Korean J Leg Med 2014;38:59-65
http://dx.doi.org/10.7580/kjlm.2014.38.2.59

원 저

미토콘드리아 DNA 데이터베이스의 통계학적 유용성 확인을 위한 파라미터 탐색

정종민[1] · 이지현[2] · 조소희[2]
이숭덕[2, 3]

[1]국립과학수사연구원 법유전자과
[2]서울대학교 의과대학 법의학교실
[3]서울대학교 의학연구원
법의학연구소

접 수 : 2014년 4월 29일
수 정 : 2014년 5월 19일
게재승인 : 2014년 5월 20일

이 논문은 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단 바이오·의료기술개발사업의 지원을 받아 수행된 연구입니다(No. 2013-057192).

책임저자 : 이숭덕
(110-799) 서울시 종로구 연건동 대학로 103번지, 서울대학교 의과대학 법의학교실
전화 : +82-2-740-8359
FAX : +82-2-764-8340
E-mail : sdlee@snu.ac.kr

Searching for Appropriate Statistical Parameters for Validation of Mitochondrial DNA Database

Chong Min Choung[1], Ji Hyun Lee[2], Sohee Cho[2], Soong Deok Lee[2, 3]
[1]Forensic DNA Division, National Forensic Service, Wonju-si, Gangwon, Korea
[2]Department of Forensic Medicine, [3]Institute of Forensic Science, Seoul National University College of Medicine, Seoul, Korea

Recently, studies on mitochondrial DNA (mtDNA) have increased rapidly. Conventional parameters, such as diversity index, pairwise comparison, are used to interpret and validate data on autosomal DNA; however, the use of these parameters to validate data from mitochondrial DNA databases (mtDNA DBs) needs to be verified because of the different transmission patterns of mtDNA. This study was done to verify the use of these conventional parameters and to test the "coverage concept" for a new parameter. The mtDNA DB is not very big; however, it is necessary to check how the change in parameters corresponds to the DB size. For this, we artificially rearranged a Korean DB into several small sub-DBs of variable sizes. The results show that the diversity in nucleotide variations and the different haplotype numbers do not vary as the size of DB increases. However, the "coverage" changed a lot. The coverage increased from 0.113 in a DB of 100 people to 0.260 in a DB of 653 people. Additionally, using the "coverage concept", we predicted how the total number of haplotypes changed with variations in the sub-DB size and compared the predicted result with final result. In conclusion, "coverage", in addition to conventional statistical parameters, can be used to check the usability of an mtDNA DB. Finally, we tried to predict the size of the whole mtDNA number in Korea using "saturation concept".

Key Words : mtDNA DB, Statistical parameter, Coverage, Phylogeny, Saturation curve

서 론

일반적으로 미토콘드리아 유전자(이하 mtDNA)는 핵 유전자와 달리 독립적으로 복제되며 모계유전을 한다. mtDNA는 핵 유전자에 비해 재조합 없이 일배체형의 형태로 모계로 유전되므로, 모계유전되는 mtDNA는 염기서열의 변이가 연속적으로 축적되어 계통학적으로 진화해왔다.[1] 따라서 여러 지역 및 인종의 mtDNA의 유전자 정보는 세대를 지나오면서 거쳐 간 모계혈통의 이주 경로를 알려준다. 이들 중 비슷한 염기서열 혹은 유전자형을 지닌 일배체형들을 모은 것을 하플로그룹(haplogroup)이라 한다.[2] mtDNA는 법의학적으로 혹은 계통발생학적으로 모계 인척 관계 여부를 확인하거나 모계 유전자의 지역적 기원을 추정하고자 사용하곤 한다.

mtDNA의 경우 핵 유전자형 분석에서 이용하는 개인식별지수를 이용할 수 없다. 이러한 경우에는 흔히 셈(counting) 방법

pISSN 1225-0589
eISSN 2287-2078

이 적용되곤 한다. 따라서 mtDNA에서의 통계란 데이터베이스(이하 DB)의 크기에 매우 의존적이다. 이러한 제한점 때문에 mtDNA에서 확률 통계적인 접근을 하려면 DB가 매우 크거나, DB의 효용성 혹은 포함하고 있는 정보의 양을 가늠하기 위한 적절한 파라미터가 있어야 한다.

