Statistics for Life Sciences - Model building

Nathan Harmston

2022-04-07

Model building

do statistical hypotheses really match up with how we really *do* science?

multiple hypothesis testing ????

p-values / hypothesis testing focuses on the arbitrary $p < 0.05$ cutoff and ignores effect sizes and how good a model is R^2

often we have a large set of candidate predictor/independent variables from which we try to identify the best predictors to include in our regression model

so we have lots of alternative models to explain our dataset .....

caveat: if you don't include that variable/model you can never evaluate it

all models are wrong, some are useful

George Box

Lets build a some models and make some predictions

Birth Weight and Smoking

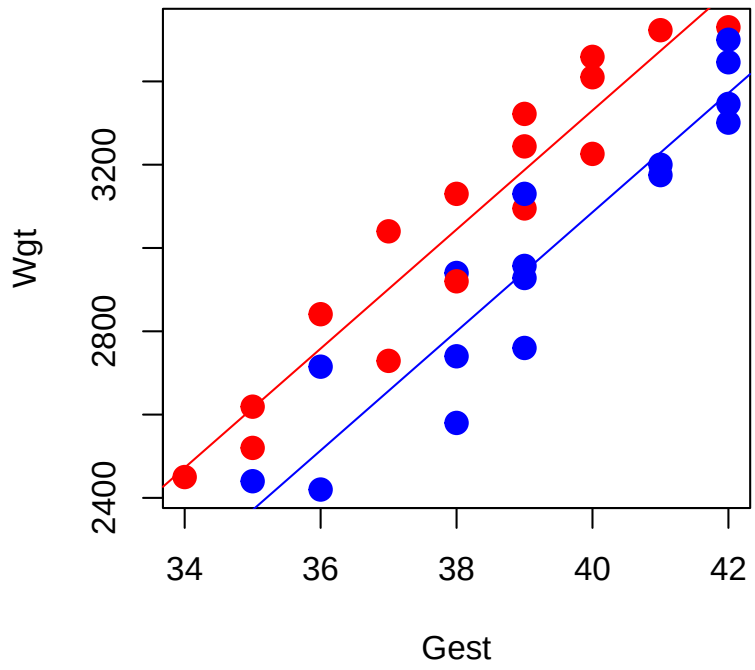
- birth weight (Weight) in grams of baby
- Smoking status (Smoke) of mother (yes or no)
- length of gestation (Gest) in weeks

Daniel, (1999)

$$Wgt = -2389.573 + 143\,Gest - 244.544\,Smoke \tag{1}$$

```
?predict
```

Predicting from a model - linear regression



Gest	Smoke	Wgt
40	Smoke	3089.894
37	Not Smoke	2905.137

Predicting from a model - logistic regression

Ovarian cancer

- logistic regression tries to predict probabilities

Age	p	resid.ds
30	0.21	no
65	0.71	yes
100	0.96	yes

Predicting from a model - Cox proportional hazards regression

- predicts the log hazards ratio

rx	resid.ds	age_group	
A	yes	old	1.446982
B	no	young	-2.996760

we need ways to ...

- evaluate which model is best
- determine which independent/predictor variables to include
- evaluate the performance of our model

Model selection - Comparing different models

different combinations of independent/predictor variables

select the model which best explains the variance in the dependent variables with the fewest number of independent variables

