Week 11

Genome-wide association studies (GWAS)

Plotting in R

R has always had some plotting capabilities. However, the number of packages that are designed to produce data visualizations has grown dramatically over the last 15 years. Today the plotting landscape is dominated by two largely incompatible ecosystems one in base R and one integrated with the package ggplot2. I use both in my own work.

Base R	**ggplot2**
Shallow learning curve	Steep learning curve
More freedom to do anything you want to do	Many good decisions are default behavior

ggplot2 (data)

wide data

time 1	time 2
1.202	1.45
1.301	1.271
0.987	0.654
2.013	2.458
1.750	1.989

long data

Rate	Time
1.202	1
1.301	1
0.987	1
2.013	1
1.750	1
1.45	2
1.271	2
0.654	2
2.458	2
1.989	2

ggplot2 (grammar)

Heath made the cool plot.

Noun	Heath	Heath	Heath
Verb	made	made	fixed
Article	the	the	the
Adjective	cool	horrible	horrible
Noun	plot	plot	plot

ggplot2 (grammar)

Grammatical elements in ggplot2

Element	Description
data	The data being plotted
aesthetics	The scales onto which we plot our data
geometries	The visual elements used for our data
facets	Splitting plots into multiples based on a variable
statistics	Ways of summarizing data
coordinates	The space on which data will be plotted
themes	Aspects unrelated to the data

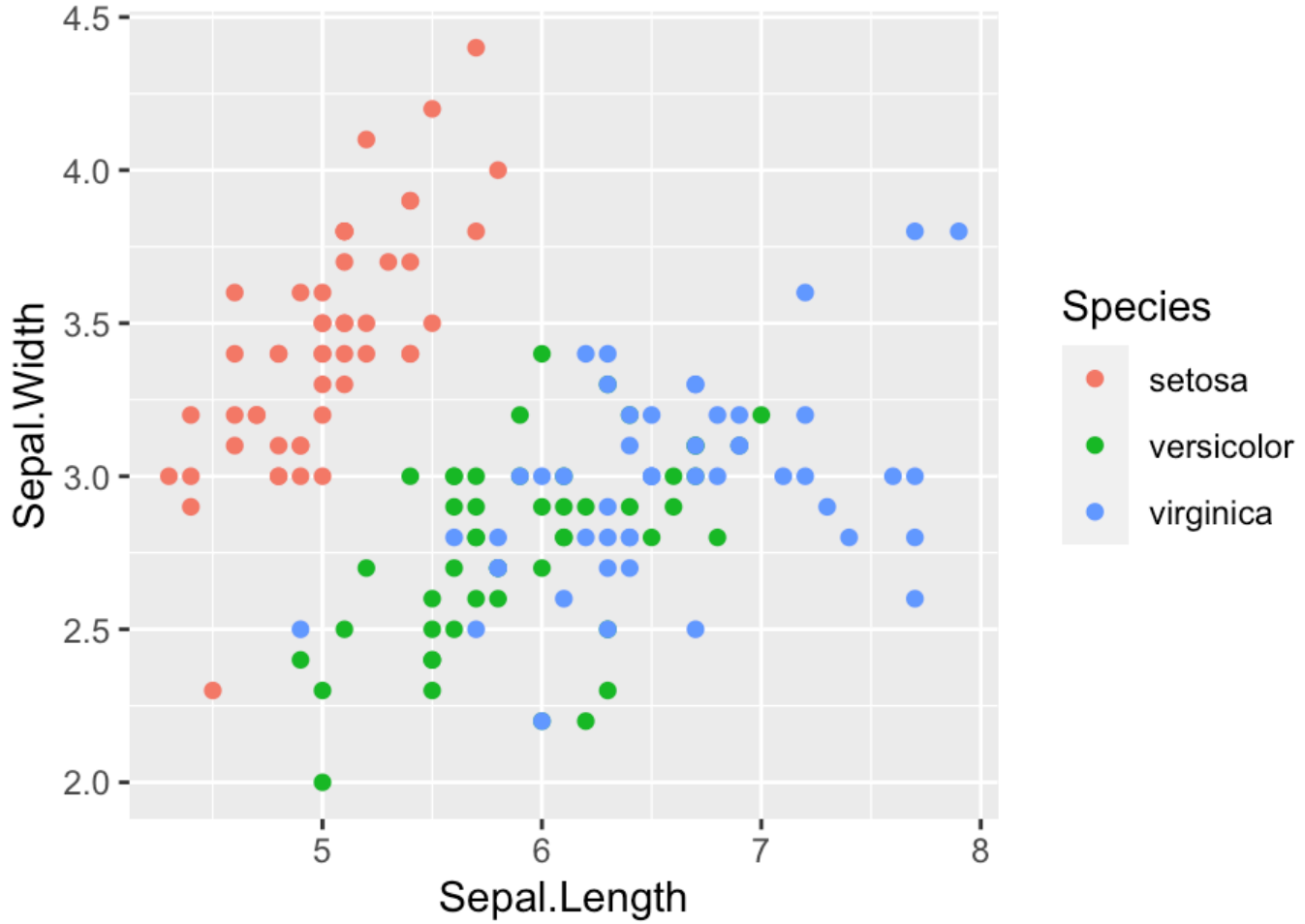
ggplot2 (simple example)

```
library(ggplot2)
data(iris)
ggplot(iris, aes(x=Sepal.Length, y= Sepal.Width, col=Species)) + geom_point()
```

In this case I wanted an XY scatter plot so these aesthetics make sense. Depending on the geometry you will you will use other things may make more or less sense to include. Some common options include: x, y, fill, col, shape, size.

