

Manhattan: ggplot2-based Manhattan plots

Boxiang Liu

2017-12-15

Contents

Introduction	1
Installation	1
Usage	2
Basic Manhattan plot	2
Customizing colors	3
Highlight and label SNPs and genes	4
Adding GWAS significance line	6
Plotting multiple GWAS studies	7
Details	8
Questions and Bugs	9

Introduction

The Manhattan plot is a specialized form of scatterplot to display genome-wide association studies (GWAS). The x-axis of a Manhattan plot is the genomic position, and the y-axis is usually the $-\log_{10}(P\text{-value})$ (although other sensible metric can be used as well).

There are many packages for making Manhattan plots, but most of them are not easily extensible. The package `ggplot2` has become increasingly popular among the **R** and bioinformatics communities. Therefore, a package that is fully compatible with `ggplot2` and customizable with its `geoms` will facilitate the use of Manhattan plots.

Introducing `manhattan`, a `ggplot2` based package for making Manhattan plots. It takes in a `data.frame` and returns a standard `ggplot` object, upon which the user can add other `geoms`. This makes it very convenient for users to build on `manhattan` and customize their plots.

Installation

To install `manhattan`, use the standard R package installation command.

```
# install.packages('manhattan')
```

If you want the latest development version, install it using `devtools`.

```
devtools::install_github("boxiangliu/manhattan")
```

The package has been tested on Linux and Mac OSX. It has not been tested on Windows.

Usage

Basic Manhattan plot

To illustrate its usage, let us plot the coronary artery disease GWAS based on *Deloukas et al*(2013). The original dataset provides nominal p-values. Since we want to plot the $-\log_{10}(P\text{-value})$, let us take the logarithm.

```
library(manhattan)
data(cad_gwas)
cad_gwas$y=-log10(cad_gwas$pval)
head(cad_gwas)

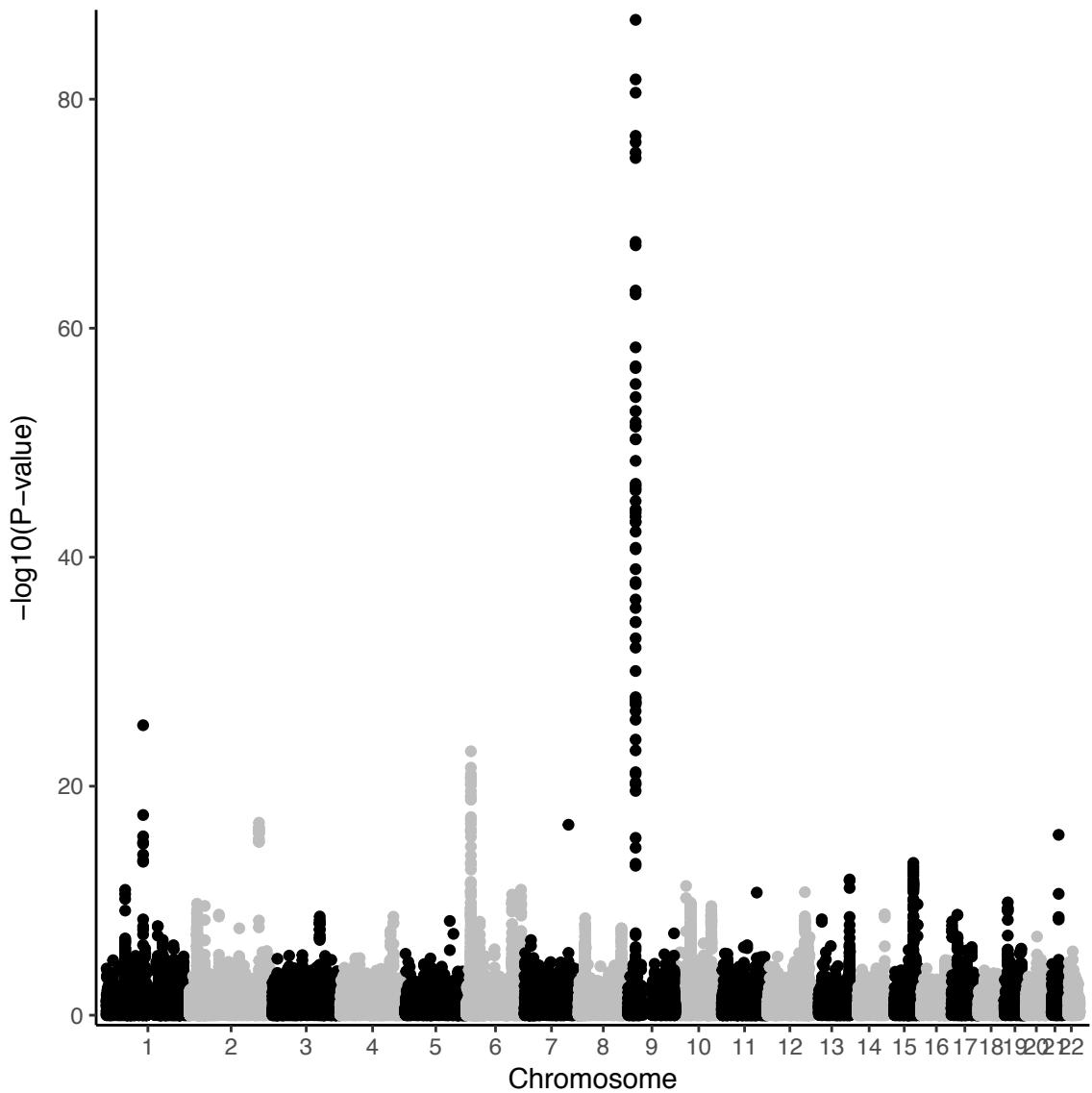
##      chrom      pos      pval      rsid      y
## 1: chr1 100098846 0.180432  rs494626 0.7436864
## 2: chr1 100128148 0.030573  rs10747505 1.5146619
## 3: chr1 100183875 0.081842  rs1541044 1.0870238
## 4: chr1 100258576 0.423634  rs531174 0.3730092
## 5: chr1 100351915 0.518362  rs2810422 0.2853668
## 6: chr1 100385263 0.190763  rs499479 0.7195059
```