pick a simple model over a more complex one

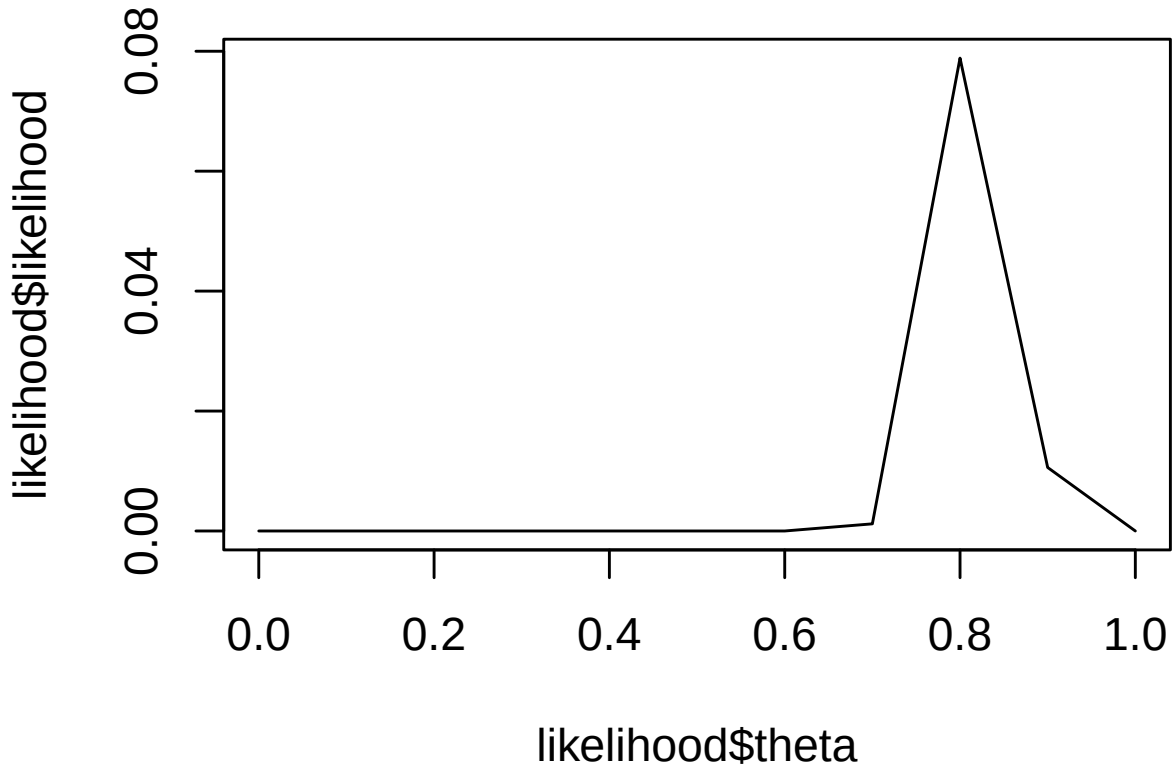
penalise complex models

how can you compare different candidate models and select the best one / least bad / most plausible?

- R^2 or adjusted R^2
- AIC
- BIC
- Mallows C_p

Likelihood

so we had this likelihood function that described the $P(D|\theta)$ after getting 83 heads from 100 coin tosses



the maximum is at 0.8 - our maximum likelihood estimate of θ is 0.8

- really our maximum likelihood estimator of θ is 0.83

log-likelihood

- a lot of the time people talk about a log-likelihood or LL
- the natural logarithm of the likelihood.
- used because sum of logs of x is the same as the products of x
- computational accuracy

why a simple model vs a complex model?

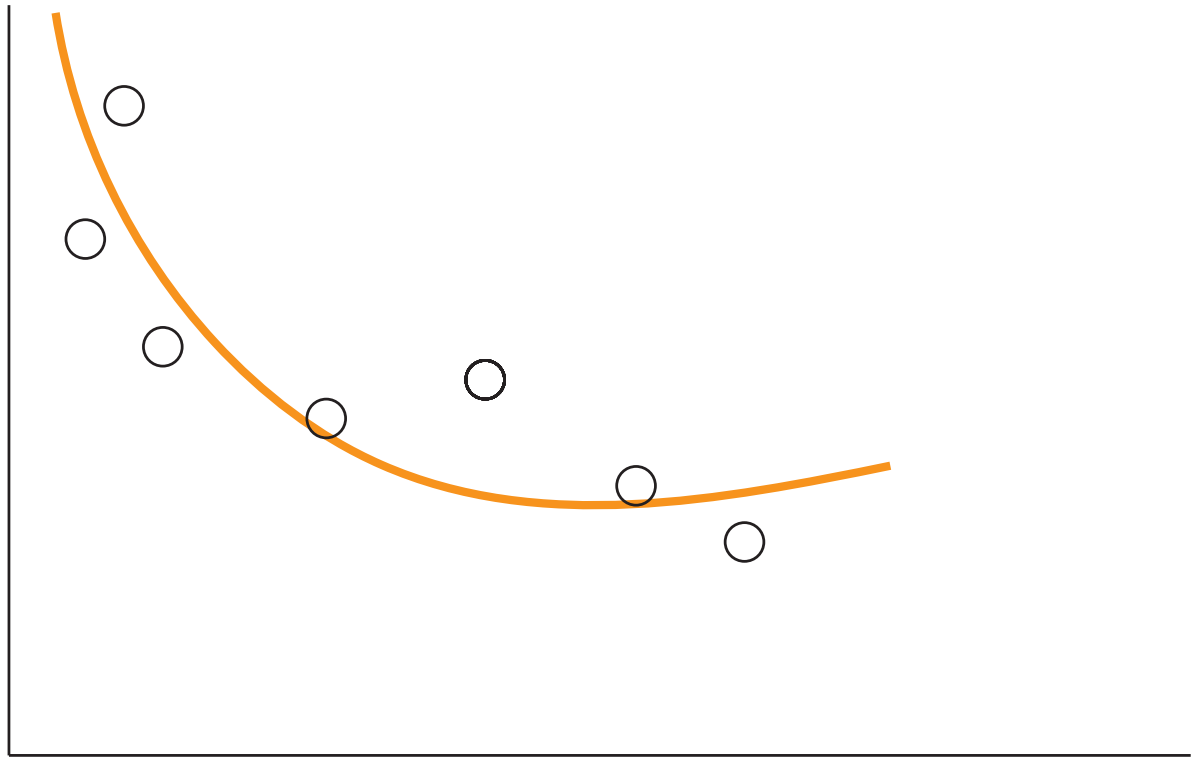
adding variables to a model will never decrease the R^2
more variables more problems - interactions, collinearity