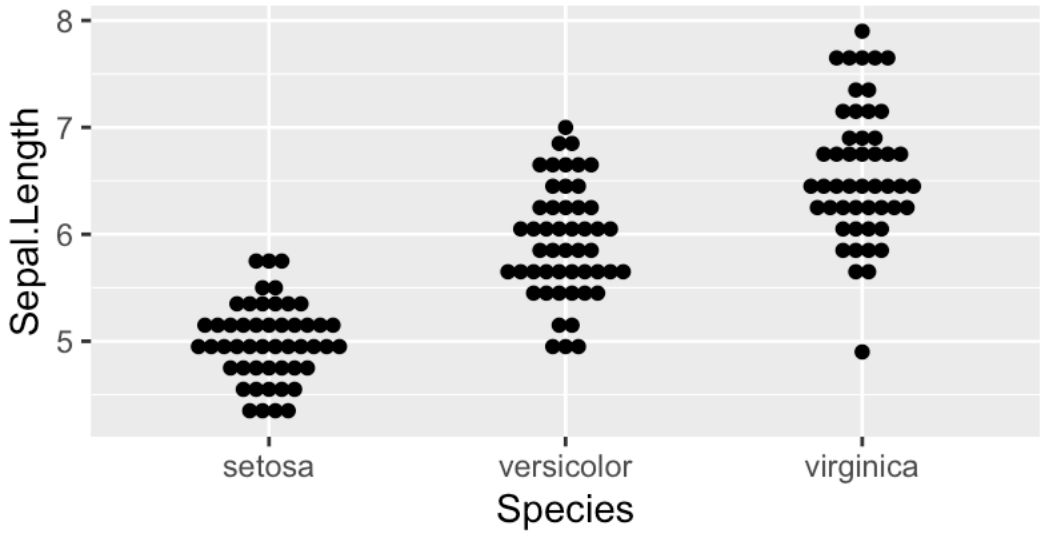
ggplot2 (simple example)

```
library(ggplot2)
data(iris)
ggplot(iris, aes(x=Sepal.Length, y= Sepal.Width, col=Species)) + geom_point()
```

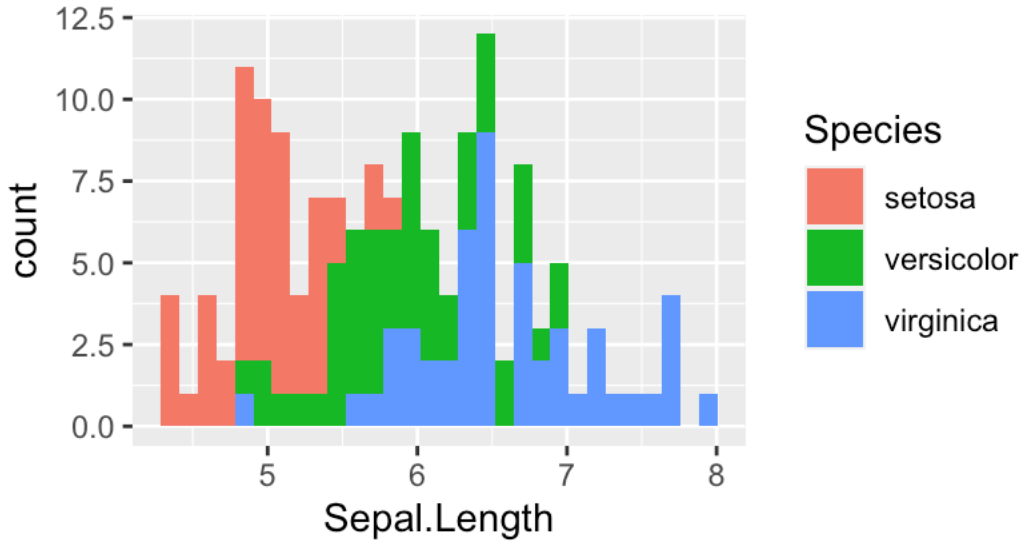


ggplot2 (simple example)

```
a <- ggplot(iris, aes(x=Species, y=Sepal.Length)) +
       geom_dotplot(binaxis = "y", stackdir = "center")
```



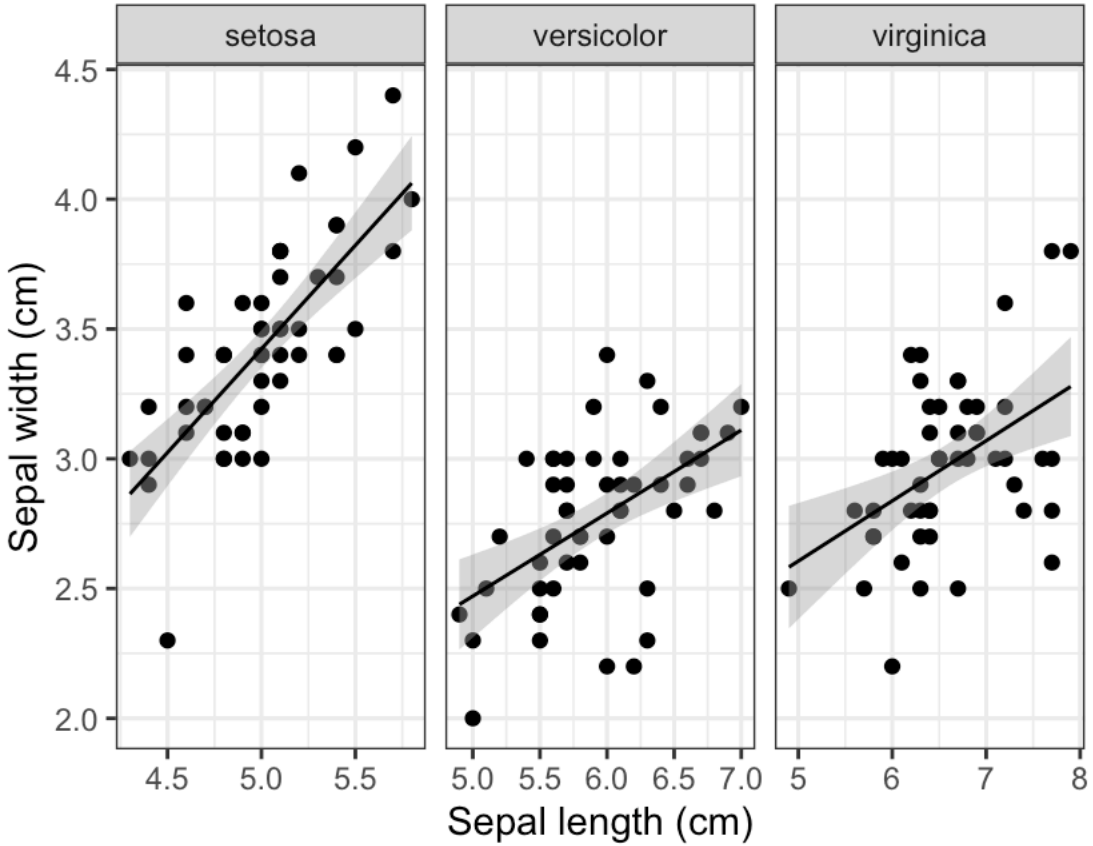

```
b <- ggplot(iris, aes(x=Sepal.Length, fill=Species)) +
       geom_histogram(position="dodge")
```




```
library(gridExtra)
grid.arrange(a,b)
```