The dataset contains five columns: `chrom`, `pos`, `rsid`, `pval`, and `y`. Three of them are required:

1. `chrom`
2. `pos`
3. `y`

The `chrom` and `pos` columns specify the genomic location (x-axis), and `y` specify the y-axis (duh!). The choice of the column name “`y`” is intentional - not every Manhattan plot uses $-\log_{10}P\text{-value}$ as the y-axis. After loading the data, we are ready to make a Manhattan plot. Notice that *Deloukas et al* uses hg18. To get the chromosome lengths correctly, we specify hg18 as an argument.

```
manhattan(cad_gwas,build='hg18')
```

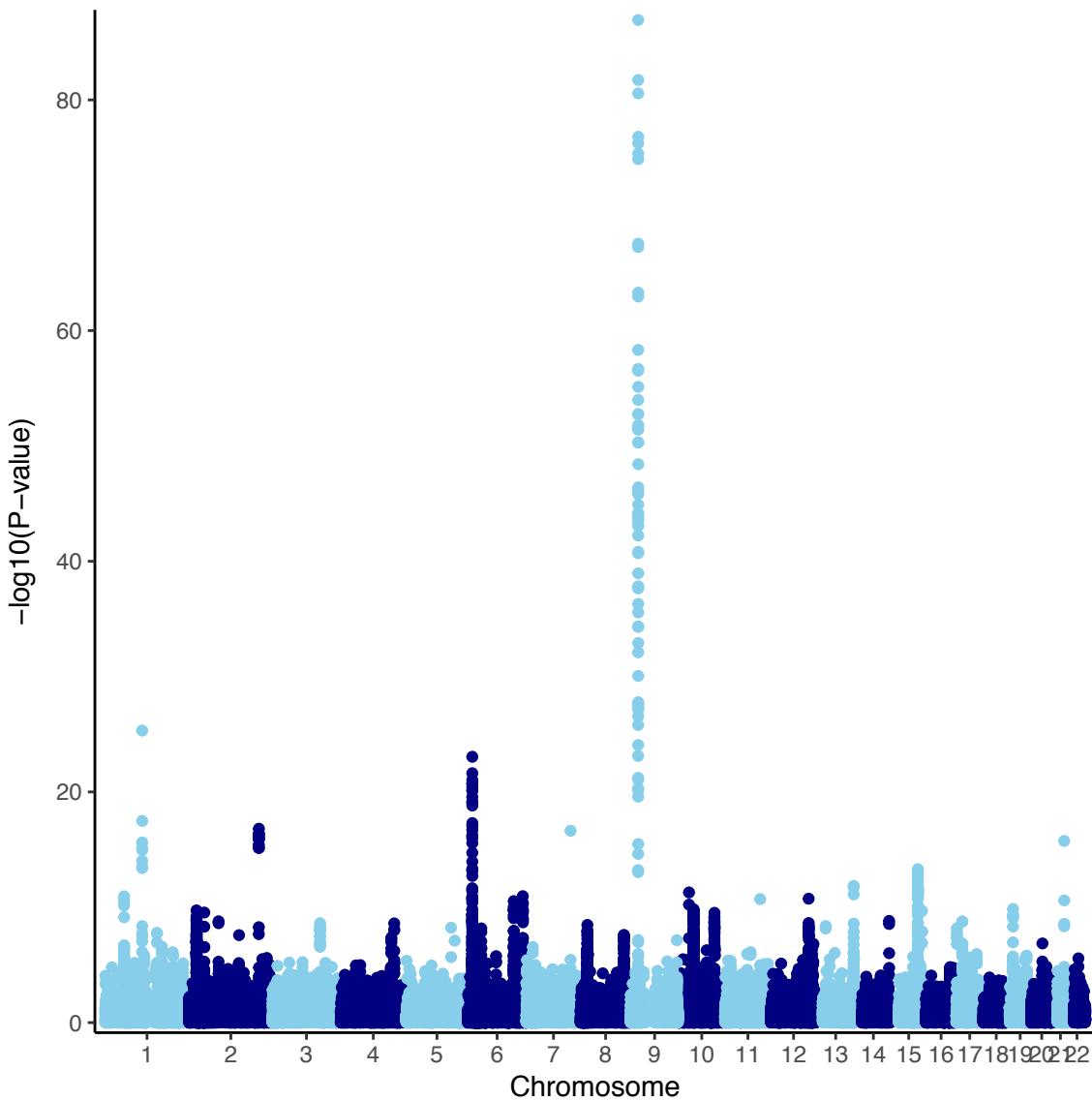


Ta-da! Our first Manhattan plot.

Customizing colors

If black and grey are dull, we can change the color of each chromosome.

```
manhattan(cad_gwas, build='hg18', color1='skyblue', color2='navyblue')
```

