REFERENCES

Alansari, A., A. Hajeer, et al. (2002). "Transforming growth factor-12 polymorphism and systemic lupus erythematosus." J Rheumatol 29(6): 1189-91.

Alansari, A., A. H. Hajeer, et al. (2001). "Two novel polymorphisms in the human transforming growth factor beta 2 gene." Genes Immun 2(5): 295-6.

Amani, D., J. Zolghadri, et al. (2004). "The promoter region (-800, -509) polymorphisms of transforming growth factor-beta1 (TGF-beta1) gene and recurrent spontaneous abortion." J Reprod Immunol 62(1-2): 159-66.

Arkwright, P. D., S. Laurie, et al. (2000). "TGF-beta(1) genotype and accelerated decline in lung function of patients with cystic fibrosis." Thorax 55(6): 459-62.

Arnett, F. C., J. D. Reveille, et al. (1997). "Autoantibodies to a nucleolar RNA helicase protein in patients with connective tissue diseases." Arthritis Rheum 40(8): 1487-92.

Arrick, B. A., R. L. Grendell, et al. (1994). "Enhanced translational efficiency of a novel transforming growth factor beta 3 mRNA in human breast cancer cells." Mol Cell Biol 14(1): 619-28.

Awad, M. R., A. El-Gamel, et al. (1998). "Genotypic variation in the transforming growth factor-beta1 gene: association with transforming growth factor-beta1 production, fibrotic lung disease, and graft fibrosis after lung transplantation." Transplantation 66(8): 1014-20.

Beranek, M., K. Kankova, et al. (2002). "Polymorphism R25P in the gene encoding transforming growth factor-beta (TGF-beta1) is a newly identified risk factor for proliferative diabetic retinopathy." Am J Med Genet 109(4): 278-83.

Bowness, P., K. A. Davies, et al. (1994). "Hereditary C1q deficiency and systemic lupus erythematosus." Qjm 87(8): 455-64.

Buckova, D., L. Izakovicova Holla, et al. (2001). "TGF-beta1 gene polymorphisms." Allergy 56(12): 1236-7.

Cambien, F., S. Ricard, et al. (1996). "Polymorphisms of the transforming growth factor-beta 1 gene in relation to myocardial infarction and blood pressure. The Etude Cas-Temoin de l'Infarctus du Myocarde (ECTIM) Study." Hypertension 28(5): 881-7.

Caserta, T. M., A. A. Knisley, et al. (2004). "Genotypic analysis of the TGF beta-509 allele in patients with systemic lupus erythematosus and Sjogren's syndrome." Ann Genet 47(4): 359-63.

de Martin, R., B. Haendler, et al. (1987). "Complementary DNA for human glioblastoma-derived T cell suppressor factor, a novel member of the transforming growth factor-beta gene family." Embo J 6(12): 3673-7.

de Souza, A. P., P. C. Trevilatto, et al. (2003). "Analysis of the TGF-beta1 promoter polymorphism (C-509T) in patients with chronic periodontitis." J Clin Periodontol 30(6): 519-23.

Derynck, R., J. A. Jarrett, et al. (1985). "Human transforming growth factor-beta complementary DNA sequence and expression in normal and transformed cells." Nature 316(6030): 701-5.

Gaffney, P. M., G. M. Kearns, et al. (1998). "A genome-wide search for susceptibility genes in human systemic lupus erythematosus sib-pair families." Proc Natl Acad Sci U S A 95(25): 14875-9.

Gaffney, P. M., W. A. Ortmann, et al. (2000). "Genome screening in human systemic lupus erythematosus: results from a second Minnesota cohort and combined analyses of 187 sib-pair families." Am J Hum Genet 66(2): 547-56.

Gentry, L. E., M. N. Lioubin, et al. (1988). "Molecular events in the processing of recombinant type 1 pre-pro-transforming growth factor beta to the mature polypeptide." Mol Cell Biol 8(10): 4162-8.

Gewaltig, J., K. Mangasser-Stephan, et al. (2002). "Association of polymorphisms of the transforming growth factor-beta1 gene with the rate of progression of HCV-induced liver fibrosis." Clin Chim Acta 316(1-2): 83-94.

Gladman, D. D. (1996). "Prognosis and treatment of systemic lupus erythematosus." Curr Opin Rheumatol 8(5): 430-7.

Gorelik, L. and R. A. Flavell (2002). "Transforming growth factor-beta in T-cell biology." Nat Rev Immunol 2(1): 46-53.

Grande, J. P. (1998). "Mechanisms of progression of renal damage in lupus nephritis: pathogenesis of renal scarring." Lupus 7(9): 604-10.

Gray-McGuire, C., K. L. Moser, et al. (2000). "Genome scan of human systemic lupus erythematosus by regression modeling: evidence of linkage and epistasis at 4p16-15.2." Am J Hum Genet 67(6): 1460-9.

Gray, J. D., M. Hirokawa, et al. (1994). "The role of transforming growth factor beta in the generation of suppression: an interaction between CD8+ T and NK cells." J Exp Med 180(5): 1937-42.

Hochberg, M. C. (1997). "Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus." Arthritis Rheum 40(9): 1725.

Hogan, B. L., M. Blessing, et al. (1994). "Growth factors in development: the role of TGF-beta related polypeptide signalling molecules in embryogenesis." Dev Suppl: 53-60.

Hoh, J., F. Matsuda, et al. (2003). "SNP haplotype tagging from DNA pools of two individuals." BMC Bioinformatics 4: 14.

Horwitz, D. A., J. D. Gray, et al. (1997). "The immunoregulatory effects of NK cells: the role of TGF-beta and implications for autoimmunity." Immunol Today 18(11): 538-42.

http://innateimmunity.net.gov/SNP/.

http://linkage.rockefeller.edu/register/.

http://snpper.chip.org.

Jakowlew, S. B., T. W. Moody, et al. (1998). "Transforming growth factor-beta expression in mouse lung carcinogenesis." Exp Lung Res 24(4): 579-93.

Kaartinen, V., J. W. Voncken, et al. (1995). "Abnormal lung development and cleft palate in mice lacking TGF-beta 3 indicates defects of epithelial-mesenchymal interaction." Nat Genet 11(4): 415-21.

Kingsley, D. M. (1994). "The TGF-beta superfamily: new members, new receptors, and new genetic tests of function in different organisms." Genes Dev 8(2): 133-46.

Kondaiah, P., M. J. Sands, et al. (1990). "Identification of a novel transforming growth factor-beta (TGF-beta 5) mRNA in Xenopus laevis." J Biol Chem 265(2): 1089-93.

Kotzin, B. L. (1996). "Systemic lupus erythematosus." Cell 85(3): 303-6.

Kulkarni, A. B. and S. Karlsson (1993). "Transforming growth factor-beta 1 knockout mice. A mutation in one cytokine gene causes a dramatic inflammatory disease." Am J Pathol 143(1): 3-9.

Letterio, J. J. and E. P. Bottinger (1998). "TGF-beta knockout and dominant-negative receptor transgenic mice." Miner Electrolyte Metab 24(2-3): 161-7.

Lindqvist, A. K., K. Steinsson, et al. (2000). "A susceptibility locus for human systemic lupus erythematosus (hSLE1) on chromosome 2q." J Autoimmun 14(2): 169-78.

Lu, L. Y., H. H. Cheng, et al. (2004). "Single-nucleotide polymorphisms of transforming growth factor-beta1 gene in Taiwanese patients with systemic lupus erythematosus." J Microbiol Immunol Infect 37(3): 145-52.

Luedecking, E. K., S. T. DeKosky, et al. (2000). "Analysis of genetic polymorphisms in the transforming growth factor-beta1 gene and the risk of Alzheimer's disease." Hum Genet 106(5): 565-9.

Madisen, L., A. Krumm, et al. (1998). "The immunoglobulin heavy chain locus control region increases histone acetylation along linked c-myc genes." Mol Cell Biol 18(11): 6281-92.

Massague, J. (1990). "The transforming growth factor-beta family." Annu Rev Cell Biol 6: 597-641.

Massague, J. (1998). "TGF-beta signal transduction." Annu Rev Biochem 67: 753-91.

Millan, F. A., F. Denhez, et al. (1991). "Embryonic gene expression patterns of TGF beta 1, beta 2 and beta 3 suggest different developmental functions in vivo." Development 111(1): 131-43.

Miyazono, K., U. Hellman, et al. (1988). "Latent high molecular weight complex of transforming growth factor beta 1. Purification from human platelets and structural characterization." J Biol Chem 263(13): 6407-15.

Morel, L., B. P. Croker, et al. (2000). "Genetic reconstitution of systemic lupus erythematosus immunopathology with polycongenic murine strains." Proc Natl Acad Sci U S A 97(12): 6670-5.

Morel, L., C. Mohan, et al. (1999). "Multiplex inheritance of component phenotypes in a murine model of lupus." Mamm Genome 10(2): 176-81.

Morel, L., U. H. Rudofsky, et al. (1994). "Polygenic control of susceptibility to murine systemic lupus erythematosus." Immunity 1(3): 219-29.

Morel, L. and E. K. Wakeland (1998). "Susceptibility to lupus nephritis in the NZB/W model system." Curr Opin Immunol 10(6): 718-25.

Moser, K. L., B. R. Neas, et al. (1998). "Genome scan of human systemic lupus erythematosus: evidence for linkage on chromosome 1q in African-American pedigrees." Proc Natl Acad Sci U S A 95(25): 14869-74.

Munger, J. S., J. G. Harpel, et al. (1997). "Latent transforming growth factor-beta: structural features and mechanisms of activation." Kidney Int 51(5): 1376-82.

Nath, S. K., J. Kilpatrick, et al. (2004). "Genetics of human systemic lupus erythematosus: the emerging picture." Curr Opin Immunol 16(6): 794-800.

Nath, S. K., B. Namjou, et al. (2004). "A candidate region on 11p13 for systemic lupus erythematosus: a linkage identified in African-American families." J Investig Dermatol Symp Proc 9(1): 64-7.

Ohtsuka, K., J. D. Gray, et al. (1998). "Decreased production of TGF-beta by lymphocytes from patients with systemic lupus erythematosus." J Immunol 160(5): 2539-45.

Ohtsuka, T., A. Yamakage, et al. (2002). "The polymorphism of transforming growth factor-beta1 gene in Japanese patients with systemic sclerosis." Br J Dermatol 147(3): 458-63.