ggplot2 (nicer example)

```
library(ggplot2)
data(iris)
ggplot(iris, aes(x=Sepal.Length, y= Sepal.Width)) +
  geom_point() +
  geom_smooth(method="lm", col="black", size=.5) +
  facet_wrap(~Species, scales="free_x") +
  theme_bw() +
  xlab("Sepal length (cm)") +
  ylab("Sepal width (cm)")
```



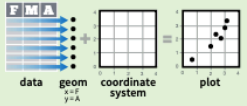
ggplot2 (cheat sheet)

Data Visualization with ggplot2 Cheat Sheet

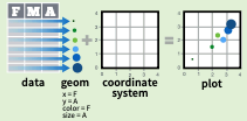
RStudio

Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same few components: a **data** set, a set of **geoms**—visual marks that represent data points, and a **coordinate system.**



To display data values, map variables in the data set to aesthetic properties of the geom like **size**, **color**, and **x** and **y** locations.



Build a graph with **qplot()** or **ggplot()**

aesthetic mappings | data | geom

qplot(x = cty, y = hwy, color = cyl, data = mpg, geom = "point"**)**

Creates a complete plot with given data, geom, and mappings. Supplies many useful defaults.

ggplot(data = mpg, **aes(**x = cty, y = hwy**))**

Begins a plot that you finish by adding layers to. No defaults, but provides more control than qplot().

data

```
ggplot(mpg, aes(hwy, cty)) +
 geom_point(aes(color = cyl)) +
 geom_smooth(method ="lm") +
 coord_cartesian() +
 scale_color_gradient() +
 theme_bw()
```

add layers, elements with +

layer = geom + default stat + layer specific mappings

additional elements

Add a new layer to a plot with a **geom_*()** or **stat_*()** function. Each provides a geom, a set of aesthetic mappings, and a default stat and position adjustment.

last_plot()

Returns the last plot

ggsave("plot.png", width = 5, height = 5)

Saves last plot as 5' x 5' file named "plot.png" in working directory. Matches file type to file extension.

Geoms - Use a geom to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.

One Variable

Continuous

a <- ggplot(mpg, aes(hwy))

a + geom_area(stat = "bin")
x, y, alpha, color, fill, linetype, size
b + geom_area(aes(y = ..density..), stat = "bin")

a + geom_density(kernel = "gaussian"**)**
x, y, alpha, color, fill, linetype, size, weight
b + geom_density(aes(y = ..county..))

a + geom_dotplot()
x, y, alpha, color, fill

a + geom_freqpoly()
x, y, alpha, color, linetype, size
b + geom_freqpoly(aes(y = ..density..))

a + geom_histogram(binwidth = 5**)**
x, y, alpha, color, fill, linetype, size, weight
b + geom_histogram(aes(y = ..density..))

Discrete

b <- ggplot(mpg, aes(fl))

b + geom_bar()
x, alpha, color, fill, linetype, size, weight

Two Variables

Continuous X, Continuous Y

f <- ggplot(mpg, aes(cty, hwy))

f + geom_blank()

f + geom_jitter()
x, y, alpha, color, fill, shape, size

f + geom_point()
x, y, alpha, color, fill, shape, size

f + geom_quantile()
x, y, alpha, color, linetype, size, weight

f + geom_rug(sides = "bl"**)**
alpha, color, linetype, size

f + geom_smooth(model = lm**)**
x, y, alpha, color, fill, linetype, size, weight

f + geom_text(aes(label = cty)**)**
x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

Discrete X, Continuous Y

g <- ggplot(mpg, aes(class, hwy))

g + geom_bar(stat = "identity")
x, y, alpha, color, fill, linetype, size, weight

g + geom_boxplot()
lower, middle, upper, x, ymax, ymin, alpha, color, fill, linetype, shape, size, weight

g + geom_dotplot(binaxis = "y", stackdir = "center"**)**
x, y, alpha, color, fill

g + geom_violin(scale = "area"**)**
x, y, alpha, color, fill, linetype, size, weight

Discrete X, Discrete Y

h <- ggplot(diamonds, aes(cut, color))

h + geom_jitter()
x, y, alpha, color, fill, shape, size

Continuous Bivariate Distribution

i <- ggplot(movies, aes(year, rating))

i + geom_bin2d(binwidth = c(5, 0.5)**)**
xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size, weight

i + geom_density2d()
x, y, alpha, colour, linetype, size

i + geom_hex()
x, y, alpha, colour, fill size

Continuous Function

j <- ggplot(economics, aes(date, unemploy))

j + geom_area()
x, y, alpha, color, fill, linetype, size

j + geom_line()
x, y, alpha, color, linetype, size

j + geom_step(direction = "hv"**)**
x, y, alpha, color, linetype, size

Visualizing error

df <- data.frame(grp = c("A", "B"), fit = 4:5, se = 1:2)
k <- ggplot(df, aes(grp, fit, ymin = fit-se, ymax = fit+se))

k + geom_crossbar(fatten = 2**)**
x, y, ymax, ymin, alpha, color, fill, linetype, size

k + geom_errorbar()
x, ymax, ymin, alpha, color, linetype, size, width (also **geom_errorbarh()**)

k + geom_linerange()
x, ymin, ymax, alpha, color, linetype, size

k + geom_pointrange()
x, y, ymin, ymax, alpha, color, fill, linetype, shape, size

