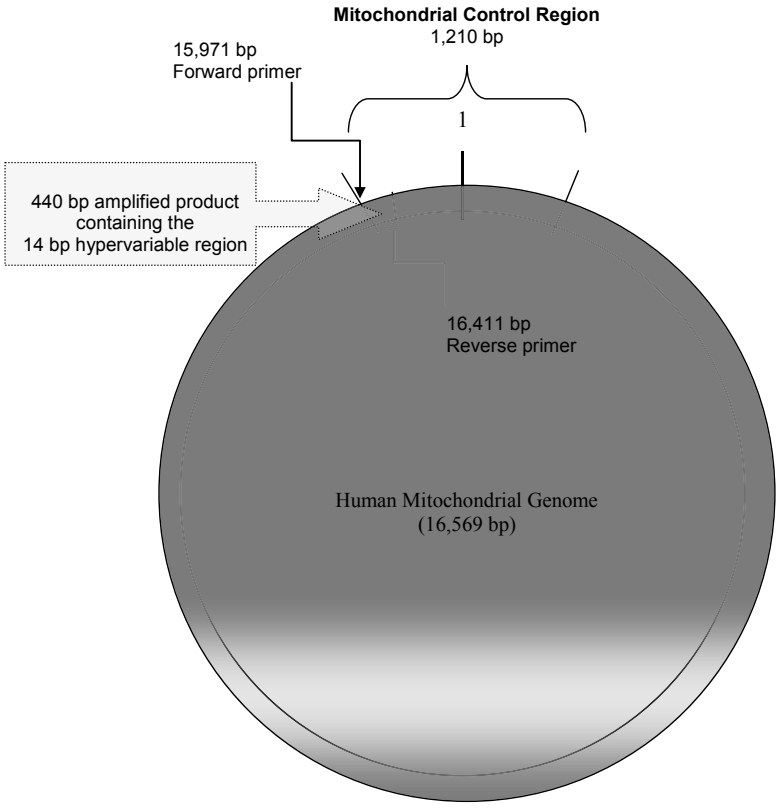

Figure 3. Human mitochondrial DNA map including the targeted hypervariable region. A 440 base pair (bp) region was amplified within the mitochondrial control region. The locations of the Forward (5'-TTAACTCCACCATTAGCACC-3') and Reverse (5'-GAGGATGGTGGTCAAGGGAC-3') primers are indicated. The researched 14 bp hypervariable region is located within the 440 bp amplified region. Graphic produced and edited by Sharon Clarke.

Material and Methods:

DNA isolation and amplification were completed using the Cold Spring Harbor Dolan Learning Center protocol (http://www.geneticorigins.org/mito/mitoframeset.htm). Briefly, each participant swished 10 mL of saline (0.9% NaCl) for 30 seconds; this solution was expelled into a collection tube. Each tube was carefully numbered and labeled accordingly (for example: Japanese 1, Japanese 2, Chinese 1, Control 1, etc..). One mL of each sample was spun in a microcentrifuge for one minute to concentrate the cheek cells within the saline solution. Once concentrated, the excess saline solution was poured off and the pellet of cheek cells was suspended in 100 µl of a Chelex solution. This mixture was boiled for 10 minutes. After boiling, each sample was spun in a microcentrifuge for 1 minute. Thirty microliters were drawn from the top of each spun sample and placed into the labeled tubes. These served as the DNA source for all of the samples. Amplification

was done according to the above mentioned protocol with the following exceptions:

- -15.0µl dH_2O, 2.5µl of 20µM forward primer, 2.5µl of 20µM reverse primer and 5µl of human DNA were added to the Ready to Go™ PCR tube. This replaced the 22.5µl pre-mixed (water, primer mixture) and 2.5µl of DNA.
- The thermal cycler was programmed for 35 cycles instead of 30 cycles.
- The annealing temperature was set to 54°C rather than 58°C.

Following DNA amplification by Polymerase Chain Reaction (PCR), 10µl of 25µl (stock) sample were removed for gel electrophoresis. Then 2.5µl of a 5x loading dye were added to each sample. The samples were then loaded into a 2% agarose gel. Each gel also received 10µl of a 100 base pair standard. Once a 440 base pair band was visually confirmed in the agarose gel for each sample, the stock sample was used to retrieve another 10µl amount which was then sent to the Dolan DNA Learning Center (according to their instructions found at: http://www.geneticorigins.org/mito/mitoframeset.htm) at Cold Spring Harbor, NY to be sequenced.