Proetzel, G., S. A. Pawlowski, et al. (1995). "Transforming growth factor-beta 3 is required for secondary palate fusion." Nat Genet 11(4): 409-14.

Ram, M. D. and G. D. Chisholm (1969). "Cystic disease of renal pyramids (medullary sponge kidney)." Br J Urol 41(3): 280-6.

Roberts, A. B. (1998). "Molecular and cell biology of TGF-beta." Miner Electrolyte Metab 24(2-3): 111-9.

Sanford, L. P., I. Ormsby, et al. (1997). "TGFbeta2 knockout mice have multiple developmental defects that are non-overlapping with other TGFbeta knockout phenotypes." Development 124(13): 2659-70.

Schmid, P., D. Cox, et al. (1991). "Differential expression of TGF beta 1, beta 2 and beta 3 genes during mouse embryogenesis." Development 111(1): 117-30.

Schotte, H., P. Willeke, et al. (2003). "The transforming growth factor-beta1 gene polymorphism (G915C) is not associated with systemic lupus erythematosus." Lupus 12(2): 86-92.

Sha, X., L. Yang, et al. (1991). "Identification and analysis of discrete functional domains in the pro region of pre-pro-transforming growth factor beta 1." J Cell Biol 114(4): 827-39.

Shai, R., F. P. Quismorio, Jr., et al. (1999). "Genome-wide screen for systemic lupus erythematosus susceptibility genes in multiplex families." Hum Mol Genet 8(4): 639-44.

Shull, M. M., I. Ormsby, et al. (1992). "Targeted disruption of the mouse transforming growth factor-beta 1 gene results in multifocal inflammatory disease." Nature 359(6397): 693-9.

Sporn, M. B., A. B. Roberts, et al. (1986). "Transforming growth factor-beta: biological function and chemical structure." Science 233(4763): 532-4.

Sugiura, Y., T. Niimi, et al. (2002). "Transforming growth factor beta1 gene polymorphism in rheumatoid arthritis." Ann Rheum Dis 61(9): 826-8.

Tan, E. M., A. S. Cohen, et al. (1982). "The 1982 revised criteria for the classification of systemic lupus erythematosus." Arthritis Rheum 25(11): 1271-7.

ten Dijke, P., P. Hansen, et al. (1988). "Identification of another member of the transforming growth factor type beta gene family." Proc Natl Acad Sci U S A 85(13): 4715-9.

Tsao, B. P. (2003). "The genetics of human systemic lupus erythematosus." Trends Immunol 24(11): 595-602.

Tsao, B. P., R. M. Cantor, et al. (1999). "PARP alleles within the linked chromosomal region are associated with systemic lupus erythematosus." J Clin Invest 103(8): 1135-40.

Tsao, B. P., R. M. Cantor, et al. (1997). "Evidence for linkage of a candidate chromosome 1 region to human systemic lupus erythematosus." J Clin Invest 99(4): 725-31.

Vyse, T. J. and B. L. Kotzin (1998). "Genetic susceptibility to systemic lupus erythematosus." Annu Rev Immunol 16: 261-92.

Wakefield, L. M., D. M. Smith, et al. (1988). "Latent transforming growth factor-beta from human platelets. A high molecular weight complex containing precursor sequences." J Biol Chem 263(16): 7646-54.

Wakeland, E., L. Morel, et al. (1997). "Speed congenics: a classic technique in the fast lane (relatively speaking)." Immunol Today 18(10): 472-7.

Wakeland, E. K., K. Liu, et al. (2001). "Delineating the genetic basis of systemic lupus erythematosus." Immunity 15(3): 397-408.

Wandstrat, A. and E. Wakeland (2001). "The genetics of complex autoimmune diseases: non-MHC susceptibility genes." Nat Immunol 2(9): 802-9.

Webb, N. R., L. Madisen, et al. (1988). "Structural and sequence analysis of TGF-beta 2 cDNA clones predicts two different precursor proteins produced by alternative mRNA splicing." DNA 7(7): 493-7.

Yamada, Y., H. Okuizumi, et al. (2000). "Association of transforming growth factor beta1 genotype with spinal osteophytosis in Japanese women." Arthritis Rheum 43(2): 452-60.

Yu, L., W. A. Border, et al. (2003). "TGF-beta isoforms in renal fibrogenesis." Kidney Int 64(3): 844-56.

Ziv, E., J. Cauley, et al. (2001). "Association between the T29-->C polymorphism in the transforming growth factor beta1 gene and breast cancer among elderly white women: The Study of Osteoporotic Fractures." Jama 285(22): 2859-63.

APPENDIXES

APPENDIX A

Reagent for agarose gel electrophoresis

1. 50X Tris-acetate buffer (TAE)

Tris base	424.0	g
Glacial acetic acid	57.1	g
0.5 m EDTA Ph 8.0	100	ml

Adjust volume to 1 liter with distilled water. The solution was mixed and sterilized by autoclaving at 121 °C for 15 min.

2. 10 mg/ml Ethidium bromide

Ethidium bromide	1.0	g
Distilled water	100	ml

Mix the solution and store in the dark at 4 °C

3. 1.5% Agarose gel

Agarose	0.3	g
1X TAE	20	ml

Dissolve by heating in microwave oven and occasional mix until no granules of agarose are visible.

4. 5X Loading buffer 100 ml

Tris HCL	0.6	g
EDTA	1.68	g
SDS	0.5	g
Bromphenol Blue	0.1	g
Sucrose	40	g

Adjust volume to 100 ml with distilled water. Mix the solution, aliquot into 1.5 microtube and store at 4 °C.

APPENDIX B

Reagent for denaturing polyacrylamide gel electrophoresis

1. 10X Tris-boric acid buffer (TBE) 1000 ml

Tris base	108.0	g
Boric acid	55	g
0.5 m EDTA Ph 8.0	40	ml

Adjust volume to 1 liter with distilled water. The solution was mixed and sterilized by autoclaving at 121 ºC for 15 min.

2. 6% denaturing polyacrylamide gel

Urea	25.2	g
40% polyacrylamide:Bis 19:1 (ml)	9	ml
10X TBE(ml)	6	ml
H_2O	26	ml
10% Ammoniumpersulphate (µl)	396	µl
TEMED (µl)	34.8	µl

Dissolve by vortex in magnetic sterer and occasional mix until no granules of ureas are visible.

3. 5'- end labeling primer (kinase reaction) Microsatellite Protocal Modified from DR. Apiwat Mutirangura

10X kinase buffer(1X)	1	µl
T4 kinase (Biolabs) 1U/1µl	1	µl
γ- P^{32}ATP (Dupont) 2U/µl	2	µl
primer 10 µmol	5	µl
H_2O	1	µl

4. 3X Loading buffer 100 ml

Formamide	100	ml
Bromphenol Blue	0.1	g
Xylene cyanol	0.1	g

Adjust volume to 100 ml with distilled water. Mix the solution, aliquot into 1.5 microtube and store at 4 °C.

APPENDIX C

Reagent for DNA extraction

1. Red Cell Lysis Buffer (RCLB)

NH_4Cl	1.875	g
Tris-HCl	0.25	g

Dissolve NH_4Cl and Tris-HCL in 500 ml of distilled water. Adjust pH to 7.2. The solution was mixed and sterilizes by autoclaving at 121 ^{o}C for 15 minutes. Keep refrigerated. Shelf life is approximately 6 months.

2. Nuclei Lysis Buffer (NLB)

1 M Tris (pH 8.0)	10	ml
5 M NaCl	0.5	ml
0.5 M EDTA (pH 8.0)	0.4	ml

Adjust volume to 100 ml with distilled water. Adjust pH to 7.2. Keep refrigerated. Shelf life is approximately 6 months.

3. 1 M Tris

Tris base	12.11	g
Distilled water	100	g

Adjust volume to 100 ml with distilled water. Adjust pH to 7.2. The solution was mixed and sterilizes by autoclaving at 121 °C for 15 minutes.

4. 5 M NaCl

NaCl	29.22	g
Distilled water	100	ml

Adjust volume to 100 ml with distilled water. The solution was mixed and sterilized by autoclaving at 121 °C for 15 minutes.

5. EDTA

EDTA	37.22	g
Distilled water	200	ml

Adjust volume to 200 ml with distilled water. The solution was mixed and sterilizes by autoclaving at 121 °C for 15 minutes. Keep refrigerated.

6. 5.3 M NaCl

NaCl	15.5	g
Distilled water	50	ml

Adjust volume to 50 ml with distilled water. The solution was mixed and sterilizes by autoclaving at 121 °C for 15 minutes. Keep refrigerated.

7. Proteinase K 10 mg/ml

Proteinase K	100	mg
Distilled water	10	ml

Mix the solution and store at -20 °C

8. 10% SDS

SDS	10	g
Distilled water	100	ml

Adjust volume to 100 ml with distilled water. The solution was mixed and sterilizes by autoclaving at 121 °C for 15 minutes.

APPENDIX D

SNP haplotype tagging from DNA pools of two individuals by Pools2 package

Pools2 package program estimates haplotype frequencies in pools of DNA from two individuals each. The DNA pooling is a technique to reduce genotyping effort while incurring only minor losses in accuracy of allele frequency estimates for single nucleotide polymorphism (SNP) markers. The following programs are distributed under the terms of the GNU General Public License. The software is available online at (http://linkage.rockefeller.edu/register/) (Hoh, Matsuda et al. 2003)

This is accomplished with the following steps:

- Based on pool phenotypes at a number of SNPs (input file), the *snp* program estimates frequencies of genotype vectors, where for a given individual a genotype vector is a set of genotypes at all SNPs. Based on these estimated frequencies, the program outputs estimated genotypes.
- For the next step of estimating haplotypes from genotypes, a number of programs may be used, for example, *EHplus* [2] or *SNPHAP* [3] or *HAPINF* [4]. We currently provide the latter program with this package.
- Based on pairs of haplotypes inferred for each individual, the *htSNP.py* program determines the smallest set of SNPs that can tag all haplotypes.