Maps

data <- data.frame(murder = USArrests$Murder, state = tolower(rownames(USArrests)))
map <- map_data("state")
l <- ggplot(data, aes(fill = murder))

l + geom_map(aes(map_id = state), map = map**) + expand_limits(**x = map$long, y = map$lat**)**
map_id, alpha, color, fill, linetype, size

Three Variables

seals$z <- with(seals, sqrt(delta_long^2 + delta_lat^2))
m <- ggplot(seals, aes(long, lat))

m + geom_contour(aes(z = z)**)**
x, y, z, alpha, colour, linetype, size, weight

m + geom_raster(aes(fill = z), hjust=0.5, vjust=0.5, interpolate=FALSE**)**
x, y, alpha, fill

m + geom_tile(aes(fill = z)**)**
x, y, alpha, color, fill, linetype, size

Graphical Primitives

c <- ggplot(map, aes(long, lat))

c + geom_polygon(aes(group = group)**)**
x, y, alpha, color, fill, linetype, size

d <- ggplot(economics, aes(date, unemploy))

d + geom_path(lineend="butt", linejoin="round', linemitre=1**)**
x, y, alpha, color, linetype, size

d + geom_ribbon(aes(ymin=unemploy - 900, ymax=unemploy + 900)**)**
x, ymax, ymin, alpha, color, fill, linetype, size

e <- ggplot(seals, aes(x = long, y = lat))

e + geom_segment(aes(xend = long + delta_long, yend = lat + delta_lat)**)**
x, xend, y, yend, alpha, color, linetype, size

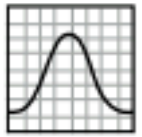
e + geom_rect(aes(xmin = long, ymin = lat, xmax= long + delta_long, ymax = lat + delta_lat)**)**
xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size

ggplot2 (cheat sheet)

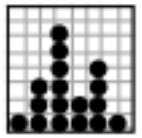
Continuous

a <- ggplot(mpg, aes(hwy))

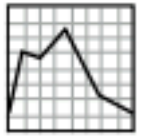
a + geom_area(stat = "bin")
x, y, alpha, color, fill, linetype, size
b + geom_area(aes(y = ..density..), stat = "bin")



a + geom_density(kernel = "gaussian"**)**
x, y, alpha, color, fill, linetype, size, weight
b + geom_density(aes(y = ..county..))



a + geom_dotplot()
x, y, alpha, color, fill



a + geom_freqpoly()
x, y, alpha, color, linetype, size
b + geom_freqpoly(aes(y = ..density..))

a + geom_histogram(binwidth = 5**)**
x, y, alpha, color, fill, linetype, size, weight
b + geom_histogram(aes(y = ..density..))

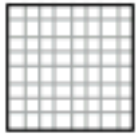
Discrete

b <- ggplot(mpg, aes(fl))

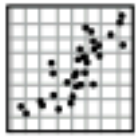
b + geom_bar()
x, alpha, color, fill, linetype, size, weight

Continuous X, Continuous Y

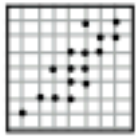
f <- ggplot(mpg, aes(cty, hwy))



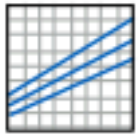
f + geom_blank()



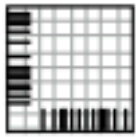
f + geom_jitter()
x, y, alpha, color, fill, shape, size



f + geom_point()
x, y, alpha, color, fill, shape, size

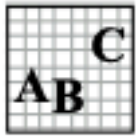


f + geom_quantile()
x, y, alpha, color, linetype, size, weight



f + geom_rug(sides = "bl"**)**
alpha, color, linetype, size

f + geom_smooth(model = lm**)**
x, y, alpha, color, fill, linetype, size, weight



f + geom_text(aes(label = cty)**)**
x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

ggraptR – a gentle transition to ggplot

ggraptR – a gentle transition to ggplot

```
ggplot(iris, aes(x=Sepal.Length, y= Sepal.Width)) +
  geom_point() +
  geom_smooth(method="lm", col="black", size=.5) +
  facet_wrap(~Species, scales="free_x") +
  theme_bw() +
  xlab("Sepal length (cm)") +
  ylab("Sepal width (cm)")
```

```
ggplot(iris, aes(y=Sepal.Width, x=Sepal.Length)) +
  geom_point(stat="identity", position="jitter", alpha=0.5, size=3) +
  geom_smooth(stat="smooth", position="identity", method="lm",
              se=TRUE, n=80, level=0.95, span=0.75) +
  facet_grid(. ~ Species, scales="free_x") +
  theme_bw() +
  theme(text=element_text(family="sans", face="plain",
                          color="#000000", size=15, hjust=0.5, vjust=0.5)) +
  scale_size(range=c(1, 3)) +
  xlab("Sepal.Length") +
  ylab("Sepal.Width")
```

GWAS

GWAS: Genome wide association study.

The goal of GWAS is to determine what genes have alleles that are responsible for a trait of interest. The trait can be any measurable trait in any organism that you wish to study. For instance, a disease in humans, an economically important trait of a crop or domestic animal, an adaptation like a certain color pattern in birds, etc.

GWAS uses existing segregating variation in a population
QTL creates a collection of crosses designed specifically for mapping the trait.

