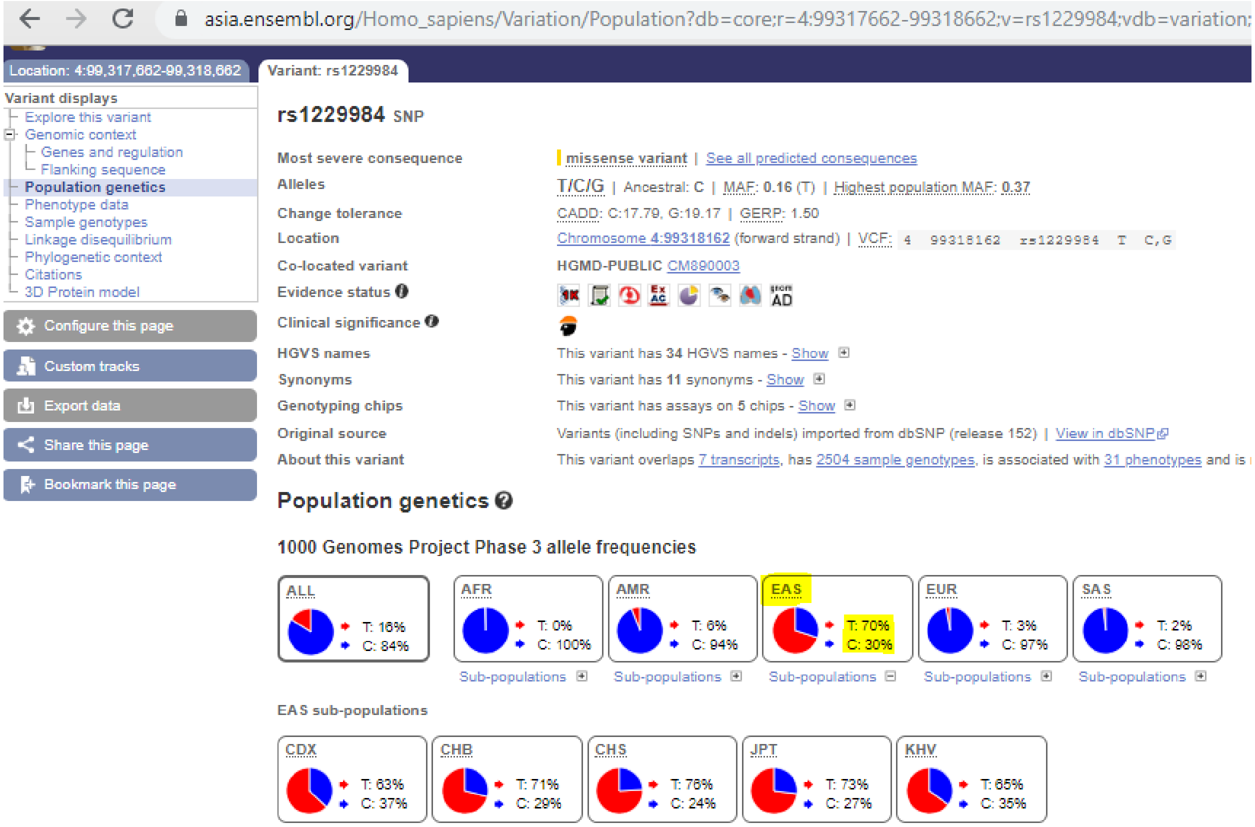
R Workshop Session 2 Exercise

Exercise 1:

The human genome is diploid, meaning that it carries two copies of an allele (one from each parent). The two alleles might be identical (homozygous genotype) or not (heterozygous).

rs1229984 is a missence variant in the ADH1B gene that has been associated with Esophageal cancer in Japanese (PMID:19698717) and with oral cavity and pharyngeal cancer in Europeans (PMID:21437268 and PMID:27749845).



asia.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=4:99317662-99318662;v=rs1229984;vdb=variation;

rs1229984 SNP

Most severe consequence	missense variant | See all predicted consequences
Alleles	T/C/G | Ancestral: C | MAF: 0.16 (T) | Highest population MAF: 0.37
Change tolerance	CADD: C:17.79, G:19.17 | GERP: 1.50
Location	Chromosome 4:99318162 (forward strand) | VCF: `4 99318162 rs1229984 T C,G`
Co-located variant	HGMD-PUBLIC CM890003
Evidence status	1K ExAC AD
Clinical significance	
HGVS names	This variant has 34 HGVS names - Show
Synonyms	This variant has 11 synonyms - Show
Genotyping chips	This variant has assays on 5 chips - Show
Original source	Variants (including SNPs and indels) imported from dbSNP (release 152) | View in dbSNP
About this variant	This variant overlaps 7 transcripts, has 2504 sample genotypes, is associated with 31 phenotypes and is

