Paper 219-2011

Exploring selective sweeps with SAS®

Kristan A. Schneider, University of Vienna, Vienna, Austria

Barbara Schneider, Medical University of Vienna, Vienna, Austria

ABSTRACT

We exemplify how SAS® can be used to perform analyses of genetic data. In particular, we perform a population genetic analysis based on a dataset containing microsatellite data in a population of *Plasmodium falciparum* (the parasite that causes the most virulent form of human malaria). The microsatellite markers are located around the *dhps* gene of *P. falciparum*, which is associated with resistance against the antimalarial drug *Sulfadoxine.* Resistance-causing mutations are selectively favored and spread in the population. A selective sweep, or the "hitchhiking" effect of an advantageous mutation, refers to the elimination of pre-existing genetic variation. Studying the pattern of genetic variation is important because by measuring the span of selective sweeps we may infer the selective pressures on drug resistance and relate it to the demographic and clinical factors specific to the population under examination. We illustrate how various components of SAS® can be used and linked together to perform the desired analysis.

INTRODUCTION

AIM

We exemplify how to use SAS® to perform a simple analysis of genetic data. We choose an example that is relevant for malaria control. The aim of the manuscript is to illustrate how to perform meaningful analyses and statistics starting from a dataset (such as an EXCEL spreadsheet) that was created in a naïve way.

WHAT IS MALARIA?

Malaria is still a threat to the public health in large areas of the developing world. It is an infectious disease produced by (unicellular eukaryotic) protozoan parasites of the genus *Plasmodium*, with *Plasmodium falciparium* being responsible for the most virulent form of human malaria (cf. Snow et al., 2005). *P. falciparum* causes high morbidity and mortality, which annually results in 200 million to 300 million infections and one to three million deaths (cf. WHO 2000). Malaria is a vector-borne disease with a complex transmission cycle having sexual phases in the mosquito vector and asexual phases in the infected host (e.g., Daily, 2006).

DRUG-RESISTANCE IN MALARIA

Many effective and widely used drugs, e.g., *Chloroquine* or *Sulfadoxine* and *Pyrimethamine*, have been rendered useless because parasites rapidly evolved resistance against them. The limited repertoire of safe, effective, and affordable anti-malarial drugs has made research on the emergence and dispersion of resistance a global health priority (see Marsh, 1998).

IDEA BEHIND THE GENETIC ANALYSIS

We need to reconstruct the past dynamics of drug resistance in a given population to understand which factors lead to the spread of resistant parasites. In the absence of reliable public-health records, such retrospective analysis may not be feasible. However, fast accumulation of parasite genome-sequence data provides a means to examine indirectly the past events of drug-resistance evolution without epidemiological data: by examining "selective sweeps" around the loci of drug resistant mutations (cf. Schneider and Kim 2010).

A selective sweep, or the "hitchhiking" effect of an advantageous mutation, refers to the elimination of pre-existing genetic variation when a particular chromosome segment carrying a favored allele sweeps through the population (Maynard Smith and Haigh, 1974; Stephan et al., 1992; Barton, 2000). The extent of this wipeout depends on how fast the favored allele increases to high frequency while meiotic recombination is constantly eroding the association between the favored allele and the surrounding chromosome segment. The chromosomal span of reduced variation thus depends on the relative reproductive advantage of resistant alleles over sensitive alleles, which determines the speed of frequency increase. Therefore, by measuring the span of selective sweeps, we may infer the selective pressures on drug resistance and relate it to the demographic and clinical factors specific to the population under examination.

Label	Dhfr	MS locus 1	MS locus 3	MS locus 3	MS locus 4
F70	3	307		101	290
	3	314			286
	3				278
	3				296