대부분의 mtDNA 연구에서는 통계 파라미터로 diversity index, pairwise 검정에 의한 집단 간 평균 변이 개수 비교법, 공통모계지수(θ) 등이 쓰였고,[3] 최근 coverage란 개념이 도입되었다. 이 개념은 원래 "지구에는 총 몇 종류의 생물이 살까" 등과 같은 개념을 해결하기 위해 도입된 것인데, 즉 개념적으로 "현재 사용하고 있는 DB가 실제 존재하는 mtDNA의 일배체형들을 어느 정도 포함하고 있는가" 혹은 "아직 발견되지 않은, 새로이 발견될 가능성이 있는 일배체형들은 어느 정도인지"와 크게 다르지 않고 결과의 해석이나 방법론적인 접근이 크게 다르지 않으리라고 예상한다.

Egelend 등은 Chao 등[4]과 Huang 등[5]이 발표한 sample coverage방법을 응용하여 기존의 DB를 토대로 하여 기존의 DB에서 발견된 일배체형 외에도 전체 인구 집단에서 존재하는 아직 발견되지 않은 일배체형을 추정하여 그 효용성을 확인하여 보기도 하였다.[6] 이들 방법은 다양한 통계 공식을 응용하는 확률적인 방법과 PCA (Principal Component Analysis)를 응용한 방법이다.

본 연구에서는 한 DB를 활용하여 여러 다양한 크기들의 작은 DB들을 새로이 생성하고, diversity index 와 coverage값을 구하고 DB 크기에 따라 변하는 양상을 관찰해서 위 파라미터들이 DB의 검증에 의미가 있는 파라미터인지를 확인해보고자 한다. 한편 coverage에 대해서는 크기 변화에 따라 예측되는 총 일배체형의 개수를 실제 관찰되는 개수와 비교하여 그 효용성을 다른 측면에서 확인하여 보고자 하였다. 추가로 mtDNA 분석에 중요한 "과연 어느 정도 다른 일배체형이 존재할 것인지"에 대한 정보를 얻기 위해 포화곡선 방법을 적용해 한국인의 전체 mtDNA DB의 크기를 가늠하여 보았다.

재료 및 방법

1. 시료 및 염기서열 분석

본 연구에서는 서로 모계 인척 관계가 없는 한국인 653명에 대한 mtDNA 염기서열 결과를 활용하였다. 위 서열은 다른 목적으로 진행된 사업을 통해 확보하였으며, 일상적으로 알려진 방법으로 시료 채취, 유전자 분리 및 미토콘드리아 과변이부위(이하 HV) 염기서열 결정 과정 등을 거쳤다(Table 1). 한편, 본 연구는 서울대학교 의과대학/서울대학교 병원의 연구윤리심의 위원회의 심의를 거쳤다.

1.1. 새로운 DB의 인위적 구성

현재 가지고 있는 DB의 크기가 충분하지 못하고, 결국 DB 크기의 변화에 따라 여러 파라미터가 어떻게 변하는지는 매우 중요하다. 이와 관련된 정보를 얻기 위해 전체 653명의 자료를 순차적으로 100명 단위로 누적시켜 작은 DB들을 새로이 구성한 후 각각의 자료에서 염기서열 다양성 및 일배체형 다양성을 계산하였다.

2. mtDNA database의 유용성 확인을 위한 자료의 분석

2.1. 기존 통계 방법 적용

먼저 얻어진 자료들에 대해 기존의 핵 유전자형에 적용되는 diversity 방법을 이용한 염기서열 다양성과 일배체형 다양성을 구하였다. 염기서열 다양성은 653명의 전체 염기 서열자료를 표준염기서열과 다른 부위를 위치별로 나타낸 것을 토대로 하여 염기서열변이의 빈도(x)를 전체 출현 횟수로 나눈 값의 제곱 합을 이용한 Genetic diversity = $1-\Sigma \times^2$에 따라서 계산하였다.[7] 일배체형 다양성은 각 개인별로 표준염기서열과 다른 mtDNA 염기서열들을 일배체형 으로 묶어서 653명 중 출현 빈도를 구해서 앞의 염기서열 다양성과 같은 방법으로 계산하였다. 집단 간 비교를 위해서는 F-검정법을 이용하였고 95%의 신뢰 수준으로 검증하였다.