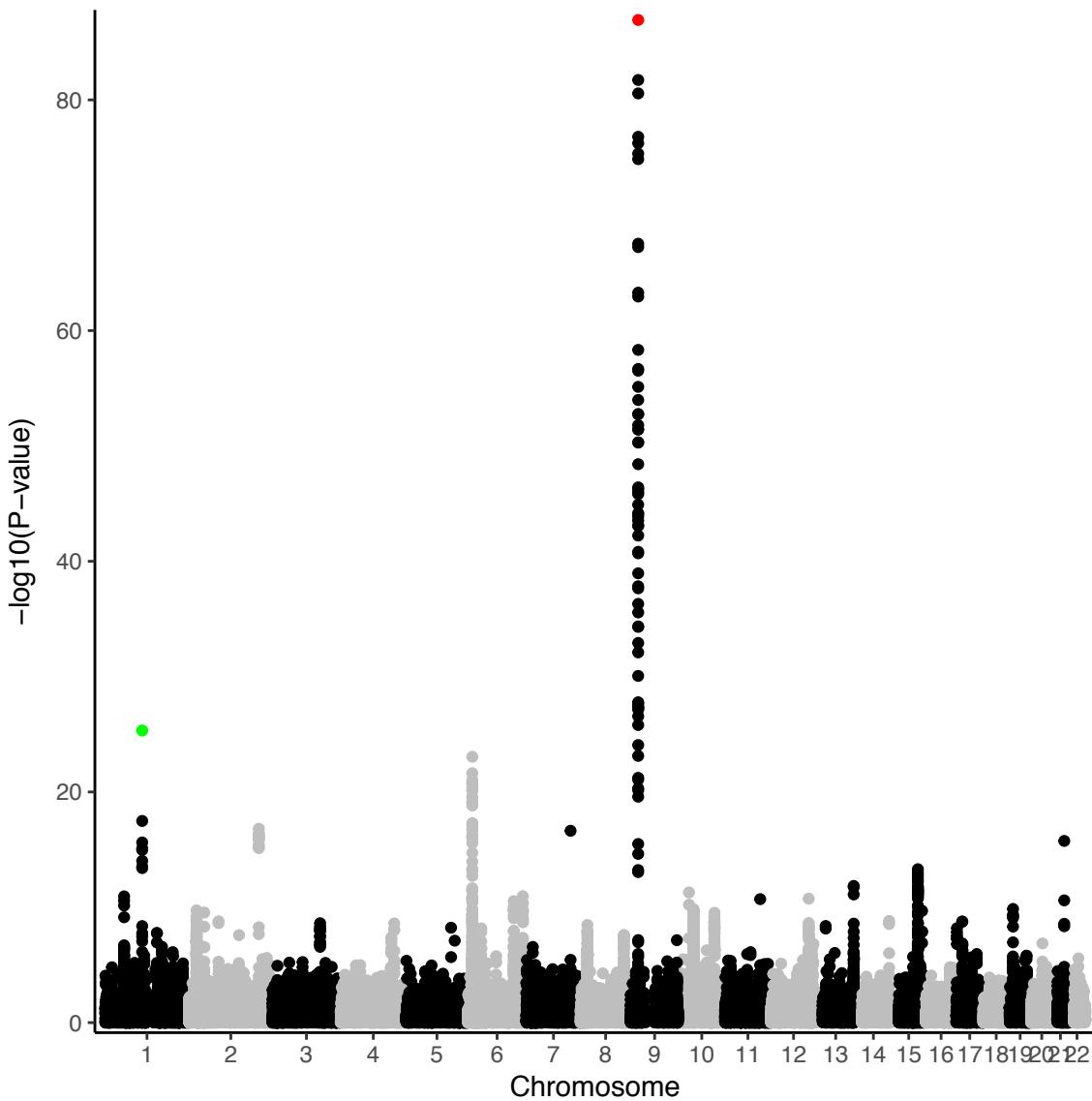


Highlight and label SNPs and genes

A common task is to highlight and annotate SNPs of interest. The package `manhattan` requires color and SNP labels to be specified in the input `data.frame` as two columns: color and label. Note that only SNPs of interest have color strings, and other SNPs **must be left as NA**.

Let us highlight two SNPs rs602633 and rs1333045.