***Contents of* pools2 package**

The following files are contained in the "pools2.zip" file (may be downloaded from http://linkage.rockefeller.edu/register/):

add2data.txt	Original dataset for *add2* gene
GNUlicense.txt	GNU general license for use of POOLS2 program package
hapinf.exe	Executable for *hapinf* program
hapinf.for	Source code for *hapinf* program
hapinf.out	Sample output file, *hapinf* program (*add2* data)

hapinf.txt	Sample input file for *hapinf* program
htdata.out	Sample output, *htSNP* program
htdata.txt	Sample input for *htSNP* program (*add2* data)
htSNP.py	Code for *htSNP* program
Readme.txt	Mentions web site with detailed explanations
ReadmeHapinf.txt	Explanations to *hapinf* program
Readme-htSNP.txt	Explanations to *htSNP* program
seed.txt	Contains seed for random number generator
snp.exe	Executable for *snp* program
snp.pas	Souce code for *snp* program
snpdata.txt	Sample input file to *snp* program
snpeh.txt	Sample output from *snp* program
snpfinal.txt	Sample output from *snp* program
snpfreepas.pas	Part of source code of *snp* program
snpout.txt	Sample output from *snp* program

***Input files for* snp program**

1. Pool phenotypes

The user must prepare an input file called "snpdata.txt" with the following specifications:

Line 1

Number of pools used (must be the same for all data in this file). Optionally this number may be followed by arbitrary text.

Line 2

Gene ID (an integer number)

SNP ID (an integer number)

Phenotype code for pool 1 (integer)

Phenotype code for pool 2, etc.

Repeat line 2 for each new SNP and for as many genes as desired. SNPs in the same gene must have same gene ID. It is ok to use the same gene ID at different places in the file. For example, if gene ID 335 is followed by gene ID 638 and then again by gene ID 335, the program will look at this as three different genes.

The data matrix now consists of a number of rows (= SNPs) and columns (= pools).

This study, considered the wxample input file, snpdata.tex with the follows:

Pool phenotype codes are as follows, with X and Y designating two SNP alleles:

Pool phenotype	*Code*	***Explanation***
XXXX	0	All alleles are of the X type
YYYY	1	All alleles Y type
XXYY	2	Equal number of X and Y alleles
XXXY	3	3 X alleles, 1 Y allele
XYYY	4	1 X allele, 3 Y alleles
XY?	5	X and Y alleles both present but in unknown numbers
XYZ	6	3 alleles, will be treated as N
N	-1	No data (unknown)

2. Random seed

For the random number generator, the user must prepare a file called "seed.txt" containing a negative integer number, i.e., the random number seed. This file will be updated with each successive run of the *snp* program.

Running the snp program

Open a Windows command window ("DOS box") and change directories until you are in the directory containing the *snp* program and the input files. Then simply type snp. The program will then impute genotypes as described in [1]. A given SNP will be considered missing entirely (and will not be processed further) if the number of missing pool phenotypes is larger than the highest frequency of any of the known phenotypes at that SNP.

The program then displays program constants, which are currently set as follows:

Max. number of pools = 100

Max. number of SNPs per gene = 200

Max. number of genotype vectors per gene = 30,000

Max. number of interations = 99. If successive iterations differ by $ss < 10^{-11}$ then iterations will stop early, where

ss = sum of squared deviations between old and new genotype vector frequencies.

As the program is running it announces each gene and the number of valid (non-missing) SNPs at that gene, also the total number of genotype vectors compatible with the pool phenotypes. If that number exceeds the maximum set by the corresponding program constant, the given SNP will be skipped (it would take too long to run anyway). Then the program reports progress at each iteration and displays the sum of squared deviations (ss) of genotype vector frequencies between the current and the previous iteration.

Output files

The *snp* program writes three output files, "snpout.txt", "snpfinal.txt", and "snpeh.txt".

1. snpout.txt

This file will contain "cleaned" data with imputed pool phenotypes and an indication of which SNPs are missing.

2. snpfinal.txt

This file contains, for each gene, a list of estimated individual genotypes. The following items (columns) will be provided for each individual (output line):

Gene number

Genotype code for SNP 1

Genotype code for SNP 2, etc.,

where codes 1, 2, and 3 represent genotypes 1/1, 1/2, and 2/2, respectively. After some modification as indicated, this file is suitable for input to the *hapinf* program (see below).

3. snpeh.txt

This file is analogous to the previous file but is formatted for input to the *EHplus* program [2]. The following items (columns) are provided for each individual (output line):

ID for individual. Example: 3.119 (meaning individual 3 at gene 119)

A fixed code of 0 meaning "control individual"

Allele 1 at SNP 1

Allele 2 at SNP 1

Allele 1 at SNP 2

Allele 2 at SNP 2, etc.

Running the hapinf program

The program has been developed by Dr. Andrew Clark [4]. He graciously allowed me to include his program with my programs in this package.

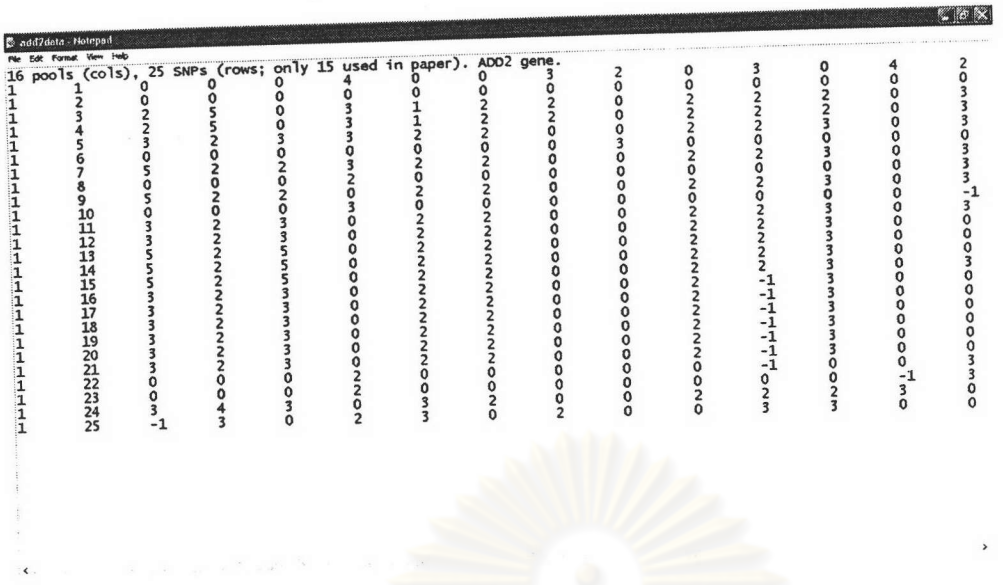
The program requires an input file called "hapinf.txt" (for an example see *Analysis of sample data*, below). Then run the *hapinf* program by typing hapinf in the command window. The program will produce an output file called "hapinf.out".

Running the *htSNP* program

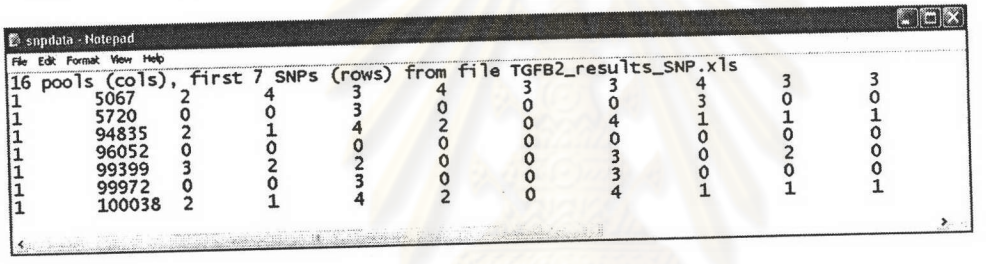
This program is written in the Python language. To run it you must have the Python system installed, for example, on your PC under Windows. See http://www.python.org/. For a brief list of commands, type htSNP.py.

Analysis of sample data (*add2* gene)

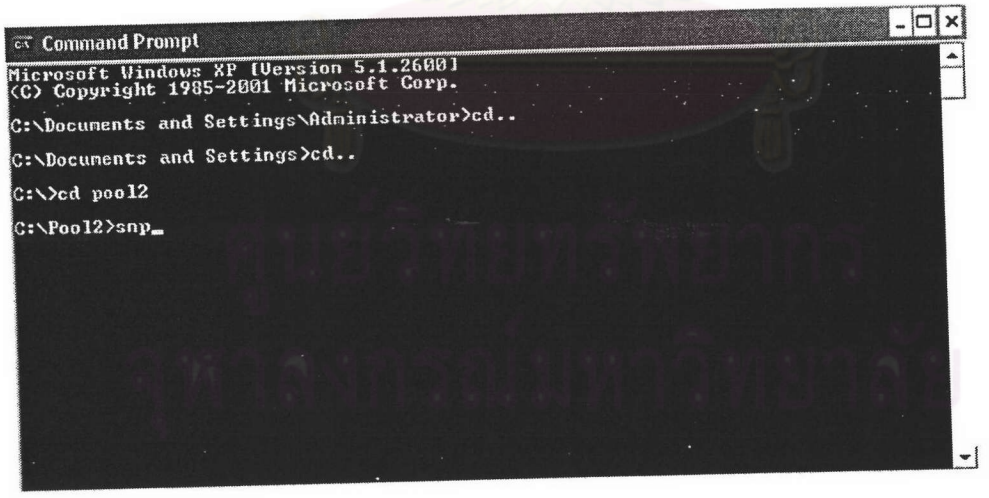
This sample dataset consists of 16 pools genotyped at 25 SNPs ("add2data.txt" file) was shown below.



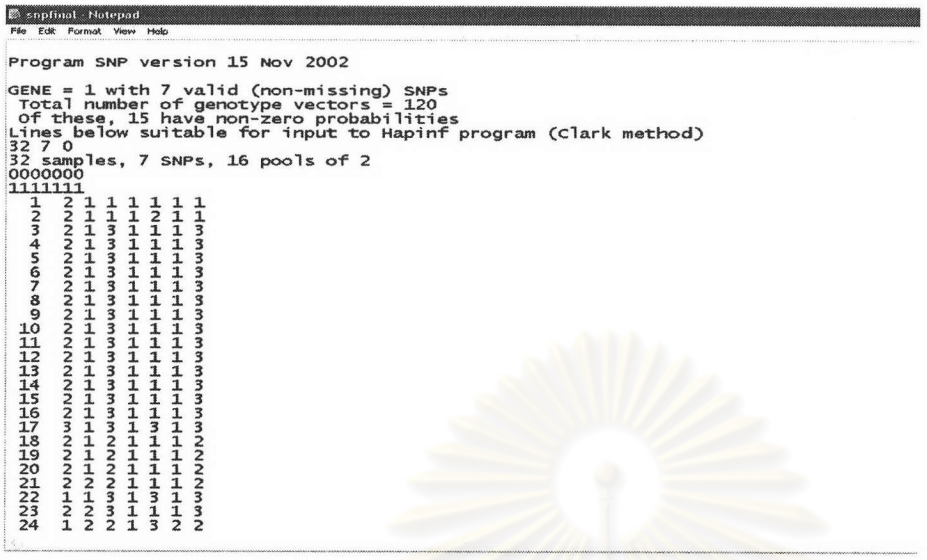
In [1] we use only the first 15 SNPs, resulting in the "snpdata.txt" file (input to *snp* program).



