

Utilize SAS®9 SGPLOT to Create Genome Wide Association Studies Plots

Huei-Ling Chen, Merck & Co., Inc., Kenilworth, NJ USA
Jialin Xu, Merck & Co., Inc., Kenilworth, NJ USA

ABSTRACT

In clinical studies, there is an increasing research interest focusing on the associations between genetic variants and quantitative or categorical traits of clinical interest, including disease status, treatment response to a specific medication, especially as the cost of the molecular technology drops dramatically. For example, one may examine whether cancer patients with a certain genetic variant can respond better to a particular oncology drug compared to cancer patients without that type of genetic variant. A major methodology to test that genetic association is by Genome Wide Association Study (GWAS) Analysis. The GWAS analysis consist three main plots to do the research step by step: Q-Q Plot, Manhattan Plot, and Regional Association Plot. The Q-Q Plot is to justify the distribution of the observed p-value for each genetic variant. The Manhattan Plot and the Regional Association Plot is to identify which particular region in the genome is associated with the clinical interest. This paper introduces the plots and utilizes SAS® version 9 SGPLOT procedure to develop three macros for each of the three GWAS plots respectively.

KEYWORDS

GWAS, Q-Q PLOT, MANHATTAN PLOT, REGIONAL ASSOCIATION PLOT, SGPLOT

INTRODUCTION

The GWAS Analysis focuses on the association between individuals' genetic variants, usually single-nucleotide polymorphisms (SNPs), with quantitative or categorical traits of clinical interest such as disease status, treatment response to the drug. The statistical significance of such association can be represented by the p-value. The p-value indicates the significance level of the hypothesis testing that a certain SNP has impact to a trait or an endpoint, e.g. respond to drug. A typical GWAS study usually involves univariate test on around 10 million SNPs. Each SNP will be assigned a p-value from its corresponding univariate test.

The first task in GWAS is to draw a Q-Q plot. The Q-Q plot is a common plot to compare two probability distributions. Each dot in the Q-Q plot represents the observed and expected p-value for a SNP from univariate test. The Q-Q plot in GWAS here is to compare the observed p-value to the expected p-value from the assumed distribution. This is to ensure whether there is a shift of the observed p-value from the expected p-value. In an ideal scenario, if the assumed distribution holds, all dots should be on the 45 degree diagonal line. A major departure from the diagonal line usually suggests confounding factors need to be added into the model.

After justify the distribution of the observed p-values, the next task in GWAS is to display the p-values based on their genome locations in each of 22 chromosomes and relative distance. The Manhattan plot helps researcher visualize clearly whether any SNPs has an peak like significant p-value passing a pre-specified threshold. It also helps researcher to visualize the adjacent regions with significant or close to significant SNPs. The identification of such SNPs is a key to GWAS.

The Regional Association Plot is also called 'Zoom-in' Manhattan Plot which is to zoom-in to the interested SNP neighborhood and further examine other SNPs in that area. This is usually triggered by a SNP with significant p-value. It helps researcher to identify the genome location and distance of adjacent significant or close to significant SNPs in details.

Q-Q PLOT

A Q-Q Plot is for comparing observed p-values and expected p-values by plotting their quantiles against each other. In addition, pre-specified p-value cutoffs may be plotted as horizontal lines for reference.

The expected p-values are derived from the observed p-values. Take the $[-\log(10)]$ form of the observed p-values. Take the $[-\log(10)]$ form of the corresponding expected p-values. Compare both values by plotting the observed values at the X axis and corresponding expected values at the Y axis.

The dataset for a Q-Q plot has one required variable: the observed P value.

Obs	pval
1	0.0004
2	0.1420
3	0.1827
4	0.1931
5	0.2528
6	0.0829

7	0.0843
8	0.1286
9	0.1314
10	0.1372
11	0.1372
12	0.2547
13	0.2560
14	0.2580
15	0.2582

Key Syntax – prepare the dataset for the plot

Step 1: Count the number of the observed p-value being examined.

```
proc sql noprint;
    select count(*) into :tot
    from pvector
    ;
quit;
```

Step 2: Sort observed p-values from smallest to largest.

```
proc sort data=pvector;
    by pval;
run;
```

Step 3: Match each observed p-value with an expected p-value, which is at the even gap position from 0 for the smallest observed p-value to 1 for the largest observed p-value.