GWAS – Discrete condition (often disease)


```
>case 1
CATACTACTACTGAACGTTTGCTCCTGCtactatctctctctctctctctttctctctctctctctCATGC
>case 2
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
>case 3
CATACTACTACTGAACGTTTGCTCCTGCtactatctctctctctctctctttctctctctctctctCATGC
>control 1
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
>control 2
CATACTACTACTGAACGTTTGCTCCTGCtactatctctctctctctctctttctctctctctctctCATGC
>control 3
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
```

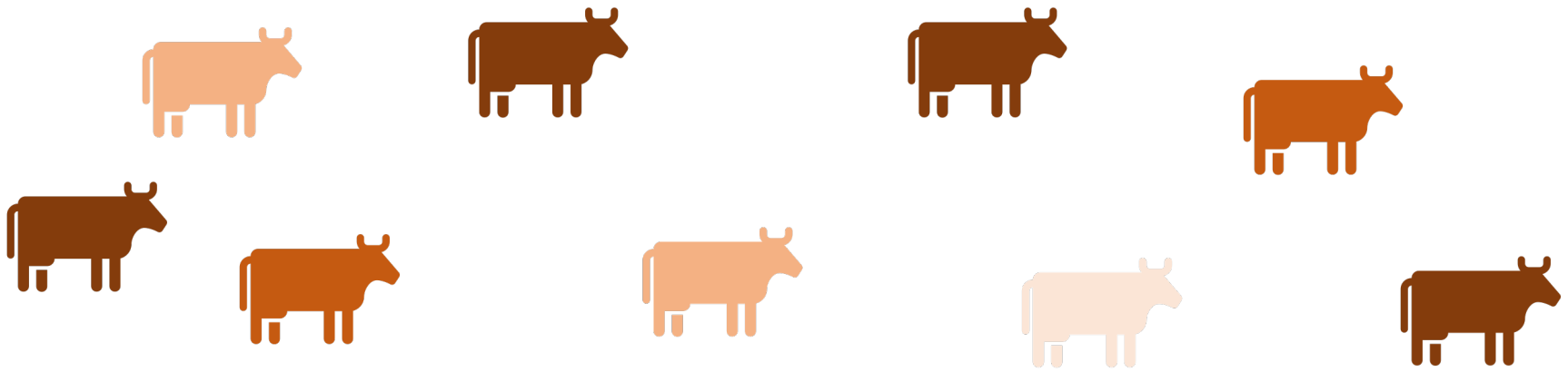
	Case	Control
Allele G	589	145
Allele C	105	487

$$odds\ ratio = \frac{589/145}{105/487} = 18.8$$

	Case	Control
Allele G	321	290
Allele C	310	210

$$odds\ ratio = \frac{321/290}{310/210} = 0.75$$

GWAS – Continuous trait



```
>sample 1
CATACTACTACTGAACGTTTGCTCCTGCtactatctctctctctctctttctctctctctctctCATGC
>sample 2
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
>sample 3
CATACTACTACTGAACGTTTGCTCCTGCtactatctctctctctctctttctctctctctctctCATGC
>sample 4
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
>sample 5
CATACTACTACTGAACGTTTGCTCCTGCtactatctctctctctctctttctctctctctctctCATGC
>sample 6
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
```

GWAS – Continuous trait

AGTTGACTACTGCATACTCGT
AGTTGACTACTGCATACTCGT

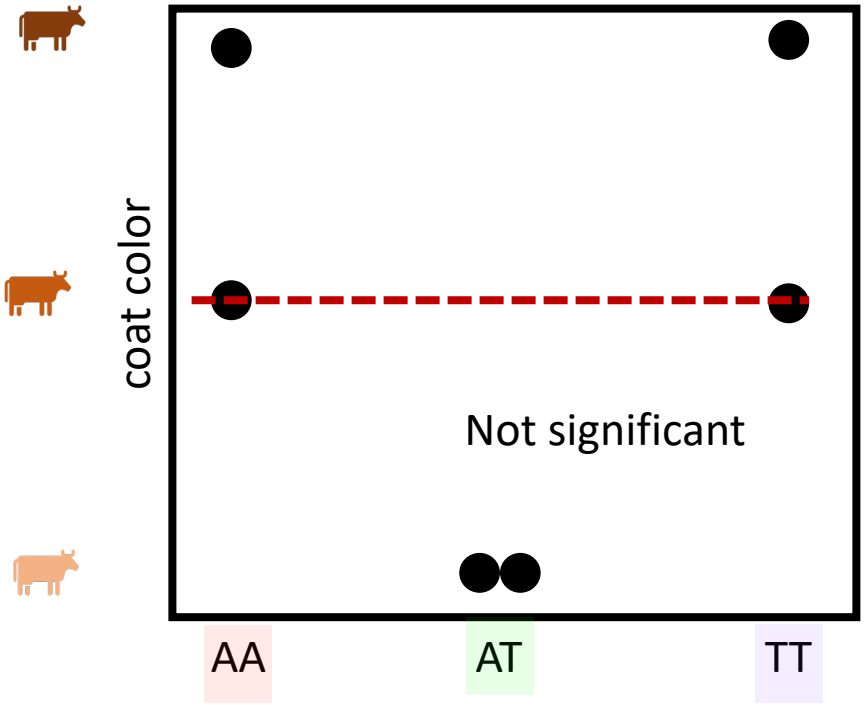
AGTTGACTACTGCATACTCGT
AGTTGACTACTGCATACTCGT

