

STATISTICAL ANALYSIS OF POPULATION-BASED CANDIDATE GENE STUDIES

ŚAUNAK SEN*

UNIVERSITY OF CALIFORNIA SAN FRANCISCO

sen@biostat.ucsf.edu

INTRODUCTION

Candidate gene studies are undertaken when there is *a priori* evidence to suggest that variation in a gene may be responsible for variation in an observable phenotype. Such studies may be preceded by genome-wide linkage studies in humans or in rodents, or there may be molecular evidence for a possible mode of action for the gene, or the gene may belong to a family of genes whose members have been previously implicated.

In this note we will go through the practical statistical aspects of analyzing data from population-based candidate gene studies. In such studies, the haplotype phase of the individuals is generally unknown and has to be statistically estimated. Performing association analysis using haplotypes has some advantages, most importantly model parsimony, when there is strong LD in the candidate gene. Using haplotypes reduces the number of comparisons to make, thereby increasing power. Sometimes, however, the analyzing the raw SNP data might be better, specially if the phenotype requires specialized modeling, as would be the case with a time to event phenotype.

We will illustrate the basic idea behind haplotype association analysis using data from a cohort study with a dichotomous outcome, and two SNPs. Then we will look at a second dataset where we have a survival (time to event phenotype), and show how we can analyze such data when there is strong LD in the gene.

We will follow a three-step process. First we will *check for data quality*. This includes examining the phenotypes, as well as the genotype data. In particular, we will check for Hardy-Weinberg equilibrium in the SNPs. The second step is to *examine the haplotype structure* in the region spanned by the SNPs. We will look for pairwise linkage disequilibrium between SNPs, and the extent of LD across all the SNPs. We will choose SNPs to focus on (haplotype tagging SNPs) depending on the haplotype structure. The final step is *haplotype reconstruction and measuring association* between the haplotype variation and phenotypes. Association methods that do not rely on haplotype reconstruction are closely-related, and may be sometimes preferable to those that rely on haplotype reconstruction. The exact choices to make depend on the data at hand, and the goals of the investigator.

SOFTWARE

What software? The statistical methods in genetic analysis are constantly evolving, and there are many software packages for performing genetic analyses. Some are available as add-on packages to existing statistical packages such as Stata and R. Many exist as stand-alone packages, often with their own data formatting requirements. Please consult your “primary care” statistician for suggestions on which software to use. In this note we will use add-on packages to Stata.

Downloading and installing packages We will use Stata packages written by Adrian Mander and David Clayton. To find out what the packages do and to install them type the following commands:[†]

*© 2006 Śaunak Sen; Last updated April 24, 2006.

[†]The `net describe` command describes a package available over the net. The `net install` command installs a package.

```

/* Packages from Adrian Mander */
net describe http://www.stata.com/stb/stb55/sbe34
net install sbe34
net describe http://www.stata.com/stb/stb57/sbe38
net install sbe38
net describe http://www.stata-journal.com/software/sj2-1/st0008
net install st0008
/* Packages from David Clayton */
net describe http://www-gene.cimr.cam.ac.uk/clayton/software/stata/genassoc
net install genassoc

```

CASE STUDY: A DICHOTOMOUS PHENOTYPE AND TWO SNPs

In a large community-based cohort study subjects were measured at two time points. If the change was more than a certain threshold, they were assessed to have “declined”, else not. Four SNPs were measured on the subjects; for this note we will focus on two SNPs – the first and the fourth. There were two racial/ethnic groups in the study which was conducted in two different sites. We first enter the data:[‡]

```

set memory 10m
insheet using decline.csv
. desc

Contains data
  obs:      3,075
  vars:       13
  size:     76,875 (99.3% of memory free)
-----
variable name  storage  display  value  variable label
              type   format   label
-----
race           byte    %8.0g
gender         byte    %8.0g
site           byte    %8.0g
changescore    byte    %8.0g
decline        byte    %8.0g
snp1           byte    %8.0g
snp2           byte    %8.0g
snp3           byte    %8.0g
snp4           byte    %8.0g
csnp1          str3    %9s
csnp2          str3    %9s
csnp3          str3    %9s
csnp4          str3    %9s
-----

```

For simplicity we will use only the males in the data, and remove all individuals for whom we had missing data for the phenotype (declined or not). This step also decreases the memory requirements for the analysis.

```

drop if gender==2
drop if decline ==.