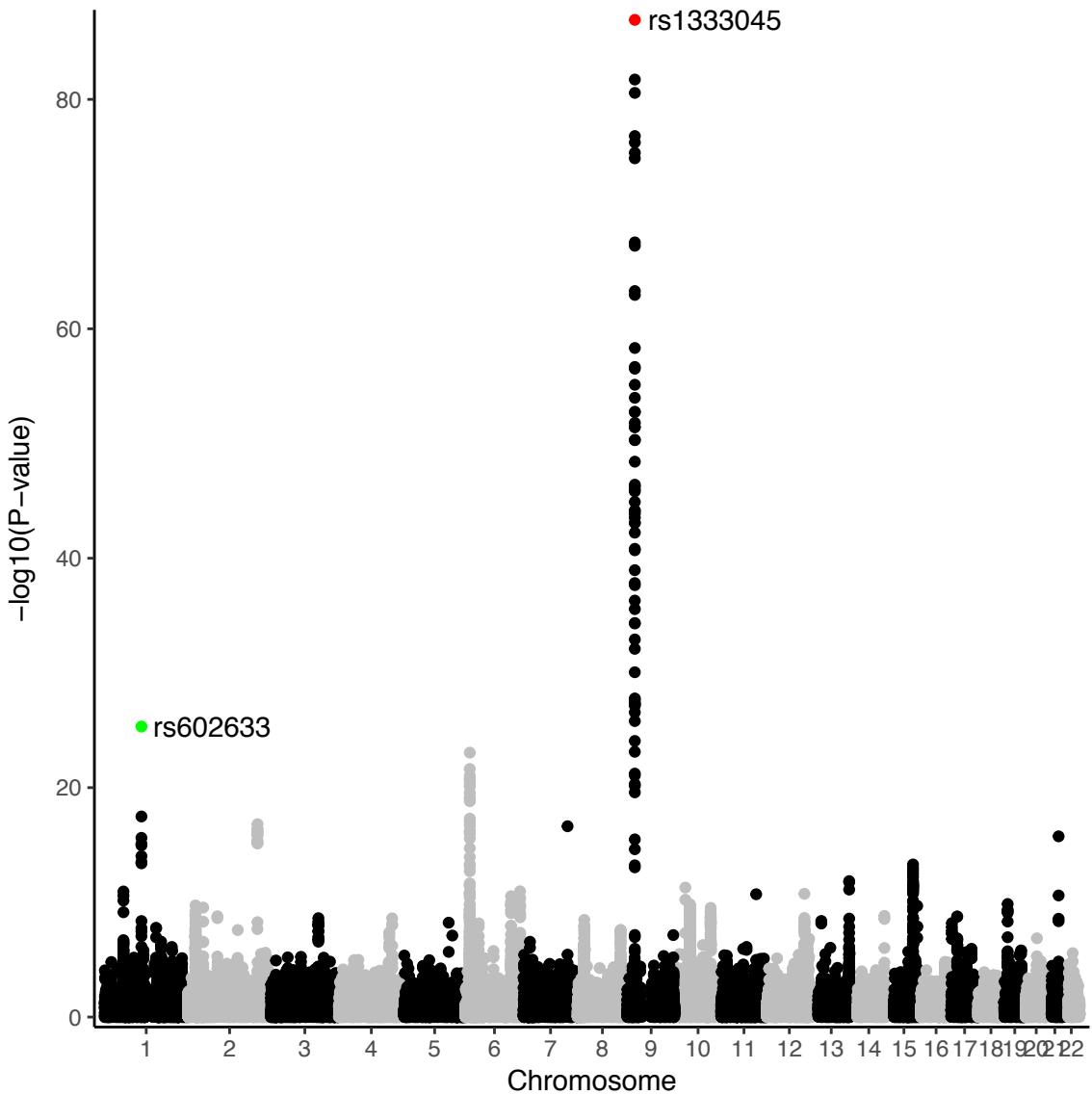
```
cad_gwas[cad_gwas$rsid=='rs602633','color']='green'
cad_gwas[cad_gwas$rsid=='rs1333045','color']='red'
manhattan(cad_gwas,build='hg18')
```



We could also label the two SNPs. Note again that only SNPs of interest have label strings, other SNPs should be left as NAs. Since `manhattan` returns a `ggplot` object, we could just add a `geom_text` layer.

```
cad_gwas[cad_gwas$rsid=='rs602633','label']='rs602633'
cad_gwas[cad_gwas$rsid=='rs1333045','label']='rs1333045'
manhattan(cad_gwas,build='hg18')+geom_text(aes(label=label),hjust=-0.1)
```

```
## Warning: Removed 79126 rows containing missing values (geom_text).
```