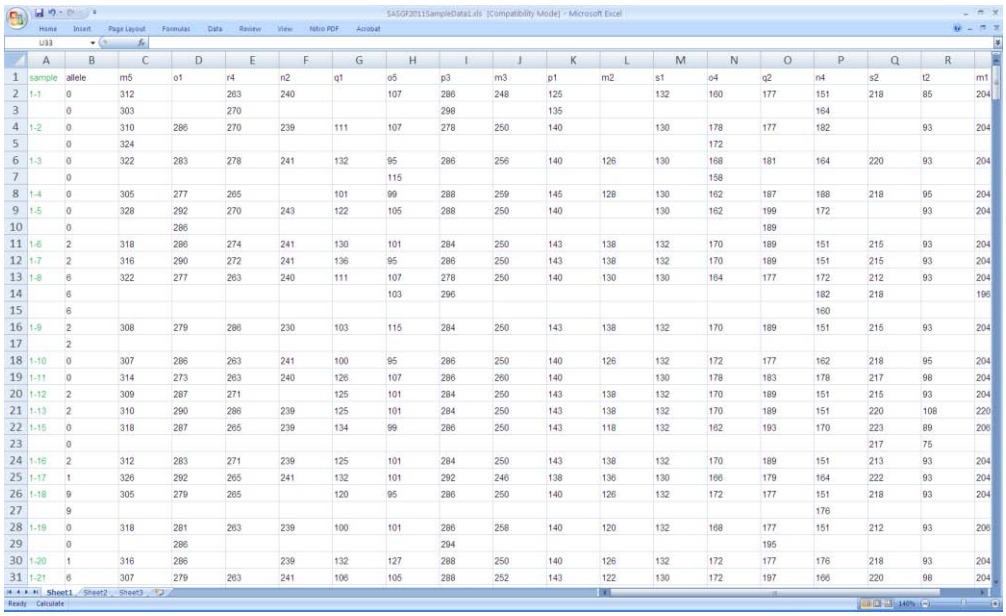
Table 1. The table shows the entry for sample "F70". The parasites have allele A_3 at the *dhps* locus. At the first MS locus, alleles with 307 and 314 repeats of the core sequence were found. The entry for MS locus 2 is empty. At MS locus 3 only alleles with 101 repeats of the cores sequence were detected, and at MS locus 4, four different alleles were found in sample 'F70'.

DATA AND ANALYSIS

For our illustration purposes, we use a simulated dataset that imitates an original dataset obtained from a study in Kenya.

Blood samples were taken from patients infected with malaria caused by *P. falciparum*. The parasites were extracted from the bold, and parts of the genome were sequenced. In particular, a gene associated with resistance against Sulfadoxine (dhps gene) was sequenced. Moreover, for both genes several microsatellite (MS) markers were sequenced. Microsatellites are highly repetitive DNA sequences, which are typically neutral.

The dataset contains the following information. One or more consecutive rows represent the genetic variation from the parasites of one blood sample (Table 1 shoes an example). The first column labels the blood sample. If more than on rows represent the same blood sample, the label is specified only in the respective first row. Column three specifies the particular allele found at the dhfr locus associated with drug resistance. Again entered only in the first row of a given sample (only such samples were included in which all parasites had the same allele at the dhfr locus). The remaining columns represent the microsatellite variation at different genome positions. For a given blood sample the entries in the rows, are the number of sequence repeats found at the various microsatellite loci in a given blood sample. Figure 1 shows a screenshot of the initial dataset in .xls format.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	sample	allele	m5	o1	r4	n2	q1	o5	p3	m3	p1	m2	s1	o4	q2	n4	s2	t2	m1
2	1-1	0	312		263	240		107	286	248	125		132	160	177	151	218	85	204
3		0	303		270				298		135					164			
4	1-2	0	310	286	270	239	111	107	278	250	140		130	178	177	182		93	204
5		0	324											172					
6	1-3	0	322	283	278	241	132	95	286	256	140	126	130	168	181	164	220	93	204
7		0						115						158					
8	1-4	0	305	277	265		101	99	288	259	145	128	130	162	187	188	218	95	204
9	1-5	0	328	292	270	243	122	105	288	250	140		130	162	199	172		93	204
10		0		286											189				
11	1-6	2	318	286	274	241	130	101	284	250	143	138	132	170	189	151	215	93	204
12	1-7	2	316	290	272	241	136	95	286	250	143	138	132	170	189	151	215	93	204
13	1-8	6	322	277	263	240	111	107	278	250	140	130	130	164	177	172	212	93	204
14		6						103	296							182	218		196
15		6														160			
16	1-9	2	308	279	286	230	103	115	284	250	143	138	132	170	189	151	215	93	204
17		2																	
18	1-10	0	307	286	263	241	100	95	286	250	140	126	132	172	177	162	218	95	204
19	1-11	0	314	273	263	240	126	107	286	260	140		130	178	183	178	217	98	204
20	1-12	2	309	287	271		125	101	284	250	143	138	132	170	189	151	215	93	204
21	1-13	2	310	290	286	239	125	101	284	250	143	138	132	170	189	151	220	108	220
22	1-15	0	318	287	265	239	134	99	286	250	143	118	132	162	193	170	223	89	206
23		0															217	75	
24	1-16	2	312	283	271	239	125	101	284	250	143	138	132	170	189	151	213	93	204
25	1-17	1	326	292	265	241	132	101	292	246	138	136	130	166	179	164	222	93	204
26	1-18	9	305	279	265		120	95	286	250	140	126	132	172	177	151	218	93	204
27		9														176			
28	1-19	0	318	281	263	239	100	101	286	258	140	120	132	168	177	151	212	93	206
29		0		286					294						195				
30	1-20	1	316	286		239	132	127	288	250	140	126	132	172	177	176	218	93	204
31	1-21	6	307	279	263	241	106	105	288	252	143	122	130	172	197	166	220	98	204

Figure 1. Screen shot of the EXCEL spreadsheet.

