Lab Report 4: Analysis of Microsatellite Genotype Data

Paola Quinones

Genetics Laboratory

April 30, 2018

Dr. Ma

**Abstract**

Simple Sequence Repeats (SSRs), also known as microsatellite DNA, are molecular markers that because of their high polymorphism, they are able to track individual chromosomes, determine parentage (paternity testing), and identify individuals (used in forensics). In this experiment, 3 distinctive microsatellite loci from chromosomes 14 and 15 were amplified using Multiplex PCR technique and later analyzed as a whole, including the data of all the other students in the class. The purpose of this experiment was to analyze if the class was in Hardy-Weinberg Equilibrium (HWE) at the D15S657 locus. The hypothesis was that the class would be in HWE. The results show that in the D143S06, allele, 234 was the most common because it had the highest frequency while the allele 198 had the lowest frequency; therefore, it was the least common. In the D15S655 locus, allele 257 was the most common while allele 275 was the least common. And, in the D15S657 loci, allele 357 was the most common while allele 376 was the least common. Moreover, the Chi-square analysis, which yielded an X2 value of 48.82 and P-value of 0.2854, shows that the data is insignificant and HWE is not rejected. These results show that microsatellite DNA is indeed diverse and that DNA profiling is a good technique to identify unknown DNA samples and see variability within a population.

**Introduction**

 DNA profiling is a flexible practice that lets researchers survey genetic variation, track individual chromosomes, & determine identity (Carlini, 2018). DNA may be taken from minor samples of cells, like the ones taken from the cheek. Microsatellite DNAs are small, simple sequence tandem repeat motifs; the Simple Sequence Repeats (SSRs) make up for 3% of the human genome (Carlini, 2017). Most microsatellite loci are not transcribed into RNA; therefore, because mutations in microsatellites do not usually affect the phenotype, microsatellite loci are very polymorphic. This high polymorphism at microsatellite loci is because of the variability in the number of copies of repeat units in an allele (Carlini, 2018). Microsatellite DNA is found in the coding and noncoding regions of the genome, but is more common in the noncoding areas. The high level of polymorphism at microsatellite loci, produces a unique DNA profile/fingerprint that can be acquired for every individual (with the exception of identical twins) if sufficient number of loci (10-15) are surveyed (Carlini, 2018).

 You can identify a human by extracting its DNA from a small sample of cells, either by saliva or blood, and combine it with a polymerase chain reaction (PCR) (Carlini, 2018). For this experiment, PCR reaction was used to profile a DNA sample collected from the saliva of all students in the class. Three specific microsatellite loci were amplified using primers matching the DNA. The three loci amplified are: D14S306, D15S655 and D15S657. Polyacrylamide gel electrophoresis was later used to see the difference in base pairs in the PCR products, which generated a band pattern. This band pattern was to compare it with a ladder of known lengths to distinguish the genotype of the 40 students. The PCR technique used is multiplex, which analyzes the three primer pairs (6 in total) being used, at the same time. In other words, it studies three different loci from a single sample tube (Carlini, 2018).

The DNA extracted from cheek cells in this experiment is amplified using PCR technology. PCR is a bio-technique where one or several copies of an isolated DNA sample can be increased into millions of copies of the target sequence. Polyacrylamide gel electrophoresis (PAGE) is used to label the obtained sequence by separating proteins by according to their size. The way it works is by having sodium dodecyl sulfate (SDS) bind to proteins, and negatively charging them, then these proteins separate according to their size because the bigger proteins take longer to travel through the gel to the positively charged end. The molecular weight marker, known as ladder, is a standard of known lengths that is used to identify lengths of protein in the sample. By doing this, it is possible to determine the genotype of that sample. These techniques were used to analyze the microsatellite genotypes at three different loci found on chromosomes 14 and 15. The data produced is used to find the allelic and genotypic frequencies at each of the loci and analyze the 15S657 locus to see if it agrees Hardy-Weinberg Equilibrium (HWE). HWE is model that states that frequencies of both alleles and genotypes stay constant throughout generations if no evolutionary mechanisms are acting in a population. The Chi-square analysis, demonstrates how the observed data fits compared to that of an expected data, which can be used to figure out if the data obtained from this experiment is under evolutionary control/is in accordance to HWE.

