ggplot

Hadley Wickham

Outline

- Introduction to the data
- Introduction to ggplot
- Supplemental statistical summaries
- Iterating between graphics and models
- Graphical margins

Intro to data

- Response of trees to gypsy moth attack
- 5 genotypes of tree: Dan-2, Sau-2, Sau-3, Wau-1, Wau-2
- 2 treatments: NGM / GM
- 2 nutrient levels: low / high
- 5 reps

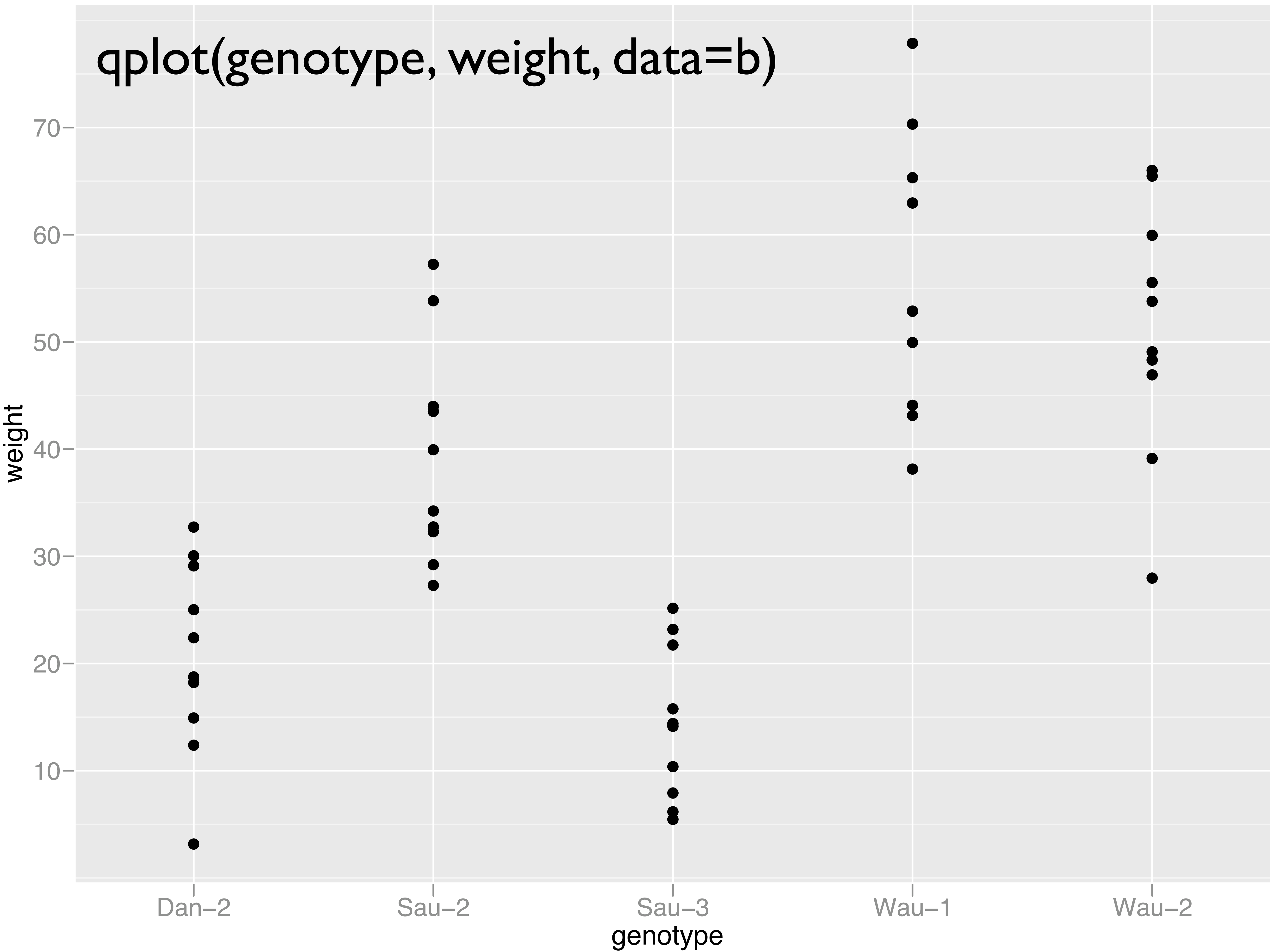
Intro to data

- 50 bugs (2 x 5 x 5)
  - Weight of living bugs
- 100 leaves (2 x 2 x 5 x 5)
  - Nitrogen (~ protein)
  - Salicylates
  - Tannins