Run the *snp* program by typing snp was shown below.



The resulting output file, "snpfinal.txt" was shown below.

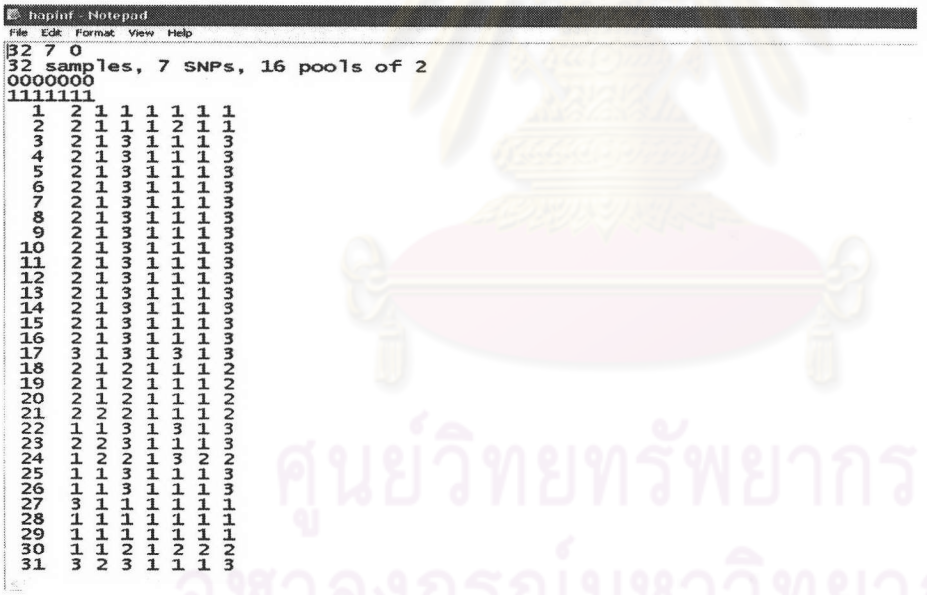


snpfinal - Notepad

File Edit Format View Help

```
Program SNP version 15 Nov 2002

GENE = 1 with 7 valid (non-missing) SNPs
 Total number of genotype vectors = 120
 of these, 15 have non-zero probabilities
Lines below suitable for input to Hapinf program (Clark method)
32 7 0
32 samples, 7 SNPs, 16 pools of 2
0000000
1111111
  1   2 1 1 1 1 1 1
  2   2 1 1 1 2 1 1
  3   2 1 3 1 1 1 3
  4   2 1 3 1 1 1 3
  5   2 1 3 1 1 1 3
  6   2 1 3 1 1 1 3
  7   2 1 3 1 1 1 3
  8   2 1 3 1 1 1 3
  9   2 1 3 1 1 1 3
 10   2 1 3 1 1 1 3
 11   2 1 3 1 1 1 3
 12   2 1 3 1 1 1 3
 13   2 1 3 1 1 1 3
 14   2 1 3 1 1 1 3
 15   2 1 3 1 1 1 3
 16   2 1 3 1 1 1 3
 17   3 1 3 1 3 1 3
 18   2 1 2 1 1 1 2
 19   2 1 2 1 1 1 2
 20   2 1 2 1 1 1 2
 21   2 2 2 1 1 1 2
 22   1 1 3 1 3 1 3
 23   2 2 3 1 1 1 3
 24   1 2 2 1 3 2 2
```

Then modified (first 7 lines deleted) and saved as "hapinf.txt", which serves as input to the *hapinf* program.

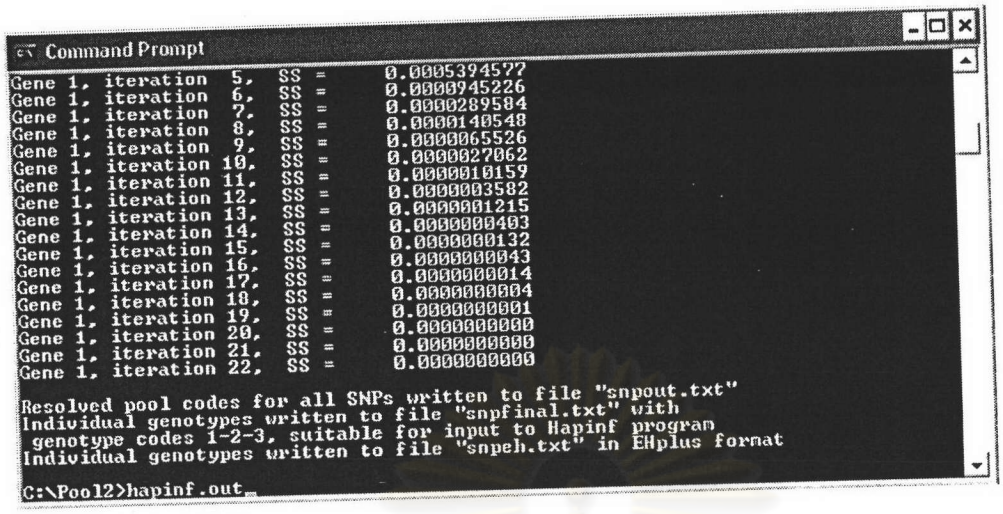


hapinf - Notepad

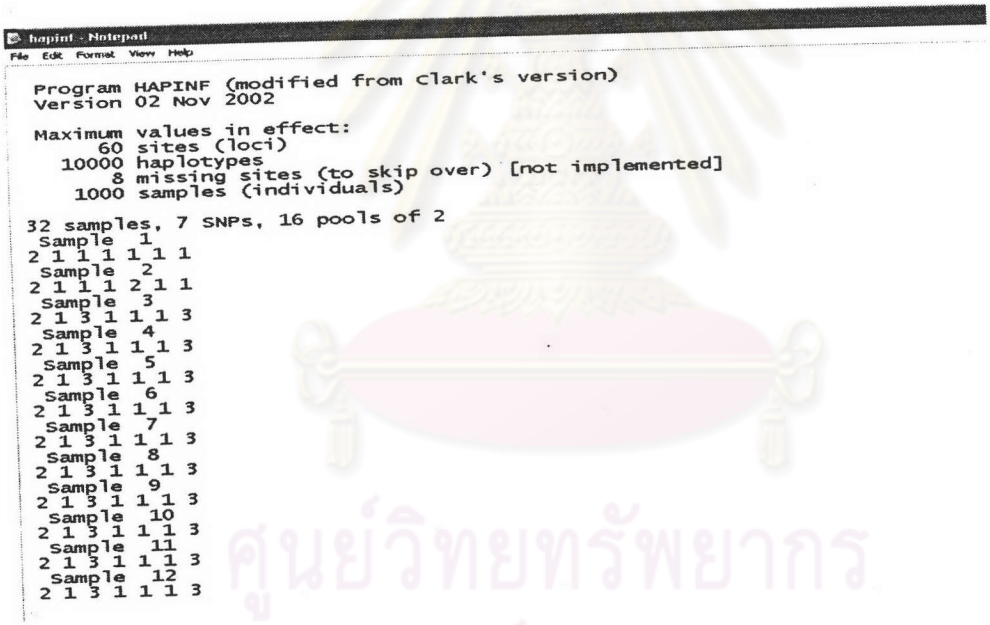
File Edit Format View Help

```
32 7 0
32 samples, 7 SNPs, 16 pools of 2
0000000
1111111
  1   2 1 1 1 1 1 1
  2   2 1 1 1 2 1 1
  3   2 1 3 1 1 1 3
  4   2 1 3 1 1 1 3
  5   2 1 3 1 1 1 3
  6   2 1 3 1 1 1 3
  7   2 1 3 1 1 1 3
  8   2 1 3 1 1 1 3
  9   2 1 3 1 1 1 3
 10   2 1 3 1 1 1 3
 11   2 1 3 1 1 1 3
 12   2 1 3 1 1 1 3
 13   2 1 3 1 1 1 3
 14   2 1 3 1 1 1 3
 15   2 1 3 1 1 1 3
 16   2 1 3 1 1 1 3
 17   3 1 3 1 3 1 3
 18   2 1 2 1 1 1 2
 19   2 1 2 1 1 1 2
 20   2 1 2 1 1 1 2
 21   2 2 2 1 1 1 2
 22   1 1 3 1 3 1 3
 23   2 2 3 1 1 1 3
 24   1 2 2 1 3 2 2
 25   1 1 3 1 1 1 3
 26   1 1 3 1 1 1 3
 27   3 1 1 1 1 1 1
 28   1 1 1 1 1 1 1
 29   1 1 1 1 1 1 1
 30   1 1 2 1 2 2 2
 31   3 2 3 1 1 1 3
```

Run it by typing hapinf.out.