```

Next we have to do some recoding for the Stata programs to work properly. Basically, we recode each SNP genotype using two variables. For example for the first SNP we will use:

[‡]The variables snp1 through snp4 are coded numerically 0, 1, or 2, while the variables csnp1 through csnp4 are the character representations, such as C/C, C/G, G/G.

```

generate gsnpl1=1 if snp1>=1 & snp1!=.
replace gsnpl1=0 if snp1==0 & snp1!=.
generate gsnpl2=1 if snp1==2 & snp1!=.
replace gsnpl2=0 if snp1!=2 & snp1!=.

```

Checking HWE This is a data quality check, which may alert us about possible strata in the data. We do this using the `gtab` command. Here we check for the strength, direction, and statistical significance of the deviation from HWE.

```
. gtab gsnpl1 gsnpl2
```

Allele	Total frequency	Homozygots Obsvd	(Expctd)	Heterozygots Obsvd	(Expctd)	Z (HWE)
0:	1721	681	(643.9)	359	(433.2)	5.8 11
1:	579	110	(72.9)	359	(433.2)	5.8 11

```

Global kappa statistic for Hardy-Weinberg equilibrium = 0.171
(Z-value = 5.811)
(p-value = 0.0000)

```

That looks pretty bad! But wait, we should stratify by ethnic group.

```
. gtab gsnpl1 gsnpl2 if race==1
```

Allele	Total frequency	Homozygots Obsvd	(Expctd)	Heterozygots Obsvd	(Expctd)	Z (HWE)
0:	1328	585	(577.1)	158	(173.8)	2.5 16
1:	200	21	(13.1)	158	(173.8)	2.5 16

```

Global kappa statistic for Hardy-Weinberg equilibrium = 0.091
(Z-value = 2.516)
(p-value = 0.0119)

```

```
. gtab gsnpl1 gsnpl2 if race==2
```

Allele	Total frequency	Homozygots Obsvd	(Expctd)	Heterozygots Obsvd	(Expctd)	Z (HWE)
0:	393	96	(100.0)	201	(192.9)	-0.8 21
1:	379	89	(93.0)	201	(192.9)	-0.8 21

```

Global kappa statistic for Hardy-Weinberg equilibrium = -0.042
(Z-value = -0.821)
(p-value = 0.4116)

```

Now it doesn't look so bad. If the allele frequencies are not in HWE, one should check the marker typing, and investigate further. Note that the kappa statistic above gives the proportion of homozygotes in excess of what would be expected under HWE. For example in the whole dataset, the excess of homozygotes was 17% which is pretty strong deviation from HWE. However for the Whites (`race==1`), although the p-value for the statistic is significant (0.011), the magnitude of deviation is only about 9% which is not so strong. In this data, the HWE deviation decreases further when we stratify by site.

The extent of linkage disequilibrium Now we check for the extent of linkage disequilibrium between the two SNPs. This can be done in many different ways, but we start with the most basic, which is a simple table. Then we will look at common measures of LD such as D' , and finally we will measure its statistical significance using a likelihood ratio test.

```
. tabulate csnp1 csnp4, chi2
```

csnp1	csnp4			Total
	C/C	C/T	T/T	
C/C	197	335	131	663
C/G	174	152	24	350
G/G	60	36	12	108
Total	431	523	167	1,121

```
Pearson chi2(4) = 66.6968 Pr = 0.000
```