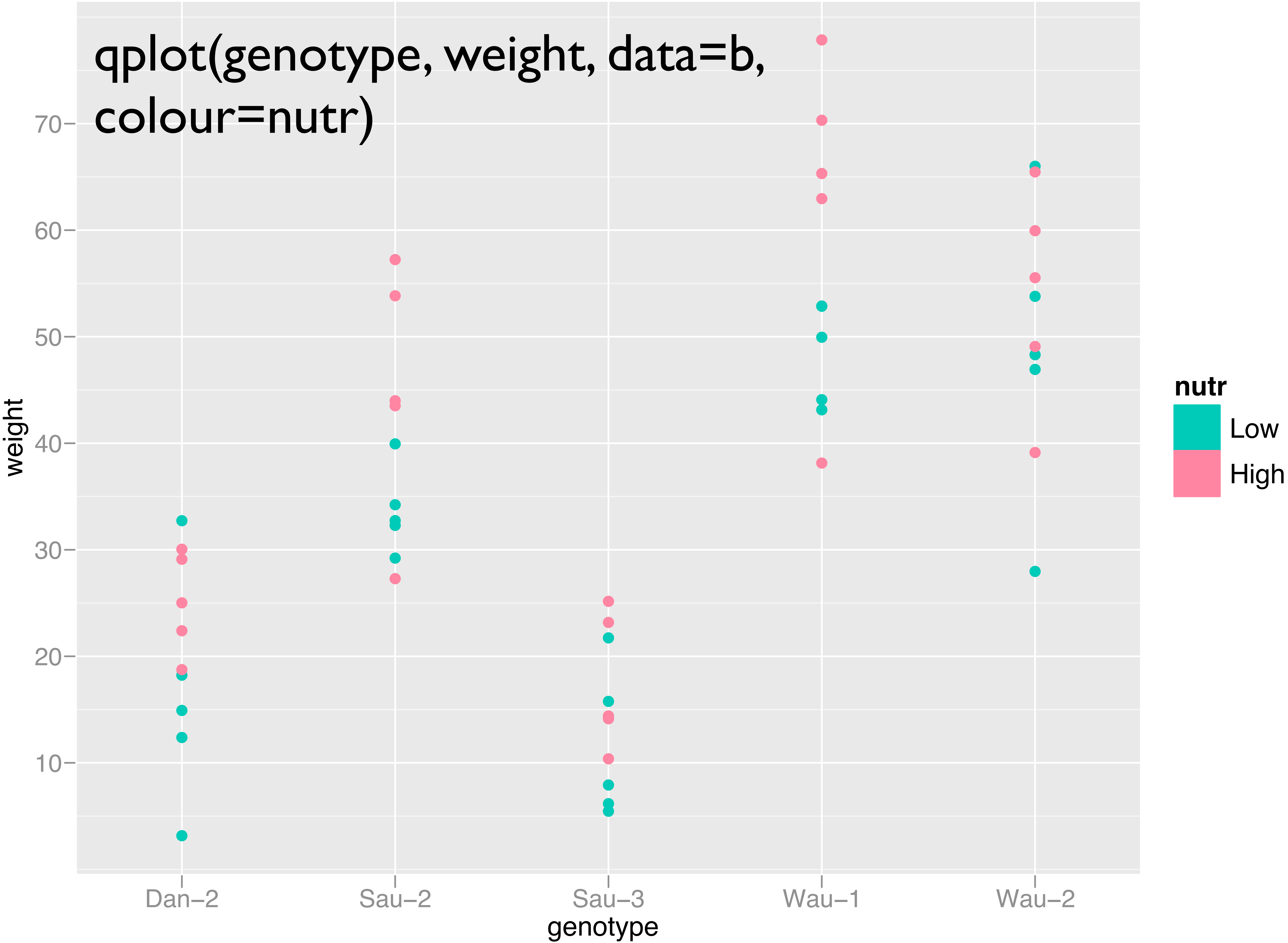
Intro to ggplot

- High-level package for creating statistical graphics - has a rich and comprehensive set of components, and a user friendly wrapper
- An implementation of "The Grammar of Graphics", Wilkinson 2005