The data was shown below.



hapinf - Notepad

File Edit Format View Help

```
 Sample  13
2 1 3 1 1 1 3
 Sample  14
2 1 3 1 1 1 3
 Sample  15
2 1 3 1 1 1 3
 Sample  16
2 1 3 1 1 1 3
 Sample  17
3 1 3 1 3 1 3
 Sample  18
2 1 2 1 1 1 2
 Sample  19
2 1 2 1 1 1 2
 Sample  20
2 1 2 1 1 1 2
 Sample  21
2 2 2 1 1 1 2
 Sample  22
1 1 3 1 3 1 3
 Sample  23
2 2 3 1 1 1 3
 Sample  24
1 2 2 1 3 2 2
 Sample  25
1 1 3 1 1 1 3
 Sample  26
1 1 3 1 1 1 3
 Sample  27
3 1 1 1 1 1 1
 Sample  28
1 1 1 1 1 1 1
 Sample  29
1 1 1 1 1 1 1
 Sample  30
```

hapinf - Notepad

File Edit Format View Help

```
 Sample  30
1 1 2 1 2 2 2
 Sample  31
3 2 3 1 1 1 3
 Sample  32
3 1 2 2 2 1 1

Homozygotes:
 sample  17
1 0 1 0 1 0 1  1
1 0 1 0 1 0 1  1
 sample  22
0 0 1 0 1 0 1  2
0 0 1 0 1 0 1  2
 sample  25
0 0 1 0 0 0 1  3
0 0 1 0 0 0 1  3
 sample  27
1 0 0 0 0 0 0  4
1 0 0 0 0 0 0  4
 sample  28
0 0 0 0 0 0 0  5
0 0 0 0 0 0 0  5
The number of homozygotes   =  7
The number of distinct homo =  5

Single-site heterozygotes:
 sample  1
0 0 0 0 0 0 0  5
1 0 0 0 0 0 0  4
 sample  3
0 0 1 0 0 0 1  3
1 0 1 0 0 0 1  6
 sample  4
0 0 1 0 0 0 1  3
```

hapinf - Notepad

File Edit Format View Help

```
1 0 1 0 0 0 1  6
 sample  5
1 0 1 0 0 0 1  6
0 0 1 0 0 0 1  3
 sample  6
1 0 1 0 0 0 1  6
0 0 1 0 0 0 1  3
 sample  7
1 0 1 0 0 0 1  6
0 0 1 0 0 0 1  3
 sample  8
0 0 1 0 0 0 1  3
1 0 1 0 0 0 1  6
 sample  9
0 0 1 0 0 0 1  3
1 0 1 0 0 0 1  6
 sample  10
0 0 1 0 0 0 1  3
1 0 1 0 0 0 1  6
 sample  11
1 0 1 0 0 0 1  6
0 0 1 0 0 0 1  3
 sample  12
1 0 1 0 0 0 1  6
0 0 1 0 0 0 1  3
 sample  13
1 0 1 0 0 0 1  6
0 0 1 0 0 0 1  3
 sample  14
1 0 1 0 0 0 1  6
0 0 1 0 0 0 1  3
 sample  15
1 0 1 0 0 0 1  6
0 0 1 0 0 0 1  3
 sample  16
```

hapinf - Notepad

File Edit Format View Help

```
 sample  16
0 0 1 0 0 0 1  3
1 0 1 0 0 0 1  6
 sample  31
1 0 1 0 0 0 1  6
1 1 1 0 0 0 1  7

The number of single hets       =  16
Count of unambiguous haplotypes =  7

List of unambiguous haplotypes:
 hap  1
1 0 1 0 1 0 1
 hap  2
0 0 1 0 1 0 1
 hap  3
0 0 1 0 0 0 1
 hap  4
1 0 0 0 0 0 0
 hap  5
0 0 0 0 0 0 0
 hap  6
1 0 1 0 0 0 1
 hap  7
1 1 1 0 0 0 1

 sample  24
0 0 1 0 1 0 1  2
0 1 0 0 1 1 0  8
 sample  30
0 0 1 0 1 0 1  2
0 0 0 0 0 1 0  9
 sample  18
0 0 1 0 0 0 1  3
1 0 0 0 0 0 0  4
```

hapinf - Notepad

File Edit Format View Help

```
 sample  19
0 0 1 0 0 0 1  3
1 0 0 0 0 0 0  4
 sample  20
0 0 1 0 0 0 1  3
1 0 0 0 0 0 0  4
 sample  21
0 0 1 0 0 0 1  3
1 1 0 0 0 0 0  10
 sample  23
0 0 1 0 0 0 1  3
1 1 1 0 0 0 1  7
 sample  2
1 0 0 0 0 0 0  4
0 0 0 0 1 0 0  11
 sample  32
1 0 0 0 0 0 0  4
1 0 1 1 1 0 0  12
All genotypes resolved

#mega
32 samples, 7 SNPs, 16 pools of 2

 hap# 1
1 0 1 0 1 0 1
 hap# 2
0 0 1 0 1 0 1
 hap# 3
0 0 1 0 0 0 1
 hap# 4
1 0 0 0 0 0 0
 hap# 5
0 0 0 0 0 0 0
 hap# 6
1 0 1 0 0 0 1
```

haplnf - Notepad

File Edit Format View Help

```
 hap# 7
1 1 1 0 0 0 1
 hap# 8
0 1 0 0 1 1 0
 hap# 9
0 0 0 0 0 1 0
 hap# 10
1 1 0 0 0 0 0
 hap# 11
0 0 0 0 1 0 0
 hap# 12
1 0 1 1 1 0 0

Print all haplotypes in original order:
(sample with ambiguous haplotypes listed more than once)

SampleID HapCode Haplotype

       1   5 0 0 0 0 0 0 0
       1   4 1 0 0 0 0 0 0

       2   4 1 0 0 0 0 0 0
       2  11 0 0 0 0 1 0 0

       3   3 0 0 1 0 0 0 1
       3   6 1 0 1 0 0 0 1

       4   3 0 0 1 0 0 0 1
       4   6 1 0 1 0 0 0 1

       5   6 1 0 1 0 0 0 1
       5   3 0 0 1 0 0 0 1

       6   6 1 0 1 0 0 0 1
       6   3 0 0 1 0 0 0 1
```

hapinf - Notepad

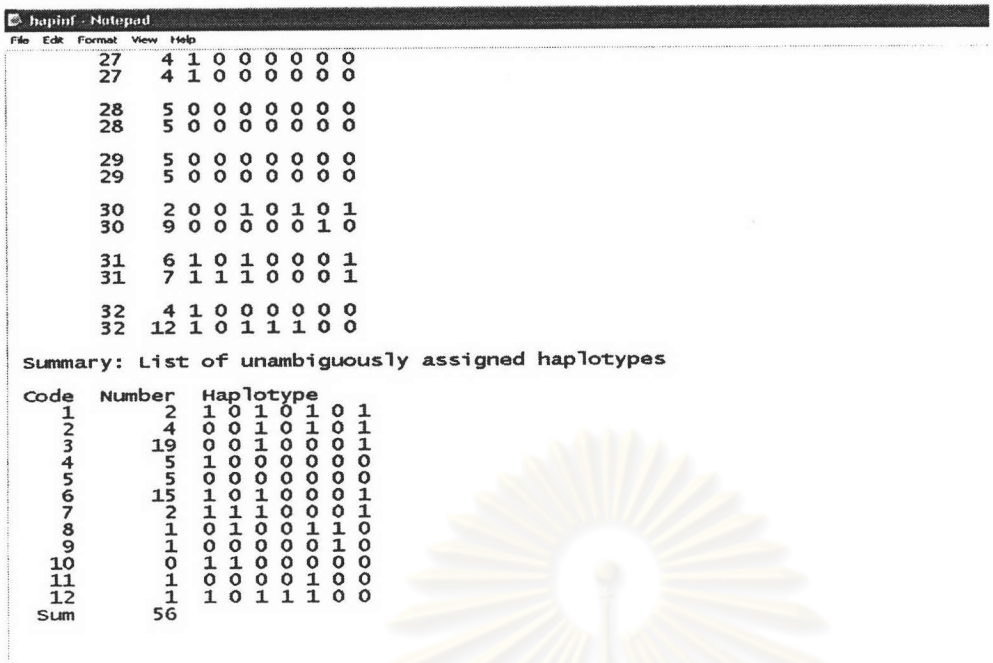
File Edit Format View Help

```
 7    6 1 0 1 0 0 0 1
 7    3 0 0 1 0 0 0 1

 8    3 0 0 1 0 0 0 1
 8    6 1 0 1 0 0 0 1

 9    3 0 0 1 0 0 0 1
 9    6 1 0 1 0 0 0 1

10    3 0 0 1 0 0 0 1
10    6 1 0 1 0 0 0 1

11    6 1 0 1 0 0 0 1
11    3 0 0 1 0 0 0 1

12    6 1 0 1 0 0 0 1
12    3 0 0 1 0 0 0 1

13    6 1 0 1 0 0 0 1
13    3 0 0 1 0 0 0 1

14    6 1 0 1 0 0 0 1
14    3 0 0 1 0 0 0 1

15    6 1 0 1 0 0 0 1
15    3 0 0 1 0 0 0 1

16    3 0 0 1 0 0 0 1
16    6 1 0 1 0 0 0 1

17    1 1 0 1 0 1 0 1
17    1 1 0 1 0 1 0 1

18    3 0 0 1 0 0 0 1
18    4 1 0 0 0 0 0 0
```

hapinf - Notepad

File Edit Format View Help

```
 7    6 1 0 1 0 0 0 1
 7    3 0 0 1 0 0 0 1

 8    3 0 0 1 0 0 0 1
 8    6 1 0 1 0 0 0 1

 9    3 0 0 1 0 0 0 1
 9    6 1 0 1 0 0 0 1

10    3 0 0 1 0 0 0 1
10    6 1 0 1 0 0 0 1

11    6 1 0 1 0 0 0 1
11    3 0 0 1 0 0 0 1

12    6 1 0 1 0 0 0 1
12    3 0 0 1 0 0 0 1

13    6 1 0 1 0 0 0 1
13    3 0 0 1 0 0 0 1

14    6 1 0 1 0 0 0 1
14    3 0 0 1 0 0 0 1

15    6 1 0 1 0 0 0 1
15    3 0 0 1 0 0 0 1

16    3 0 0 1 0 0 0 1
16    6 1 0 1 0 0 0 1

17    1 1 0 1 0 1 0 1
17    1 1 0 1 0 1 0 1

18    3 0 0 1 0 0 0 1
18    4 1 0 0 0 0 0 0
```

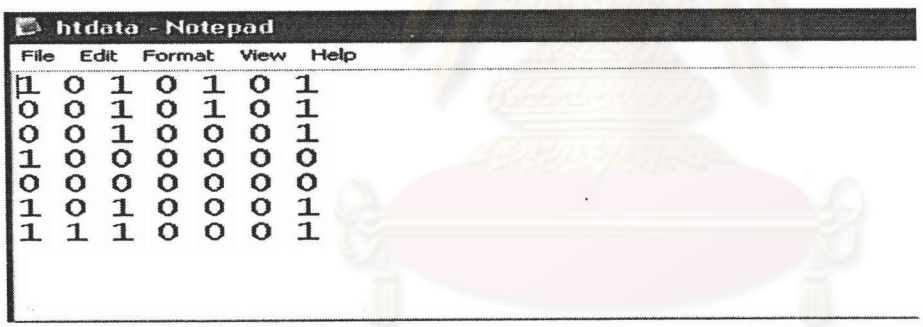


hapinf - Notepad

File Edit Format View Help

```
       27    4 1 0 0 0 0 0 0
       27    4 1 0 0 0 0 0 0

       28    5 0 0 0 0 0 0 0
       28    5 0 0 0 0 0 0 0

       29    5 0 0 0 0 0 0 0
       29    5 0 0 0 0 0 0 0

       30    2 0 0 1 0 1 0 1
       30    9 0 0 0 0 0 1 0

       31    6 1 0 1 0 0 0 1
       31    7 1 1 1 0 0 0 1

       32    4 1 0 0 0 0 0 0
       32   12 1 0 1 1 1 0 0

Summary: List of unambiguously assigned haplotypes

Code   Number   Haplotype
   1        2   1 0 1 0 1 0 1
   2        4   0 0 1 0 1 0 1
   3       19   0 0 1 0 0 0 1
   4        5   1 0 0 0 0 0 0
   5        5   0 0 0 0 0 0 0
   6       15   1 0 1 0 0 0 1
   7        2   1 1 1 0 0 0 1
   8        1   0 1 0 0 1 1 0
   9        1   0 0 0 0 0 1 0
  10        0   1 1 0 0 0 0 0
  11        1   0 0 0 0 1 0 0
  12        1   1 0 1 1 1 0 0
 Sum       56
```

To determine haplotype-tagging SNPs, save that section of text, for example, as "htdata.txt", but delete the first two columns (*Code* and *Number*). The data was shown below.