The sequenced samples were accessed using the Dolan Learning Center database (http://www.bioservers.org/html/sequences/sequences). Analysis required locating the 14-base pair hypervariable region within the 440 sequenced bases. The 14-base pair hypervariable region immediately follows the sequence CACATC, which is located approximately 182 bases within the 440 sequenced bases.

Sequence comparisons were determined using CLUSTAL W (http://align.genome.jp/). This program determined the number of nucleotide differences amongst the sequences.

The phylogenetic tree was created using on-line San Diego Supercomputer Center (SDSC) Biology Workbench program (http://workbench.sdsc.edu/).

Results:

From the 75 mtDNA samples analyzed, there were 23 different 14-base pair sequences (Table 1). The conserved 14-base pair hypervariable region was found in 46% of the Japanese population, 59% of the Chinese population, and 75% of the control group. This conserved region is represented by AAAACCCCCTCCCC (Table 1). In addition to this, there were two other sequences (AAAACCCCCCCCCC, AAAACCCTCCCCCC) found in all three sample groups, although they occurred in far less frequency than the conserved region. It was also noted that each population had unique mutations within the 14-base pair sequence. The Japanese sample group had twelve other 14-base pair sequences in addition to the conserved one, the Chinese samples had six, and the Control group had two additional and distinctive 14-base pair sequences (Table 1).

Table 1. Observed 14 base pair hypervariable sequences found within the tested Japanese, Chinese, and Control samples

		Number of Subjects			
	14 bp mtDNA Sequences	**Control**	**Japanese**	**Chinese**	**Total**
1	AAAACCCCCTCCCC*	15	13	16	44
2	AAAACCCCGTCCCC	1			1
3	AAAACCCCCCCCCC	2	1	2	5
4	AACACCCCGCCCCC	1			1
5	AAAACCCTCCCCCC	1	2	1	4
6	AACCCCCCCCCCC		1		1
7	AAAACCCCCCTCCCCC		1		1
8	AAACCCCCCTCCCC		1		1
9	AAAACCCCCTCGCC		1		1
10	AAACCCCCTCCCG		1		1
11	AACCCCCCCCCCCCC		1		1
12	ACAACCCCCTCCCC		1		1
13	AAACACCCCTCTC		1		1
14	AAAAACCCCCCCCCCC		1		1
15	AAATAACCCCCCCTCC		1		1
16	AAATCCCCCCCTCC		1		1
17	AAACCCTC		1		1
18	AAACCCCCCCCCCCC			2	2
19	AAAACCCCCTCCTC			2	2
20	AAAACTCCCCCCC			1	1
21	AAACCCCCCCCCCC			1	1
22	AAAATCCCCCCCCC			1	1
23	AACCCCCCCCCCCC			1	1
	Total	**20**	**28**	**27**	**75**

* conserved sequence.

It was noted that the Japanese group had the greatest number (48) of total mutations (Table 2). In addition, they also had the greatest number of mutations in each of the categories (Table 2). This included insertions, deletions, transversion, and transition point mutations. All three groups tested had a variety of point mutations. Finally, the Control group had no deletions and insertions, and just a small number of transversion and transition point mutations. The most common mutation for all three groups was a transition from a T to a C (Table 2).

Table 2. Observed number of nucleotide substitutions, deletions and insertions found with the three tested populations.

Mutation	Control	Japanese	Chinese
Transition			
T→C	4	11	7
C→T	1	6	4
Total	5	17	11
Transversion			
C→G	2	2	0
C→A	0	3	0
A→C	1	4	0
Total	3	9	0
Deletion	0	0	0
A→_	0	6	4
C→_	0	4	
Total	0	10	4
Insertion			
_→A	0	1	0
_→T	0	1	0
_→C	0	10	7
Total	0	12	7

Discussion

The experimental results supported the data in the published literature. The experimental conserved 14-base pair region was very close to the published conserved 14-base pair region (Horai et al., 1993). There were two sequences, in addition to the conserved sequence, that were common in all three sampled groups (AAAACCCCCCCCCC and AAAACCCTCCCCCC in Table 1).
The Japanese samples were seen to have a higher mutation rate than the Chinese or Control group sequences and as a result had a greater variety of 14-base pair sequences (Table 1). This may be due to the fact that most of the Japanese samples were from students from heavily populated regions of Japan (including Tokyo). It is a possibility that pollution, or other epigenetic factors associated with heavily populated areas could affect the DNA by causing mutations. Also, it should be noted that a number of the Japanese students are descendents of family members who lived close to where the atomic bombs were detonated during World War II. The radiation could have induced additional mutations. However, neither one of these hypotheses were supported through this research.