overfitting

interpretability

Overfitting

the more complex the model the better you can fit the training data

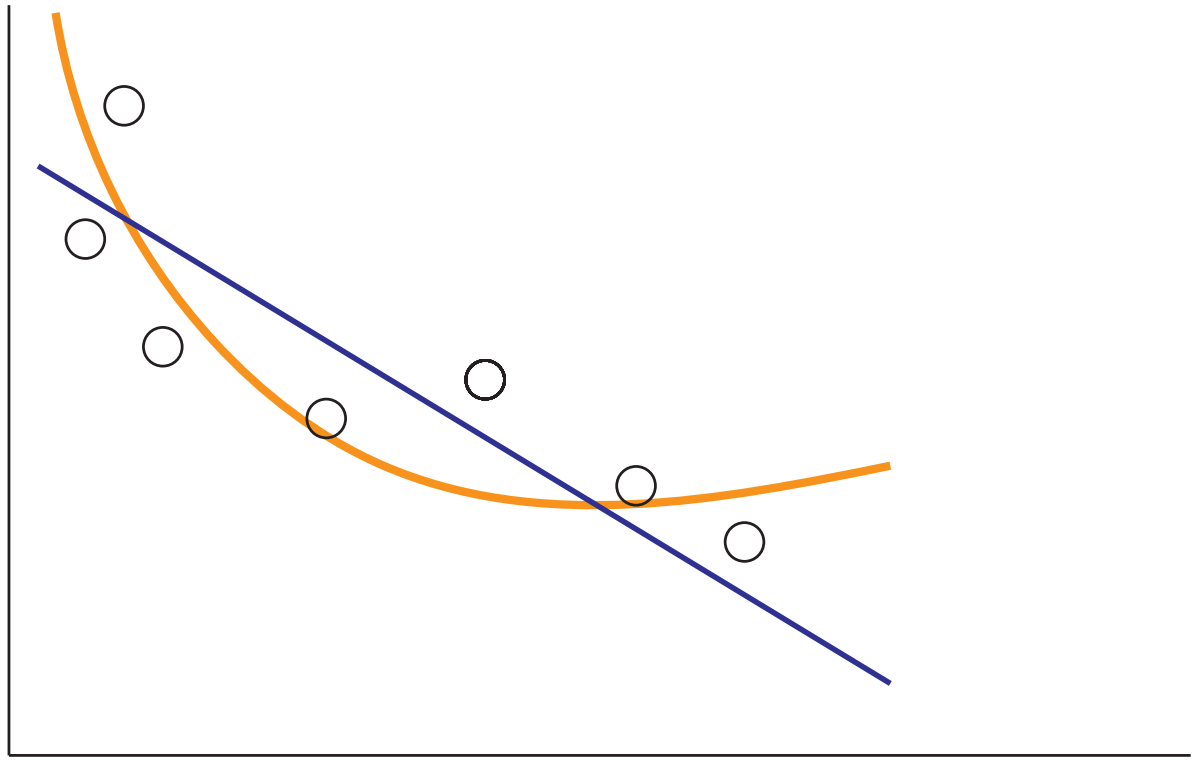


"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk"