IMPORTING THE DATASET

First, we import the EXCEL spreadsheet into SAS®. This can be done by using the import wizard or by the following code:

```
PROC IMPORT OUT= sasgf.dhps1
            DATAFILE="C:\Documents and
Settings\User\SASGF2011\SASGF2011SampleData.xls"
            DBMS=EXCEL REPLACE;
     RANGE="microsatellites$";
     GETNAMES=YES; * column names created from first row;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;
```

This code stores the spreadsheet as 'dhps' in the SAS® library 'SASGF' (see Figure 2 for an illustration of the imported data).

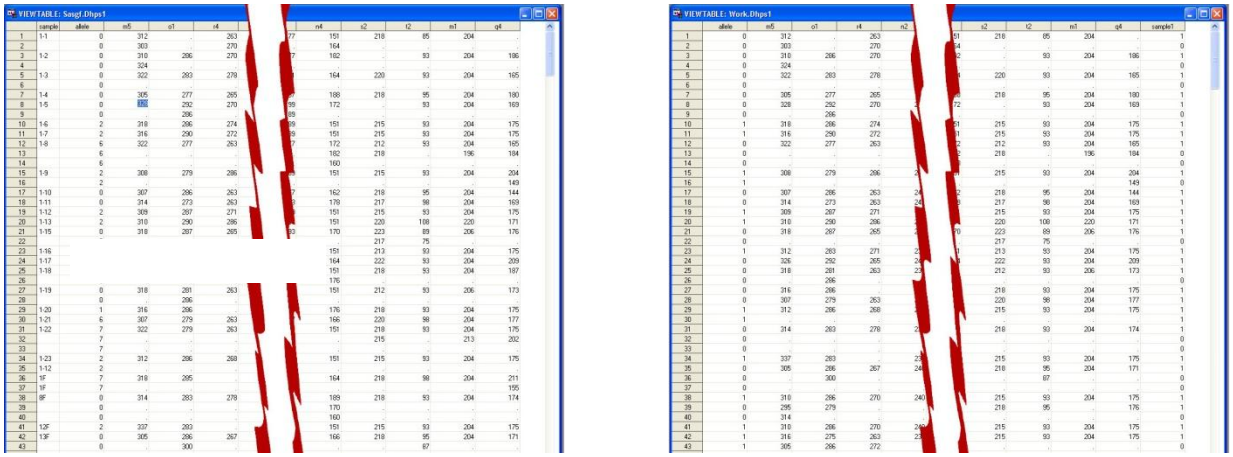


Figure 2. Screenshot of dataset 'sasgf.dhps1' (left), and 'work.dhps1' after performing the two data steps described in step 1. The first column 'sample' was deleted, and the 0-1 variable 'sample1' was added.

CALCULATING MEASURES FOR GENETIC VARIABILITY

We are interested in the genetic variability at each MS locus. A measure for the genetic variability is the heterozygosity. It is calculated as

$$H = \frac{1}{n-1}\left(1 - \sum_{i=1}^{n} {p_i}^2\right).$$

For a given MS locus *n* refers to the number of different alleles found among all samples at this locus. Moreover, p_i is the relative frequency of MS allele A_i (calculated among all samples). We want to calculate the heterozygosity separately for the blood samples that contain resistant and sensitive parasites, respectively.

For a given MS locus, some blood samples contain various MS alleles. When calculating the relative frequencies of the MS alleles we have to take this into account. Therefore, for each blood sample, we weigh the MS alleles reciprocal to the number of alleles found in this sample. In the example of Table 1, we weigh the alleles at MS locus 1 by ½, that of MS locus 3 by 1, and those of MS locus 4 by 1/4. At MS locus 2 the entries are missing, so they are not counted at all.

Step1: Manipulating the data

Within the sample there are various alleles at the *dhps* locus, A_1, A_2, ... Parasites carrying the allele A_2 are resistant whereas all others are sensitive. Since we want to distinguish only between sensitive and resistant, we replace the

column allele by a 0-1 variable, where '1' codes for resistant and '0' for sensitive. Entries with missing value for 'allele' are deleted since they are uninformative. This is done with a basic data step:

```
data work.dhps1;
set sasgf.dhps1;
if allele=0 then allele=0;
if allele=1 then allele=0;
if allele=2 then allele=1;
if allele=3 then allele=0;
if allele=. then delete;
run;
```

The first problem that arises with our data is that not every row contains a label referring to the respective blood sample. The second problem is that the labels are not in numeric format. We overcome the two problems in two steps.