If the two SNPs were not in LD, then their genotypes would be independent of each other. In other words, the χ^2 test for independence in the 3×3 contingency table can be used to detect existence of LD. Although this test should dispel any doubts about the existence of LD, but we can also use some other measures, such as D' .

```
. pwld snp1 snp4, gtype
```

Off-diagonal elements are estimates of Lewontin's D' (assuming H-W equilibrium)
Diagonal elements are relative frequencies of allele 2

```
      snp1  snp4
snp1  0.25
snp4  0.52  0.38
```

Yet another way of measuring LD is with a likelihood ratio test using the `hapipf` command which we will use extensively later. This command fits the observed genotype counts to different models.

```
. hapipf gsnp11 gsnp12 gsnp41 gsnp42, ipf(11+12) mv model(0) display
Alleles at locus 11 are contained in variables (gsnp11 gsnp12)
Alleles at locus 12 are contained in variables (gsnp41 gsnp42)
```

```
EXPANDING MISSING DATA....
```

```
There are 146 missing values at locus 1
There are 384 missing values at locus 2
```

```
Iteration 1 loglhd = -2453.741630392266
Iteration 2 loglhd = -2451.37576027427
Iteration 3 loglhd = -2451.365793904624
Iteration 4 loglhd = -2451.365657217473
Iteration 5 loglhd = -2451.365545290318
```

```
Haplotype Frequency Estimation by EM algorithm
```

```
-----
No. loci          = 2
Log-Likelihood    = -2451.365545290318
DF                = 1
No. parameters    = 3
```

No. cells = 4

Imputed Frequencies

Haplo	freq	eprob
0.0	1087.0587	.444423
0.1	752.15696	.30750489
1.0	432.55643	.17684237
1.1	174.22796	.07122975

Expected Frequencies

Haplo	freq	eprob
0.0	1142.6402	.46714645
0.1	696.57475	.28478117
1.0	376.97441	.15411873
1.1	229.81062	.09395365

TOTAL FREQ is 2446

This command fits the observed genotype counts to a model which says that the two loci are segregating independently (i.e. there is no LD between the loci). There are several things to note in the output. The option `ipf(11+12)` specifies that we are fitting a model where the two loci are independent. The option `mv` specifies that we will be filling in (or imputing) missing genotype data if any genotype data are missing, using information from other genotypes. The `model(0)` option stores the output from the command for comparisons with other models (see below). The “imputed” frequencies are the observed haplotype frequencies, while the “expected” ones are those predicted by the model. Note the proportionality of the expected frequencies; this is because of the independent loci model.

Next we fit a model that incorporates the dependence of the genotype of the two SNPs on each other, and compare the two models using a likelihood ratio test.

```
. hapipf gsnp11 gsnp12 gsnp41 gsnp42, ipf(11*12) mv model(1) lrtest(1,0) display  
Alleles at locus 11 are contained in variables (gsnp11 gsnp12)
```

```
Iteration 1 loglhd = -2433.574990843922  
Iteration 9 loglhd = -2424.479804979907
```

Haplotype Frequency Estimation by EM algorithm

```
-----  
No. loci = 2  
Log-Likelihood = -2424.479804979907  
Df = 0  
No. parameters = 4  
No. cells = 4
```

Imputed Frequencies

Haplo	freq	eprob
0.0	1042.4952	.4262041
0.1	795.72823	.32531816
1.0	475.22991	.1942886
1.1	132.54662	.05418913

Expected Frequencies

Haplo	freq	eprob
0.0	1042.5267	.42621697
0.1	795.69763	.32530565
1.0	475.19992	.19427633
1.1	132.57578	.05420105

TOTAL FREQ is 2446.0001

Likelihood Ratio Test Comparing Model 11+12 to 11*12

```
-----
llhd2 (df2)          = -2451.3655 1
llhd1 (df1)          = -2424.4798 0
```

```
-2*(llhd2-llhd1)    = 53.771481
Change in df        = 1
p-value             = 2.252e-13
Accept Model 11*12 at 5% significance level
```