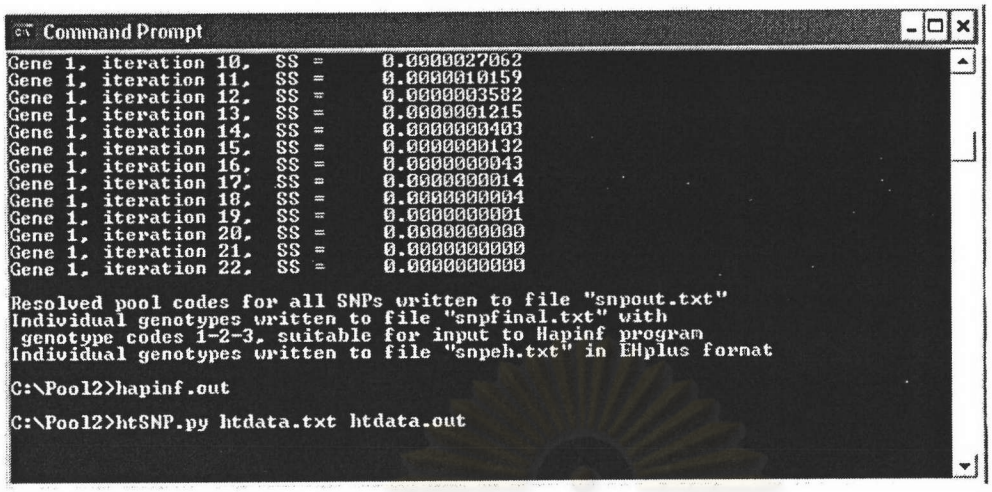


htdata - Notepad

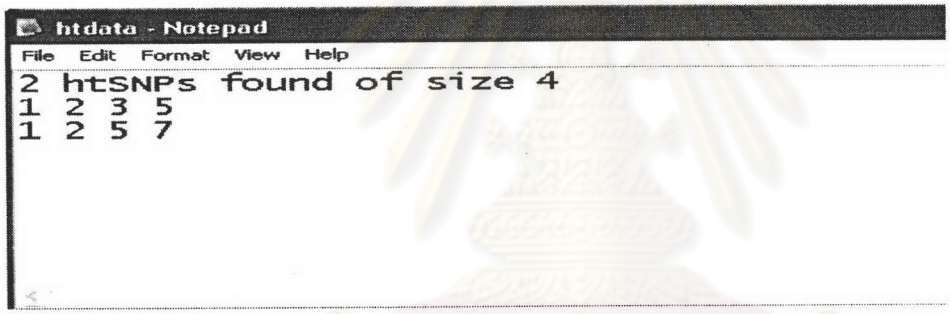
File Edit Format View Help

```
1 0 1 0 1 0 1
0 0 1 0 1 0 1
0 0 1 0 0 0 1
1 0 0 0 0 0 0
0 0 0 0 0 0 0
1 0 1 0 0 0 1
1 1 1 0 0 0 1
```

Then type the command htSNP.py htdata.txt htdata.out. The data was shown below.



The "htdata.out" file then shows the smallest number of haplotype tagging SNPs. The data was shown below.



APPENDIX E

Haplotype analysis

PHASE program is produced by the Mathematical Genetics Group, University of Oxford, Oxford, UK. The software is available online at http://www.stats.ox.ac.uk/mathgen/.

Input file format

The input file is supplied by the user to specify how many individuals there are to be analysed, how many loci/sites each individual has been typed at, what sort of loci/ sites these are (SNP or microsatellite), and the genotypes for each individual. The default format input file, as illustrated in the accompanying file test.inp. The default structure for the input file can be represented as follows:

Number of Individuals
Number of Loci
P Position(1) Position(2) Position(Number of Loci)
LocusType(1) LocusType(2) ... LocusType(Number of Loci)
ID(1)
Genotype(1)
ID(2)
Genotype(2)
:
:
ID(Number of Individuals)
Genotype(Number of Individuals)

Where the quantities above are as follows:

- Number of Individuals An integer specifying the number of individuals who have been genotyped.

- Number of Loci An integer specifying the number of loci or sites at which each individual has been typed.

- P The character 'P' (upper case, without quotation marks).

- Position(i) A number indicating the position of locus i, relative to some arbitray reference point (typically in units of base pairs, but any units can be used: if you use a unit other than base pairs, see the documentation on the -R option). The loci must be in their physical order along the chromosome (ie these Positions must be increasing).

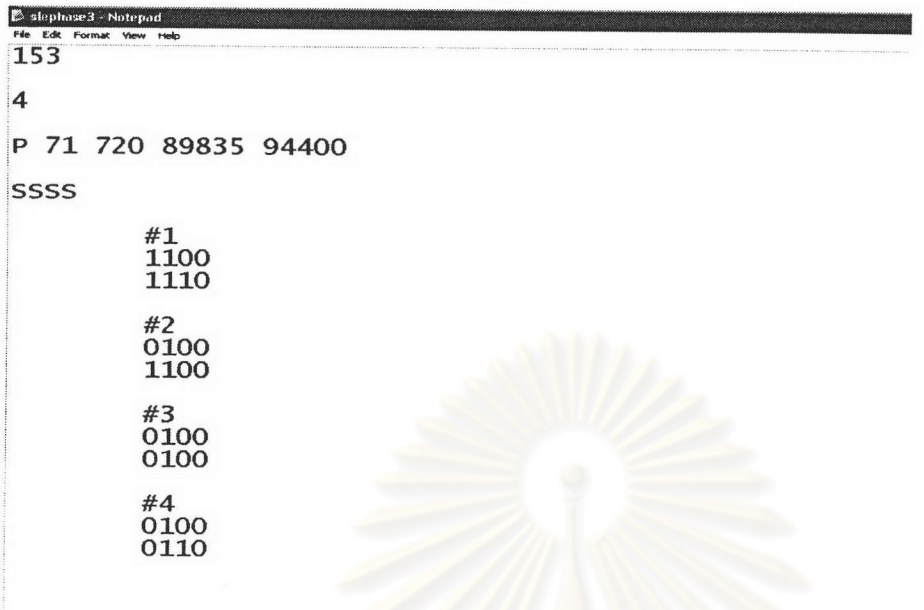
- LocusType(i) A letter indicating the type of locus i. The options are (a) S for a biallelic (SNP) locus, or biallelic site in sequence data. (b) M for microsatellite, or other multi-allelic locus (eg triallelic SNP, or HLA allele). The default assumption is that this denotes a microsatellite locus with stepwise mutation mechanism.

- ID(i) A string, giving a label for individual i.

Genotype(i) The genotypes for the ith individual. This is given on two consecutive rows. At each locus, one allele is entered on the first row, and one on the second row. It does not matter which allele is entered on each row. For biallelic loci, any two characters (e.g. A/C, G/T, 0/1) can be used to represent the two alleles, and they do not need to be separated by a space. Missing alleles at SNP loci should be entered as ?. For multiallelic loci a positive integer must be used for each allele (representing the number of repeats at microsatellite loci), and data for each locus should be separated by aspace. Missing alleles at multiallelic loci should be represented by -1.

This study, considered the example input file, slephase3.inp, which is as follows:

The example of input file (slephase3.inp) was shown below.



Running of PHASE program was shown below

Output file

When run, the program initially outputs the data it has read from input file. The program produces a number of output files. The first, which has the user-specified name, and a similar format to previous versions of PHASE, contains a summary of the individual haplotype estimates for each individual.

The example of output file (slephase3.inp) was shown below.

slephase3 - Notepad

File Edit Format View Help

index	haplotype	E(freq)	S.E
1	1100	0.147738	0.008863
2	1101	0.016235	0.004422
3	1110	0.140009	0.008812
4	1111	0.004614	0.002143
5	1000	0.032839	0.004453
6	1001	0.001134	0.001556
7	1010	0.010372	0.003803
8	0100	0.336445	0.009316
9	0101	0.054241	0.006363
10	0110	0.246288	0.009570
11	0111	0.005410	0.005433
12	0000	0.002805	0.002371
13	0001	0.000066	0.000458
14	0010	0.001804	0.001943

APPENDIX F

Linkage disequilibrium analysis

The values of linkage disequilibrium (LD) were determined using LDplotter. The LDPlotter tool allowed conversation of a Lickerson Lab Prettybase format file into aplot showing pairwise LD of various types (r^2, r, D' and absolute D).

Example

This example will walk through the generation of a plot like the one below:

I. Genename

The gene name can automatically be included in the plot title, and is also used to label the optional SNP Map which is drawn along the diagonal of the LD Plot. You can change the gene name in the gray box below to anything you like:

II. Prettybase

Input is a standard prettybase file. This study, consider the example input file, total 4 positions (71_72insACAA , + 720(T/G) , + 89835(A/G) and 94400_94401 insA)

III. LD Type

The software is available online at (http://innateimmunity.net.gov/SNP/.)