Take the $[-\log(10)]$ form of the p-values.

```
data dataout;
    set pvector;
    retain x_ ;

    gap=1/&tot;
    ini=gap/2;

    if _n_=1 then x_=ini;
    else x_=x_ + gap;

    x = -log10(x_);
    y = -log10(pval);
run;
```

Dataset to feed into the plot

Obs	pval	y	x_	x
1	0.0004	3.39794	0.03333	1.47712
2	0.0829	1.08145	0.10000	1.00000
3	0.0843	1.07417	0.16667	0.77815
4	0.1286	0.89076	0.23333	0.63202
5	0.1314	0.88140	0.30000	0.52288
6	0.1372	0.86265	0.36667	0.43573
7	0.1372	0.86265	0.43333	0.36318
8	0.1420	0.84771	0.50000	0.30103
9	0.1827	0.73826	0.56667	0.24667
10	0.1931	0.71422	0.63333	0.19837
11	0.2528	0.59722	0.70000	0.15490
12	0.2547	0.59397	0.76667	0.11539
13	0.2560	0.59176	0.83333	0.07918
14	0.2580	0.58838	0.90000	0.04576
15	0.2582	0.58804	0.96667	0.01472

Key Syntax – PROC SGPLOT to create plot

This paper uses the PROC SGPLOT procedure SCATTER statement to plot the X axis value corresponding to the Y axis value. In addition, users get to decide which reference lines are placed in the plot. The color pattern for each horizontal reference line is also user's choice.

See Figure 1 below.

- A diagonal reference line $Y=X$
- Multiple horizontal reference lines, $Y = -\log_{10}(pval)$

```
proc sgplot data=dataout noautolegend sganno=anno;
  scatter x=x y=y / MARKERATTRS = (size=3 );
  xaxis label = " " values=(0 to &max by 1) ;
  yaxis label = ' ' values=(0 to &max by 1) ;

  lineparm x=0 y=0 slope=1 /lineattrs=(color=red);
  %do e=1 %to &c %by 1;
    refline &&ref&e
      /axis=y
      label=" "
      Lineattrs = (color=%sysfunc(scan(&refcolor,&e)));
  %end;
```

