SNPAnalyzer: web-based workbench for the SNPs analysis

Jinho Yoo[1], Bonghee Seo[1], Yangseok Kim[1*]

[1] Bioinformatics Unit, Istech Inc., #704 Hyundai Town Vill, 848-1 Janghang, Ilsan, Goyang, Gyeonggido, 411-380, Korea

[*]To whom correspondence should be addressed. E-mail: yskim@istech21.com

Abstract

Summary: The analysis of human genetic variation is one of the key issues for the understanding of the different drug response among individuals and many programs are developed for this purpose. However, current publicly available programs have so many limitations such as time complexity problem for the analysis of large amount of alleles or SNPs, difficult manipulation for installation, data import, and usage, and low-quality visual output.

Here we present workbench for SNP anlaysis, SNPAnalyzer. SNPAnalyzer consists of 3 main modules: 1) Hardy-Weinberg Equilibrium, 2) Haplotype Estimation, and 3) Linkage Disequilibrium. Each module has several different widely-used algorithms for the extensive analysis and can handle large amount of alleles and SNPs with simple format. Analysis results are displayed in user-friendly formats such as table, graph and map. SNPAnalyzer is developed using C and C++ and users can easily access through web-interface.

Availability: SNPAnalyzer can be freely implemented at http://www.istech.info/istech/board/login_form.jsp

Introduction

인간 유전자 변이와 집단 유전학을 연구하는데 이용되는 프로그램은 현재 많이 나와 있다. EH 프로그램 (Terwilliger and Ott, 1994; Ott, 2003) 과 Haplotyper (Niu et al., 2002) 는 genotype 데이터에서 haplotype 데이터를 통계적인 방법으로 추정하는 프로그램으로 현재 가장 많이 이용되고 있는 text-based 프로그램들이다. 이러한 프로그램들의 가장 큰 단점은 분석을 위한 데이터 변환 과정과 분석 결과를 해석하기가 까다롭다는 점이다. 또한, 각 프로그램은 단 한가지만의 분석 알고리즘을 적용하고 있기 때문에 서로 다른 알고리즘간의 비교, 분석이 어렵다. 본 연구에서는 이러한 기존 프로그램들의 단점을 개선한 프로그램인 SNPAnalyzer 를 개발하였다. SNPAnalyzer 는 입력 데이터 형식을 실험적으로 나오는 genotype 서열 데이터 형식과 거의 동일하게 하여 데이터 변환과정이 거의 필요 없게 하여 사용자의 편의성을 증대 시켰다. Haplotype estimation 을 위해서는 Clark's algorithm (Clark, 1990) 과 EM-based algorithm (Excoffier and Slatkin, 1995),

This work is supported by Ministry of Science and Technology

Gibbs sampling-based algorithm (Stephens et al., 2001; Niu et al., 2002) 의 세가지 알고리즘을 적용하여 각 알고리즘을 통해 나오는 결과를 비교, 분석할 수 있게 하였다. Linkage disequilibrium 측정을 위해서는 D, D', |D'|, Δ, Δ^2 의 5가지 index (Hill and Robertson, 1968; Hill and Weir, 1994; Lewontin, 1964; Devlin and Risch, 1995) 와 Fisther's Exact P-value 를 이용하였으며, four gamete test 결과도 병행하여 나타내었다. 또한, 웹상에서 실행 가능하게 하고 분석 결과를 표와 그래프 형태로 나타내어 한눈에 파악할 수 있게 하는 등 사용자의 편의를 도모하였다. 알고리즘은 C와 C++ 를 이용하여 개발하였으며 인터넷 브라우저를 통해 실행할 수 있다.