Note the use of the `ipf(11*12)` option which instructs Stata to fit a model of dependence between the loci. The comparison of the two models using the χ^2 test indicates that the second model incorporating dependence between the two SNPs fits better. Therefore we conclude that the two SNPs are in LD.[§] We can also see the frequencies of the four haplotypes:

	Hapl	freq	eprob
CC	0.0	1042.5	.426
CT	0.1	795.7	.325
GC	1.0	475.2	.194
GT	1.1	132.6	.054

Notice that when the first SNP is C, then the odds that the second SNP is C is about 4:3. When the first SNP is G, the odds for the second SNP to be C is much higher, about 4:1.

Association between haplotype and outcome We know that there is association between the two SNPs, since they are in LD. So we will fit a model that assumes that the SNPs are associated with each other, but the phenotype (decline) is not. We do this as follows:

```
. hapipf gsnpl1 gsnpl2 gsnpl41 gsnpl42, ipf(11*12+decline) mv model(2) display
```

Haplotype Frequency Estimation by EM algorithm

```
-----
No. loci          = 2
Log-Likelihood    = -3740.424619557014
Df                = 3
No. parameters    = 5
No. cells         = 8
```

Expected Frequencies

```
-----+
```

[§]We have neglected to stratify the data by ethnic group; this discussion is for illustration of the basic concepts only.

Haplo	decline	freq	eprob
0.0	0	803.84522	.32863664
0.0	1	238.68151	.09758034
0.1	0	613.52642	.25082846
0.1	1	182.17115	.07447717
1.0	0	366.40517	.1497977
1.0	1	108.79475	.04447864
1.1	0	102.22319	.04179198
1.1	1	30.35259	.01240907

TOTAL FREQ is 2446

Notice that because we are fitting a model where the SNPs depend on each other, but are independent of the phenotype, the ratio of controls to cases in each haplotype class is constant and about 3.5:1. Next we fit a model that assumes that the decline phenotype is associated with the haplotypes. We then compare the fit with the previous no association model.

```
hapipf gsnpl1 gsnpl2 gsnpl41 gsnpl42, ipf(l1*l2*decline) mv model(3) lrtest(3,2) display
```

Haplotype Frequency Estimation by EM algorithm

```
-----
No. loci           = 2
Log-Likelihood     = -3731.124926922354
DF                 = 0
No. parameters     = 8
No. cells          = 8
```

Expected Frequencies

Haplo	decline	freq	eprob
0.0	0	824.38544	.33703411
0.0	1	215.00064	.08789887
0.1	0	625.22327	.2556105
0.1	1	172.51465	.07052929
1.0	0	356.38052	.14569932
1.0	1	121.57411	.04970324
1.1	0	80.010757	.03271086
1.1	1	50.910591	.02081382

TOTAL FREQ is 2446

Likelihood Ratio Test Comparing Model l1*l2+decline to l1*l2*decline

```
-----
llhd2 (df2)       = -3740.4246 3
llhd1 (df1)       = -3731.1249 0

-2*(llhd2-llhd1)  = 18.599385
Change in df     = 3
p-value          = .00033082
Accept Model l1*l2*decline at 5% significance level
```

The χ^2 test has a p-value of 0.00033, which implies that the decline phenotype is associated with the SNP haplotypes. In other words, we have demonstrated the association between the candidate gene and the phenotype. It is helpful to examine the haplotype frequencies and their covariation with the decline phenotype as follows:

	Haplo	decline	freq	eprob	Ctrls	Cases	Haplo
CC	0.0	0	824.4	.337			
CC	0.0	1	215.0	.088	824	215	CC
CT	0.1	0	625.2	.256			
CT	0.1	1	172.5	.071	625	172	CT
GC	1.0	0	356.4	.146			
GC	1.0	1	121.6	.050	356	122	GC
GT	1.1	0	80.0	.033			
GT	1.1	1	50.9	.021	80	51	GT

Which haplotype appears to be associated with the decline phenotype? It appears to be the GT haplotype. What is the odds ratio for declining for the GT haplotype relative to the CC haplotype? It is about 3. (Why?)