Table 1. Primers used to Sequence the D-loop of mtDNA

Primer	Sequence			5' → 3'			
F15971	TTA	ACT	CCA	CCA	TTA	GCA	CC
F16291	AGG	ATA	CCA	ACA	AAC	CTA	C
F049	CTC	ACG	GGA	GCT	CTC	CAT	GC
R16410	GAG	GAT	GGT	GGT	CAA	GGG	AC
R16493	GAA	GTA	GGA	ACC	AGA	TGT	CGG
R159	ATA	TTG	AAC	GTA	GGT	GCG	AT
R408	CTG	TTA	AAA	GTG	CAT	ACC	GCC
R921	CTT	TAC	GCC	GGC	TTC	ATA	TG

http://dx.doi.org/10.7580/kjlm.2014.38.2.59

2.2. coverage 적용

2.2.1 일상적인 coverage의 결정

일배체형의 출현 빈도를 이용한 coverage를 적용하는 방법은 몇 가지 방법이 제시되어 있다. 소개된 방법들은 약간씩 다르나 기본적으로

$$C=\sum_{i=1}^{N} p_i I(X_i>0) \quad (1)$$

$$N_1^*=D/C^* \quad (2)$$

$$C^*=1-f_1/n. \quad (3)$$

n : total number of haplotypes in the database

D : the number of different, unique haplotypes

Pi : the probability that a haplotype belongs to the ith class, I = 1.....,N

Xi : the number of elements of the ith class in the sample

C*: Coverage

f1: the number of haplotypes observed once

공식을 기초로 한다.[6] 이 공식은 관련 연구자들에 따르면 대다수의 일배체형이 드물게 발견되는 경우에 잘 맞을 것으로 제시하였다.

Pereira등은 포화곡선(Saturation curve)를 이용해 집단의 크기가 커졌을 때 관찰 가능한 일배체형의 개수를 추정하는 방법을 제시하였다.[8, 9] 본 연구에서는 이들 방법에 따라서 앞서 만든 100명 단위의 자료로 구한 값을 토대로 하여 Curve expert professional 1.6.5 (www.curveexpert.net)을 이용하여 포화곡선을 적용 시 발견 가능한 일배체형의 개수를 추정해 보았다.

2.2.2 coverage 개념의 추가적인 점검 – 일배체형 수의 예측과 실제 관찰 수치와의 비교

이를 위한 방법 또한 위에서 언급된 논문에 따랐는데, 좀 더 구체적으로 언급하여 보자면 다음과 같다.

Egeland 등은 좀 더 공식을 간편화시켜서

$$\gamma^*=\max\left\{N_0\sum_{i=1}^{\infty} i(i-1)f\Big/[n(n-1)]-1,0\right\} \quad (4)$$

와

$$\gamma^{\sim}=\max\left\{\gamma^*\left(1+f_1\sum_{i=1}^{\infty} i(i-1)f\Big/[n(n-1)]-1\right),0\right\}. \quad (5)$$

의 두 가지 값으로 제시하였다.[3]

Hass[10]와 Huang[5]등의 연구자들은 관찰된 일배체형을 abundant와 rare의 두 그룹으로 나누어서

$$D_{abun}+D_{rare}/C^*_{rare}=D_{abun}+N_0 \quad (6)$$

D : the number of different, unique haplotypes

C*: Coverage

의 공식으로 coverage 값을 구하였다. 여기서 관련 연구자들은 abundant와 rare의 기준값으로 K 상수를 도입하였으며 이 값을 10으로 설정해두었으나 실제 유전자 자료에 도입하는 것에는 주의를 당부하였다. 이에 반해 Mao는 fraction t라는 상수를 도입하여 집단의 크기가 커졌을 때 예상되는 일배체형의 개수를 추정하는 공식을 제안하였다. Mao에 의한 공식은

$$\Delta(t)=f_1t-f_2t^2+f_3t^3-\ldots. \quad (7)$$

f1 means the number of haplotype unique

f2 means the number of haplotype which are shared of two

f3 means the number of haplotype which are shared of three, and so on…

으로 Mao에 의한 공식을 적용하기 위해 사용한 t 값은 0에서 1 사이의 범위로 1, 0.5, 0.25였으며 t = 1일 때 두 배의 크기를 적용시켰으며, t = 0.5일 때는 1.5배, t = 0.25일 때는 1.25배 크기의 집단에서 공식을 이용하여 관찰 가능한 전체 일배체형의 개수를 추정하였다.[11] 본 연구에서는 위에 제시된 2가지 방법의 공식들(기본 공식, Mao에 의한 공식)을 이용하여 coverage를 실제 자료에 적용해보고자 하였다. 위 1.1에 적용한 DB 방식을 적용하였고 이에 따른 결과가 기존 DB에서 임의로 추출하는 과정에서 통계적으로 유의한 차이가 발생하는지를 확인하기 위해 두 번에 걸쳐 시뮬레이션하였다. 임의 추출 방법은 microsoft사의 Excel 프로그램을 응용하였으며 100명에서 600명까지 Simulation−1 (sim−1), Simulation−2 (sim−2)의 자료를 만들어 coverage를 계산하였다. 이렇게 계산된 값은 회귀분석과 F−검정법을 이용해 상관관계 및 통계적으로 유의한 차이가 있는지를 95% 신뢰도 내에서 검사하였다.