Methods

Data

분석을 위한 데이터는 traditional serological alleles 가 molecular (SNP) 레벨에서 정의된 human HLA region 의 haplotype 데이터를 이용하였다. 표 1은 분석에 이용된 데이터를 설명하고 있다. 이 데이터는 NCBI 의 dbSNP 에서 구하였으며, United States 내에서 자원자를 대상으로 African American, Asian American, Caucasian, Latin American, Native American 의 5개의 ethnic group 으로 구분한 후 haplotyping 한 결과이다. 본 연구에서는 African American 과 Asian American 의 두 집단을 대상으로 분석하였다. 각 집단은 각각 24, 25 개의 가족데이터로 구성되어 있으며, 각 가족은 부, 모, 자식의 3 명으로 구성되어 있다. 본 연구에서는 가족력을 고려하지 않은 분석 방법을 이용하였기 때문에 African American 집단은 72명의 개인, Asian American 집단은 75명의 개인으로 구성된 서로 독립적인 두개의 집단으로 보고 분석하였다. Haplotype 은 모든 집단에서 동일하게 31개의 SNP 로 이루어져 있다. 그러나, 각 집단을 대상으로 SNP 들의 서열을 확인한 후 biallelic 한 SNP 만을 대상으로 하여 haplotype 을 재정의하였으며, 재정의한 haplotype 쌍을 이용해 각 집단을 구성하는 개개인들의 genotype 을 구성하였다. 이렇게 구성된 genotype 은 haplotype estimation algorithm을 통해 역으로 haplotype 의 서열과 빈도를 추정하여 실제 haplotype 과의 비교를 하였다.

표 1. Ethnic group 별 sample 수와 ethnic group 별 haplotype 을 재정의하는데 사용된 SNP 개수

Ethnic Group	Afr	Asi
Sample 수	72	75
SNP 개수	22	22

Note: Afr 은 African American, Asi 는 Asian American 을 각각 의미한다.

Development Environment

SNPAnalyzer 는 웹기반에서 실행되는 SNP 분석 프로그램이다. Microsoft Windows 98 이상의 운영체제와 Explorer 6.0 이상의 웹 브라우저가 탑재된 pc 에서 최적으로 실행된다. SNPAnalyzer 의 알고리즘은 C와 C++ 을 이용해 개발하였다. SNPAnalyzer 를 사용하기 위해서는 간단한 로그인 과정을 거친 후 웹써버에서 실행파일을 한번 다운 받으면 된다. 실제 프로그램 실행은 시용자가 실행파일을 다운받은 pc 에서 ActiveX 를 이용해 실행이 된다. 그림 1은 SNPAnalyzer 의 실행방식과 내부 구성 모듈을 나타내고 있다.

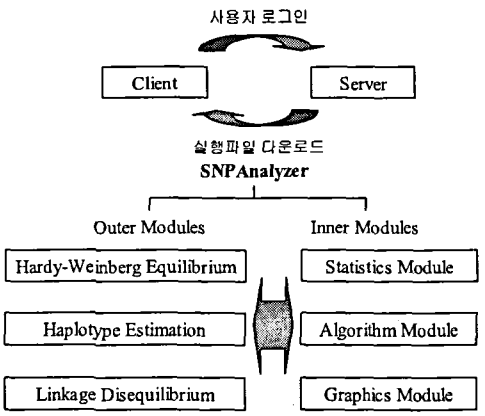

그림 1.. SNPAnalyzer 의 실행방식과 구성모듈

Hardy-Weinberg Equilibrium

개체군내 대립유전자형의 분포가 일정한 상태를 유지하는지를 확인하기 위해 아래와 같은 HWE (Hardy-Weinberg Equilibrium) 테스트를 한다.

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

여기서, Oi 는 i번째 genotype 의 관측된 빈도이고 Ei 는 기대치이다. HWE 가 성립하기 위한 조건에는 다음과 같은 것들이 있다. 1) 교배는 완전히 무작위적이어야 한다. 2) 돌연변이는 있을 수 없다. 3) 이입과 이출이 있을 수 없다. 4) 대립유전자는 멘델의 제 1 법칙에 따라 분리되어야 한다. 5) 기대값은 개체군과 표본집단의 크기가 대단히 클 때에 한해서만 정확하다. 6) 개체군에는 선택이 작용하지 않는다. SNPAnalyzer 는 genotype 데이터의 모든 SNP 에 대한 HWE 를 동시에 테스트한 후에 χ^2 값과 p-value 값을 테이블 형태로 보여준다.