Exercise: Haplotype and genotype data We could have performed a SNP by SNP analysis using logistic regression as below. Is the SNP analysis consistent with the haplotype analysis? Which one is better? We can check this by fitting three models. The first two are logistic regressions testing for association with each individual SNP. The third model has both SNPs in the model as additive effects.

```
. logit decline snp1
```

```
Logit estimates                               Number of obs =      1150
                                             LR chi2(1)      =       10.58
                                             Prob > chi2     =       0.0011
Log likelihood = -606.91745                 Pseudo R2      =       0.0086
```

decline	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
snp1	.3355784	.1020168	3.29	0.001	.1356291	.5355276
_cons	-1.42298	.0926556	-15.36	0.000	-1.604582	-1.241379

```
. logit decline snp4
```

```
Logit estimates                               Number of obs =      1141
                                             LR chi2(1)      =        0.96
                                             Prob > chi2     =       0.3269
Log likelihood = -603.21741                 Pseudo R2      =       0.0008
```

decline	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
snp4	.1006319	.1024733	0.98	0.326	-.1002122	.301476
_cons	-1.333483	.1076788	-12.38	0.000	-1.544529	-1.122436

```
. logit decline snp1 snp4
```

```
Logit estimates                               Number of obs =      1121
                                             LR chi2(2)      =       13.83
                                             Prob > chi2     =       0.0010
Log likelihood = -587.99874                 Pseudo R2      =       0.0116
```

decline	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
snp1	.3836686	.1062135	3.61	0.000	.1754939 .5918433
snp4	.1905169	.1064727	1.79	0.074	-.0181657 .3991996
_cons	-1.605997	.1365936	-11.76	0.000	-1.873716 -1.338279

HAPLOTYPE RECONSTRUCTION USING THE EM ALGORITHM

By now you must have noticed the haplotype reconstruction program mention of the “EM Algorithm”. So what is it and how does it work?

The Expectation-Maximization (EM) Algorithm is a widely-used and very general method for obtaining maximum likelihood estimates when there is missing data. We will illustrate how it works in the context of inferring haplotypes using unphased genotype data.

Consider two SNPs that are in LD with each other. The first SNP has two alleles A, and T. The second SNP has alleles C and G. Then we have four possible haplotypes. Assume that the haplotypes are in Hardy-Weinberg Equilibrium in the population, and the population frequencies are as follows:

Haplotype	Frequency (%)
AC	10
AG	10
TC	70
TG	10

If we observe genotypes from 500 individuals, who would have 1000 haplotypes in all. For all possible genotype combinations we can unambiguously infer haplotypes except when an individual is heterozygous at both SNPs.

SNP1 genotype	SNP2 genotype		
	CC	CG	GG
AA	A-C/A-C	A-C/A-G	A-G/A-G
AT	A-C/T-C	*	A-G/T-G
TT	T-C/T-C	T-C/T-G	T-G/T-G

In that case, the individual may either be A-C/T-G or A-G/T-C. On the other hand, if we know (or are willing to assume) that the haplotypes are in HWE, and their frequencies, we can make a good guess. The probability that an individual has the A-C/T-G diplotype is $2 \times 0.1 \times 0.1 = 0.02$, and the A-G/T-C diplotype is $2 \times 0.7 \times 0.1 = 0.14$. Thus we expect 1/8 of the heterozygous individuals to have the A-C/T-G diplotype and the rest to have the A-G/T-C diplotype. If the haplotype frequencies are unknown, then we can follow the following algorithm.

EM algorithm for haplotype reconstruction

Initialize. Begin by making an assumption on haplotype frequencies (say all of them being equally likely).