결 과

1. 통계 파라미터의 적용 결과

1.1. Diversity index 적용 결과

각 자료별 염기서열 변이의 개수는 HV I 부위에서는 100명일 때 82개에서 600명이었을 때는 135개로 꾸준히 증가하였으나, 개체 수가 400명이었을 때 125개, 500명이었을 때 129개, 600명 일 때 135개로 자료 크기가 증가할수록 새롭게 추가되는 염기서열 변이개수가 그리 많지 않았다. HV II 부위에서는 100명 일 때 46개에서 600명일 때 75개로 증가하였으나 HV I 부위에서 보다 추가되는 변이의 개수는 작았다. 분석 범위를 HV I 과 HV II 부위를 모두 고려했을 때 100명일 때 128개에서 154개로 증가하였으며, 자료의 크기가 커질수록 증가하는 변이의 개수도 HV I, HV II 부위만 고려했을 때보다 많았

다. 그러나 자료의 크기가 400명 이상 되었을 때는 HV I, HV II 부위일 때와 마찬가지로 증가 폭은 감소하였다.

염기서열 다양성은 분석범위를 HV I으로 했을 때 염기서열 다양성은 0.9454였으며, HV II일 때는 0.8763, HV I과 HV II를 모두 고려했을 때는 0.9533 근처에서 크게 증가하지 않았다. 분석범위를 D-loop 전체로 확대했을 때 염기서열 다양성은 자료 크기와 상관없이 0.9636 수준이었다(Table 2).

자료의 누적 양상에 따라 관찰되는 일배체형의 경우, 100명일 때부터 600명까지 관찰되는 개수는 꾸준히 증가하나, 다양성을 나타내는 수치는 0.9894에서 0.9979로 큰 변화가 없었다(Table 3).

1.2. Coverage 적용 결과

1.2.1. 기본 공식을 적용해 본 결과

순차적으로 누적시켜 만든 자료에서 발견되는 고유한 일배체형의 개수를 이용하여 산출한 coverage값은 100명일 때 0.113, 200명일 때 0.118, 300명일 때 0.166, 400명일 때 0.205, 500명일 때 0.252, 600명일 때 0.260으로 꾸준히 증가하였다(Table 3).

또한, 653명의 자료를 무작위로 섞은 뒤, 100명 단위로 추출하여 각각의 coverage값을 두 번(Sim-1, Sim-2)에 걸쳐 계산하였다. 그리고 이 값들을 F-검정법을 이용하여 앞서 순차적으로 누적시킨 값(Non selected)과 비교하였다. 그 결과 100명일 때 Sim-1 및 Sim-2에서 산출한 coverage는 각각 0.05 및 0.02로 순차적으로 누적시켰을 때 구한 값인 0.113보다 매우 작았다. 그러나 200명부터는 세 집단 간에 산출한 coverage값은 통계적으로 유의한 차이는 없었으며 꾸준히 증가하는 패턴을 보였다(Table 4).

1.2.2. coverage 의 검증 - Mao에 의한 공식 적용 결과

100명에서 600명까지 순차적으로 누적시킨 자료에서 관찰

Table 2. Variation of Sequence Diversity and Number of Sequence Changes at each Sample Sizes*

	HV Ⅰ		HV Ⅱ		HV Ⅰ+HV Ⅱ		D-loop	
	n‡	H†	n	H	n	H	n	H
N§ = 100	82	0.9419	46	0.8685	128	0.9509	154	0.963
N = 200	96	0.0944	47	0.8750	143	0.9522	172	0.9627
N = 300	113	0.9438	60	0.8737	173	0.9521	210	0.9630
N = 400	125	0.9465	71	0.8693	196	0.9523	237	0.9632
N = 500	129	0.9465	72	0.8730	201	0.9523	240	0.9629
N = 600	135	0.9460	75	0.8718	210	0.9525	258	0.9632
N = 653	139	0.9454	84	0.8763	223	0.9533	266	0.9636

*: Data was grouped by randomly labeled number for making small DB from original 653 people DB.
†: Sequence diversity calculated from 1-$\Sigma \times^{2.10}$)
‡: number of nucleotides that show difference when compared to rCRS
§: DB size