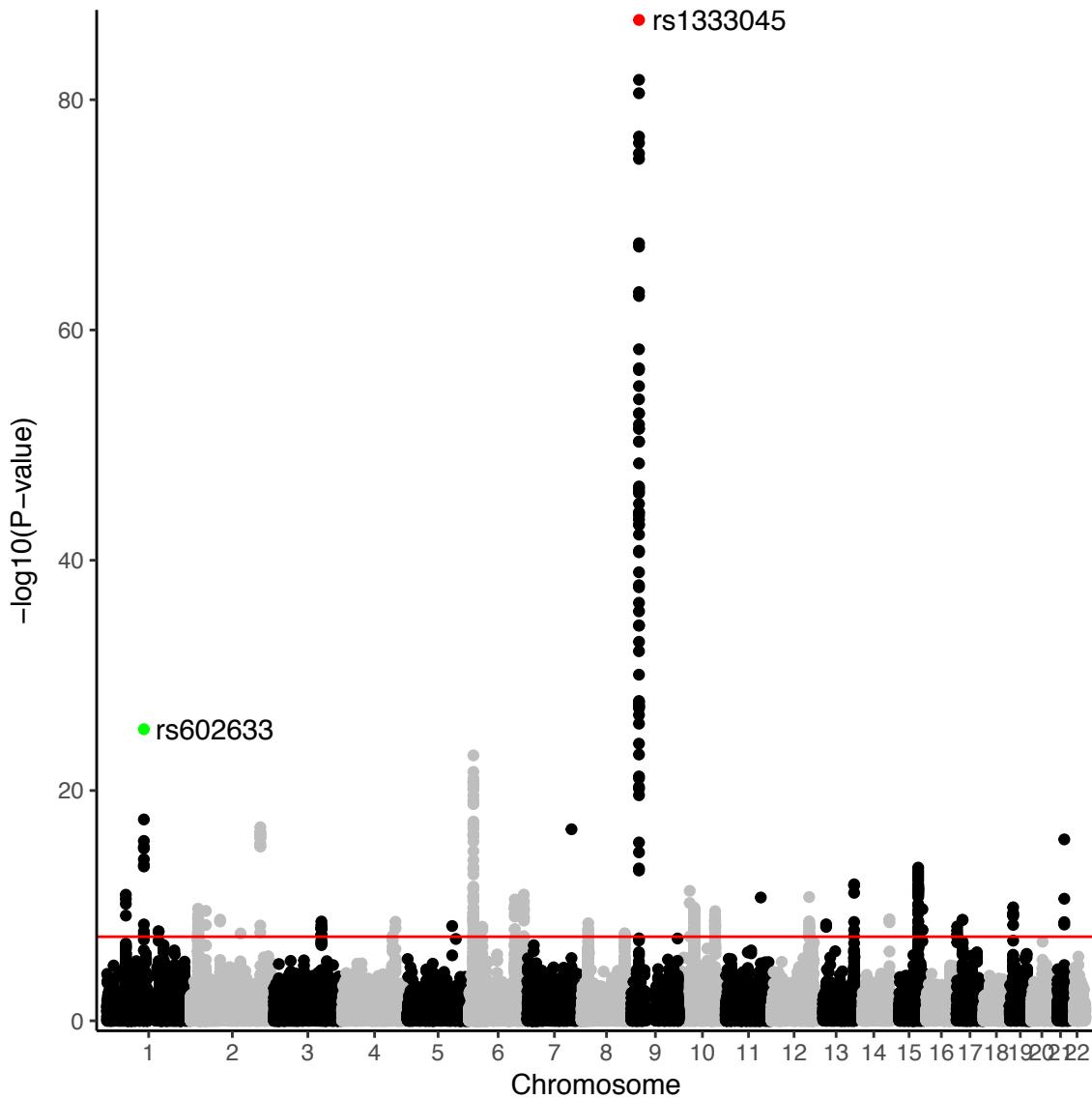


Adding GWAS significance line

It is worth noting that any standard `geoms` can be used with `manhattan`. For instance, let's add a line indicating genome-wide significant threshold.

```
manhattan(cad_gwas, build='hg18') + geom_text(aes(label=label), hjust=-0.1) + geom_hline(yintercept=-log10(5e-
```