von Neumann

Overfitting

the more complex the model the better you can fit the training data

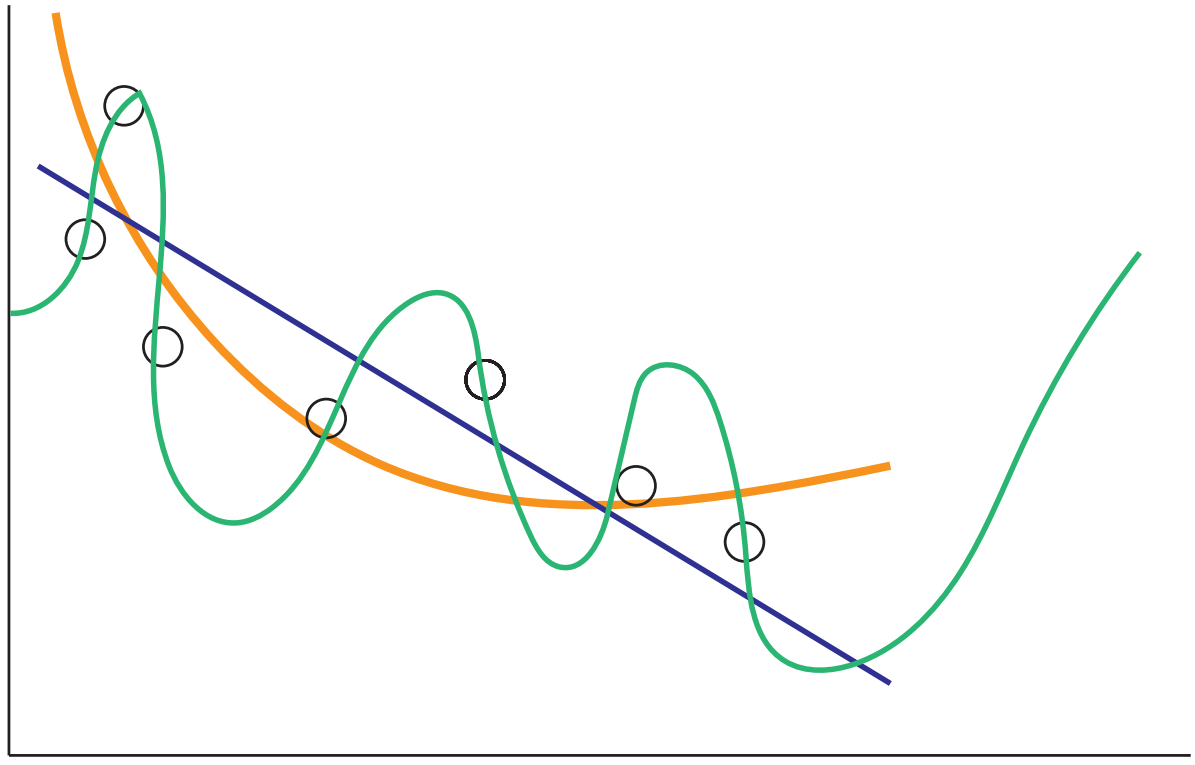


"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk"

von Neumann

Overfitting

the more complex the model the better you can fit the training data



"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk"

von Neumann

Akaike information criterion

AIC

- sample size (n)
- maximum log-likelihood
  - how well the model captures the variance of the dependent variable
- number of parameters (K) in the model + error term

$$AIC = -2 * logLik(m) + 2 * K \quad (2)$$

Akaike et al 1973

Akaike information criterion

so if two models explain the variance in the dependent just as well (i.e. have the same log likelihood) the one with the lowest number of parameters will have the lowest AIC and be preferred

AIC only means something compared to another AIC

```
?AIC
```

AICc - Akaike information criterion - corrected

AICc is corrected for small sample size
AICc converges to AIC as sample size (n) increases

$$AICc = -2 * logLik(m) + 2 * K + \frac{2K(K+1)}{n-k-1} \tag{3}$$

```
library('MuMIn')
?AICc
```

Akaike information criterion

Bayesian information criterion

function of posterior probability that the model is TRUE lower BIC - better model - more likely to be true / better

related to AIC - but has a stronger penalty

$$BIC = -2 * logLik(m) + Klog(n) \tag{4}$$

```
?BIC
```

Model building techniques

so I've collected lots of data on the relationship between a dependent variable of interest and lots of independent variables

- forward regression
- backwards regression
- step regression
- best subsets regression

Forward selection

starts with no predictors in the model

iteratively adds the predictors which contribute the most to explaining the dependent variable

stops when the improvement is no longer statistically significant.