E-step. Count the expected number of haplotypes of each type using the haplotype frequency estimate to resolve ambiguous haplotype assignments.


```

. stem survtime

Stem-and-leaf plot for followup

followup rounded to nearest multiple of .1
plot in units of .1

 0* | 1234567799
 1* | 00023455566668999
 2* | 000112223333346667777888899999
 3* | 00122223333334444555566667788999999
 4* | 00001112222233344444455555666788888899999
 5* | 00011122223333445667899
 6* | 112444566777899
 7* | 0124456788
 8* | 111256779
 9* | 123359
10* | 001345999
11* | 0136
12* | 129
13* | 259
14* | 9
15* |
16* | 448
17* |
18* |
19* |
20* | 1
21* |
22* |
23* | 8
24* |
25* |
26* |
27* |
28* |
29* | 2

```

The distribution of age at dialysis has a long left tail, with mode around 60 years. The survival time since dialysis has a long right tail, as one would be expected for survival times.

Haplotype structure Next we look at the haplotype distribution of the 7-SNP haplotypes.

Haplotype distribution of 7 SNPs (with 4% or more frequency):

snp	abcdefg	hap.freq
	AAAAATG	0.046
	AAAAGCA	0.404
	AAGGATG	0.121
	TAAAGCA	0.057
	TAGGATG	0.257

Total		0.884

Three SNPs (first, fourth, and seventh) describe this haplotype variation completely. Using those SNPs only we get the following haplotype distribution. It matches the haplotype distribution of the 7-SNP haplotypes quite well.

Haplotype distribution of 3 SNPs (with 4% or more frequency):

```

snp  adg  hap.freq
AAA  0.449
AAG  0.065
AGG  0.125
TAA  0.084
TGG  0.261
-----
Total      0.985

```

Our subsequent analysis uses these three “tagging” SNPs.[¶]

Association analysis We simplified our predictor variables by choosing 3 SNPs out of the 7 typed. Now we try to show that there is association between genetic variation in the gene (as measured by the three SNPs) and the phenotype. We do this by using Cox proportional hazards regression on each SNP in turn, also adjusting for ethnicity and age at dialysis.

```

. generate lifetime = followup
. stset lifetime, failure(died)

. /* after other commands */
. /* store base model */
. stcox ageatdial raceasian raceafamr racehispn

      failure _d:  died
      analysis time _t:  lifetime

Iteration 0:   log likelihood = -395.98549
Iteration 1:   log likelihood = -374.39245
Iteration 2:   log likelihood = -372.95193
Iteration 3:   log likelihood = -372.88244
Iteration 4:   log likelihood = -372.88165
Iteration 5:   log likelihood = -372.88165
Refining estimates:
Iteration 0:   log likelihood = -372.88165

Cox regression -- Breslow method for ties

No. of subjects =          229          Number of obs =          229
No. of failures =           84
Time at risk    = 1220.832873

Log likelihood = -372.88165          LR chi2(4) =          46.21
                                      Prob > chi2 =          0.0000

-----
      _t | Haz. Ratio  Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
 ageatdial |  1.045046   .0096831    4.76   0.000    1.026239    1.064198
 raceasian |  .6782949   .196576    -1.34   0.180    .384355    1.197029
 raceafamr |  .810036    .2138194   -0.80   0.425    .4828572    1.358908
 racehispn |  .1825651   .1349006   -2.30   0.021    .0428993    .7769372
-----

. est store Null

. /* do cox regression for each snp */

```

[¶]We could have used other methods for obtaining tagging SNPs.