Indicates which measure of LD you would like to plot:

☐ $r^2 = D^2/P_A * P_B * (1 - P_A) * (1 - P_B)$

☐ $r = sqrt(r^2)$

☐ $D' = D/D_{max}$

☐ $abs(D) = | P_{AB} - (P_A * P_B) |$

IV. Configure Populations

This text area allows you to configure how the LD Plotter will split the samples in the prettybase file into different populations. The field should have one line for each population represented in your dataset. Each line should be a population identifier, followed by a colon, followed by a description of the population. Leaving this textarea empty will indicate that you do not wish to partition your sample set, but instead, would like to consider all of the samples to be part of a single population.

V. Plot Title

The plot title has two modes of operation, one for simple use, and the other for more advanced purposes for users who are familiar with Python format strings, and would to specify exactly what the plot title should look like.

VI. Miscellaneous

- **Miniraf:** You can set an arbitrary threshold for minimum allele frequence. The default value of 0.0 will not exclude SNPs based on rare frequency. If you specify a higher value, SNPs with frequency for the rare allele below this threshold in any for the rare population will NOT appear in the plot.
- **Color Scheme:** Several color schemes can be used to indicate the extend of LD between two loci.

The output file: The output file (total 4 positions.txt) was shown below.

output tgfb2 - Notepad

File Edit Format View Help

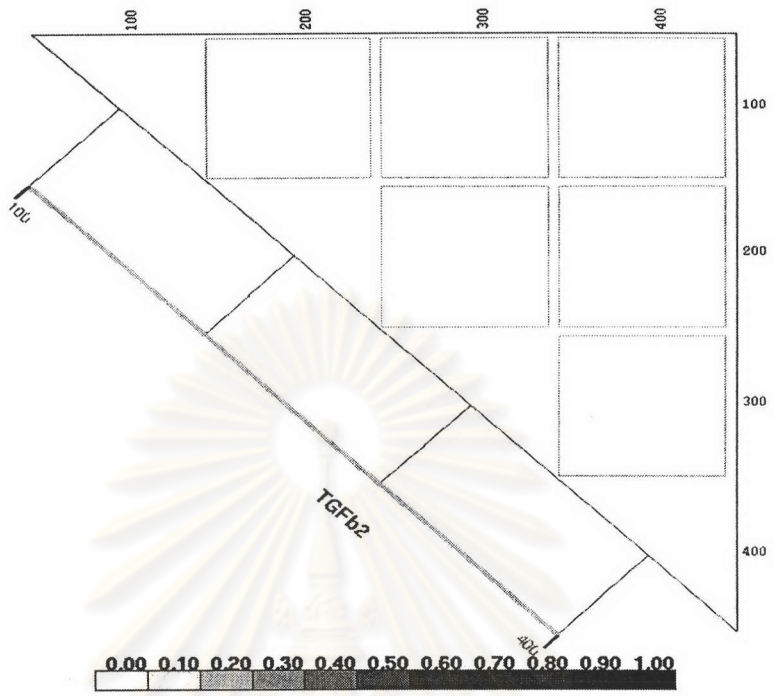
```
Pairwise LD values for sample = .*
M1   M2 all al2     N df(0) df(1)   ChiSq  pvalue delta2     r2
2  1  2  2  153   151   150   0.0   0.0   0.0649   0.0649
3  1  2  2  153   151   150   0.0   0.0   0.0051   0.0051
3  2  2  2  153   151   150   0.0   0.0   0.0155   0.0155
4  1  2  2  153   151   150   0.0   0.0   0.0094   0.0094
4  2  2  2  153   151   150   0.0   0.0   0.0040   0.0040
4  3  2  2  153   151   150   0.0   0.0   0.0290   0.0290

Pairwise LD values for sample = .*
M1   M2 all al2     N df(0) df(1)   ChiSq  pvalue delta2     r
2  1  2  2  153   151   150   0.0   0.0   0.0649   0.2548
3  1  2  2  153   151   150   0.0   0.0   0.0051   0.0716
3  2  2  2  153   151   150   0.0   0.0   0.0155   0.1245
4  1  2  2  153   151   150   0.0   0.0   0.0094   0.0968
4  2  2  2  153   151   150   0.0   0.0   0.0040   0.0632
4  3  2  2  153   151   150   0.0   0.0   0.0290   0.1702

Pairwise LD values for sample = .*
M1   M2 all al2     N df(0) df(1)   ChiSq  pvalue delta2     D'
2  1  2  2  153   151   150   0.0   0.0   0.0649   0.8467
3  1  2  2  153   151   150   0.0   0.0   0.0051   0.0788
3  2  2  2  153   151   150   0.0   0.0   0.0155   0.6597
4  1  2  2  153   151   150   0.0   0.0   0.0094   0.4610
4  2  2  2  153   151   150   0.0   0.0   0.0040   1.0000
4  3  2  2  153   151   150   0.0   0.0   0.0290   0.7360

Pairwise LD values for sample = .*
M1   M2 all al2     N df(0) df(1)   ChiSq  pvalue delta2     |D|
2  1  2  2  153   151   150   0.0   0.0   0.0649   0.0264
3  1  2  2  153   151   150   0.0   0.0   0.0051   0.0169
3  2  2  2  153   151   150   0.0   0.0   0.0155   0.0132
4  1  2  2  153   151   150   0.0   0.0   0.0094   0.0120
4  2  2  2  153   151   150   0.0   0.0   0.0040   0.0035
4  3  2  2  153   151   150   0.0   0.0   0.0290   0.0216
```

Linkage disequilibrium coefficients ($|D'|$ and r^2) among TGFß2 SNP

$\|D'\|$ / r^2	+ 71_72 ins	+ 720	+ 89835	+ 94400_94401ins
+ 71_72 ins	-	0.8467	0.0788	0.4610
+ 720	0.0649	-	0.6597	1.0000
+ 89835	0.0051	0.0155	-	0.7360
+ 94400_94401ins	0.0094	0.0042	0.0290	-

In this study, linkage disequilibrium coefficients ($|D'|$ and r^2) among TGFß2 SNP at position + 71_72 ins(ACAA), + 720 (T/G), + 89835 (A/G) and + 94400_94401ins(A)

The output figure r^2: The output figure r^2 (total 4 position.txt) was shown below.



APPENDIX G

Exact Hardy-Weinberg Equilibrium Test

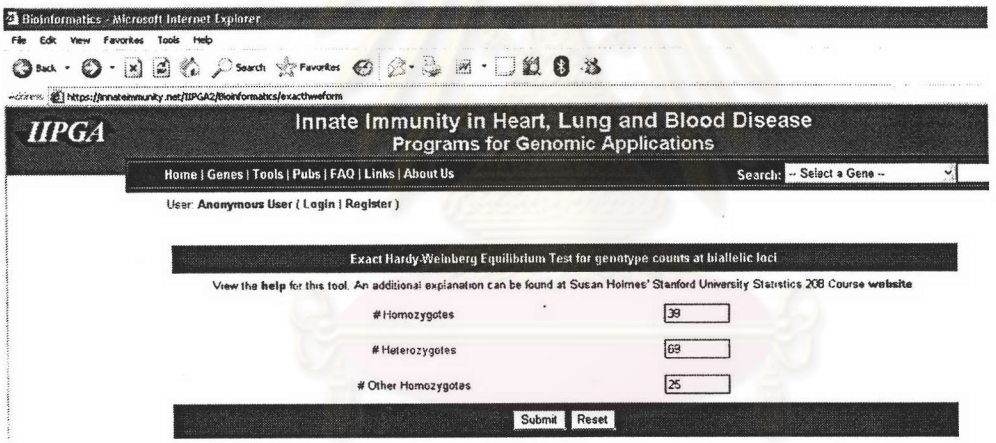
Exact Hardy-Weinberg Equilibrium Test for genotype counts, biallelic locus. This program will estimate the statistical significance of a test that the genotype counts are in Hardy-Weinberg equilibrium.

The software is available online at (http://innateimmunity.net.gov/SNP/.)

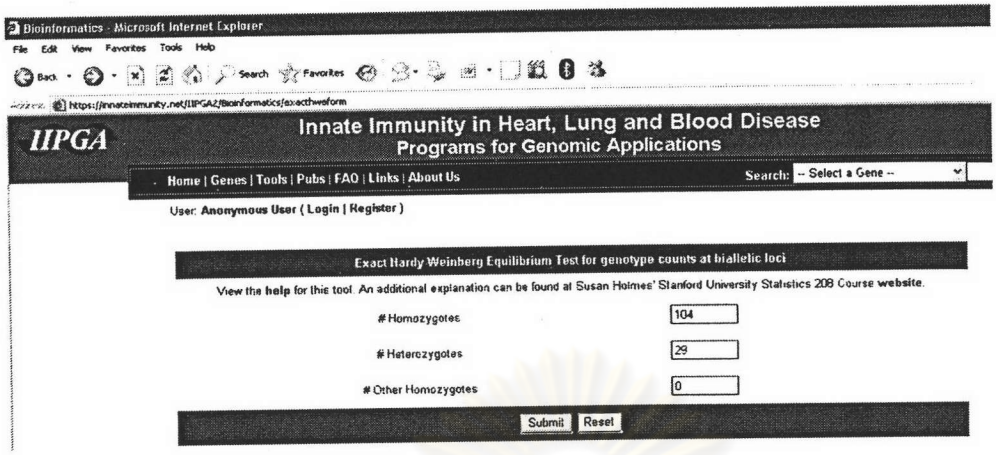
I Counts

The input to this tool is simply three integer counts representing the number of heterozygotes, common homozygotes, and rare homozygotes at a given locus in your data set.