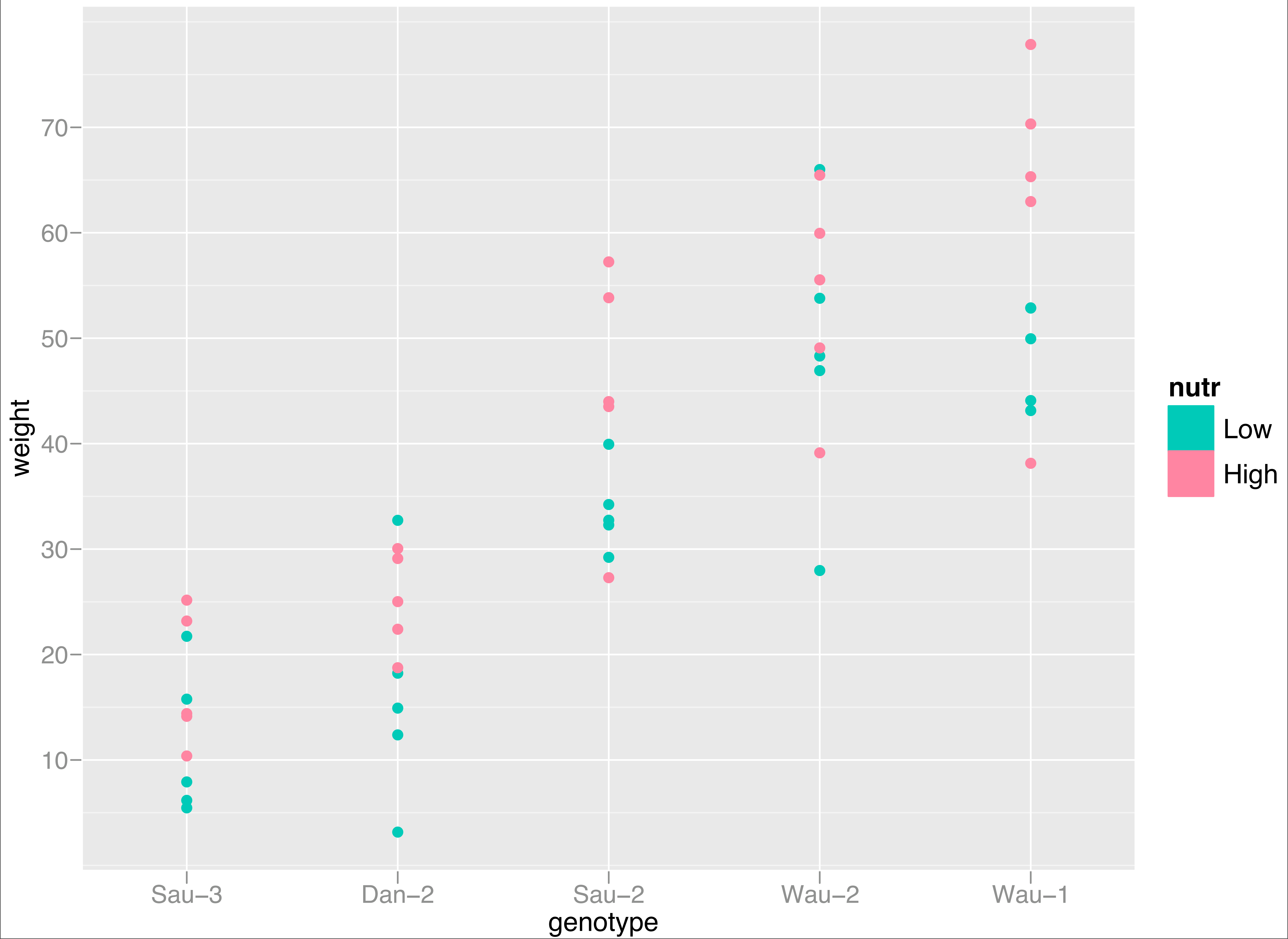
- Find out more at http://had.co.nz/ggplot2



```
qplot(genotype, weight, data=b)
```



```
qplot(genotype, weight, data=b,
colour=nutr)
```