Diagonal Reference
Line: $Y=X$

Multiple horizontal
reference lines
specified with
designated colors

The Q-Q plot output ---

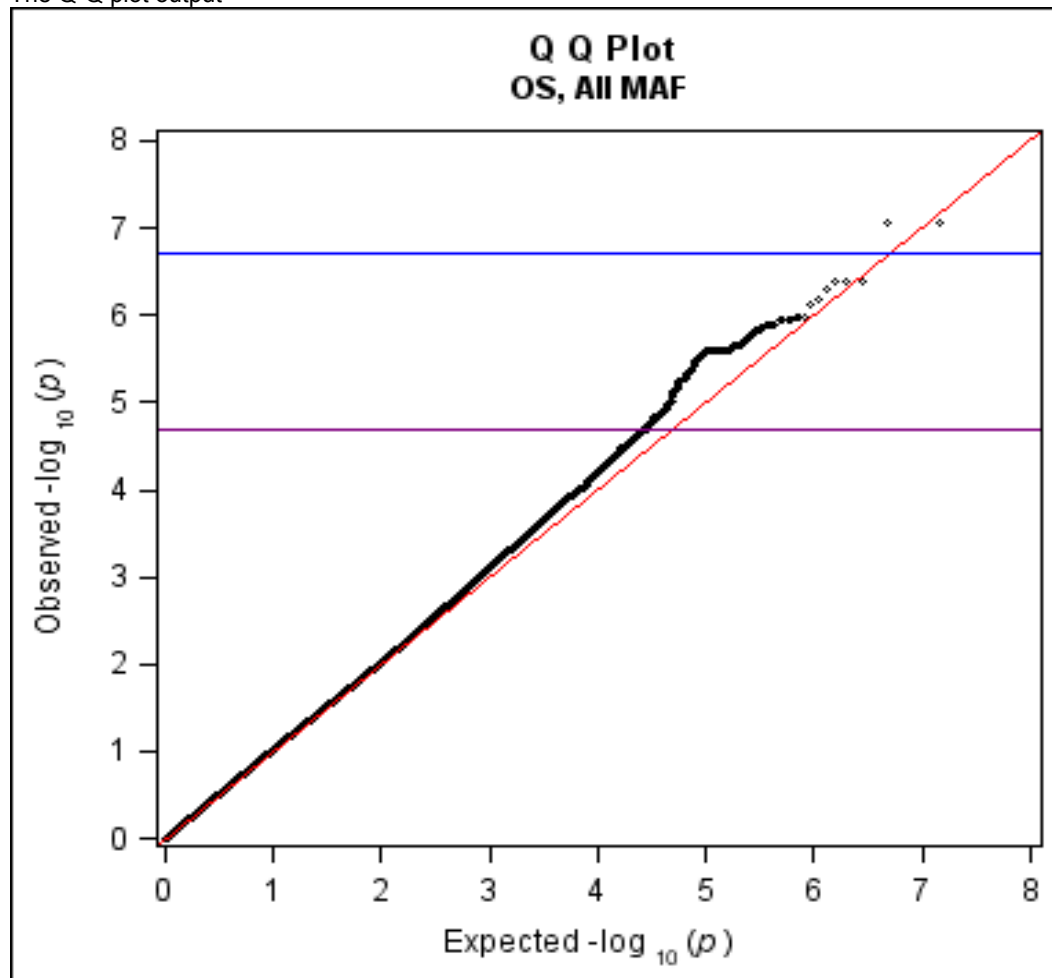


Figure 1: Q-Q Plot

MANHATTAN PLOT

A Manhattan Plot is a scatter plot with the Y axis for the negative logarithm form of the p-values for each single nucleotide polymorphism (SNP), and the X axis for the genomic positions in each chromosome. The lower the p-value, the higher the negative logarithm of the p-value. A relatively extreme low p-value is like a skyscraper above the lower level 'buildings'. The SNPs can be classified into 22 chromosomes.

The dataset for a Manhattan plot has three required variables: chromosome, P-value, and genomic position.

Input Dataset

chromosome	position	pvalue
13	94326747	0.029941
13	94341096	0.630603
1	39542129	0.509956
13	94364399	0.157736
13	94383970	0.620993
13	94389413	0.667778
13	94391588	0.002061
13	94398942	0.332334
...		

Take the $[-\log(10)]$ form of the p-values.

Take the genomic position (gene sequence) for each record. Create the associated X axis position accordingly. Based on the Sanger Institute's human genome information in the Vertebrate Genome Annotation (VEGA) database, the total base of each chromosome is below:

Chromosome	Total bases
1	247,199,719
2	242,751,149
3	199,446,827
4	191,263,063
5	180,837,866
6	170,896,993
7	158,821,424
8	146,274,826
9	140,442,298
10	135,374,737
11	134,452,384
12	132,289,534
13	114,127,980
14	106,360,585
15	100,338,915
16	88,822,254
17	78,654,742
18	76,117,153
19	63,806,651
20	62,435,965

21	46,944,323
22	49,528,953

Define the X axis position based on a cumulative sum from the previous chromosomes total base. In case the latest position from the data is larger than the total base, the cumulative sum takes the latest position from the data. For example, according to the above table, chromosome 1 total base is 247199719 and chromosome 2 total base is 242751149. In the data, the latest chromosome 1 position is 248451965, which is larger than the total base 247199719. Take the latest position rather than the total bases. One chromosome 3 record with position 135627433 then has x axis position 626830547 (sum of 135627433 and chromosome 1 latest position 248451965 and chromosome 2 total bases 242751149).