Population genetics

1000 Genomes Project Phase 3 allele frequencies

ALL	AFR	AMR	EAS	EUR	SAS
T: 16% C: 84%	T: 0% C: 100%	T: 6% C: 94%	T: 70% C: 30%	T: 3% C: 97%	T: 2% C: 98%
	Sub-populations	Sub-populations	Sub-populations	Sub-populations	Sub-populations

EAS sub-populations

CDX	CHB	CHS	JPT	KHV
T: 63% C: 37%	T: 71% C: 29%	T: 76% C: 24%	T: 73% C: 27%	T: 65% C: 35%

Looking this SNP up on Ensembl (see above) shows the frequency of T-allele is 70% and C-allele is 30% in East Asians. If you were to screen 100 East Asians randomly, how many people do you expect to carry the:

a) TT genotype

b) CT genotype

c) CC genotype

SOLUTION:

The trick here is to realize that there are two possibilities to get the CT genotype.

From mother	From Father	Genotype	Probability
T	T	TT	0.49 (= 0.7 x 0.7)
C	T	CT	0.21 (= 0.3 x 0.7)
T	C	CT	0.21 (= 0.7 x 0.3)
C	C	CC	0.09 (= 0.3 x 0.3)

Total probability = 0.49 + 0.21 + 0.21 + 0.09 = 1

. . .

So you should expect

- 49 people to carry the TT allele
- 42 people to carry the CT allele
- 9 people to carry the CC allele

We will generalize this to the binomial distribution in Session 3.

Exercise 2:

Background:
The data in CO2_uptake.csv comes from an experiment where carbon dioxide uptake in grass plants from two origins was measured with and without cold exposure.

a. Set your working directory, load your packages and clean your workspace. Then read in your data

SOLUTION:

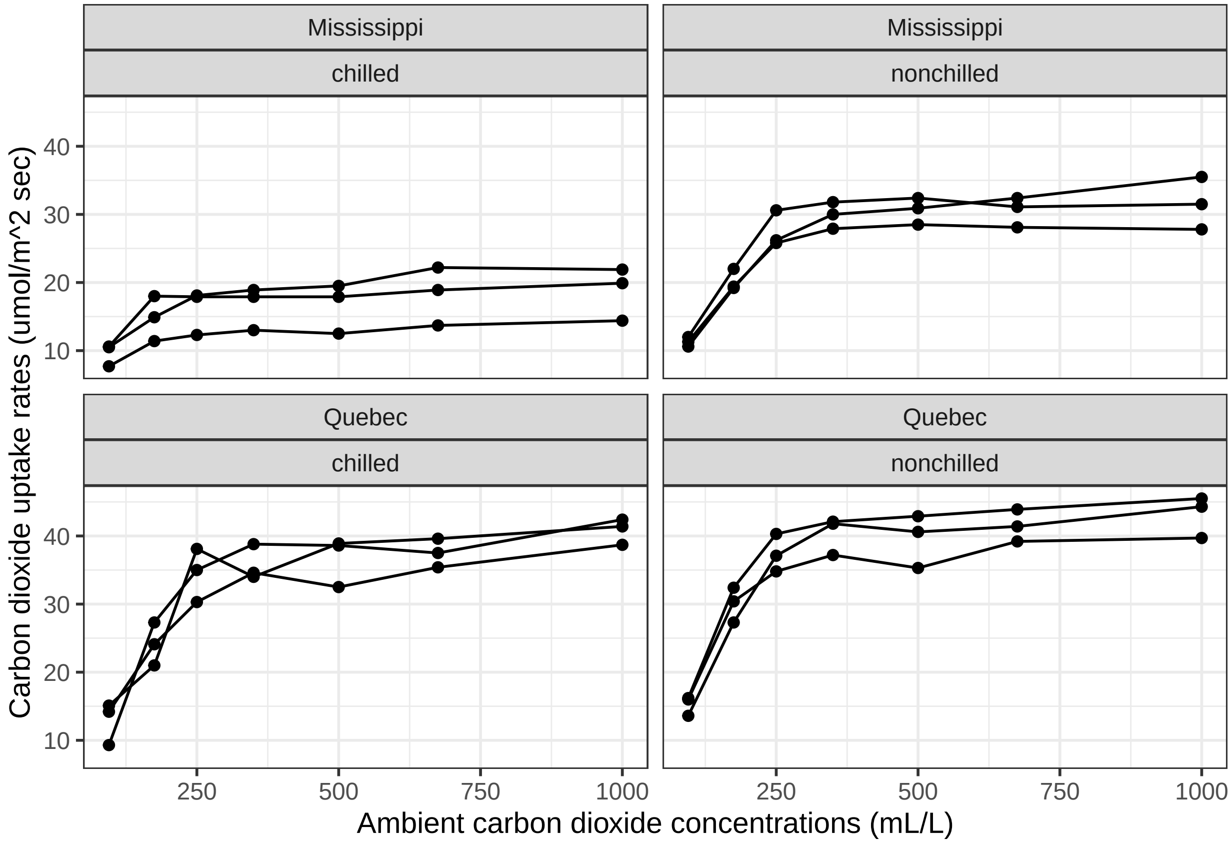
```
setwd("C:/Users/oguzg.A-GIS/Desktop/R_workshop")
pacman::p_load(tidyverse, readxl, janitor,skimr, data.table)
rm(list=ls())
```

```
co2_uptake <- read_csv("data/CO2_Uptake.csv")
```