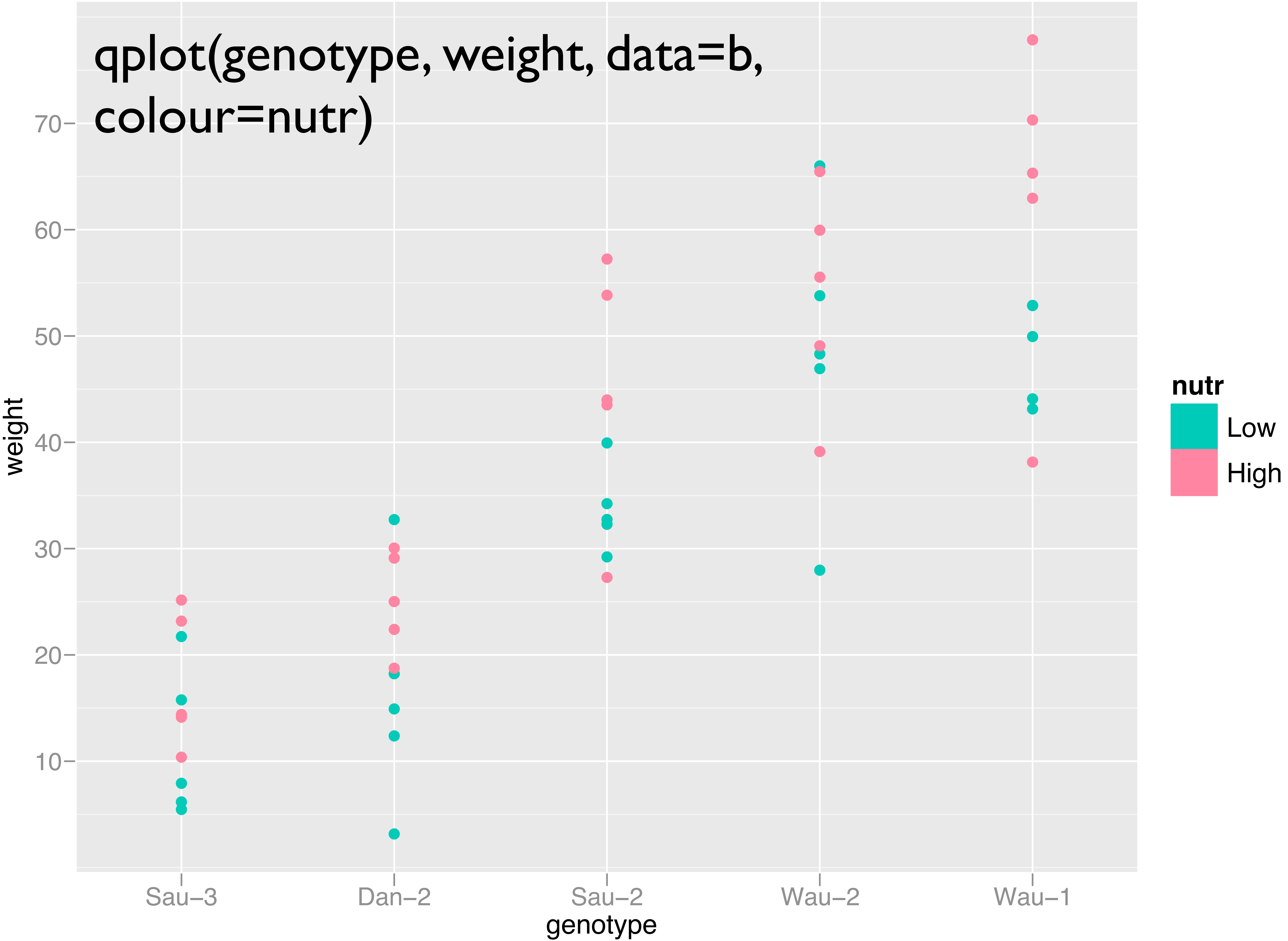

weight
70
60
50
40
30
20
10
Sau−3
Dan−2
Sau−2
Wau−2
Wau−1
genotype
nutr
Low
High

Comparing means

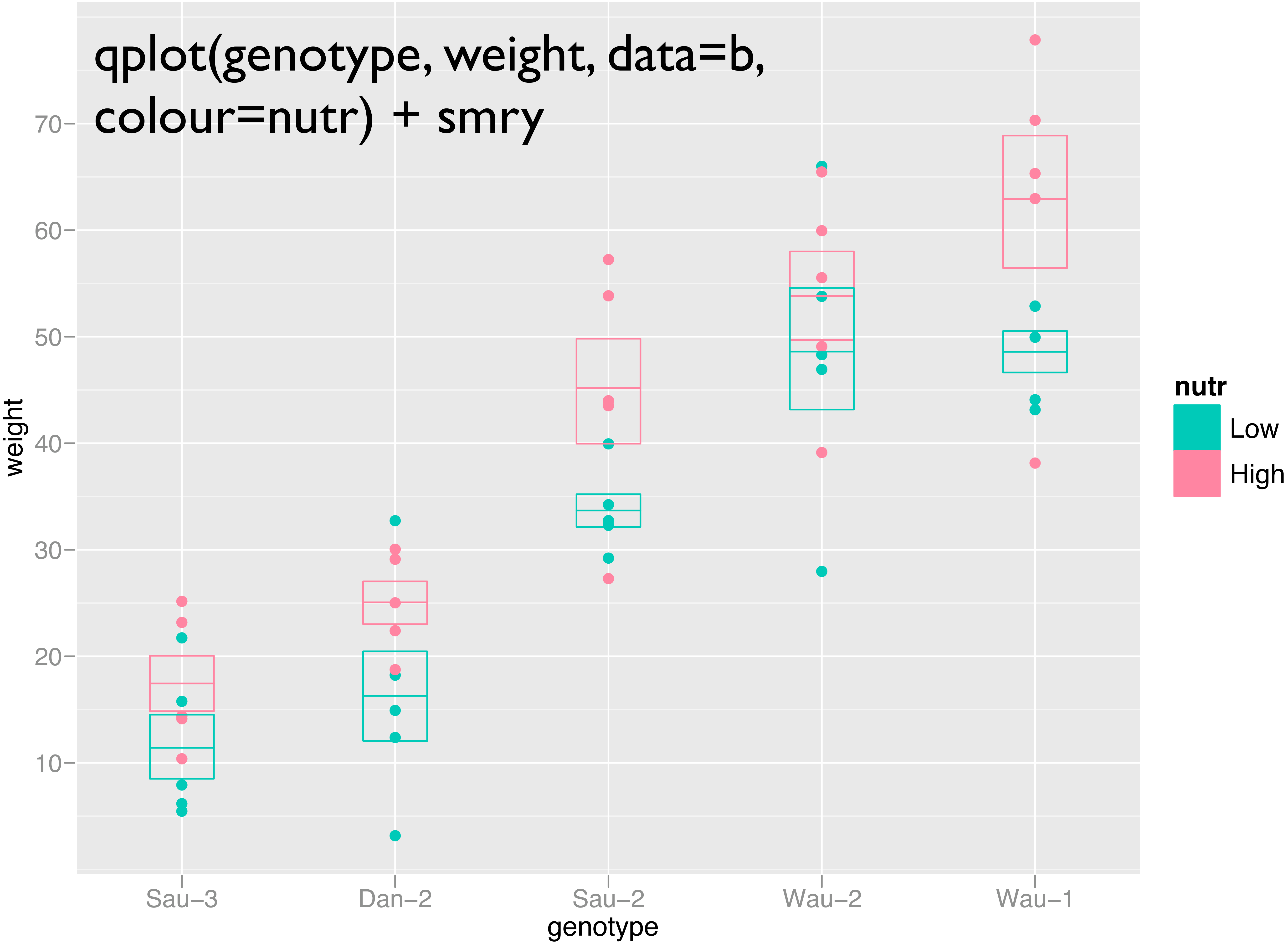
- Actually interested in comparing the means of the groups
- But hard to do visually - eyes naturally compare ranges
- What can we do? - Visual ANOVA