Key Syntax – prepare the dataset for the plot

Step 1: Based on the Sanger Institute's human genome information in the Vertebrate Genome Annotation (VEGA) database, create a dataset with the total base of each chromosome.

```
data chr_ ;
  c=1; pos_ =247199719; output;
  c=2; pos_ =242751149; output;
  c=3; pos_ =199446827; output;
  ...
  c=22; pos_ =49528953; output;
run;
```

Step 2: Find the last position for each chromosome in the dataset.

```
proc summary data=input nway;
  class c ;
  var pos ;
  output out=chr_m max=;
run;
```

Step 3: Define the X axis position based on a cumulative sum from the maximum value of the previous chromosomes total base vs. the latest position in the dataset.

```
data chr_max;
  merge chr_ chr_m;
  by c;
  max=max(pos, pos_);
  maxpos_ + max;
  maxpos = lag (maxpos_);
run;

data manhattan ;
  merge input chr_max;
  by c;
  logp = -log10(p) ;
  if c = 1 then x = pos ;
  if c > 1 then x = pos + maxpos ;
  x = x / 10000;
run;
```

Define the cumulative sum from the previous chromosomes total bases.

Define the X axis position.

Key Syntax – PROC SGPLOT to create plot

In the Manhattan plot, there are 22 chromosomes specified by different colors. The color pattern for each chromosome is the user's choice – a user defined macro parameter.

This paper uses PROC TEMPLATE to generate symbol color for plots with alternating colors specified by macro parameter &ColorPattern.

```

%let m=%eval(%sysfunc(countw(&ColorPattern,' ')));
%let n=%eval(%sysfunc(countw(&ColorPattern,' '))-1);

proc template;
  define style styles.mystyle;
    parent=styles.listing;
  %do j=1 %to 12 %by &m;
    style GraphData%eval(&j) from GraphData%eval(&j) /
      contrastcolor=%sysfunc(scan(&ColorPattern,1))
      linestyle=1;
    %do k=1 %to &n %by 1;
      style GraphData%eval(&j+&k) from GraphData%eval(&j+&k) /
        contrastcolor=%sysfunc(scan(&ColorPattern,%eval(&k+1)))
        linestyle=1;
    %end;
  %end;
end;
run;

```

The color pattern is usually specified by the user to differentiate different chromosomes. The GROUP option in the SCATTER statement would search the STYLE defined in the PROC TEMPLATE procedure. This is how the user can decide the color for each chromosome.

Multiple horizontal reference lines $Y = -\log(10)$ form of a user defined macro parameter `&&REF&E` are also added in the code. These reference lines in the Manhattan plot helps researchers visualize clearly whether any SNPs has a peak like significant p-value passing a pre-specified threshold.

```

ODS GRAPHICS on / RESET=index IMAGENAME = "&outfile" IMAGEFMT =PNG
width=1600px height=1000px ;
ODS LISTING GPATH = "&outfilepath" style=styles.mystyle;

proc sgplot data=manhattan noautolegend sganno=anno
  pad=(bottom=15% top=10%);
  scatter x=x y=logp / group = c
    MARKERATTRS = (size = &symbolsize
      symbol = circlefilled);
  axis display=(nolabel novalues noticks) offsetmin=0.02 offsetmax=0;
  yaxis label = '-Log10(p)' values=(&YValues.) ;

  %do e=1 %to &c %by 1;
    refline &&ref&e /axis=y
      label=""
      lineattrs=(color=%sysfunc(scan(&refcolor,&e)));
  %end;
run;