TGTTGACTACTGCATACTCGT
AGTTGACTACTGCATACTCGT

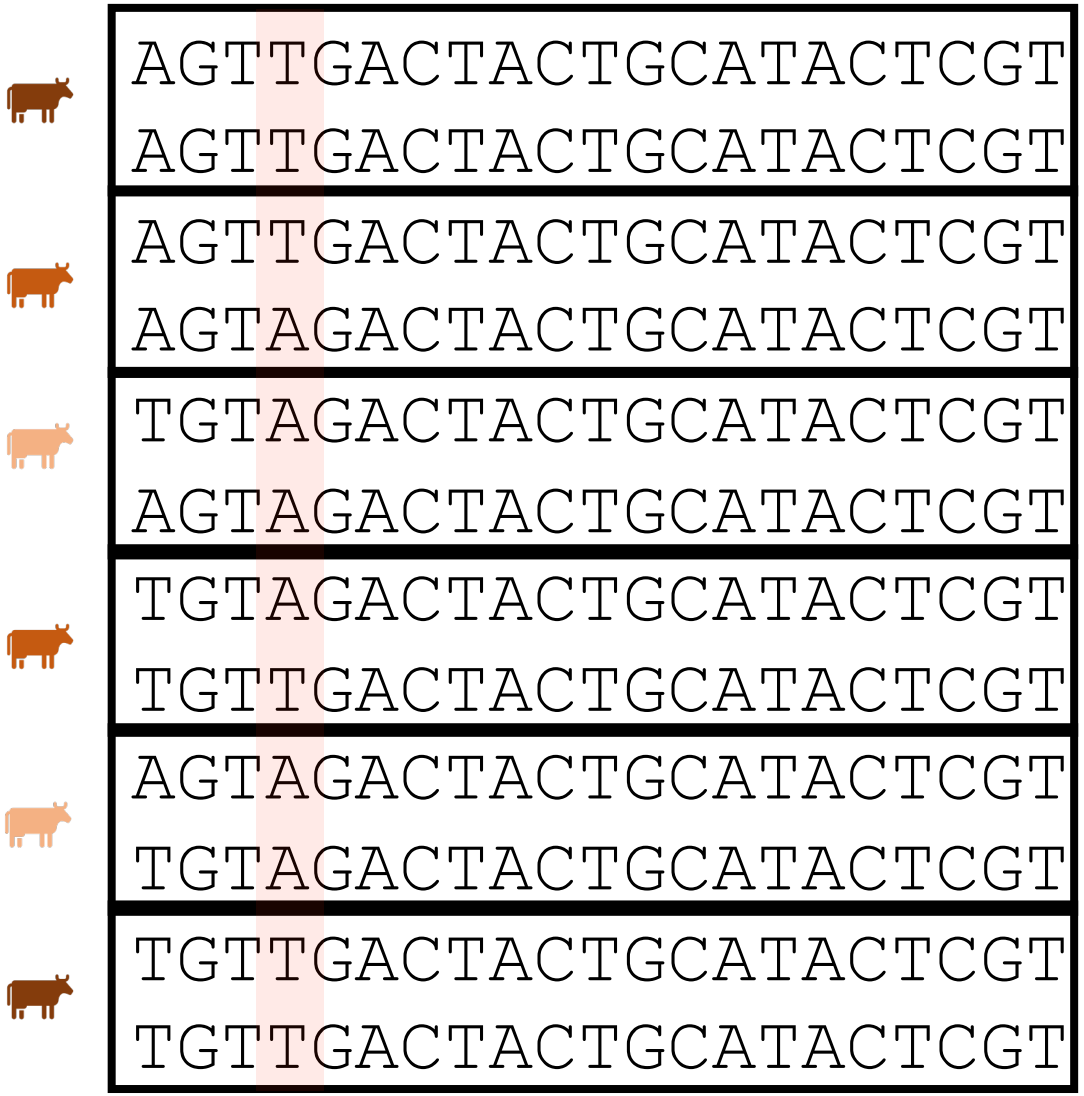
TGTTGACTACTGCATACTCGT
TGTTGACTACTGCATACTCGT

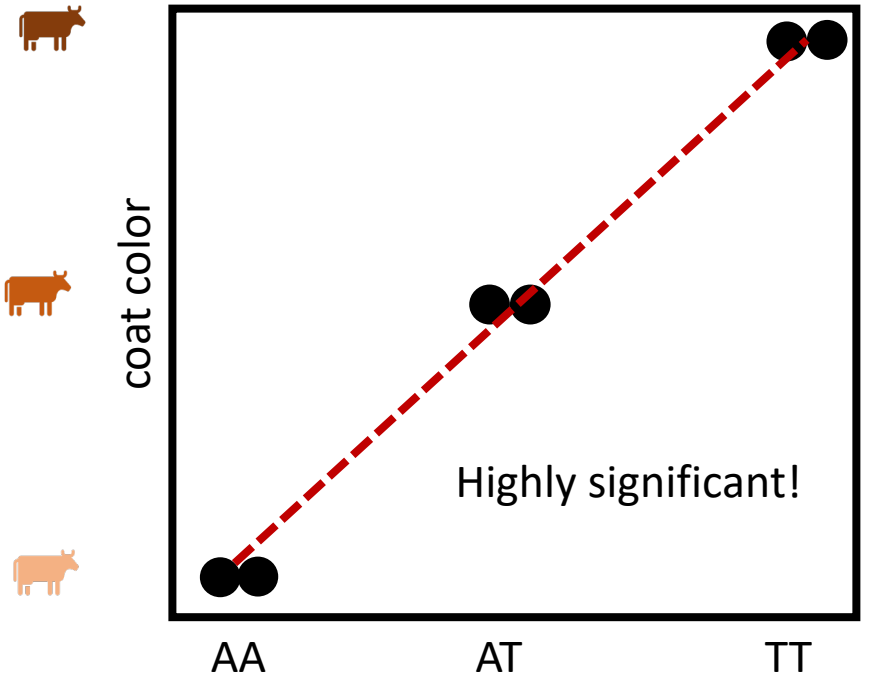
AGTTGACTACTGCATACTCGT
TGTTGACTACTGCATACTCGT