We add a column 'sample1' to 'SASGF.dhps', with an entry '1' if a row contains a label for the sample, and '0' if it does not contain a label. Moreover, we delete the column 'sample'.

```
data work.dhps1;
set dhps1;
if sample=' ' then sample1=0;
else sample1=1;
drop sample;
run;
```

The output is illustrated in Figure 2Figure 3.

Hypothetically, one can overwrite the column 'sample' instead of deleting it and creating a new one. However, the column 'sample' has character format. Therefore, it cannot be read into SAS/IML® as a numeric matrix.

Next, we label the samples. More precisely, we replace the 0-1 variable 'sample1' by proper labels (1, 2, 3, …) for the samples. We use SAS/IML® for this purpose:

```
proc iml;
use dhps1;
read all into A;
close dhps1;
call delete(dhps1);
B=A;
B[,ncol(A)]=J(nrow(A),1,1);
inc=1;
label=1;
do k=1 to nrow(A);
  if A[k,ncol(A)]=0 then do;
    inc=inc+1;
  end;
  else do;
    B[k,ncol(A)]=label;
    if inc>1 then do;
       B[k-inc:k-1,ncol(A)]=J(inc,1,label-1);
       inc=1;
    end;
    label=label+1;
  end;
end;
create dhps1 from B;
edit dhps1;
append from B;
quit;
```

Figure 3 shows a screenshot of the manipulated dataset.

Now, we eliminated the problems with the initial dataset, and are ready to rearrange the data such that the heterozygosity can be calculated.

	COL4	COL5	COL6	COL7	COL8	COL9	COL10	COL11	COL12	COL13	COL14	COL15	COL16	COL17	COL18	COL19	COL20
1	263	240		107	286	248	125		132	160	177	151	218	85	204		1
2	270				298		135					164					1
3	270	239	111	107	278	250	140		130	178	177	182		93	204	186	2
4										172							2
5	278	241	132	95	286	256	140	126	130	168	181	164	220	93	204	165	3
6				115						158							3
7	265		101	99	288	259	145	128	130	162	187	188	218	95	204	180	4
8	270	243	122	105	288	250	140		130	162	199	172		93	204	169	5
9											189						5
10	274	241	130	101	284	250	143	138	132	170	189	151	215	93	204	175	6
11	272	241	136	95	286	250	143	138	132	170	189	151	215	93	204	175	7
12	263	240	111	107	278	250	140	130	130	164	177	172	212	93	204	165	8
13				103	296							182	218		196	184	8
14												160					8
15	286	230	103	115	284	250	143	138	132	170	189	151	215	93	204	204	9
16																149	9
17	263	241	100	95	286	250	140	126	132	172	177	162	218	95	204	144	10
18	263	240	126	107	286	260	140		130	178	183	178	217	98	204	169	11
19	271		125	101	284	250	143	138	132	170	189	151	215	93	204	175	12
20	286	239	125	101	284	250	143	138	132	170	189	151	220	100	220	171	13
21	265	239	134	99	286	250	143	118	132	162	193	170	223	89	206	176	14
22													217	75			14
23	271	239	125	101	284	250	143	138	132	170	189	151	213	93	204	175	15
24	265	241	132	101	292	246	138	136	130	166	179	164	222	93	204	209	16
25	263	239	100	101	286	258	140	120	132	168	177	151	212	93	206	173	17
26					294						195						17
27		239	132	127	288	250	140	126	132	172	177	176	218	93	204	175	18
28	263	241	106	105	288	252	143	122	130	172	197	166	220	98	204	177	19
29	268	238	112	99	290	267	143	138	132	170	189	151	215	93	204	175	20
30					284	250											21
31	270	232	129	109	286	242	138	87	130	176	191	189	218	93	204	174	22
32			113					108				170					22
33												160					22

Figure 3. Screenshot of dataset 'sasgf.dhps1' after the SAS/IML® procedure. The last column was replaced, by the labels of the samples. (Note that the column labels are replaced.)