Backward selection/elimination

starts with all predictors in the model

iteratively remove the predictors which contribute the least to explaining the dependent variable

stops when the improvement is no longer statistically significant.

Stepwise selection - sequential replacement

starts with no predictors in the model

a combination of forward and backward selections.

All subsets

try every single permutation of models - brute force

test everything

Picking the set of predictors

forward, backwards, step, all subsets

- no guarantee you'll get the best model

```
  ?step
?dredge
```

"Let the computer find out" is a poor strategy and usually reflects the fact that the researcher did not bother to think clearly about the problem of interest and its scientific setting

(Burnham and Anderson, 2002).

Histone modification levels are predictive for gene expression

Rosa Karlić[a,b,1], Ho-Ryun Chung[a,1,2], Julia Lasserre[a], Kristian Vlahovičekb,c], and Martin Vingron[a]

[a]Max-Planck-Institut für Molekulare Genetik, Department of Computational Molecular Biology, Ihnestraße 73, 14195 Berlin, Germany; [b]Bioinformatics Group, Division of Biology, Faculty of Science, Zagreb University, Horvatovac 102a, 10000 Zagreb, Croatia; and [c]Department of Informatics, University of Oslo, P.O. Box 1080, Blindern, NO-0316 Oslo, Norway

Histone modification levels are predictive for gene expression

Karlic *et al.* 2010

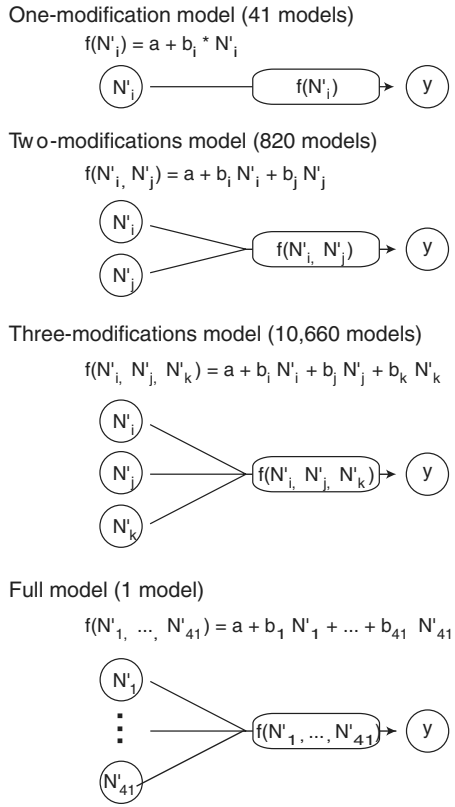


Fig. 1. Modeling framework. Models are equations that linearly relate the levels of histone modifications to the measured expression value. $\boldsymbol{N}'_i$ corresponds to a vector of length L (the number of promoters), where the components are the transformed levels of a histone modification i ($\boldsymbol{N}'_i = \log(\boldsymbol{N}_i + \alpha_i)$, with $\boldsymbol{N}_i$ representing the number of tags in each promoter), a is the y intercept , and the b_i to the slope associated with $\boldsymbol{N}'_i$. $\boldsymbol{y}$ denotes a vector of length L whose components are the expression values. In the one-modification models, i can be any of the 39 modifications or two control IgG antibodies. In the two-modifications models, i and j are chosen to cover all combinations of two modifications without repetition. In the three-modifications models, i, j, and k are chosen to cover all combinations of three modifications without repetition. The full model incorporates all 41 variables.