Haplotype Estimation

실험을 통해 밝혀지는 것은 일반적으로 반수체 (haploid) 가 아닌 배수체 (diploid) 서열 정보인 genotype 데이터이다. Genotype 데이터에는 phase 에 대한 정보가 없기 때문에, 집단 내에 분포하는 haplotype 서열 정보와 haplotype 빈도는 확률과 통계적인 방법으로 추정 (estimation) 할 수 밖에 없다. 이러한 haplotype estimation 에 사용되는 알고리즘으로는 Clark's algorithm, EM-based algorithm, Gibbs sampling-based algorithm 등이 있다. 본 연구에서는 Clark's algorithm 과 EM-based algorithm, Gibbs sampling-based algorithm 을 모두 적용하여 genotype 데이터로부터 haplotype 서열과 빈도를 추정하였고, 추정된 결과와 실제 haplotype 데이터와의 비교를 통해 분석의 정확도를 수치화 하였다.

Clark's Algorithm

Clark's algorithm (Clark, 1990) 은 rule-based 방식을 적용하여 집단을 구성하는 개개인의 haplotype 을 하나씩 차례로 재구성 해나가는 방법이다. 즉, 모든 site 가 homozygous 한 기준 genotype 들을 정렬한 후 각 sample 마다 haplotype 을 구성하고, 이렇게 구성된 haplotype 들의 빈도는 다음 sample의 haplotype 구성을 위해 이용하는 방법이다. Clark's algorithm 은 알고리즘이 상당히 간단하고 각 sample 마다 haplotype 을 재구성할 수 있다는 장점이 있지만, 모든 site 가 homozygous 하거나 단 1개의 site 만 heterozygous 하여 haplotype 을 오차 없이 완전히 구성할 수 있는 sample 이 없는 경우에는 적용할 수가 없다. 또한 heterozygous 한 site 많은 경우 알고리즘 특성상 haplotype 을 구성할 수 없는 sample 이 생길 가능성이 크다. 일반적으로 Clark's algorithm 은 다음에 설명할 EM-based

algorithm 이나 Gibbs sampling-based algorithm 에 비해서는 정확도가 떨어지는 단점이 있다

EM-based Algorithm

EM-based algorithm (Excoffier and Slatkin, 1995) 은 likelihood-based 방식을 이용하여 집단에 분포하는 haplotype 의 빈도를 추정하는 방법으로, 알고리즘이 비교적 쉽고 분석 결과의 정확도도 신뢰할 만하지만, 집단 내에 존재 가능한 모든 haplotype 에 대한 빈도를 추정하는 것이기 때문에 컴퓨터의 용량에 따라 분석할 수 있는 SNP 수와 sample 수가 매우 제한적이다. 현재 공개된 EH 프로그램 (Terwilliger and Ott, 1994; Ott, 2003) 의 경우 10개 이하의 SNP 를 가지는 genotype 데이터만 분석 가능하다. EM-based algorithm 은 다음과 같은 E-step 과 M-step 의 두가지 과정으로 이루어진다:

1. E-step: Expectation formulation of Q,

$Q(\theta,\theta^i) = E[\log p(\theta \mid Z,Y) \mid \theta^i, Y]$

Y : observed data

Z : unobserved data

θ : parameter concerned with Y.

2. M-step: Maximization of function Q,

maximizatin of Q with respect to θ.

Gibbs Sampling-based Algorithm

Gibbs sampling-based algorithm (Stephens et al., 2001; Niu et al., 2002) 은 일종의 MCMC (Markov Chain Monte Carlo) 알고리즘으로, 조건부 확률분포와 반복 sampling 을 이용하여 각 개인의 haplotype 을 재구성하고, 재구성된 haplotype 의 신뢰도를 추정하는 방법이다. 아래는 Stephens et al. (2001) 이 제시한 PGS (Pseudo Gibbs Sampler) 알고리즘이다:

1. Choose an individual, i, uniformly and at random from all ambiguous individuals (i.e., individuals with more than one possible haplotype reconstruction).