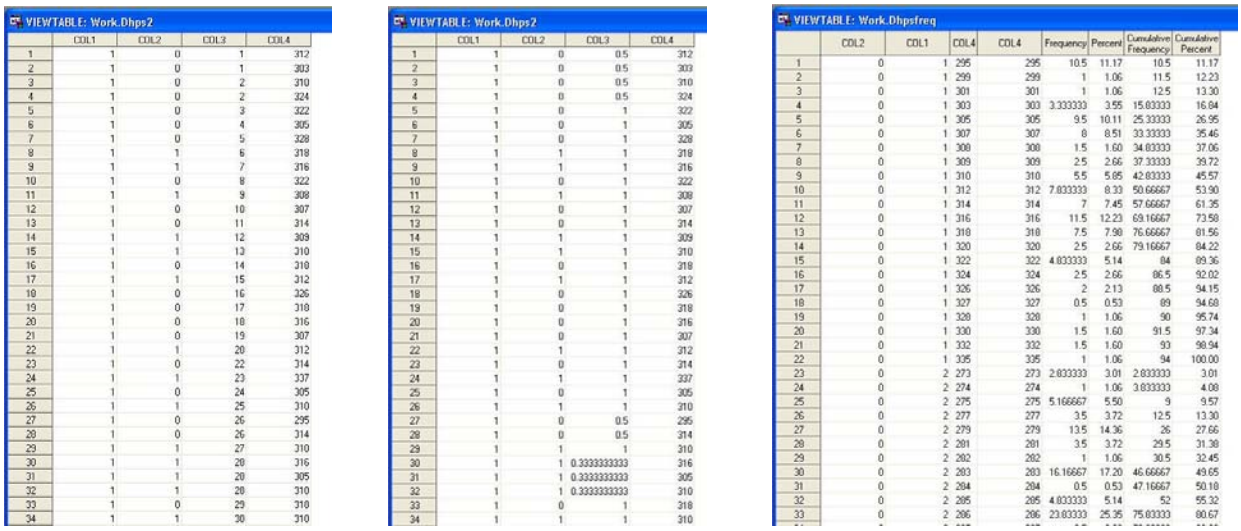
Step2: Rearranging the data

To calculate the heterozygosity at the various MS loci, we rearrange the dataset in a more convenient way. Namely, the first column specifies whether the sample carries sensitive or resistant parasites (0=sensitive, 1=resistant). The second column specifies the MS locus, in the third column the MS alleles are entered, and the fourth column specifies the weight.

```
proc iml;
use dhps1;
read all into A;
n=ncol(A)-2;
m=nrow(a) ;
B=J(m*n,4);
do k=1 to n;
  B[1+(k-1)*m:k*m,1]=J(m,1,k);
  B[1+(k-1)*m:k*m,2]=A[1:m,1];
  B[1+(k-1)*m:k*m,3]=A[1:m,ncol(A)];
  B[1+(k-1)*m:k*m,4]=A[1:m,k+1];
end;
create dhps2 from B;
edit dhps2;
append from B;
quit;
```

In the next step, we delete all rows with missing entries, because they are no longer needed. The outcome is illustrated in Figure 4.

```
data dhps2;
set dhps2;
if Col4=. then delete;
run;
```



VIEWTABLE: Work.Dhps2

	COL1	COL2	COL3	COL4
1	1	0	1	312
2	1	0	1	303
3	1	0	2	310
4	1	0	2	324
5	1	0	3	322
6	1	0	4	305
7	1	0	5	328
8	1	1	6	318
9	1	1	7	316
10	1	0	8	322
11	1	1	9	308
12	1	0	10	307
13	1	0	11	314
14	1	1	12	309
15	1	1	13	310
16	1	0	14	318
17	1	1	15	312
18	1	0	16	326
19	1	0	17	318
20	1	0	18	316
21	1	0	19	307
22	1	1	20	312
23	1	0	22	314
24	1	1	23	337
25	1	0	24	305
26	1	1	25	310
27	1	0	26	295
28	1	0	26	314
29	1	1	27	310
30	1	1	28	316
31	1	1	28	305
32	1	1	28	310
33	1	0	29	310
34	1	1	30	310

VIEWTABLE: Work.Dhps2

	COL1	COL2	COL3	COL4
1	1	0	0.5	312
2	1	0	0.5	303
3	1	0	0.5	310
4	1	0	0.5	324
5	1	0	1	322
6	1	0	1	305
7	1	0	1	328
8	1	1	1	318
9	1	1	1	316
10	1	0	1	322
11	1	1	1	308
12	1	0	1	307
13	1	0	1	314
14	1	1	1	309
15	1	1	1	310
16	1	0	1	318
17	1	1	1	312
18	1	0	1	326
19	1	0	1	318
20	1	0	1	316
21	1	0	1	307
22	1	1	1	312
23	1	0	1	314
24	1	1	1	337
25	1	0	1	305
26	1	1	1	310
27	1	0	0.5	295
28	1	0	0.5	314
29	1	1	1	310
30	1	1	0.3333333333	316
31	1	1	0.3333333333	305
32	1	1	0.3333333333	310
33	1	0	1	318
34	1	1	1	310