Supplements

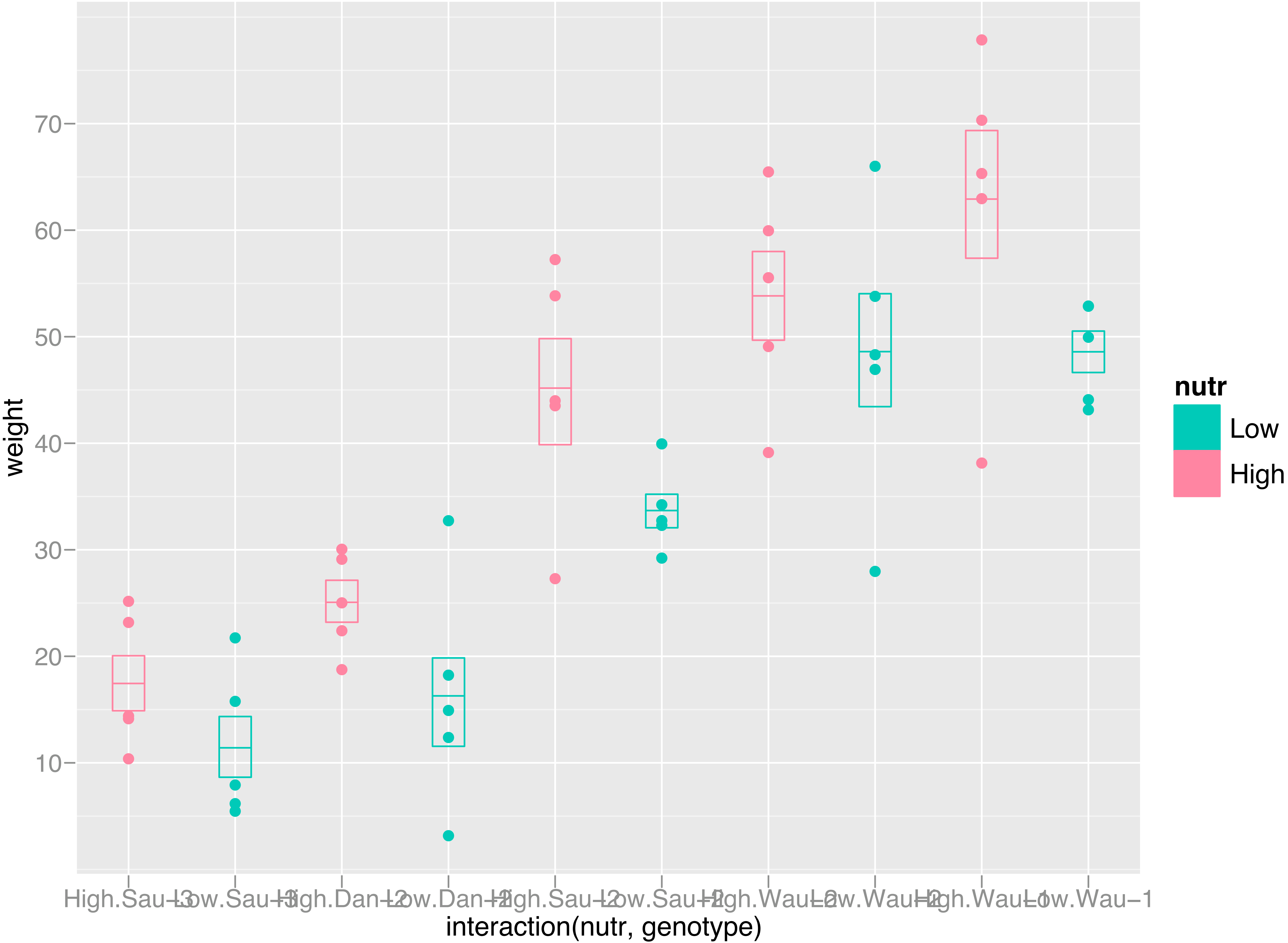
- smry <- stat_summary(fun=stat_mean_cl_boot, geom="crossbar", conf.int=0.68, width=0.3)
- Add another layer with summary statistics - mean and boot strap estimate of sem
- (from Frank Harrell's Hmisc package)



```
qplot(genotype, weight, data=b,
colour=nutr)
```



```
qplot(genotype, weight, data=b,
colour=nutr) + smry
```


weight
70
60
50
40
30
20
10
High.Sau−3
Low.Sau−3
High.Dan−2
Low.Dan−2
High.Sau−2
Low.Sau−2
High.Wau−2
Low.Wau−2
High.Wau−1
Low.Wau−1
interaction(nutr, genotype)
nutr
Low
High