2. Sample $H_i^{(t+1)}$ from $\Pr(H_i \mid G, H_{-i}^{(t)})$, where $H_{-i}^{(t)}$ is the set of haplotypes excluding individual i.

3. Set $H_j^{(t+1)} = H_j^{(t)}$ for j=1,...,n, j≠i.

여기서, G 는 모든 sample 들의 genotype 이고, H_i 는 i번째 sample 의 haplotype 이다. Gibbs sampling 을 적용한 프로그램중의 하나인 Haplotyper (Niu et al., 2002) 는 Clarks' algorithm 이나 EM-based algorithm 과 비교할 때 정확성 측면에서 상대적으로 좋고 100개 이상의 SNP 분석이 가능하다는 장점이 있지만 반복 sampling 을 하기 때문에 계산시간이 늦다는 단점이 있다.

공개되어 있는 SNP 분석 프로그램들의 다른 단점은 haplotype 분석을 위해 genotype 데이터를 각 프로그램의 특성에 맞게 변형시켜주어야 한다는 점이다. 결국 데이터 변환을 해주는 부가적인 프로그램이 필요하다. 또한, 분석되어 나온 결과도 서열 정보로의 변환이 필요하기 때문에 사용자 측면에서는 결과의 해석이 쉽지가 않다는 단점이 있다. SNPAnalyzer 는 Clark's algorithm, EM-based algorithm, Gibbs sampling-based algorithm 세가지를 하나의 interface 에서 모두 구현하고 있다. 따라서, 동일한 genotype 데이터에 각 알고리즘을 적용해 나온 추정 결과를 서로 비교 분석 할 수 있다는 장점이 있다. 또한, 분석 결과를 바로 서열정보로 출력하여 결과 해석이 용이하게 하였다. 이밖에, 재구성된 haplotype 분포를 히스토그램으로 표현해

haplotype 간의 빈도차를 쉽게 알아볼 수 있게 하였으며 phylogenetic tree 를 이용하여 재구성된 haplotype 들간의 서열차이를 시각적으로 파악하기 쉽게 하였다. 그림 2는 SNPAnalyzer 를 이용하여 haplotype estimation 을 하였을 때의 분석 결과 화면이다.

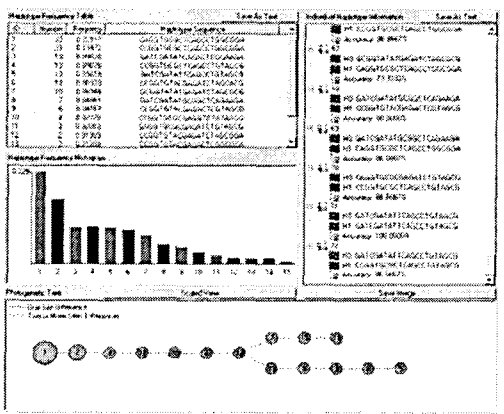

그림 2. SNPAnalyzer 로 haplotype estimation 한 결과. 22개의 SNP 으로 재정의된 haplotype 을 가지는 African American 72 명을 대상으로 Gibbs sampling-based algorithm 을 적용한 결과이다. 그림 왼쪽은 추정된 haplotype 서열과 분포를 나타내며 오른쪽은 72명 각각에 대해 재구성된 haplotype 서열과 신뢰수준을 나타낸다. 아래쪽 그래프는 추정된 haplotype 들간의 phylogenetic tree 를 보여주고 있다.