VIEWTABLE: Work.Dhpsfreq

	COL2	COL1	COL4	COL4	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	0	1	295	295	10.5	11.17	10.5	11.17
2	0	1	299	299	1	1.06	11.5	12.23
3	0	1	301	301	1	1.06	12.5	13.30
4	0	1	303	303	3.333333	3.55	15.83333	16.84
5	0	1	305	305	9.5	10.11	25.33333	26.95
6	0	1	307	307	8	8.51	33.33333	35.46
7	0	1	308	308	1.5	1.60	34.83333	37.06
8	0	1	309	309	2.5	2.66	37.33333	39.72
9	0	1	310	310	5.5	5.85	42.83333	45.57
10	0	1	312	312	7.833333	8.33	50.66667	53.90
11	0	1	314	314	7	7.45	57.66667	61.35
12	0	1	316	316	11.5	12.23	69.16667	73.58
13	0	1	318	318	7.5	7.98	76.66667	81.56
14	0	1	320	320	2.5	2.66	79.16667	84.22
15	0	1	322	322	4.833333	5.14	84	89.36
16	0	1	324	324	2.5	2.66	86.5	92.02
17	0	1	326	326	2	2.13	88.5	94.15
18	0	1	327	327	0.5	0.53	89	94.68
19	0	1	328	328	1	1.06	90	95.74
20	0	1	330	330	1.5	1.60	91.5	97.34
21	0	1	332	332	1.5	1.60	93	98.94
22	0	1	335	335	1	1.06	94	100.00
23	0	2	273	273	2.833333	3.01	2.833333	3.01
24	0	2	274	274	1	1.06	3.833333	4.08
25	0	2	275	275	5.166667	5.50	9	9.57
26	0	2	277	277	3.5	3.72	12.5	13.30
27	0	2	279	279	13.5	14.36	26	27.66
28	0	2	281	281	3.5	3.72	29.5	31.38
29	0	2	282	282	1	1.06	30.5	32.45
30	0	2	283	283	16.16667	17.20	46.66667	49.65
31	0	2	284	284	0.5	0.53	47.16667	50.18
32	0	2	285	285	4.833333	5.14	52	55.32
33	0	2	286	286	23.83333	25.35	75.83333	80.67

Figure 4. Left: Screenshot of dataset 'sasgf.dhps2' after performing the SAS/IML® procedure and the data step in step 2. The labels of the blood samples are in COL3. Middle: Screenshot of dataset 'sasgf.dhps2' after performing the SAS/IML® procedure in step 3. COL3 was replaced by the weights. Right: Screenshot of dataset 'dhpsfreq'.

Step3: Adding weights

In order to correctly calculate the heterozygosity we need to weight the data (see above). Again, SAS/IML® serves to complete this task. In dataset 'dhps2', we replace the column 3, which contains the labels, by a column containing the weights. The column containing the labels is necessary to calculate the weights but not necessary to calculate the heterozygosity once the weights are known. We use the following code:

```
proc iml;
use dhps2;
read all into A;
close dhps2;
call delete(dhps2);      /* deletes file 'work.dhps2' */
B=J(nrow(A),ncol(A),.);
B[,1:ncol(A)]=A[,];
BB=J(nrow(A),1,1);
inc=1;
do k=2 to nrow(A);
  if A[k,ncol(A)-1]=A[k-1,ncol(A)-1] then do;
    inc=inc+1;
  end;
  else do;
    if inc>1 then do;
       BB[k-inc:k-1,1]=J(inc,1,inc);
       inc=1;
    end;
  end;
end;
B[,ncol(A)-1]=1/BB;
create dhps2 from B; /* creates new file 'work.dhps2' from matrix B */
edit dhps2;
append from B;
quit;
```

Step4: Calculating the allele frequencies

We need to calculate the alleles frequencies at the various MS loci to be able to derive the heterozygosity. From the dataset 'dhps2', it is easy to calculate the relative frequencies of the various alleles at each MS locus among the samples containing resistant and sensitive parasites, respectively. We use 'proc freq' to perform a table analysis grouped by the MS loci and sensitive/resistant samples. We further export the output of the table analysis into a SAS® dataset using SAS/ODS®. This dataset contains the desired frequencies to calculate the heterozygosity.