Iterating graphics and modelling

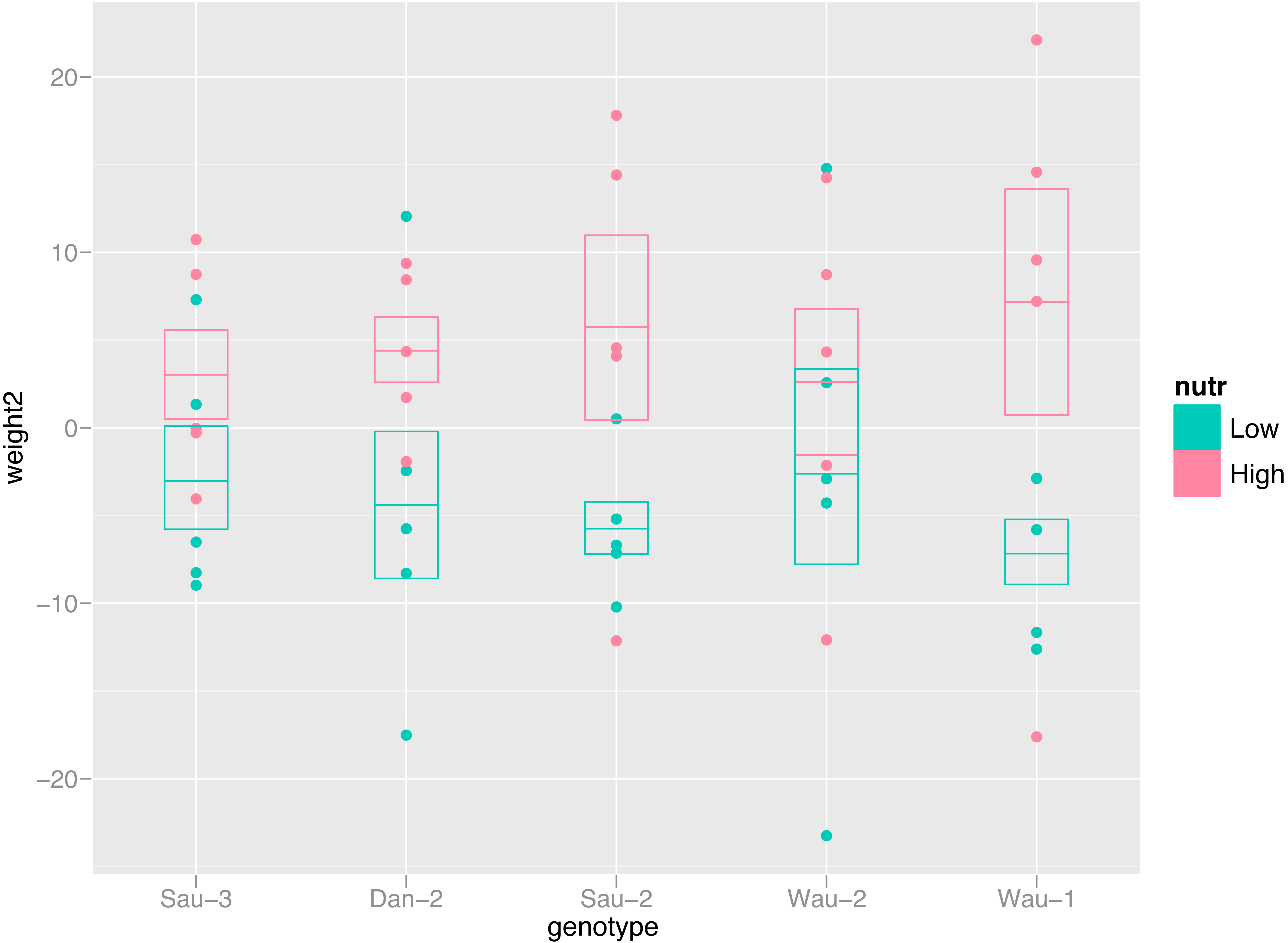
- Strong genotype effect
- Is there a nutr effect? Is there a nutr-genotype interaction?
- Hard to see from this plot - what if we remove the nutr main effect?
- (Old idea of Tukey's)