```

The “**group = c**” specified different colors for different chromosome

Multiple horizontal reference lines specified with designated colors

The Manhattan plot output ---

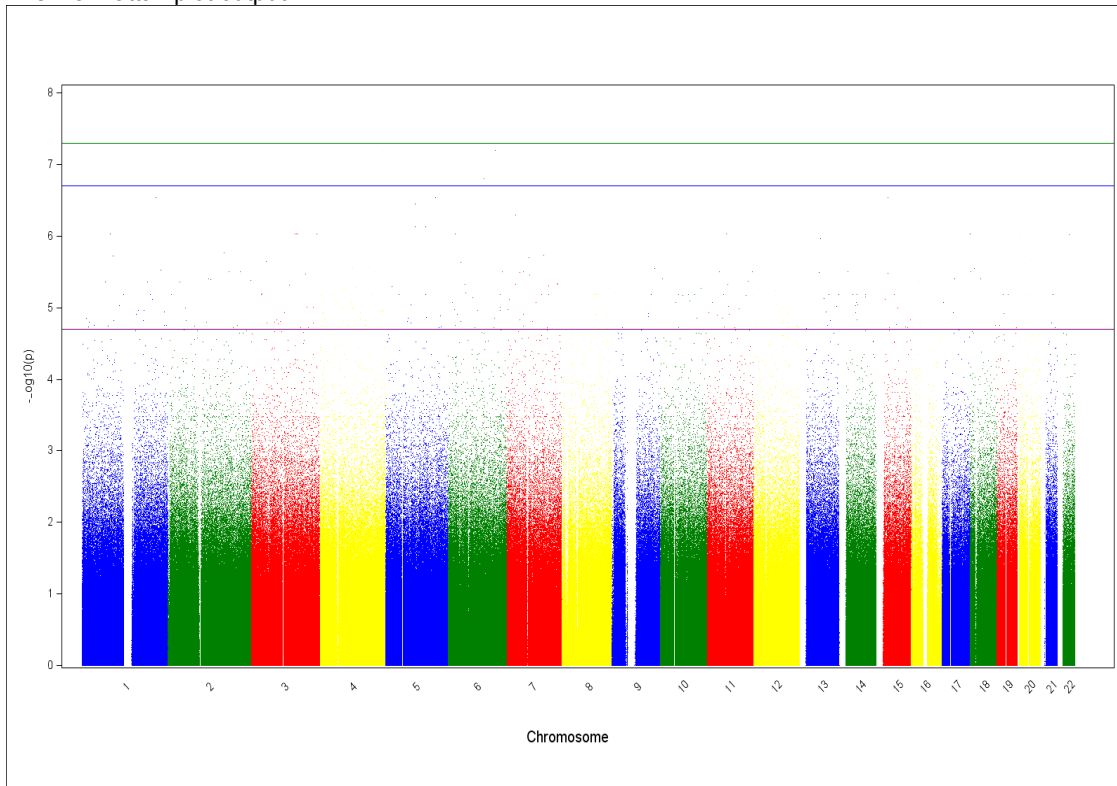


Figure 2: Manhattan Plot (Blue Green Red Yellow)