**Methods & Materials**

 To extract DNA, 10mL of saline solution (0.9% NaCl) was swished around in mouth for 30 seconds and spitted into a paper cup (Carlini, 2018). 1.5mL of the solution in the cup was pipetted into 1.5mL tube (Carlini, 2018). The tube was then put into a microcentrifuge and spun for 5 minutes (Carlini, 2018). The supernatant was removed and the cells were resuspended through an in and out suction motion (Carlini, 2018). Then we transferred the cells to a 0.5mL microcentrifuge tube with 100μL of 10% Chelex solution (Carlini, 2018). The tube was vortexed and placed into a 100°C thermocycler for 10 minutes (Carlini, 2018). The tube was placed in a balanced microcentrifuge and put to spin for 1 minute (Carlini, 2018). 50μL of the supernatant was transferred to a different 1.5mL tube, which is the genomic DNA. (Carlini, 2018).To prepare the multiplex PCR, 0.50x and 0.25x dilutions were prepared from the genomic DNA that was secluded from the previous steps. Then, these reactions were placed into a thermocycler. The Acrylamide Gel Electrophoresis was set up by the TA for each student to transfer their respective DNA from the previous procedure into the designated well. For the gel analysis to find the individual’s genotype, the size of the polymorphism from the PCR amplification products were used. The relationship between distance and size were measured and used to find the size of the bands in the sample. The best-fit line was made on a linear regression in Excel with all of the data points measured. The size of the alleles were calculated using the linear regression equation and the measured lengths of student bands (Carlini, 2018). Later on, the genotype for each individual was determined.

There were extensive calculations that needed to be done in order to get the important data. The total number of different sized bands was found using the table the TA gave us. The number of possible genotypes was calculated using the following equation:

 $\sum\_{i=1}^{i=n}i$

Where n is the # of alleles at a specific locus. Then, we had to calculate the allelic frequency for the 3 loci. In order to find this, the # of observed alleles was divided by the total # of chromosomes in the class (40 students). The allelic frequencies were used find the genotypic frequencies for the 15S657 locus and to verify HWE using chi-squared. To find the genotypic frequencies, the following equation was used:  (*p*+*q*+*r*+*s*+*t*+*u*+*v*+*w*+*x*)2 =

*p*2 + *q*2 + *r*2 + *s*2 + *t*2 + *u*2 + *v*2 + *w*2 + *x*2 +

2*pq* + 2*pr* + 2*ps* + 2*pt* + 2*pu* + 2*pv* + 2*pw* + 2*px* +

2*qr* + 2*qs* + 2*qt* + 2*qu* + 2*qv* + 2*qw* + 2*qx* +

 2*rs* + 2*rt* + 2*ru* + 2*rv* + 2*rw* + 2*rx* +

2*st* + 2*su* + 2*sv* + 2*sw* + 2*sx* +

2*tu* + 2*tv* + 2*tw* + 2*tx* +

2*uv* + 2*uw* + 2*ux*

+ 2*vw* + 2*vx* +

2*wx*

=  1

The obtained frequency was multiplied by the total number of students in the class to find the expected number of students with the genotype. The observed number of students was recorded and a chi-square test was found using the following equation:

 $X=\sum\_{}^{}\frac{(O-E)^{2}}{E}$

O is the observed number of students with a particular genotype, E is the expected number of students with a particular genotype and once calculated, the P value was found using an online database as well as the significance of this value.

**Results**

**Table 1: Frequencies of Each Allele at Each of the Three Loci**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Loci D143S06 |  | Loci D15S655 |  | Loci D15S657 |  |
| Allele | Frequency | Allele | Frequency | Allele | Frequency |
| 198 | 0.0125 | 257 | 0.2875 | 344 | 0.0250 |
| 202 | 0.0375 | 248 | 0.0250 | 352 | 0.2125 |
| 206 | 0.0375 | 254 | 0.2375 | 348 | 0.0375 |
| 210 | 0.3125 | 251 | 0.0500 | 356 | 0.2250 |
| 222 | 0.1375 | 266 | 0.0750 | 368 | 0.1375 |
| 214 | 0.1500 | 269 | 0.1375 | 360 | 0.1625 |
| 218 | 0.1875 | 260 | 0.0875 | 364 | 0.1500 |
| 226 | 0.1000 | 272 | 0.0500 | 372 | 0.0375 |
| 234 | 1.0125 | 275 | 0.0125 | 376 | 0.0125 |
| 238 | 0.0125 | 263 | 0.0125 |  |  |
|  |  | 278 | 0.0250 |  |  |