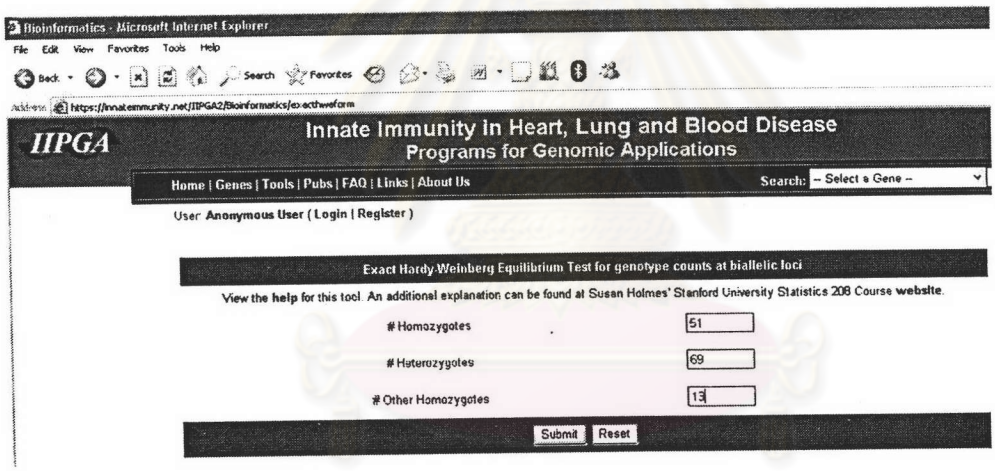
This study, the genotype counts of TGFß2 (+ 71_72insACAA) was shown



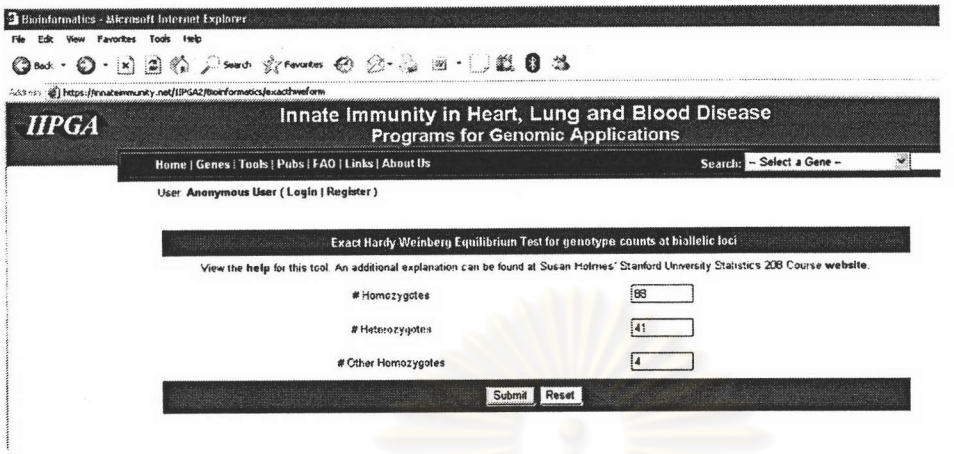
The genotype counts of TGFß2 (+ 720T/G) was shown



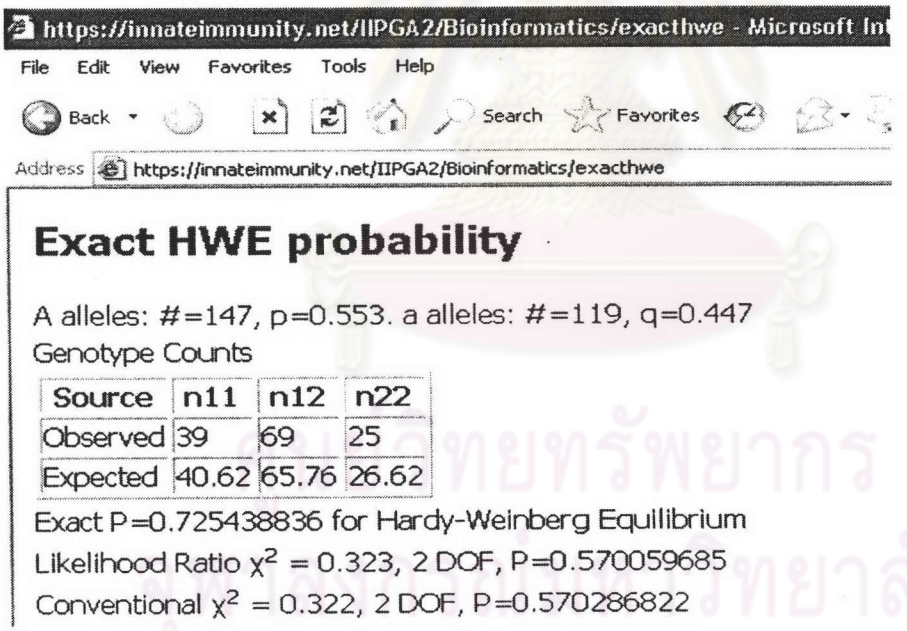
The genotype counts of TGFß2 (+ 89835A/G) was shown



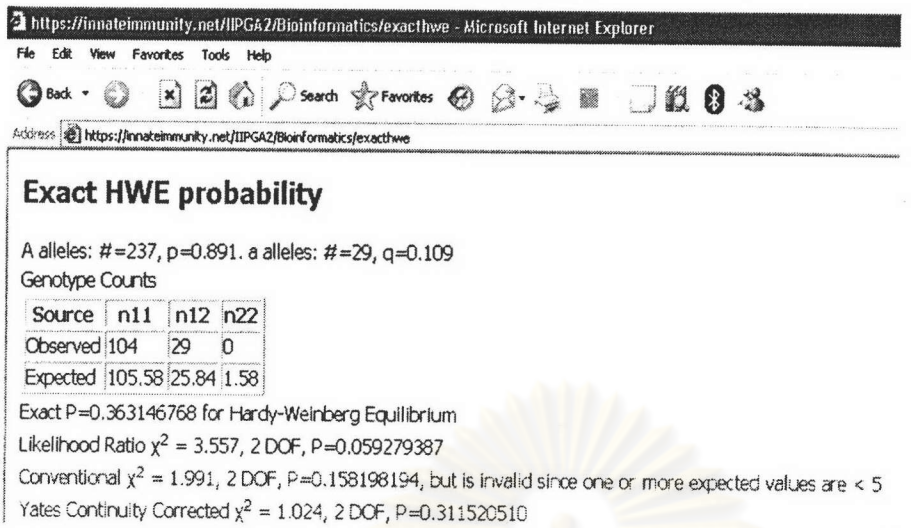
The genotype counts of TGFß2 (+ 89835A/G) was shown



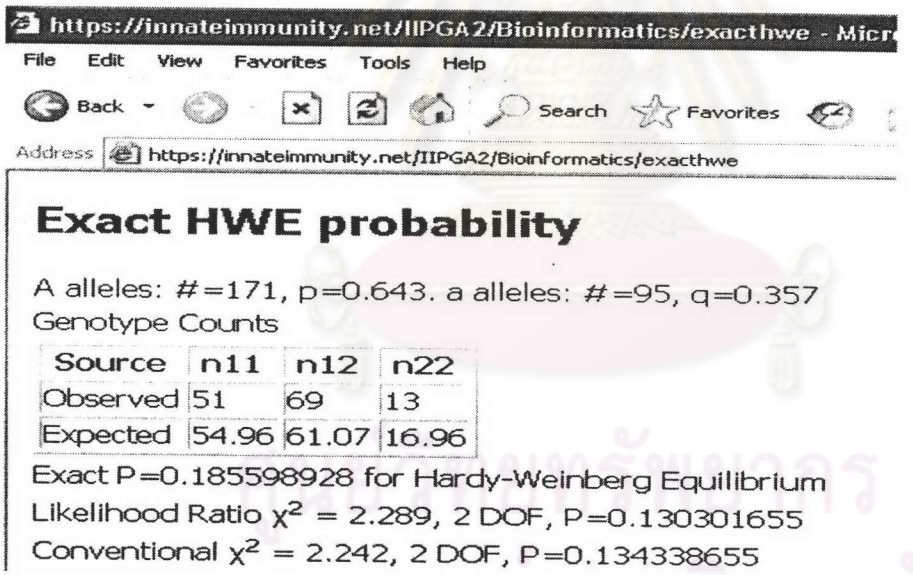
When run, the program will estimate the statistical significance of a test that the genotype counts are in Hardy-Weinberg equilibrium. The output file of TGFß2 (+71_72 ACAA) was shown below.



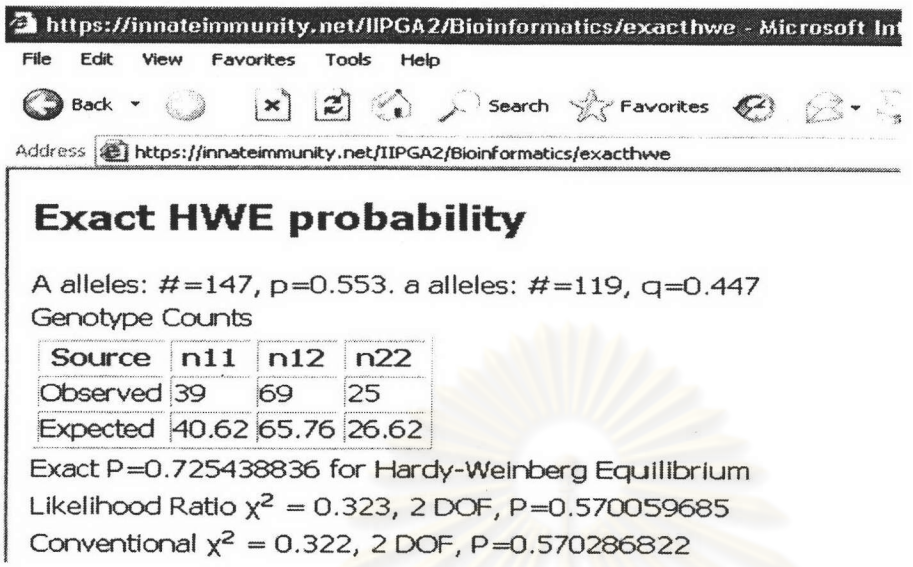
The output file of TGFß2 (+ 720T/G) was shown below.



The output file of TGFß2 (+ 89835A/G) was shown below.



The output file of TGFß2 (+ 94400_94401insA) was shown below.



Exact HWE probability

A alleles: #=147, p=0.553. a alleles: #=119, q=0.447

Genotype Counts

Source	n11	n12	n22
Observed	39	69	25
Expected	40.62	65.76	26.62

Exact P=0.725438836 for Hardy-Weinberg Equilibrium

Likelihood Ratio χ^2 = 0.323, 2 DOF, P=0.570059685

Conventional χ^2 = 0.322, 2 DOF, P=0.570286822

BIOGRAPHY

Miss. Krongkamol Hemwijit was born on February 5, 1980 in Roi-Et, Thailand. She graduated with the Bachelor degree of Science in Botany Major Genetic from Chulalongkorn University in 2003 and then attended to particulate in Medical Science program, Graduate School, Chulalongkorn University for her master degree.