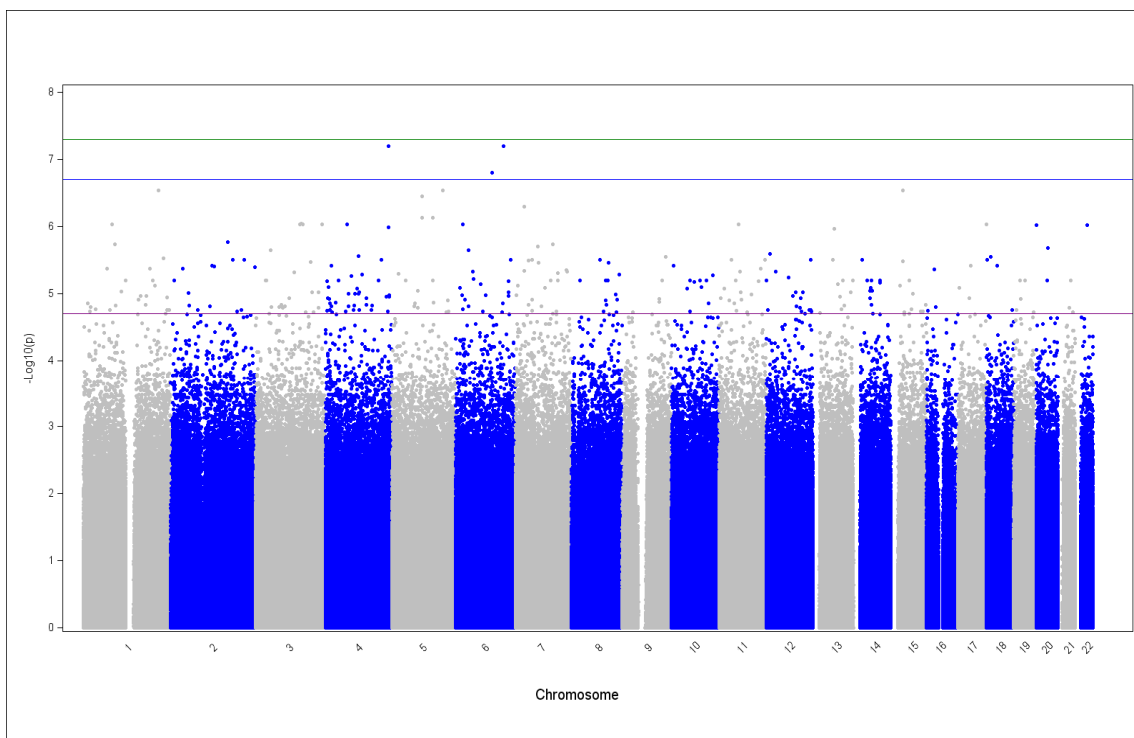


Figure 3: Manhattan Plot (Grey Blue)

REGIONAL ASSOCIATION PLOT

The main focus of the Regional Association Plot is to highlight the selected single nucleotide polymorphism (SNP) with statistical significance of an association to the clinical interests, for example, disease or clinical study endpoints. The plot displays the association results, p-value, for that highlighted SNP as well as the surrounding markers. Y axis is the negative logarithm of the p-values for each (SNP). X axis is to coordinate the genomic position. This program also provides capability to distinguish assayed SNPs from imputed SNPs using separate symbols.

This macro reads in a SAS dataset in which there are five required variables: chromosome, P-value, genomic position, SNP variant name, and a flag variable to indicate whether a SNP is "genotyped" or "imputed".

chromosome	pvalue	pos	RSID	Assay
13	0.3995307823	94326747	Affx.10000032_T	FALSE
13	0.4909448027	94341096	Affx.10000241_A	FALSE
1	0.931230645	39542129	Affx.10000335_A	FALSE
13	0.7430195079	94364399	Affx.10000479_G	FALSE
13	0.259844276	94383970	Affx.10000731_A	FALSE
13	0.6828772157	94389413	Affx.10000805_C	FALSE
13	0.2870648023	94391588	Affx.10000833_A	FALSE
13	0.7183251962	94398942	Affx.10000932_A	FALSE
13	0.1502206025	94408256	Affx.10001050_G	FALSE
13	0.494075671	94408506	Affx.10001052_A	FALSE
13	0.1850166333	94409637	Affx.10001066_G	FALSE
1	0.5344962921	39549146	Affx.10001206_G	FALSE

Users decide two criteria: statistic significant level for the p-value and the neighbor window for the surrounding variants. With the user-defined criteria, first identify the SNPs which meet the user-defined statistic significant association criteria. Next plot the identified SNPs along with the neighboring biomarkers within the user-defined window positions.

Key Syntax – prepare the dataset for the plot

Step 1: Filter the data to identify the SNPs which meet the user-defined statistic significant association criteria &peak. Count the number of SNPs which will be generated into multiple individual Regional Association Plots.

```
data datain1;
    set datain;
    where .< pvalue < &peak ;
run;

proc sql noprint;
    select count(*) into :tot
    from datain1
    ;
quit;

%let tot=&tot;
```

Step 2: For each identified SNP, retrieve the neighboring biomarkers within the user-defined window positions &window. The macro allows the user to decide the symbol color, shape, and size for the peak SNP, Assayed SNPs, and Imputed SNPs. Notice that these symbols are represented by different variables. Code below generates the negative logarithm form of the p-value (Y) into different variables, Y1, Y2, and Y3.

```

%do i=1 %to &tot;
  data zoomin;
    set datain ;
    if chromosome=%eval(&&chr&i) and abs(pos-
%eval(&&pos&i))<=%window;
    y= -log10(pvalue);
    x= pos/1e3;
    if put(y,8.2)=&&y&i and &pos=%eval(&&pos&i) then y1=y;
    if &assay='TRUE' and y1 eq . then y2=y;
    if &assay='FALSE' and y1 eq . then y3=y;

    label y1="Peak SNP"
          y2="Assayed SNPs"
          y3="Imputed SNPs";

  run;

proc sort data=zoomin;
  by x;
run;

...
%end;

```

Each Symbol
has its own
variable.