**Table 2: Observed & Expected Genotypic Frequencies & Test of HWE at D15S657 locus**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Genotype**  | **Expected Frequency**  | **Expected # of Students**  | **Observed # of students**  | **(O-E)^2/E** |
| 344/344 | 0.000625 | 0.025 | 0 | 0.025 |
| 344/352 | 0.010625 | 0.425 | 2 | 5.836764706 |
| 344/348 | 0.001875 | 0.075 | 0 | 0.075 |
| 344/356 | 0.01125 | 0.45 | 0 | 0.45 |
| 344/368 | 0.006875 | 0.275 | 0 | 0.275 |
| 344/360 | 0.008125 | 0.325 | 0 | 0.325 |
| 344/364 | 0.0075 | 0.3 | 0 | 0.3 |
| 344/372 | 0.001875 | 0.075 | 0 | 0.075 |
| 344/376 | 0.000625 | 0.025 | 0 | 0.025 |
| 352/352 | 0.0451 | 1.804 | 1 | 0.358323725 |
| 352/348 | 0.0159 | 0.636 | 1 | 0.208327044 |
| 352/356 | 0.0956 | 3.824 | 5 | 0.361656904 |
| 352/368 | 0.0584 | 2.336 | 3 | 0.188739726 |
| 352/360 | 0.069 | 2.76 | 2 | 0.209275362 |
| 352/364 | 0.0637 | 2.548 | 2 | 0.117858713 |
| 352/372 | 0.0159 | 0.636 | 0 | 0.636 |
| 352/376 | 0.0053 | 0.212 | 0 | 0.212 |
| 348/348 | 0.0014 | 0.056 | 0 | 0.056 |
| 348/356 | 0.0168 | 0.672 | 1 | 0.160095238 |
| 348/368 | 0.0103 | 0.412 | 0 | 0.412 |
| 348/360 | 0.0121 | 0.484 | 1 | 0.550115702 |
| 348/364 | 0.01125 | 0.45 | 0 | 0.45 |
| 348/372 | 0.0028 | 0.112 | 0 | 0.112 |
| 348/376 | 0.000937 | 0.03748 | 0 | 0.03748 |
| 356/356 | 0.0506 | 2.024 | 1 | 0.518071146 |
| 356/368 | 0.061875 | 2.475 | 5 | 2.576010101 |
| 356/360 | 0.073125 | 2.925 | 2 | 0.292521368 |
| 356/364 | 0.0675 | 2.7 | 4 | 0.625925926 |
| 356/372 | 0.016875 | 0.675 | 0 | 0.675 |
| 356/376 | 0.005625 | 0.225 | 0 | 0.225 |
| 368/368 | 0.01890625 | 0.75625 | 0 | 0.75625 |
| 368/360 | 0.0446875 | 1.7875 | 2 | 0.025262238 |
| 368/364 | 0.04125 | 1.65 | 0 | 1.65 |
| 368/372 | 0.0103125 | 0.4125 | 1 | 0.836742424 |
| 368/376 | 0.0034375 | 0.1375 | 0 | 0.1375 |
| 360/360 | 0.02640625 | 1.05625 | 0 | 1.05625 |
| 360/364 | 0.04875 | 1.95 | 3 | 0.565384615 |
| 360/372 | 0.0121875 | 0.4875 | 1 | 0.538782051 |
| 360/376 | 0.0040625 | 0.1625 | 0 | 0.1625 |
| 364/364 | 0.0225 | 0.9 | 2 | 1.344444444 |
| 364/372 | 0.01125 | 0.45 | 0 | 0.45 |
| 364/376 | 0.00375 | 0.15 | 0 | 0.15 |
| 372/372 | 0.00140625 | 0.05625 | 0 | 0.05625 |
| 372/376 | 0.000937 | 0.03748 | 1 | 24.71837648 |
| 376/376 | 0.000156 | 0.00624 | 0 | 0.00624 |
|  |  |  | Total=**X2 =**  | 48.82314791 |
|  |  |  | P value= | 0.2854 |

**Discussion**

The data from Table 1 shows that for the D143S06, allele 234 was the most common because it had the highest frequency while the allele 198 had the lowest frequency; therefore, it was the least common. In the D15S655 locus, allele 257 was the most common while allele 275 was the least common. Lastly, in the D15S657 loci, allele 357 was the most common while allele 376 was the least common.

Table 2 shows the Chi-square value of the data collected, which was 48.82. The p-value is 0.2854, which means that the data is insignificant and HWE is not rejected. The small sample size and possible calculation errors could be attributed to deviation from expectation. However, the fact that alleles differ in size, there is a chance of misidentification or limitations in the technology/methods used.

The microsatellite loci studied in this experiment would help identify students if a random DNA sample was obtained and needed to be attributed to someone in the class because every student has a distinct genotype when all three loci are combined. Therefore, an individual can use this sequencing to know to whom the unknown sample DNA belongs to. The more loci to sequence there are, the more specific each genotype will be because some individuals share a genotype in 2 loci so if there had not been a third, one would have to assume they are twins even if they are not even related. This technique is essential today in fields like forensics, paternal testings, and to survey genetic variations within a population.

**References:**

Carlini, David. (2017). Weeks 6 & 7 Labs. American University.