b. Let's visulize the data. Please reproduce the following graph:

SOLUTION:

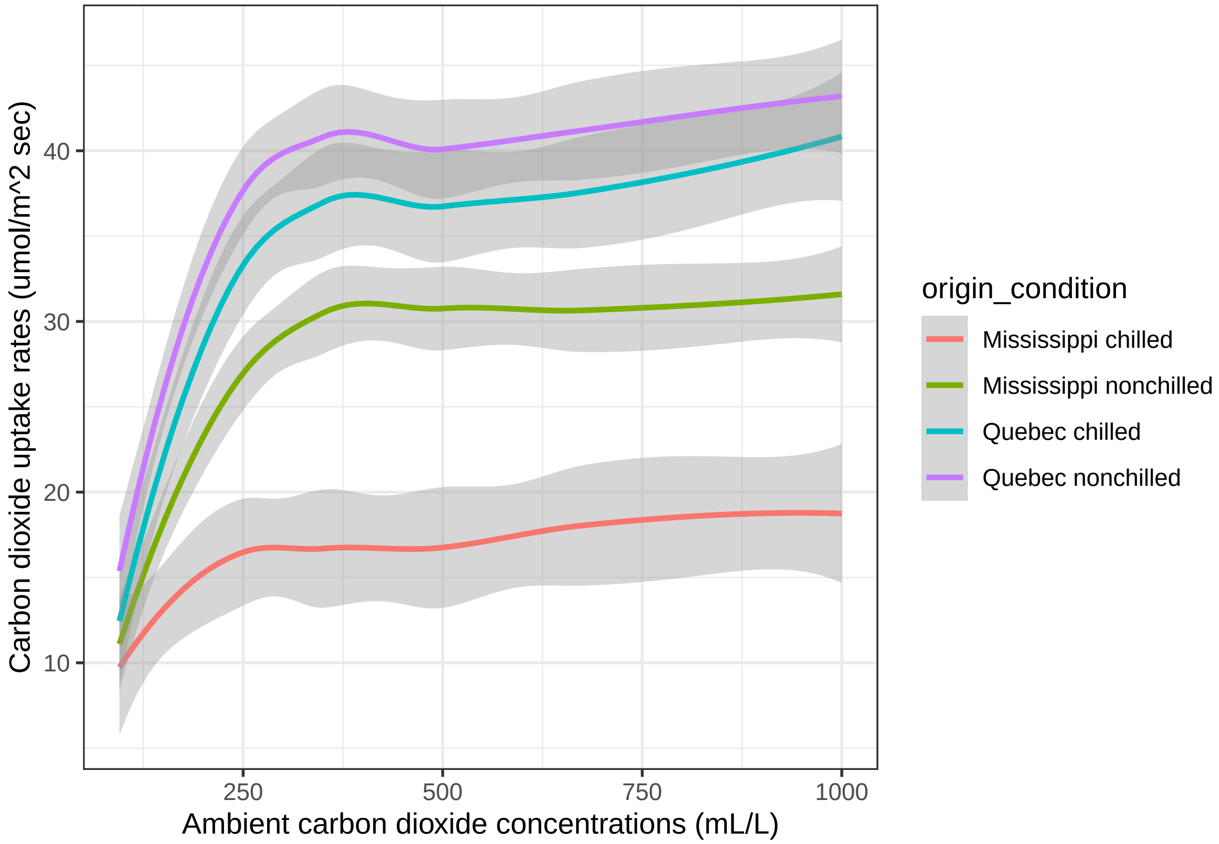
```
mylabs <- labs(x="Ambient carbon dioxide concentrations (mL/L)",
               y="Carbon dioxide uptake rates (umol/m^2 sec)")

ggplot(co2_uptake, aes(x=conc, y=uptake, group=Plant)) +
  geom_point() + geom_line() +
  facet_wrap(Type ~ Treatment) +
  theme_bw() + mylabs
```