TGTTGACTACTGCATACTCGT
TGTTGACTACTGCATACTCGT

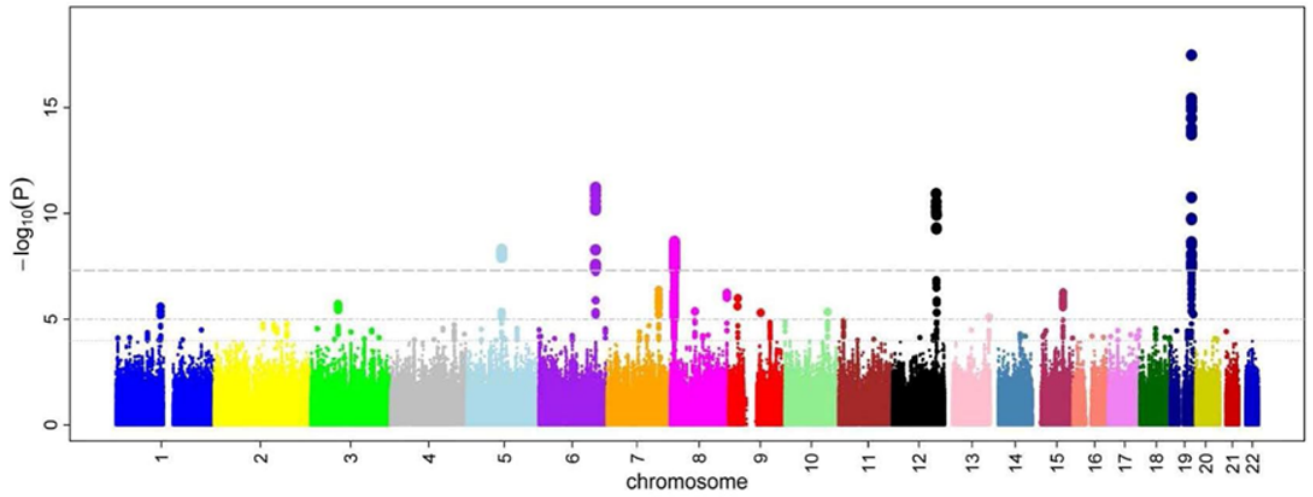


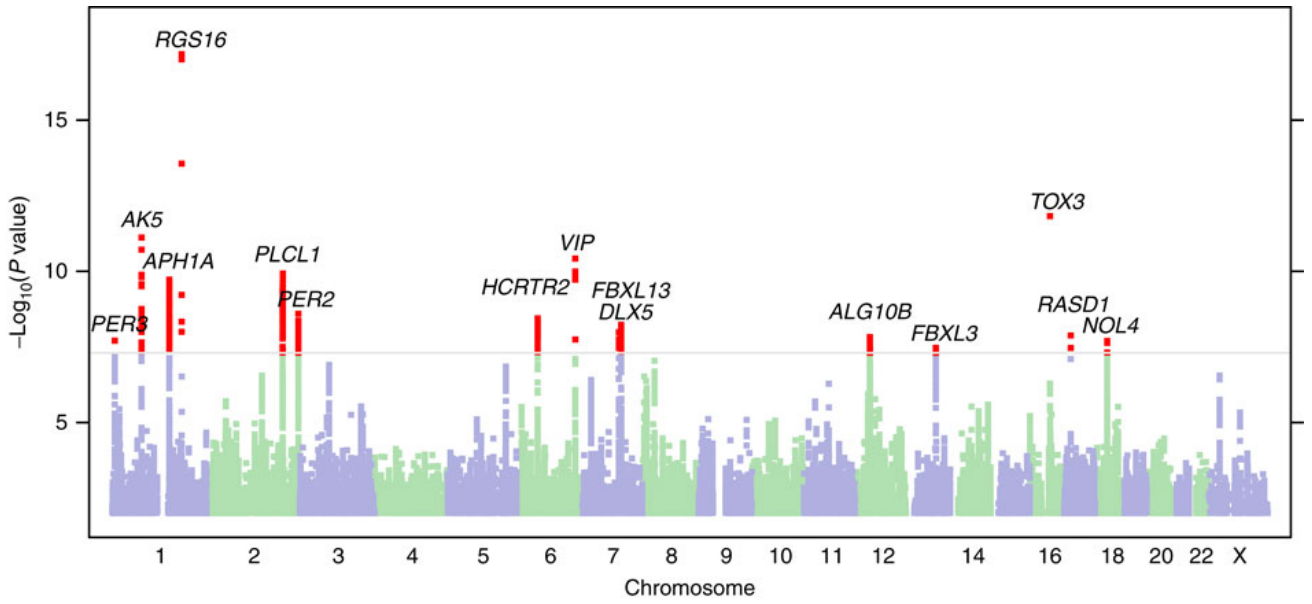
GWAS – Continuous trait

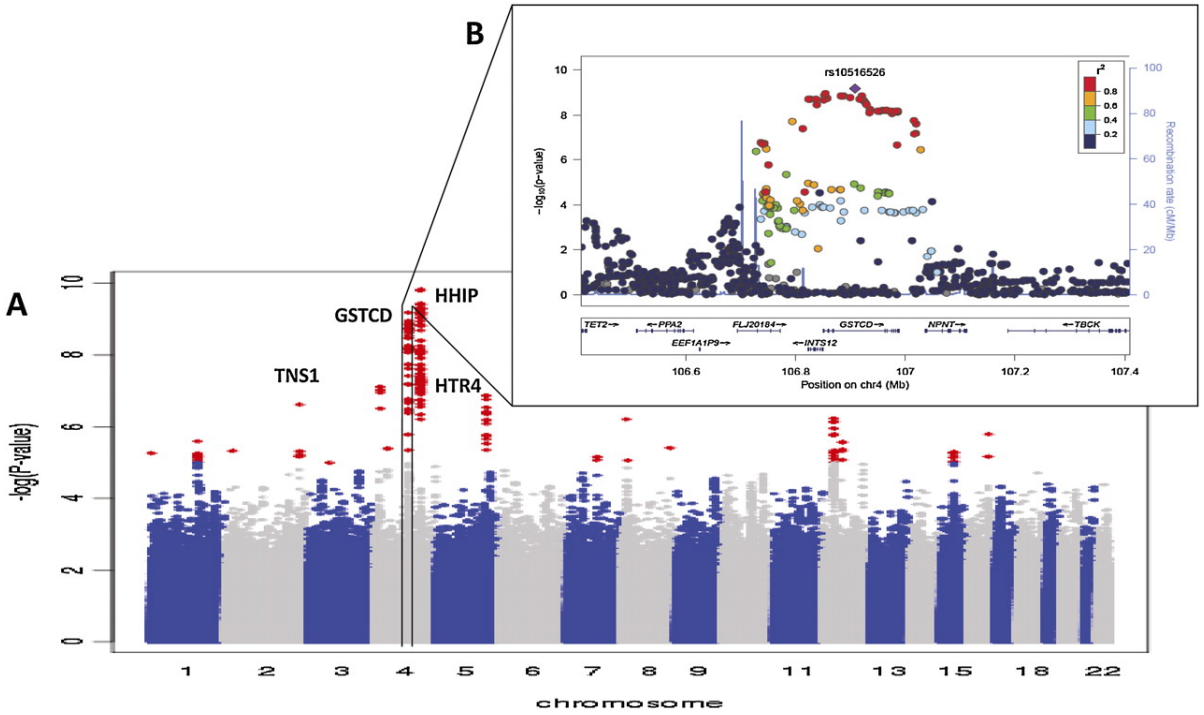




GWAS







R examples

GWAS - Problems

What is the problem with doing this across the whole genome?

Multiple tests lead to more false positives!

1) require a higher level of significance $5x10^{-8}$
2) only look at the very most significant
3) lots of more complicated approaches too!

What is one of the most basic requirements of almost all statistical tests?

Tests normally assume independence of the data points!

1) Samples from a population will be related to each other due to ancestry (trees!)
2) DNA sequencing is not done equally in all groups of people (western samples are usually over represented

GWAS - Problems

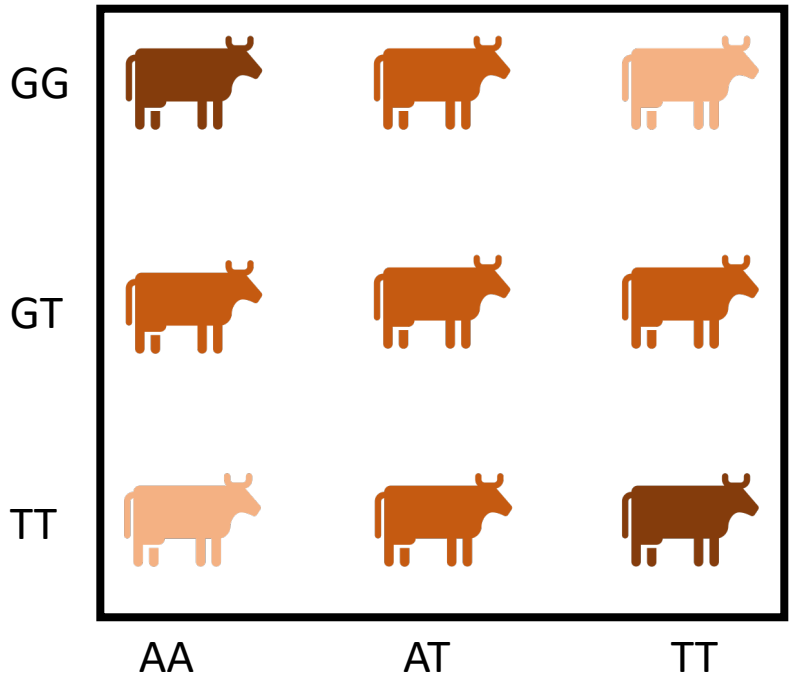
What is epistasis?

It is the case where the impact of a genotype at one locus depend on the genotype at another locus!

To find this type of effect you would need to look at all pairs of genotypes. How many tests would we need to do then?

About 6 orders of magnitude more tests would be required.

Humans have around 4,000,000 sites that are variable like this which equates to 7,999,998,000,000



GWAS - Problems

What about the environment?

Many diseases have a strong environmental component (heart disease, diabetes, cancer, etc.)

If these are left out of the study often what is discovered is actually genetic variation that happens to coincide with environmental factors?

If a disease is more common in Europeans than Africans or Asians but it is because of a lifestyle characteristic any genetic variation that is common in Europeans but rare in Africans and Asians could appear associated with the disease.

Applications and benefits of GWAS

Widely used in agriculture and domestication. For instance, you can do a GWAS on wild strains of rice (which have lots of variation in things we care about like grain size, growing time, etc.) This GWAS can tell you what variants at what locations in the genome should be introgressed into domestic varieties in hopes of introducing favorable traits.