We use the following SAS® code:

```
proc sort data=work.dhps2 out=work.dhps2; /*sorts the data in a proper way*/
 by COL2 COL1;
run;

ods listing close;
proc freq data=work.dhps2;
by COL2 COL1;
    tables COL4;
    weight COL3;
    ods output    OneWayFreqs=dhpsfreq;
run;
ods listing;

data work.dhpsfreq;
set  work.dhpsfreq;
drop Table;  /* deletes the unnecessary variable 'Table' */
run;
```

The newly created dataset 'work.dhpsfreq' contains the allele frequencies at the respective MS loci.

Step 5: Calculating the heterozygosity

Now, we calculate the heterozygosity at each MS locus grouped by sensitive/resistant. We use SAS/IML® to create a dataset in which the first column specifies sensitive/resistant, the second column the MS locus, and the third column the value of the heterozygosity at the respective MS locus. We use the following code.

```
proc iml;
call delete(dhps2); /* deletes the now unnecessary dataset 'dhps2' */
use dhpsfreq;
read all into A;
B=A[1:nrow(a)-1,2]-A[2:nrow(a),2];
index=J(1,1,1);
ind=1;
do k=1 to nrow(B);
  if B[k,1]^=0 then do;
     ind= k+1;
     index=index//ind;
  end;
end;
ind=nrow(A)+1;
index=index//ind;

A[,5]=(A[,5]/100);
n=nrow(index);
C=J(n-1,3,.);
do k=1 to n-1;
   vec=A[index[k,1]:index[k+1,1]-1,5];
   vec1=vec##2;
   m1=index[k+1,1]- index[k,1];
   if m1=1 then do;
      He=0;
   end;
   else do;
```

```
        He=(m1/(m1-1))*(1- vec1[+,]);
    end;
    C[k,1:2]=A[index[k,1],1:2];
    C[k,3]=He;
end;
create Het from C;
edit Het;
append from C;
quit;
```

Figure 5 shows a screenshot of the dataset 'work.het'.

Figure 5. Screenshot of dataset 'work.het'.

Step6: Plotting the results

Finally, we can plot the dataset using 'proc gplot'. First, we specify a new variable that will be used for the plot legend and specify several goptions.

```
data het;
label   col='a0'x;
set  sasgf.het;
if col1=0 then col="sensitive";
if col1=1 then col="resistant";
run;

axis1 value=( height=1.5  angle=-30 rotate=-0
"-72.7"  "-34.5"  "-18.7"  "-11"  "-7.4" "-2.8" "-1.5" "-0.132" "0.034" "0.5" "1.4"
"6.4" "9" "16.3" "22.8" "36" "49.5" "66.1")
     label=(angle=0 h=2 color=black "distance from dhps in kb" )
     major=(number=18 height=.1 cm  )
     minor=( );

axis2 value=(height=1.5) label=(angle=90 h=2 color=black "Heterozygosity" )
      major=(height=.1 cm)
      minor=( );

LEGEND1 value=(height=1.5);
```

```
symbol1 i =std1mt mode=include interpol=join
        value=dot
        ;
symbol2 interpol=join
        value=dot
        ;
```

Now, we can plot the heterozygosity. This is done with the following code. Figure 6 shows the plot.

```
proc gplot data=het;
   plot col3*col2=col / haxis=axis1  vaxis=axis2 Legend=Legend1;
run;
```

It is apparent from the plot that the heterozygosity among resistant parasites is reduced around the *dhps* gene compared with the heterozygosity among sensitive parasites. This is a pattern of a selective sweep. It indicates that selection (induced by drug pressure) is acting to increase the frequency of resistant parasites.