Table 3. Number of observed Haplotypes and Comparison of Statistic Parameters*

	No. of Observed Haplotypes											
	Total	H†	C‡	CI§	f1‖	f2‖	f3‖	f4‖	f5‖	f6‖	f7‖	f8‖
N = 100	93	0.9894	0.113	(0.065~0.159)	86	6	1	0	0	0	0	0
N = 200	183	0.9939	0.123	(0.083~0.151)	171	11	0	0	0	0	1	0
N = 300	269	0.9954	0.168	(0.130~0.195)	246	18	4	0	0	0	1	0
N = 400	351	0.9970	0.207	(0.177~0.233)	313	30	5	2	0	0	0	1
N = 500	433	0.9978	0.223	(0.196~0.247)	382	36	11	2	0	1	0	1
N = 600	509	0.9983	0.252	(0.227~0.275)	443	47	13	4	0	0	0	2
N = 653	546	0.9979	0.260	(0.239~0.281)	477	48	14	3	2	0	0	2

*:, Data were not selected depending on size and simply grouped with serial number.
†: Haplotype diversity calculated from 1-$\Sigma \times^{2.10}$)
‡: Calculated coverage
§: Confidential interval 95%
‖: Number of the observed haplotypes
f1 means the number of haplotype unique, f2 means the number of haplotype which are shared of two, f3 means the number of haplotype which are shared of three, and so on….

http://dx.doi.org/10.7580/kjlm.2014.38.2.59

된 일배체형의 개수와 Mao에 의한 공식을 적용해 얻은 결과를 비교해 보았다(Table 5). 그 결과 t = 1일 때 실제 자료를 이용해 200명일때와 400명, 600명, 800명, 1,000명 및 1,200명일 때 구한 관찰 가능한 일배체형의 추정값은 각각 174명, 344명, 502명, 636명, 786명 및 953명이었다. 그리고 t = 0.5일 때 구한 값인 300명일 때와 600명일 때의 값은 265.76명과 500.5명이었다. t = 0.25일 때는 400명일 때 구한 실제 자료 값을 토대로 하여 400명의 1.25배수인 500명일 때 추정값은 427.45명이었다. 본 연구에서 실제로 관찰한 일배체형의 개수와 공식에 적용해서 구한 추정값에 통계적으로 유의한 차이가 있는지를 알아보기 위해 유의 수준 5%에서 카이스퀘어 검정을 실시한 결과 P = 0.97로 유의수준 0.05보다 큰 값으로 두 자료 간에 차이가 없는 것으로 판단되었다.

Table 4. Comparison of Coverage by the Way Simulated and Non-selected DB Showing no Significant Selection Effect

	Sim-1*	Sim-2*	Non-selected†
N = 100	0.050	0.020	0.113
N = 200	0.132	0.130	0.118
N = 300	0.138	0.145	0.166
N = 400	0.180	0.200	0.205
N = 500	0.219	0.247	0.222
N = 600	0.253	0.272	0.252
N = 653	–	–	0.26

*: Data from randomly selected by random shuffling.
†: Data from simply selected by serial number.