Linkage Disequilibrium

하나의 chromosome 상에 위치한 두 loci 간의 유전적 연관관계는 두 loci 간의 recombination fraction 을 이용하여 추정한다 (Wu and Zeng, 2001). SNPAnalyzer 는 현재 많이 이용되고 있는 linkage disequilibrium index 중 5개의 index (D, D', |D'|, Δ, Δ^2) 를 이용하여 유전적 연관관계를 나타내었다. 표 2는 biallelic 한 두개의 loci 로 구성되는 haplotype 빈도에 대한 확률적인 기대값을 나타내고 있으며, 이들을 이용하여 두개 loci 에 존재하는 linkage disequilibrium 을 측정할 수가 있다. 아래는 표 2를 이용해서 계산한 linkage disequilibrium index 의 예이다 (Devlin and Risch, 1995).

$$D = (\pi_{11} - \pi_{1+}\pi_{+1} =)\pi_{11}\pi_{22} - \pi_{12}\pi_{21}$$

$$D' = \begin{cases} \dfrac{D}{\min(\pi_{1+}\pi_{+2}, \pi_{+1}\pi_{2+})} & D > 0 \\ \dfrac{D}{\min(\pi_{1+}\pi_{+1}, \pi_{+2}\pi_{2+})} & D < 0 \end{cases}$$

$$\Delta = \frac{D}{(\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2})^{1/2}}$$

SNPAnalyzer 는 이들 index 와 더불어 Fisher's Exact P-value 와 four gamete test 결과도 제시하여 다른 분석 결과와의 비교가 가능하게 하였다. 그림 3은 loci 들간의 linkage disequilibrium 측정 결과를 나타내고 있다.

표 2. *Haplotype* 과 *marker allele* 빈도의 2x2 테이블

Marker		Locus 2		
	Allele	Allele 1	Allele 2	Total
Locus 1	Allele 1	π_{11}	π_{12}	π_{1+}
	Allele 2	π_{21}	π_{22}	π_{2+}
	Total	π_{+1}	π_{+2}	1

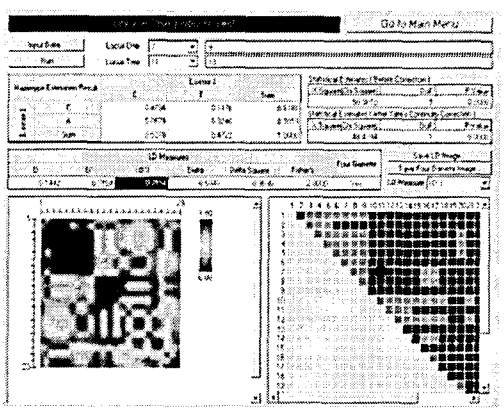

그림 3. Linkage disequilibrium 결과. 각 index 와 Fisher's Exact P-value, 카이제곱 테스트 결과를 테이블 형식으로 나타내었다. 화면 아래는 왼쪽부터 linkage disequilibrium mapping 과 four gamete test 결과를 나타낸다.

Results

Hardy-Weinberg Equilibrium

표 3은 African American, Asian American 집단의 biallelic 한 SNP 에 대해 SNPAnalyzer 를 이용해 HWE 테스트를 한 결과 중 일부를 보여주고 있다. 표 안의 굵은 글씨로 된 부분이 p-value 가 0.05 이하인 결과이다. rs2308479 와 rs2308484 에서는 유의수준을 0.05로 잡았을 때, African American 의 경우 p-value 가 0.004 이므로 HWE 가 성립하지 않음을 알 수 있다. Asian American 의 경우에는 rs2308673 에서 HWE 가 성립하지 않으며 p-value 는 0.029 이다. 이상의 결과로 해당 SNP 에서 유전적 변이가 있음을 추정할 수 있으며 또한, 같은 위치의 SNP 의 allele 유전적 분포가 두 집단에서 상이함을 알 수 있다. 다른 위치의 SNP 의 경우에는 모두 p-value 가 0.05 이상을 보인다.

표 3. HWE 테스트 결과

		Ethnic Group	
SNP No	ID	Afr	Asi
21	rs2308479	**0.004**	0.607
22	rs2308484	**0.004**	0.607
25	rs2308673	0.97	**0.029**

Note: ID 는 NCBI SNP Cluster ID 를 의미한다.