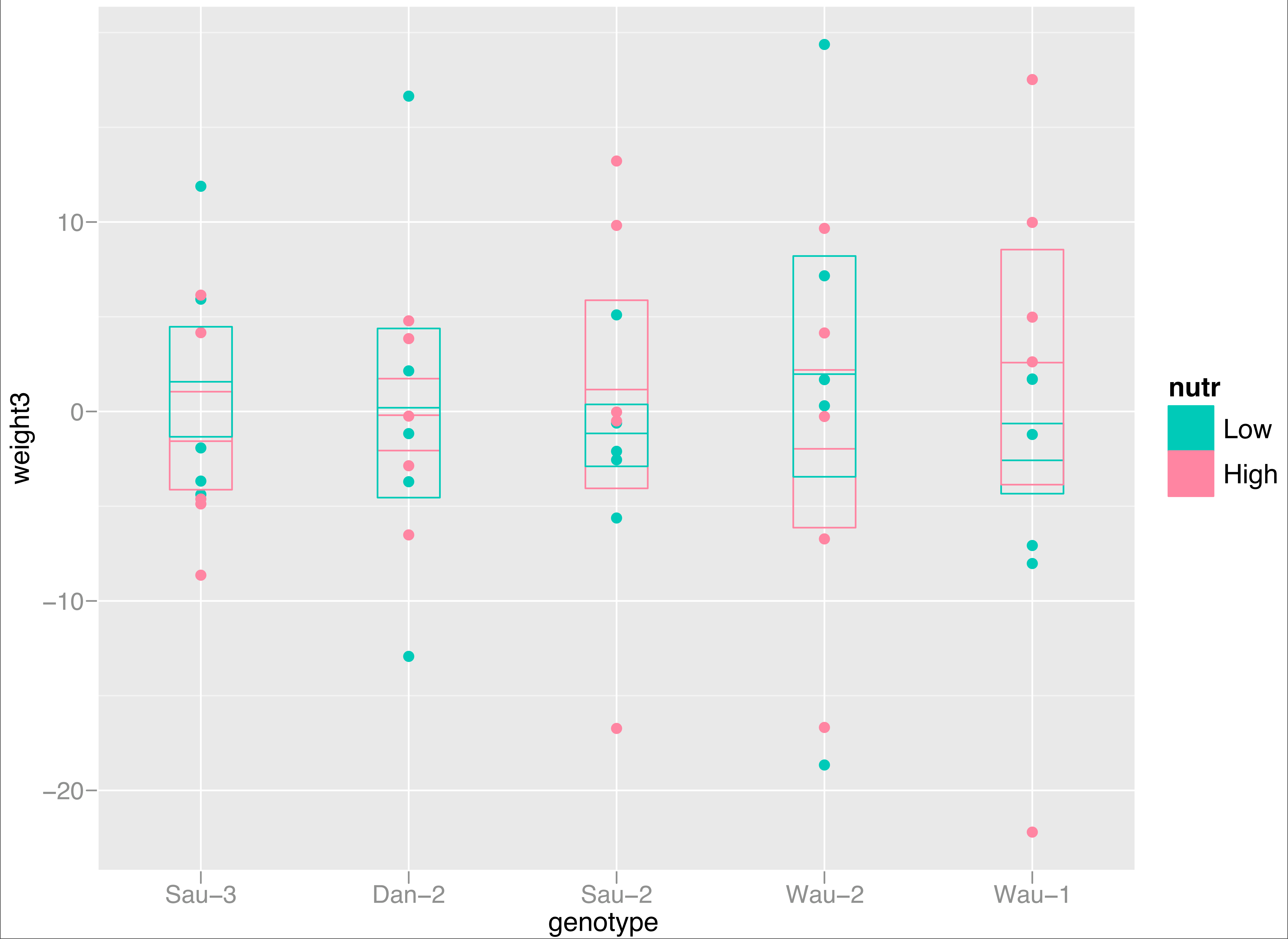
```
b$weight2 <- resid(lm(weight ~ genotype, data=b))

qplot(genotype, weight2, data=b, colour=nutr) +
smry


b$weight3 <- resid(lm(weight ~ genotype + nutr,
data=b))

qplot(genotype, weight3, data=b, colour=nutr) +
smry
```


20
10
0
−10
−20
weight2
Sau−3
Dan−2
Sau−2
Wau−2
Wau−1
genotype
nutr
Low
High


10
0
−10
−20
weight3
Sau−3
Dan−2
Sau−2
Wau−2
Wau−1
genotype
nutr
Low
High

anova(lm(weight ~ genotype * nutr, data=b))

```
              Df Sum Sq Mean Sq F value  Pr(>F)    
genotype       4  13331    3333   36.22 8.4e-13 ***
nutr           1   1053    1053   11.44  0.0016 ** 
genotype:nutr  4    144      36    0.39  0.8141    
Residuals     40   3681      92                    
```

```
p <- qplot(genotype, weight, data=b,
colour=nutr) + smry


p

p + aes(y = weight2)

p + aes(y = weight3)
```

Graphical margins

- Often interested in marginal, as well as conditional, relationships
- Or comparing one subset to the whole, rather than to other subsets
- Like in contingency table, we often want to see margins as well