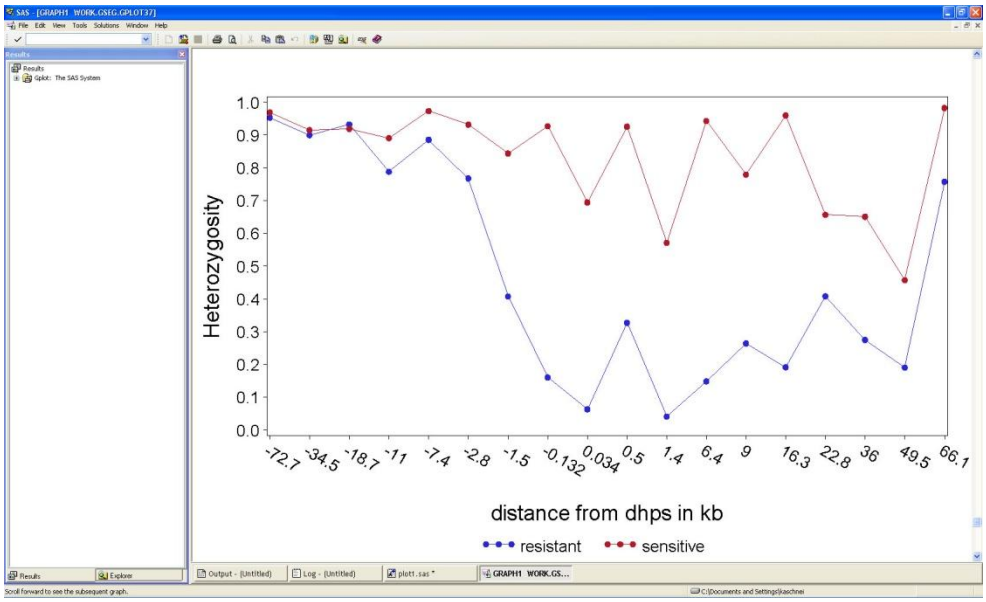


Figure 6. Screenshot of the gplot.

CONCLUSION

We provided a simple example how SAS® can be used to explore genetic data. We provided a rather simple and basic example for a genetic analysis. However, it gives some ideas how various SAS® components can be combined to perform genetic analyses. Of course, our example can be extended to perform more comprehensive analyses of genetic data.

ACKNOWLEDGMENTS

This work was partly funded by the National Institute of Health grant R01GM084320.

RECOMMENDED READING

- WHO, 2000. WHO Expert Committee on Malaria. World Health Organ Tech Rep Ser. 892, pp. 1_74. URL: http://www.genetics.org/cgi/content/abstract/160/2/765
- Snow, R.W., Guerra, C.A., Noor, A.M., Myint, H.Y., Hay, S.I., 2005. The global distribution of clinical episodes of Plasmodium Falciparum malaria. Nature 434 (7030), 214_217. URL: http://dx.doi.org/10.1038/nature03342
- Daily, J.P., 2006. Antimalarial drug therapy: the role of parasite biology and drug resistance. Journal of Clinical Pharmacology 46 (12), 1487_1497. URL: http://jcp.sagepub.com/cgi/content/abstract/46/12/1487.

- Marsh K (1998) Malaria disaster in Africa. Lancet 352(9132):924–924. URL: http://www.sciencedirect.com/science/article/B6T1B-4FWV357-JX/2/c9facab67b868b8fddd289f050bfae19
- Maynard Smith, J., Haigh, J., 1974. The hitch-hiking effect of a favourable gene. Genetics Research 23 (01), 23_35. URL: http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=1754360&fulltextType=RA&fileId=S0016672300014634
- Stephan, W., Wiehe, T.H.E., Lenz, M.W., 1992. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theoretical Population Biology 41 (2), 237_254. URL: http://www.sciencedirect.com/science/article/B6WXD-4F1Y9N0-3M/2/1245281bba0c6b542457fdd75c343edf
- Barton, N.H., 2000. Genetic hitchhiking. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 355 (1403), 1553_1562. URL: http://rstb.royalsocietypublishing.org/content/355/1403/1553.abstract
- Kristan A. Schneider and Yuseob Kim (2010). An Analytical Model for Genetic Hitchhiking in the Evolution of Antimalarial-Drug Resistance. Theor. Popul. Biol. 78(2), 93-108. URL: http://linkinghub.elsevier.com/retrieve/pii/S0040580910000626
- SAS® Language: Reference, Version 9.2
- Getting Started with the SAS® System, Version 9.2
- User's Guide, Version 9.2,

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Kristan Schneider
Enterprise: Department of Mathematics, University of Vienna
Address: Nordbergstr. 15, UZA 4
City, State ZIP: A-1090, Vienna Austria
Work Phone: +43 1 4277 507 74
Fax: +43 1 4277 506 20
E-mail: kristan.schneider@univie.ac.at
Web: http://homepage.univie.ac.at/kristan.schneider/

Or

Name: Kristan Schneider
Enterprise: CEMI/Biodesign Institute at Arizona state University
Address: 1001 S. McAllister Ave.
PO Box 875001,
City, State ZIP: Tempe, AZ 85287-5001
Work Phone: +1 (480) 965-993
Fax: +1 (480) 727-6947
E-mail: kristan.schneider@asu.edu
Web: http://www.public.asu.edu/~kaschnei/

Name: Barbara Schneider
Enterprise: Center for Medical Statistics, Informatics and Intelligent Systems
Section for Medical Statistics
Address: Spitalgasse 23
City, State ZIP: A-1090, Vienna Austria
Work Phone: +43 1 40400 7479
Fax: +43 1 40400 7477
E-mail: barbara.schneider@meduniwien.ac.at
Web: http://homepage.univie.ac.at/barbara.schneider/