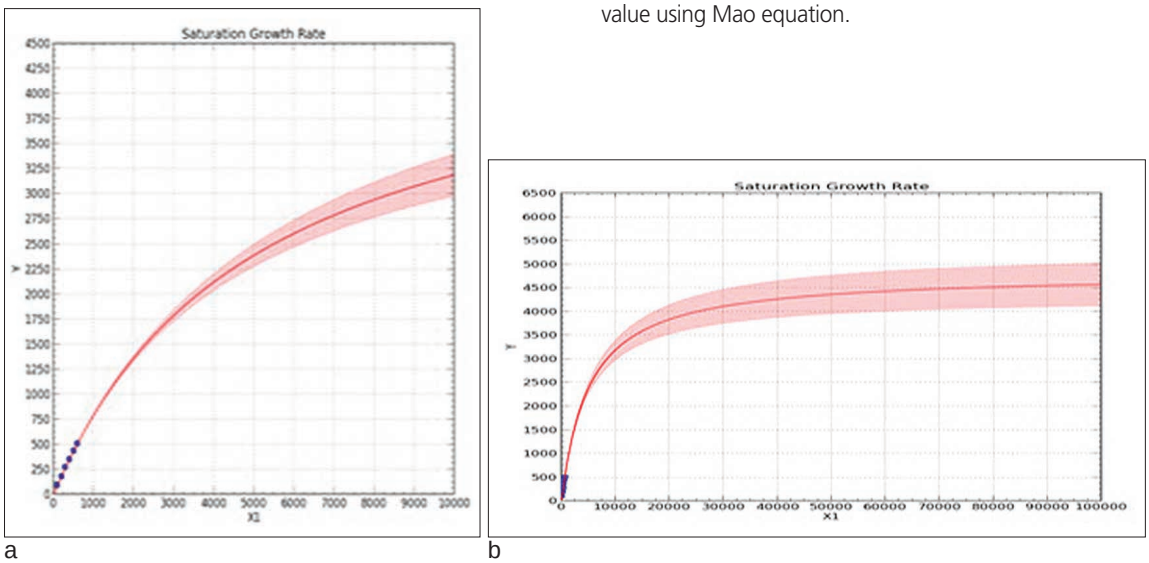
2. 총 일배체형의 예측 – 포화곡선 적용 결과

100명부터 600명까지 관찰된 일배체형의 개수를 입력한 후 Curve expert professional 1.6.5를 이용하여 포화곡선을 그려보았다. 이때 적용한 함수는 f(x) = ax/(b+x)이며, 그래프는

Table 5. Unique Haplotype Comparison between Observed One and Estimated One using Mao Equation

	Estimated value			
	Total	M (t = 1)	M (t = 0.5)	M (t = 0.25)
N = 100	93			
N = 200	183	**174**		
N = 300	269		**265.76**	
N = 400	351	**344**		
N = 500	433			**427.45**
N = 600	509	**502**	**500.5**	
N = 653	546			
N = 800		**636**		
N = 1,000		**786**		
N = 1,200		**953**		
P value			**0.97**	

The estimated values were calculated from the equation
$\Delta(t) = f_1 t - f_2 t^2 + f_3 t^3 - \ldots$.
This formula is valid t ≦ 1 and t = 1 corresponds to a doubling size. The column M (t = 1) are the estimated values when the expected number of the doubled size, M (t = 0.5) are the 1.5 times, M (t = 0.25) are the 0.5 times. The result of the qui-square test, p value was 0.97 (α = 0.05). For N = 100 ~ N = 653, total means sum of the number of observed haplotypes within each data size. For N ≥ 800 only estimated one was presented. The bolds mean estimated value using Mao equation.



a b

Fig. 1. Saturation curves of expanded sample sizes.
a. Expanded up to 10,000 people
b. Expanded up to 100,000 people
A result of examining the number of possible observed haplotypes when group size increased up to 10,000, 100,000, the final expected number of haplotypes was 4,500 over. The shaded portion of the graph is the confidence interval upper and lower limits.

회귀곡선으로 분류된다. 프로그램에서 계산된 상수인 a값과 b값은 각각 4,803.78과 5,059.36이었으며, 표준 오차 범위는 95% 신뢰 상한 구간에서 a값은 4304.7~5,302.9였으며, b값은 4,484.4~5,634.3이었다. 그래프를 각각 10,000명, 100,000명까지 확대해본 결과, 신뢰 상한 구간은 집단 크기가 커질수록 범위가 넓어 졌다(Fig. 1). 6개의 실제 관찰 값으로 그린 포화곡선에서는 약 60,000명에서부터 포화수준에 진입하였으며 100,000명에서 일배체형의 개수가 4,500여 개로 포화지점에 이르는 것을 관찰할 수 있었다(Fig. 2). 이 그래프에서 Mao의 공식에 의해 구한 예상 값을 비교해 본 결과 800명일 때 640여 개, 1,000명일 때 790여 개, 1,200명일 때 950여 개로 각각 Mao의 추정값인 636개, 786개 및 953개에서 크게 벗어나지 않았다(Fig. 1).

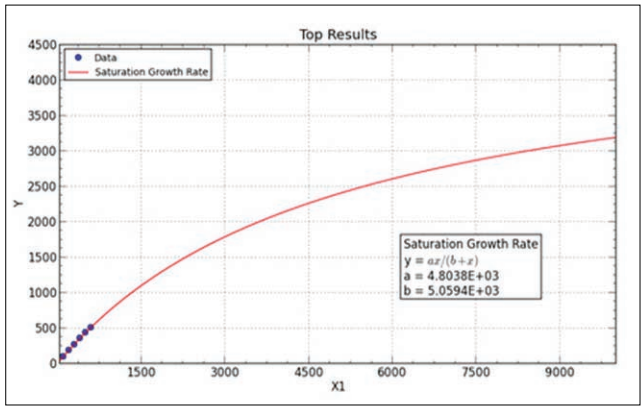


Fig. 2. Result of simulated saturation curve from N. of observed haplotypes.
Graph is obtained by curve expert professional 1.6.5 version. The fit converged to a tolerance of 1e-006 in 5 iterations. No weighting is used.