Haplotype Estimation

African American 과 Asian American 두 개의 집단에 대해서 Clark's algorithm, EM-based algorithm, Gibbs sampling-based algorithm 을 적용하여 haplotype estimation 을 하였다. 표 4는 두개의 결과 중 African American 집단에 대해서 haplotype estimation 한 결과를 나타내고 있다. Gibbs sampling-based algorithm 과 EM-based algorithm 으로 추정한 결과 두 집단 모두 총 15개의 실제 haplotype 에서 14개의 똑 같은 haplotype 을 추정하였으며, 실제 haplotype 빈도와 추정된 빈도와의 차이도 그리 크지 않았다. 또한 집단을 구성하는 각 개인의 haplotype 을 비교하였을 때 두 집단 모두 2 명만 서열이 다른 결과를 보였다. 그러나, Clark's algorithm 은 African American, Asian American 두 집단에 대해 각각 3개, 2 개의 haplotype 을 잘못 추정하였으며, 실제 haplotype 빈도와 추정 빈도와의 차이도 다른 두개 알고리즘에 비해 비교적 컸다. 또한, 집단을 구성하는 각 개인의 haplotype 도 African American 의 경우에는 16명, Asian American 의 경우에는 22명을 잘못 추정하였다. Haplotype estimation 의 정확도를 수치화 하기 위해서 다음과 같은 두가지 범주를 고려하였다 (Stephens et al., 2001).

1. 집단을 구성하는 각 개인의 haplotype reconstruction (Clark, 1990).
2. 집단 내에 분포하는 haplotype 빈도 추정 (Excoffier and Slatkin, 1995).

첫번째 범주는 error rate 으로 정확도를 수치화 하는데, 집단 내 전체 sample 수에 대해 haplotype 이 잘못 재구성된 sample 수의 비율을 나타내는 수치이다. 두번째 범주는 discrepancy 로 집단내의 실제 haplotype 빈도와 추정된 haplotype 빈도와의 차이를 나타낸다. 아래는 discrepancy 를 구하는 식이다.

$$D = \frac{1}{2}\sum | \hat{f}_j - f_j | .$$

여기서, $\hat{f}_j$ 는 j번째 sample 의 추정된 haplotype frequency 이고 f_j 는 실제 haplotype frequency 이다 (Stephens et al., 2001).

표 4. NCBI dbSNP 의 African American 72 명 집단내의 실제 haplotype 빈도와 추정된 haplotype 빈도

Haplotype No	dbSNP	Gibbs		EM		Clark	
	freq	freq	correct	freq	correct	freq	correct
1	0.215	0.229	Y	0.222	Y	0.132	Y
2	0.174	0.16	Y	0.167	Y	0.257	Y
3	0.09	0.09	Y	0.09	Y	0.069	Y
4	0.09	0.09	Y	0.091	Y	0.083	Y
5	0.09	0.09	Y	0.09	Y	0.09	Y
6	0.083	0.083	Y	0.083	Y	0.069	Y
7	0.069	0.069	Y	0.067	Y	0.042	Y
8	0.049	0.049	Y	0.048	Y	0.042	Y
9	0.042	0.042	Y	0.041	Y	0.035	Y
10	0.028	0.028	Y	0.028	Y	0.007	Y
11	0.021	0.021	Y	0.019	Y	0.014	Y
12	0.014	0.014	Y	0.013	Y	-	N
13	0.014	0.014	Y	0.014	Y	0.014	Y
14	0.014	-	N	-	N	-	N
15	0.007	0.007	Y	0.007	Y	-	N

Note: Haplotype No 는 실제 haplotype 에 대해 일련번호를 매긴 것이다. Gibbs 는 Gibbs sampling-based algorithm, EM 은 EM-based algorithm, Clark 은 Clark's algorithm 을 의미한다. freq 는 전체 집단에 대한 haplotype 빈도를 의미하며, correct 는 실제 haplotype 이 추정된 결과에 있으며 "Y", 없으면 "N" 로 나타낸다.