Dataset to feed into the plot

chr	RSID	p_w	pos	x	y1	y2	y3	y
15	rs12148391_A	0.0002252245	84145428	84145.43	.	.	3.64738	3.64738
15	rs11629958_G	0.0482821047	84147713	84147.71	.	.	1.31621	1.31621
15	rs17511735_C	0.0477772824	84149712	84149.71	.	.	1.32078	1.32078
15	rs1844859_A	0.3555077003	84149756	84149.76	.	.	0.44915	0.44915
15	rs28552876_A	0.5099712393	84149803	84149.80	.	.	0.29245	0.29245
15	rs78057470_C	0.0569138511	84149984	84149.98	.	.	1.24478	1.24478
15	rs1844861_T	0.0569138511	84149985	84149.99	.	.	1.24478	1.24478
15	rs12442717_G	0.1460816455	84150322	84150.32	.	.	0.83540	0.83540
15	rs79042299_G	5.1689992E-7	84150408	84150.41	6.28659	.	.	6.28659
15	rs12902523_C	0.0548680959	84151258	84151.26	.	.	1.26068	1.26068
15	Affx.12078565_G	0.716539291	84159181	84159.18	.	.	0.14476	0.14476
15	rs77409533_G	0.7570940889	84159181	84159.18	.	0.12085	.	0.12085
15	rs2036015_T	0.3746458166	84160927	84160.93	.	.	0.42638	0.42638
15	rs60711848_T	0.5050423032	84161002	84161.00	.	.	0.29667	0.29667
15	Affx.12078590_G	0.0320738065	84161111	84161.11	.	.	1.49385	1.49385
15	rs7166783_G	0.0330470099	84161111	84161.11	.	1.48087	.	1.48087

Key Syntax – PROC SGPLOT to create plot

In the Manhattan plot, there are 22 chromosomes specified by different colors. The color pattern for each chromosome is user's choice – a user defined macro parameter.

This paper uses PROC TEMPLATE to generate symbol color for plots with alternating colors specified by macro parameter &ColorPattern.

```

proc sgplot data=zoomin noautolegend pad=(bottom=15% top=10%);

scatter x=x y=y1 /MARKERATTRS=(symbol=&SymbolPeak
                                color=&SymbolPeakColor
                                size=&SymbolPeakSize) name='sca1';

scatter x=x y=y2 /MARKERATTRS=(symbol=&SymbolAssayed
                                color=&SymbolAssayedCoor
                                size=&SymbolAssayedSize) name='sca2';

scatter x=x y=y3 /MARKERATTRS=(symbol=&SymbolImputed

```

```

                                color=&SymbolImputedColor
                                size=&SymbolImputedSize) name='sca3';
series x=x y=y / LINEATTRS = (pattern = shortdash);

xaxis display=(nolabel ) offsetmin=0.1 offsetmax=0.1;

yaxis label = '-Log10(p)' values=(&YValues.) ;

keylegend "sca1" "sca2" "sca3" / location = inside position=topleft across=1 noborder ;

inset "Window size=&wsize kb" / position=topright noborder textattrs = (size=12) ;

%do e=1 %to &c %by 1;
    refline &&ref&e /axis=y
                label=""
                lineattrs=(color=%sysfunc(scan(&refcolor,&e)));
%end;

run;