Now we perform the likelihood ratio tests that compare the base model (without SNPs) to three models, each with a single SNP.

```

. /* make likelihood ratio tests */

. lrtest Null A

likelihood-ratio test                    LR chi2(2) =    11.89
(Assumption: Null nested in A)         Prob > chi2 =    0.0026

. lrtest Null D

likelihood-ratio test                    LR chi2(2) =     7.22
(Assumption: Null nested in D)         Prob > chi2 =    0.0270

. lrtest Null G

likelihood-ratio test                    LR chi2(2) =     6.04
(Assumption: Null nested in G)         Prob > chi2 =    0.0487

```

Since we made three comparisons, we use a Bonferroni adjustment. We take the smallest p-value (0.0026) and multiply it by 3, to get a multiplicity adjusted p-value of 0.0078. Thus we have sufficient evidence that genetic variation in this gene is associated with the outcome. Notice that all SNPs show a strong protective effect of a recessive nature. For SNPA the T/T genotype has a protective hazard ratio of 0.3 with the 95% confidence interval stretching from 0.14 to 0.71.^{||} The wide confidence interval is due to the small sample size.

```
. list snpa snpd snpg if snpa=="T/T"
```

	snpa	snpd	snpg
8.	T/T	G/G	G/G
10.	T/T	G/G	G/G
12.	T/T	A/A	A/A
17.	T/T	G/G	G/G
18.	T/T	G/G	G/G
25.	T/T	G/G	G/G
34.	T/T	G/G	G/G
40.	T/T	G/G	G/G
41.	T/T	G/G	G/G
46.	T/T	A/G	A/G
90.	T/T	G/G	G/G
98.	T/T	G/G	A/G
99.	T/T	G/G	G/G
108.	T/T	G/G	G/G
120.	T/T	G/G	G/G
130.	T/T	G/G	G/G
133.	T/T	A/A	A/A
136.	T/T	A/A	A/A

^{||}The 98% confidence interval, which may be considered Bonferroni corrected, is from 0.12 to 0.82.

145.	T/T	G/G	G/G
150.	T/T	G/G	G/G

159.	T/T	G/G	G/G
174.	T/T	G/G	G/G
187.	T/T	G/G	G/G
193.	T/T	A/A	A/A
195.	T/T	A/A	A/A

196.	T/T	A/G	A/G
199.	T/T	G/G	G/G
200.	T/T	A/G	A/G
201.	T/T	A/G	A/G
225.	T/T	A/G	A/G

226.	T/T	G/G	G/G
227.	T/T	G/G	G/G

Most of those homozygous T/T for SNPA have the TGG/TGG diplotype. The only other haplotype represented in the subset is the TAA haplotype. This leads us to suspect that the TGG haplotype may be the causal variant.

Figure 1: Survival curves stratified by SNPA genotype

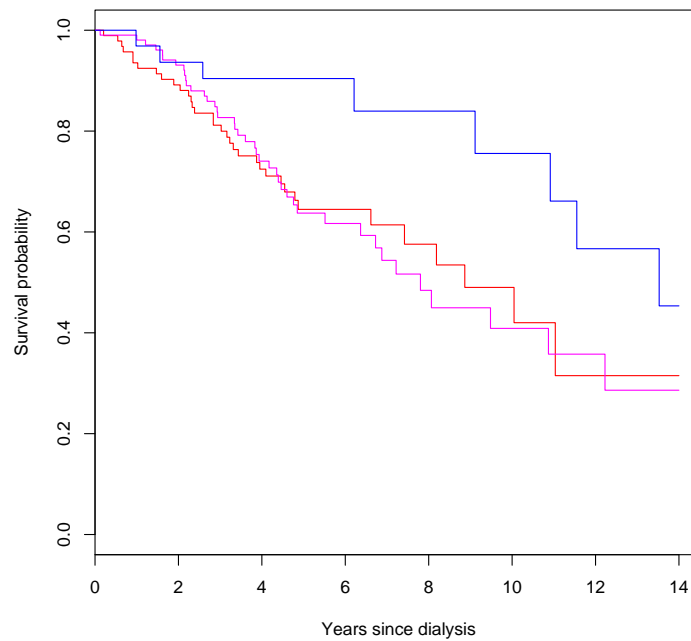


Figure 2: Fitted proportional hazard survival curves by SNPA genotype