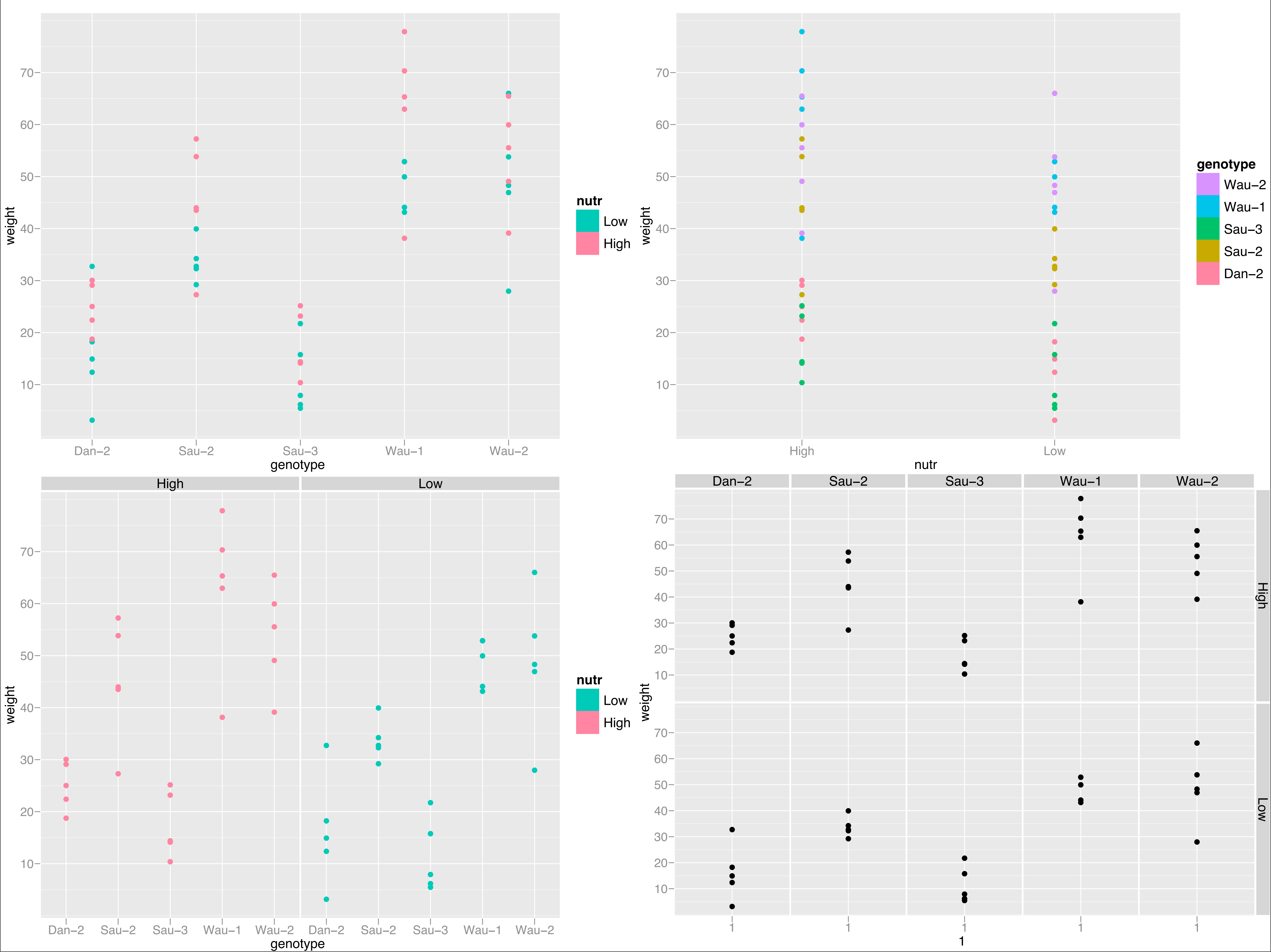

Dan−2
Sau−2
Sau−3
Wau−1
Wau−2
High
Low
n
4.5
4.0
3.5
3.0
2.5
2.0
GM
NGM
trt
NGM
GM


Dan–2
Sau–2
Sau–3
Wau–1
Wau–2
(all)
High
Low
(all)
4.5
4.0
3.5
3.0
2.5
2.0
n
GM
NGM
trt
trt
NGM
GM

Arranging plots

- Facilitate comparisons of interest
- Small differences need to be closer together (big difference can be far apart)

- Connections to model?


weight
10
20
30
40
50
60
70
Dan−2
Sau−2
Sau−3
Wau−1
Wau−2
genotype
nutr
Low
High
weight
High
Low
nutr
genotype
Wau−2
Wau−1
Sau−3
Sau−2
Dan−2
High
Low
weight
Dan−2
Sau−2
Sau−3
Wau−1
Wau−2
Dan−2
Sau−2
Sau−3
Wau−1
Wau−2
genotype
nutr
Low
High
Dan−2
Sau−2
Sau−3
Wau−1
Wau−2
weight
High
Low
1
1
1
1
1
1

Conclusions

- Three useful graphical techniques:
  - Supplement with statistical summaries
  - Iterate graphics and modelling
  - Graphical margins
- Graphics packages should get out of your way, and let you focus on creating the graphics you need

had.co.nz/ggplot2