표 5는 African American 과 Asian American 두 집단에 대해서 haplotype estimation 한 결과와 실제 데이터와의 정확도를 비교한 것이다. Gibbs sampling-based algorithm 과 EM-based algorithm 의 경우에는 AER 과 DIS 두 가지 모두 0.04 미만을 나타내고 있다. 이것은 실제 haplotype 과 추정된 haplotype 사이에 별 차이가 없음을 의미한다. 그렇지만 Clark's algorithm 으로 haplotype estimation 을 한 경우에는 AER, DIS 모두 0.15 이상의 큰 값을 보이고 있어 haplotype estimation 의 신뢰도가 나머지 두 알고리즘에 비해 떨어진다고 할 수 있다.

표 5. Haplotype estimation 의 정확도

Ethnic Group	Afr		Asi	
Error	AER	DIS	AER	DIS
Gibbs	0.021	0.028	0.017	0.027
EM	0.018	0.028	0.033	0.027
Clark	0.156	0.222	0.162	0.293

Note: AER 은 Average Error Rate, DIS 는 Discrepancy 를 의미한다.

Linkage Disequilibrium

Linkage disequilibrium 테스트를 하기 위해 EM-based algorithm 을 적용하여 두개의 SNP 로 구성되는 haplotype 을 추정하였다. 그림

4는 African American 에 대해 linkage disequilibrium 측정한 결과 중 |D'| 값과 four gamete test 결과를 나타낸 것이다. Linkage disequilibrium 측정 결과는 GOLD 프로그램 (Abecasis and Cookson, 2000) 에서 표현한 방식과 비슷하게 2-D mapping 형식으로 나타내었다. LD mapping 에서 붉은색으로 표시될수록 해당 SNP 들 사이에서 linkage disequilibrium 정도가 커짐을 나타낸다. 전반적으로 보았을 때, 22개의 SNP 중 2번째부터 9번째 SNP 사이에 비교적 큰 linkage disequilibrium block 이 생성되어 있음을 알 수 있다. 10번째와 14번째 SNP 사이에도 소규모의 linkage disequilibrium block 이 생성되어 있음을 알 수 있다. 이러한 linkage disequilibrium 측정 결과는 four gamete test 결과와 상당히 비슷한 패턴을 보여주고 있다.

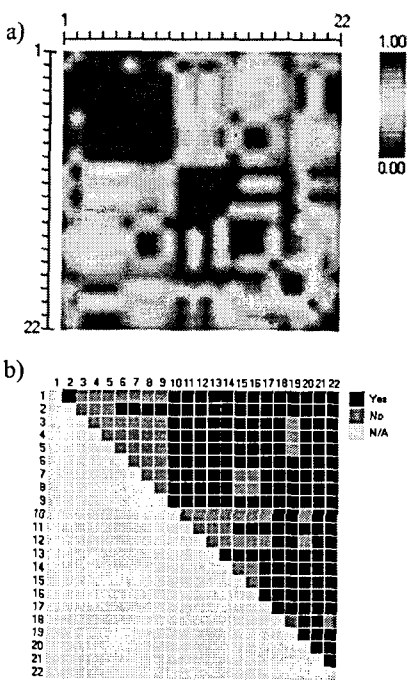

그림 4. NCBI dbSNP 의 African American 집단내 22 SNP 간의 연관관계. a) Linkage Disequilibrium mapping by |D'|. b) Four gamete test