고 찰

위 결과에서도 확인할 수 있듯이 염기서열(Sequence) 다양성 및 일배체형다양성 모두 자료 크기와 상관없이 같은 수준을 유지함을 보였다. 이는 다양성이 가지는 의미가 한 집단에 존재하는 유전적 변이의 다형성(Polymorphism)을 알아보는 척도[12]이기 때문으로 보이며, 결국 DB 자체의 유용성을 파악하기 위해서는 적절한 파라미터가 될 수 없음을 시사한다. 즉 DB가 커질수록 새로운 일배체형은 계속 발견되기 마련이고, 따라서 diversity index와 같은 기존의 자료들을 이용해 개체식별을 위한 mtDNA DB의 유용성을 언급하기에는 적절하지 않다고 생각한다. 결국, 새로운 통계적인 접근이 필요하다. 몇몇 통계 유전학자들을 포함한 관련 연구자들은 mtDNA DB의 유용성 확인을 위해 기존의 집단 유전학에서 쓰이던 통계 파라미터를 응용하여 몇 가지 공식을 제안하였다.[5, 7, 10] 이들이 제안한 통계 파라미터들은 coverage 개념을 응용한 것이다, coverage란 주어진 DB를 통해 전체 집단에 존재하는 대립유전자의 수를 추정하는 개념이다. 본 연구에서는 실제 mtDNA 자료를 크기 별로 구성하여 Egeland 등[6]이 제안한 coverage를 구해본 결과, 이 값은 100명일 때 0.113에서 600명일 때 0.260으로 증가하여 DB의 크기에 의존적임을 알 수 있었다(Table 3).

일반적으로 coverage 값은 집단 내에서 존재하는 모든 일배체형이 발견되었을 때를 1.0으로 가정한다. 이를 본 연구 결과에 적용하여 해석해보면 100명일 때 이 값이 0.113이라는 것은 아직 발견되지 않은 일배체형이 존재할 확률은 coverage 개념을 이용하면 0.887이란 수치로 표현될 수 있다. 본 연구에서 분석한 전체 653명에서 구한 값은 0.260으로 이는 1.0에 비해 매우 낮은 수치이다. 이는 현재의 DB는 일부에 국한한 분포로 일배체형의 존재 여부 혹은 분포양상만을 알 수 있을 뿐이며, 결국 600명 정도의 DB는 법의학 영역에서 여러 확률적 제고를 하기에는 충분하지 않음을 의미한다. 다만 coverage는 연구자가 가지고 있는 mtDNA DB가 전체 모집단에 비해서 어느 정도의 위치에 있는지를 파악하는 데 있어서 유용한 파라미터가 될 수 있으나, '어느 정도 수준의 coverage라면 충분한가'를 판단하기 위한 정확한 기준점 마련을 위해서는 좀 더 추가적인 연구가 필요하다고 생각된다. 물론 이러한 기준점들은 예를 들자면 일치 여부를 위한 것인지 혹은 법과학 영역에서와 같이 시료의 계통발생학적 연구에서 집단적 분포 비교를 위한 것인지 등과 같이 DB의 활용 목적에 따라 차이가 있을 것으로 예상한다. 즉, 개체식별의 목적이 아니고 집단 간 계통분류적인 차이를 넓게 보기 위해서는 우리가 유사한 일배체형으로 묶어 일배체그룹 분포를 비교하여 볼 수 있겠고, 이 경우 묶음의 과정에서 coverage의 변화가 오기 때문이다. 결국, 개인식별이냐 혹은 집단 간 비교나 등의 목적에 따라 유용한 DB의 크기는 적지 않게 차이가 있을 수 있다고 본다.

또 다른 관점에서 coverage 개념을 점검하여 보고자 하였고, Mao가 시도하였던 것과 같이 집단 내에서 공유하는 일배체형의 빈도를 이용하여 t라는 상수를 도입하여 보았다. Mao는 t = 1일 때 기존 집단 자료의 두 배 크기에서 추정 가능한 일배체형으로 가정하였고,[11] 본 연구에서는

$$\Delta(t) = f_1 t - f_2 t^2 + f_3 t^3 - \ldots. \quad (7)$$

공식에 대입하여 각각의 값을 구해보았다. 실제 자료를 이용한 값과 공식에 의해 추정한 값은 유의수준 5%에서 통계적 검증을 한 결과 유의한 차이가 없었다(Table 5). 이를 활용하면 Mao의 공식을 통해 연구자가 가지고 있는 DB를 이용하여 더 커진 DB에서 관찰 가능한 일배체형의 개수를 추정할 수 있을 것으로 예상한다. 본 연구 대상 DB의 크기가 크지 않아 1,200명에 해당하는 크기에 대해서만 추정값을 제시할 수 있었는데, 다른 자료와 비교 분석하면 이와 같은 접근의 효용성을 가늠하

http://dx.doi.org/10.7580/kjlm.2014.38.2.59

여 볼 수 있지 않나 생각한다.