The Chinese samples were also observed to have a higher mutation rate than the Control group. This could be due to the fact that a majority of the individuals sampled came from the eastern part of China where industrial development is heavy. Here again, pollution could be an epigenetic factor, and may have caused mutations in the DNA sequence.

As mentioned, part of the research findings were very close to those in the Horai et al. (1993) publication. Both studies had the same conserved 14-base pair sequence. In the Horai et al. (1993) report, 55.8% of the combined Asian and Control samples had the conserved sequence, while this investigation found the conserved 14-base pair sequence in 58.7% of the samples tested (Table 1). In the Horai report, investigators found 19 different 14-base pair sequences within the mitochondrial hypervariable region, while this study ended with 23 different

14-base pair sequences (Table 1). The additional four sequences were found within the Japanese test group. Out of the 23 different sequences found, 12 of these sequences were found in the Japanese population alone.

Educational Components:

This investigation indicated the usefulness of bioinformatics in order to show both relationships and difference in the sequences tested. This included the CLUSTAL program which showed conserved regions and another program that can be used (but was not included in this article) was a phylogenetic tree program which shows relationships of DNA sequences. Also, the experiment demonstrated how a research project in a college genetics course can be built on previous research and publications for the purpose of comparing experimental results to those already published (Horai et al, 1993). All of these mentioned comments address the Nature of Science:

- Ask a question
- Explore and discover
- Test ideas
- Analyze the outcomes
- Consider the next step

The 14 base pair hypervariable region is one of the easiest DNA sequences to acquire mutations. From the results, the sequences showed similarities of different ethnic groups at the genetic level, especially similarities between the Japanese and Chinese individuals. In addition to isolating and amplifying the DNA, it was also important for students to analyze and present their data which they did at both a state and national conference.

Not all biology curricula can provide students an opportunity to isolate and amplify DNA, but there is still an opportunity to engage students in data mining. This allows them to compare sequences already in databases using bioinformatics. For example, the 14-base pair hypervariable region studies in this inquiry based experience can be examined within other ethnic groups found listed in the Dolan DNA Learning Center database. Within this database alone are sequences from African, European and Native American populations for the mitochondrial control region. Comparisons can be made using the same approach as was done in this study. Even though one may not have the opportunity to conduct a research project, using literature and present databases for data mining can still provide a very meaningful learning experience.

Acknowledgements

Thanks to all the students that donated their DNA so this research could be done. Also, a special thanks to Becky Folsom and Juniper Networks for their funding of this project.

References

ANDERSON, S., BANKIER, A.T., BARREL, B.G., DE BRUIJIN, M. H., COULSON, A.R., SANGER, F., SCHREIER, P. H., SMITH, A.J.H., STADEN, R. AND YOUNG, G. (1981). Sequence and organization of the human mitochondrial genome. *Nature* 290:457-465.

CLUSTAL W. San Diego Supercomputer Center (SDSC) Biology Workbench, Accessed from http://seqtool.sdsc.edu/CGI/BW.cgi on September 2008. Used for creating Phylogenetic tree.

DOLAN DNA LEARNING CENTER 2001 Annual Report. Accessed from http://www.dnalc.org/ddnals/about/annreppdf/annrep2001.pdf on July 5th, 2009

DOLAN DNA LEARNING CENTER, *Sequence Server*. Accessed from http://www.bioservers.org/html/sequences/sequences.html on September 23rd, 2008. Used for obtaining sequences.

HORAI, S., K. HAYASAKA. 1990. Intraspecific Nucleotide Sequence Differences in the Major Noncoding Region of Human Mitochondrial DNA. *American Journal of Human Genetic* 46:828-842.

HORAI, S., KONDO, R., NAKAGAWA-HATTORI, Y., HAYASHI, S., SONODA, S., *et al.* 1993. Peopling of the Americas, Founded by Four Major Lineages of Mitochondrial DNA. *Molecular Biology and Evolution*. 10(1): 23-47.

Kyoto University Bioinformatics Center, Multiple Sequence Alignment. Accessed from http://align.genome.jp/ on September 2008. Used for creating CLUSTAL sequences.

LEWIS *et al*. 2007. Mitochondrial DNA and the peopling of South America. *Human Biology.* April 2007 v 79(2): 159-178.

MULTIPLE SEQUENCE ALIGNMENT BY CLUSTAL W program. Accessed from http://align.genome.jp/ on September 13, 2008. Used for creating Phylogenetic tree.

WALKER, JERILYN A., RANDALL K. GARBER, DALE J. HEDGES, GALE E.KILROY, JINCHUAN XING, *et al*. 2003. Resolution of Mixed DNA Samples Using Mitochondrial DNA Sequence Variants. *Analytical Biochemistry* 325:171-173