Discussion

SNPAnalyzer 개발의 주 목적은 두가지 이다. 하나는 인간 유전자 변이와 집단유전학을 연구하는데 사용되는 여러 프로그램 기능의 통합적인 적용이며, 이를 위해 phase 에 대한 정보가 없는 genotype 데이터를 입력데이터로 받아들여 HWE 테스트, LD 측정, haplotype estimation 등 여러 분석을 동시에 수행할 수 있게 하였다. 다른 하나는 분석을 수행하는 과정을 최소한도로 줄이고 사용자 입장에서 분석결과를 쉽고 다양하게 해석할 수 있게 하는 것이다. 두번째 목적을 위해서, 다른 프로그램에서 SNP 분석을 위해 필수적인 데이터 변환 과정이 거의 필요없게 하였다. 또한, 프로그램을 웹상에서 실행 가능하도록 하여 복잡한 설치 절차가 필요없게 하였으며 여러 개의 분석 알고리즘을 적용하였고, 분석 결과를 다양한 표와 그래프 형태로 나타내었다. 분석 결과의 정확도를 평가하기 위해서는 NCBI 의 dbSNP 에서 구한 5개의 ethnic group 중 African American 72명, Asian American 75 명에 대한 haplotype 데이터와의 비교를 수행하였다. 정확도 비교 결과 Gibbs sampling-based algorithm 과 EM-based algorithm 은 실제 haplotype 데이터와 상당히 근접한 결과를 보였지만 Clark's algorithm 은 비교적 신뢰도가 떨어지는 결과를 보였다. LD 측정을 위해서는 현재 많이 사용되고 있는 D, D', |D'|, Δ, Δ^2 의 5 가지 index (Hill and Robertson, 1968; Hill and Weir, 1994; Lewontin, 1964; Devlin and Risch, 1995) 를 이용하였으며, 각 계산 결과는 2-D mapping 형태로 나타내어 각 SNP 간의 연관성과 LD block 정도를 쉽게 파악 할 수 있게 하였다. 또한 four gamete test 도 동시에 수행하도록 하여 LD

mapping 결과와의 비교가 가능하도록 하였다.

SNPAnalyzer 는 현재 각 loci 가 두개의 형질만을 가지는 SNP 데이터 분석만 가능하다. 추후에는 각 loci 가 두개 이상의 형질을 가지는 데이터도 분석할 수 있는 기능을 추가할 예정이다. 또한, 유전자형 데이터만을 가지고 분석하는 것 이외에 sample 이 가지는 trait 과 loci 들간의 관계를 추정하는 QTL analysis 모듈도 추가해서 개발할 예정이다. Control-case 연구를 할 수 있는 환경을 개발하고 가족력 데이터가 있는 경우 이를 이용한 분석 알고리즘도 새로 개발할 예정으로 있다.

References

[1] Abecasis,G..R. and Cookson,W.O.C. (2000) GOLD-graphical overview of linkage disequilibrium. Bioinformatics Application Note, 16, 182-183.

[2] Clark,A.G. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol., 7, 111-122.

[3] Devlin,B. and Risch,N. (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics, 29, 311-322.

[4] Excoffier,L. and Slatkin,M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol., 12, 921-927.

[5] Hill,W.G. and Robertsoon,A. (1968) Linkage disequilibrium in finite populations. Theor. Appl. Genet., 38, 226-231.

[6] Hill,W.G. and Weir,B.S. (1994) Maximum-likelihood estimation of gene location by linkage disequilibrium. Am. J. Hum. Genet. 54, 705-714.

[7] Lewontin,R.C. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics, 49, 49-67.

[8] Niu,T., Qin,Z.S., Xu,X., and Liu,J.S. (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am. J. Hum. Genet., 70, 157-169.

[9] Ott,J. (2003) User's guide to the EH program. Rockefeller University, New York.

[10] Stephens,M., Smith,N.J., and Donnelly,P. (2001) A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet., 68, 978-989.

[11] Terwilliger,J., and Ott,J. (1994) Handbook of human genetic linkage. Johns Hopkins University Press, Baltimore.

[12] Wu,R., and Zeng,Z.B. (2001) Joint linkage and linkage disequilibrium mapping in natural populations. Genetics, 160, 899-909.

[13] Xu,C.F., Lewis,K., Cantone,K.L., Khan,P., Donnelly,C., White,N., Crocker,N., Boyd,P.R., Zaykin,D.V., Purvis,I.J., (2002) Effectiveness of computational methods in haplotype predictioin. Hum. Genet., 110, 148-156.

Web Site References

http://www.ncbi.nlm.nih.gov/SNP/; dbSNP