우리나라 민족에서의 총 일배체형의 숫자와 관련하여, 결과에서 보듯이 Pereira 등이 제시한 포화곡선[8]으로 100명부터 600명 단위로 예상 값을 구하여 Mao의 공식[11]을 비교한 결과 큰 차이는 없었다. 자료의 크기가 확대될수록 신뢰 상한선의 폭도 넓어지지만 약 100,000명일 때 예상되는 일배체형의 개수는 4,000~5,000여 개 수준으로 포화지점에 이르는 것을 관찰할 수 있었다. 다만 앞서 coverage를 구해서 본 결과와 비교하여, 겨우 600명의 자료가 변하는 양상을 이용하여 100,000여 명의 자료를 추정하는 것은 오차의 범위가 클 가능성을 배제할 수 없다.[6, 13]

종합하여 보자면 653명의 한국인 mtDNA DB로는 개체식별을 위한 확률 통계적인 근거를 제시하기 어렵고, 결국 제한적인 DB 크기를 고려해 통계적인 해석에는 좀 더 유의해야 한다고 생각한다. 그러나 집단 사이에 대략적인 계통 비교를 위해서라면 한국인에 대해 200명 이상이면 하플로 그룹의 전체적인 분포 양상을 확인할 수 있겠고, 결국 목적에 따라 DB 크기의 유용성은 다르게 판단하여야 한다. 그리고 새로운 개념인 coverage의 경우, DB 유용성에 중요한 파라미터일 가능성을 확인할 수 있었다. 다른 민족들과의 비교를 위해서는 크기와 HV의 범위와 같은 다른 조건들이 같아야 하겠지만, 전반적인 자료들과 비교하자면 한국인 자료에서의 coverage는 유럽인 자료들에 비해 낮았다. Egeland 등이 구한 자료 중 본 연구자료와 그 크기가 가장 비슷한 540명의 포르투갈 DB에서 구한 coverage 값과 비교해 본 결과, 포르투갈 DB에서 구한 값은 0.58로 한국인 DB에서 구한 수치인 0.223 (N = 500)와 0.252 (N = 600)의 두 배에 가깝다.[6] 이는 한국인이 모계 유전의 측면에서 다양성이 높기 때문으로 볼 수 있겠다. coverage가 어느 정도의 시점에 이르면 충분한 자료라고 말할 수 있는지 기준점을 정하는 것은 좀 어렵다고 본다. 결국, 이러한 coverage 개념을 적극적으로 활용하기 위해서는 좀 더 많은 수의 한국인에 대한 자료를 통한 검증이 필요할 것이다.

참 고 문 헌

1. Torroni A, Achilli A, Macaulay V, et al. Harvesting the fruit of the human mtDNA tree. Trends Genet 2006;22:339-45.
2. Torroni A, Schurr TG, Cabell MF, et al. Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet 1993;53:563-90.
3. Egeland T, Bøvelstad HM, Storvik GO, et al. Inferring the most likely geographical origin of mtDNA sequence profiles. Ann Hum Genet 2004;68:461-71.
4. Chao A, Lee SM. Estimating the number of classes via sample coverage. JASA 1992;87:210-7.
5. Huang SP, Weir BS. Estimating the total number of alleles using a sample coverage method. Genetics 2001;159:1365-73.
6. Egeland T, Salas A. Estimating haplotype frequency and coverage of databases. PLoS one 2008;3:e3998.
7. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989;123:585-95.
8. Pereira L, Cunha C, Amorim A. Predicting sampling saturation of mtDNA haplotypes: an application to an enlarged Portuguese database. Int J Legal Med 2004;118:132-6.
9. Pfeiffer H, Brinkmann B, Hühne J, et al. Expanding the forensic German mitochondrial DNA control region database: genetic diversity as a function of sample size and microgeography. Int J Legal Med 1999;112:291-8.
10. Haas PJ, König C. A bi-level Bernoulli scheme for database sampling. In proceedings of the 2004 ACM SIGMOD international conference on Management of data. ACM 2004;275-86.
11. Mao CX. Predicting the conditional probability of discovering a new class. JASA 2004;99:1108-18.
12. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 1979;76:5269-73.
13. Bunge J, Fitzpatrick M. Estimating the number of species: a review. JASA 1993;88:364-73.