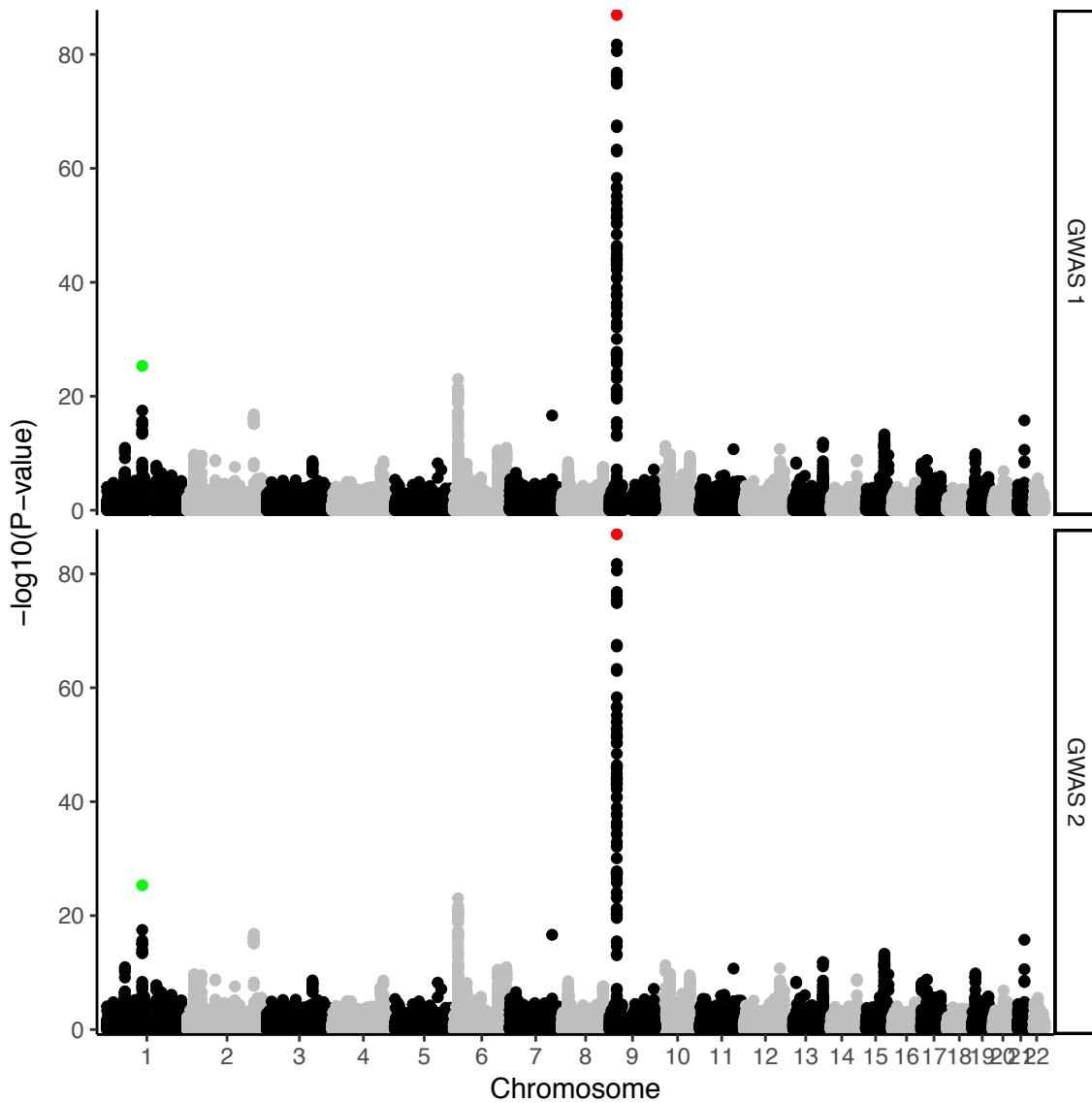
```
## Warning: Removed 79126 rows containing missing values (geom_text).
```



Plotting multiple GWAS studies

We can go even further to use facets with `manhattan`. For illustration, let us pretend that we have two GWAS studies by duplicating `cad_gwas`, and plot two GWAS studies on top of each other.

```
cad_gwas_2=rbind(cbind(cad_gwas,study='GWAS 1'),cbind(cad_gwas,study='GWAS 2'))
manhattan(cad_gwas_2,build='hg18')+facet_grid(study~.)
```