```

Regional Association Plot ---

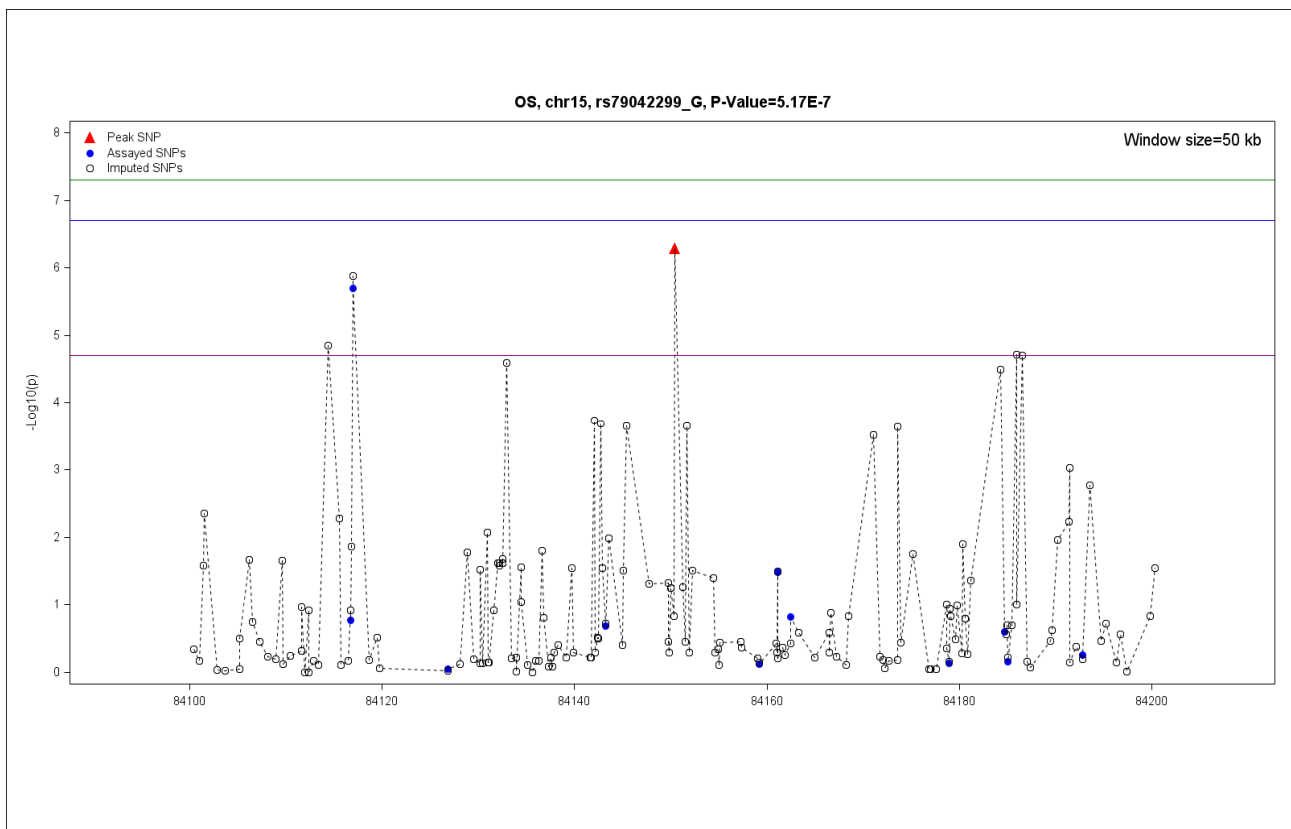


Figure 4: Regional Association Plot

CONCLUSION

Statisticians / researchers usually use R code to do the GWAS analysis. There are user-developed functions in an open share website for people to download the R code and use it to generate Q-Q plots, Manhattan Plots, and Regional Association Plots. However, without viewing the complete syntax of these R functions, when entering wrong parameter values, it could cause a risk and take the function to an unexpected application.

This paper developed three SAS macros for these GWAS plots. The details of the deriving logic can be seen and reviewed. The SAS output plots also have been compared to the plots from R code.

REFERENCES

https://en.wikipedia.org/wiki/Genome-wide_association_study
http://www.sascommunity.org/wiki/Creating_Manhattan_Plots_with_SAS/Graph

ACKNOWLEDGMENTS

The authors would like to thank the management team of Merck Research Laboratories for their advice on this paper/presentation.

Contact Information

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

Huei-Ling Chen
 Merck & Co., Inc.
 126 Lincoln Avenue
 P.O. Box 2000
 Rahway, NJ 07065
 Phone: 732-594-2249
 e-mail: Huei-Ling_Chen@merck.com

Jialin Xu
 c/o Merck & Co., Inc.
 MAILSTOP UG1C-36
 351 North Sumneytown Pike
 North Wales, PA 19454
 Phone: 267-305-8119
 e-mail: jjalin.xu@merck.com

TRADEMARK

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.