Mississippi
chilled
Mississippi
nonchilled
Quebec
chilled
Quebec
nonchilled
Carbon dioxide uptake rates (umol/m^2 sec)
10
20
30
40
250
500
750
1000
Ambient carbon dioxide concentrations (mL/L)

c. There seems to be very little variation between the plants in each condition. Let us simplify the graph to show the smoothed trendline representing the average and standard errors. Please reproduce the following graph:

SOLUTION:

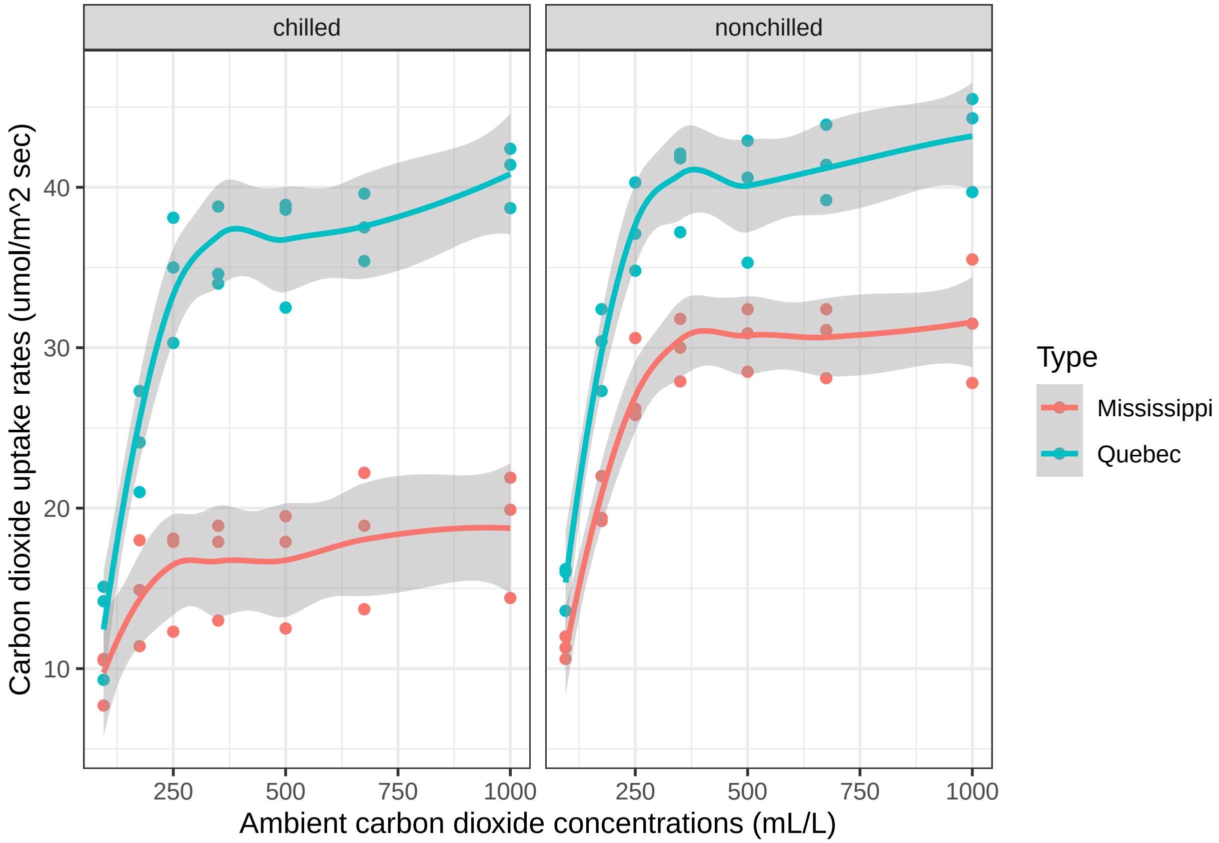
```
co2_uptake <- co2_uptake %>%
  mutate(origin_condition=paste(Type, Treatment))

ggplot(co2_uptake, aes(x=conc, y=uptake, col=origin_condition)) +
  geom_smooth() +
  theme_bw() + mylabs
```



d. Let's try another view of the same data. Please reproduce the following graph:

SOLUTION:

```
ggplot(co2_uptake, aes(x=conc, y=uptake, col=Type)) +
  geom_point() + geom_smooth() + facet_wrap(~Treatment) +
  theme_bw() + mylabs
```

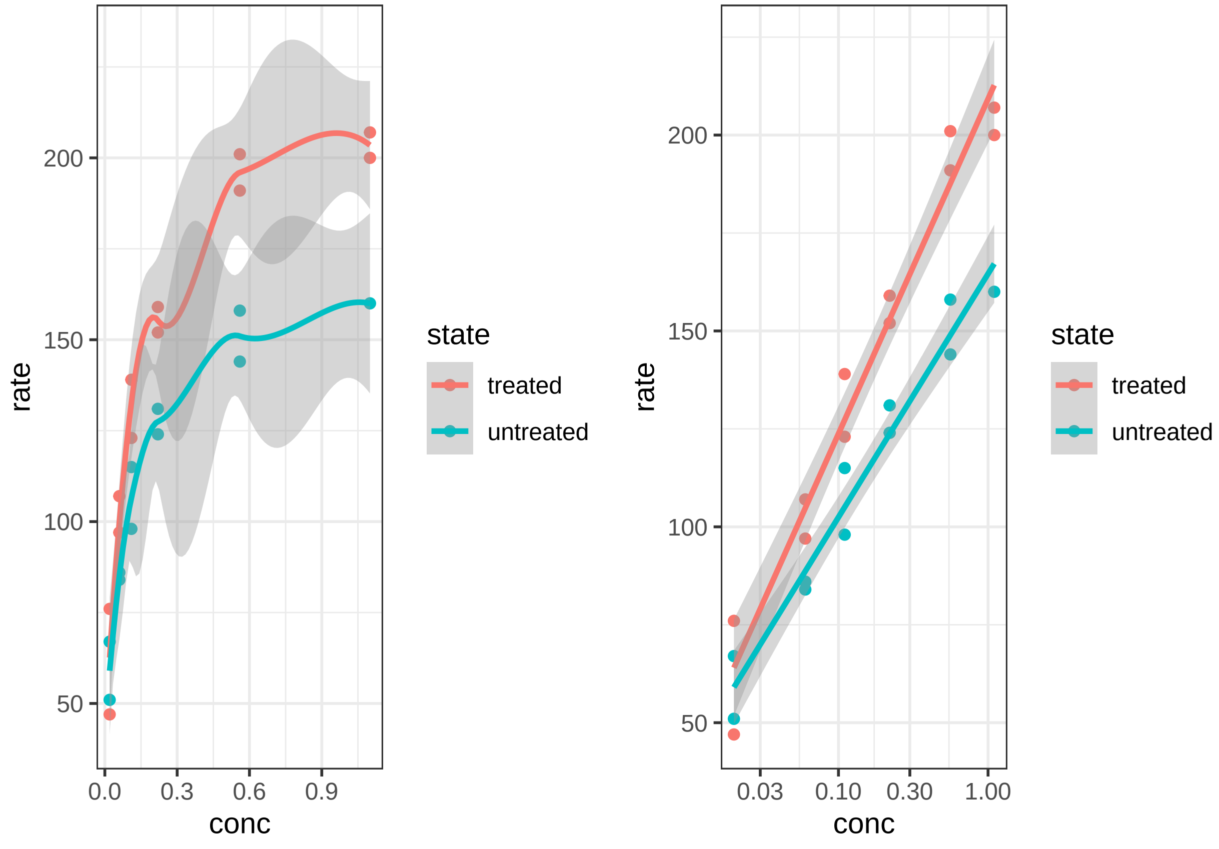


Exercise 3:

Background: The Puromycin data shows the reaction velocity versus substrate concentration in an enzymatic reaction involving untreated cells or cells treated with Puromycin.

a. Read in the *puromycin_reaction.csv*
b. Plot the reaction rate against the substrate concentration
c. Repeat step b but with the concentration on a log10 scale

```
Puro <- read_csv("data/puromycin_reaction.csv")
## Warning: Missing column names filled in: 'X1' [1]

g <- ggplot(Puro, aes(x=conc, y=rate, col=state)) +
       geom_point() + theme_bw()

g1 <- g + geom_smooth()
g2 <- g + geom_smooth(method="lm") + scale_x_log10()

gridExtra::grid.arrange(g1, g2, nrow=1)
```



```
cat("I will show you in Session 3 classroom, why using the log explicitly inside aes or coord_trans is n
```

I will show you in Session 3 classroom, why using the log explicitly inside aes or coord_trans is not optimal.