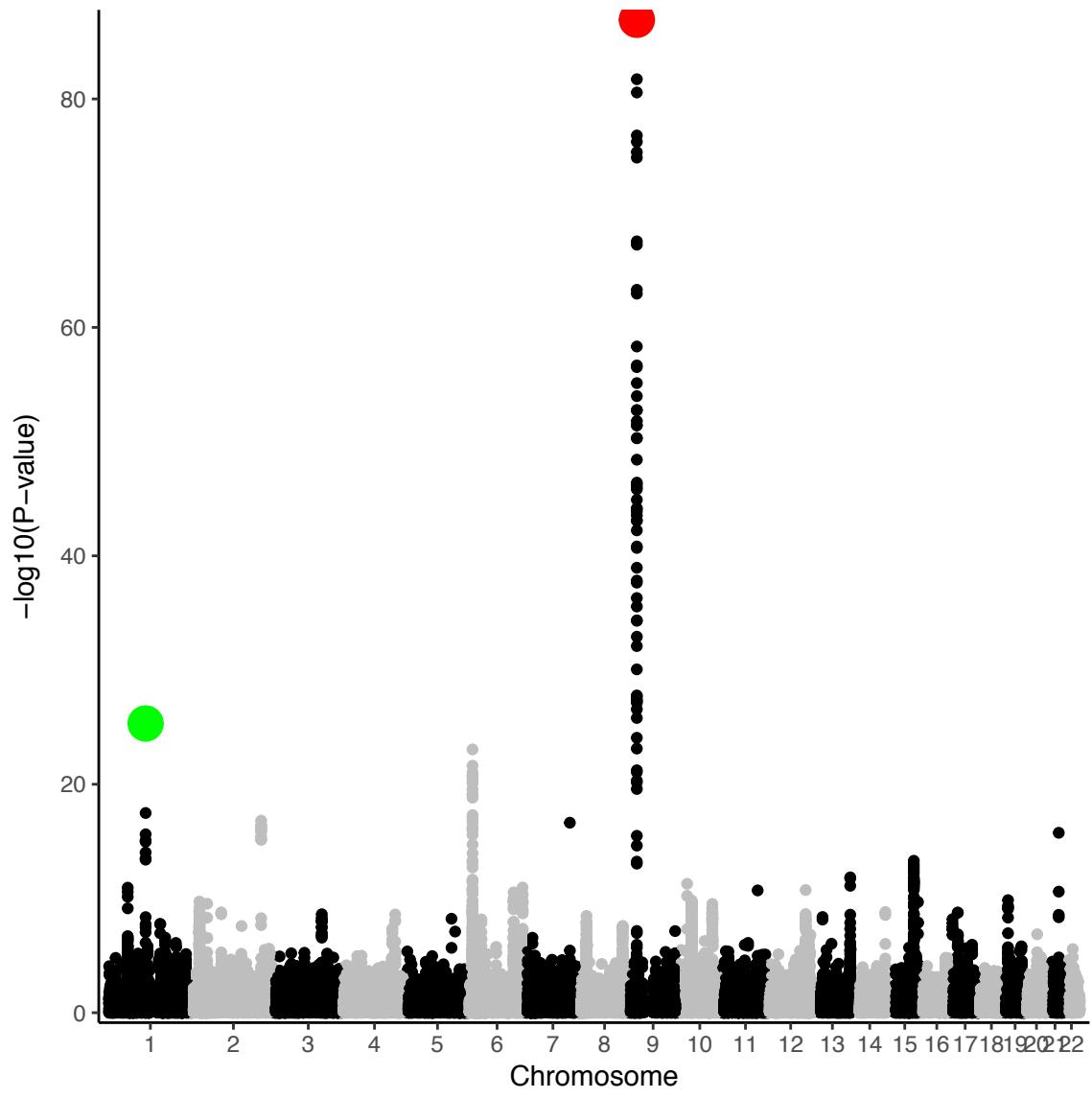


Details

Behind the scene, `manhattan` is no more than a wrapper around `ggplot`, with a few tricks to transform a genomic axis to a scatterplot axis. In brief, `manhattan` transforms the chrom:pos pairs to cumulative positions. For instance, chr2:1 would be the length of chromosome 1 plus 1, chr3:1 would be chromosome 1 plus chromosome 2 plus 1, so on and so forth. Therefore, it is important to specify the genomic build (e.g. hg19) so that `manhattan` can make the correct transformation. It then positions the chromosome labels on the x-axis according to these transformations.

Again, it is important to note that `manhattan` is no more than a wrapper around `ggplot`, which makes `manhattan` highly customizable. For instance, we can change the size of the SNPs of interest by adding a `geom_point` layer.

```
cad_gwas$size=ifelse(cad_gwas$rsid%in%c('rs602633','rs1333045'),5,2)
manhattan(cad_gwas,build='hg18')+geom_point(aes(color=color,size=size),show.legend=FALSE)
```



Questions and Bugs

If you have any question or want to report a bug, please open a github issue [here](#).