Widely used in medicine to identify the genes responsible for disease. This is often the first step necessary in being able to create an animal model for a disease that will then allow researchers to study the disease and develop pharmaceutical interventions.

Many diseases have multiple different underlying genetic causes and treatments that are available may only work on some forms of the diseases. Thanks to GWAS studies we now know what these different causal genes are. Now patients can be genotyped at these causative loci and medication can be tailored to their version of the disease. (precision or personalized medicine)

Applied in a more limited fashion to inform patients of risk that they will develop more severe versions of a disease. For instance depending on your genotype at a gene you may choose a more aggressive form of treatment.

Complex Disease / Complex Phenotypes

Many diseases are what we call complex diseases. There is no one gene responsible for the disease. Instead the disease can manifest due to variations in 10 or 100s of different genes in the genome acting in concert with the environment. GWAS is less insightful (though still important) for diseases like this.

Schizophrenia: Not really a clearly delineated disease like say COVID, type 1 diabetes, or cystic fibrosis. Instead it is a constellation of symptoms that individuals exhibit to varying degrees.

From studies of multiple generations of families we know that 20-30% of risk is inherited (genetic). The other 70-80% of your risk of developing the disease is environmental and is poorly understood. Massive GWAS studies with 1000s of individuals have studied the genetic component. These studies have identified more than 100 different loci in the genome that seem to have some predictive power. However, these variations are associated with an increase in risk and there are many people who carry many alleles that increase risk but that never develop the disease.

Phylo-GWAS


c
latitude/
seasonality
A A T T A A A A T A T A T T T T T T T T
Genes: PSI RCIII, HCF243
heavy
metals
A T A A A T T T A
Genes: HMA, Lac-2