Histone modification levels are predictive for gene expression

Karlic *et al.* 2010

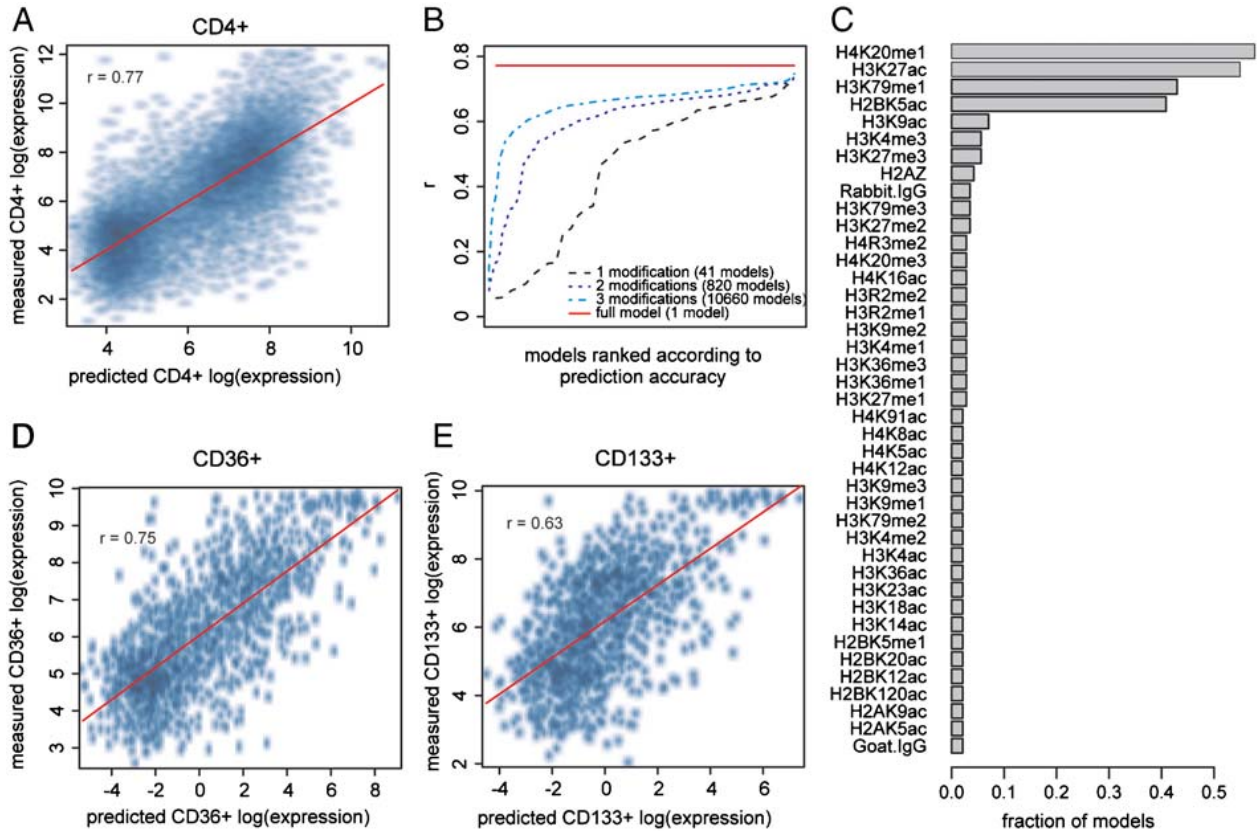


Fig. 2. Quantitative relationship between histone modification levels and expression. (*A*) Scatterplot with the predicted expression value in CD4+ T-cells using the full linear model on the *x* axis and the measured expression value in CD4+ T-cells on the *y* axis. The shades of blue indicate the density of points; the darker color, the more points. The red line indicates the linear fit between predicted and measured expression ($y = 0.99x + 0.02$), which are highly correlated ($r = 0.77$), indicating a quantitative relationship between levels of histone modifications at the promoter and gene expression levels. (*B*) Comparison of prediction accuracy between all possible one-modification, two-modifications, three-modifications models, and the full model for CD4+ T-cells. Models are sorted by ascending prediction accuracy along the *x* axis. The best models using only a small subset of modifications almost reach the prediction accuracy of the full linear model. (*C*) Bar plot showing the frequency of appearance of different histone modifications in best scoring three-modifications models (142 models) for CD4+ T-cells. Best scoring models are defined as reaching at least 95% of prediction accuracy of the full linear model. (*D*, *E*) Expression values of genes, which were at least 5-fold up or down regulated in CD36+ and CD133+ cells with respect to CD4+ T-cells, predicted using model parameters trained on data from CD4+ T-cells. The predicted and measured expression values are highly correlated for both CD36+ (*D*) ($r = 0.75$; 1,412 genes) and CD133+ (*E*) ($r = 0.63$; 1,243 genes) cells. The equations of the regression line for both CD36+ and CD133+ cells ($y = 0.43x + 6.04$ and $y = 0.53x + 6.17$, respectively) show a high value of the intercept and a slope different from one due to the fact that the levels of the histone modifications were not normalized across cell types.

Histone modification levels are predictive for gene expression

Karlic *et al.* 2010

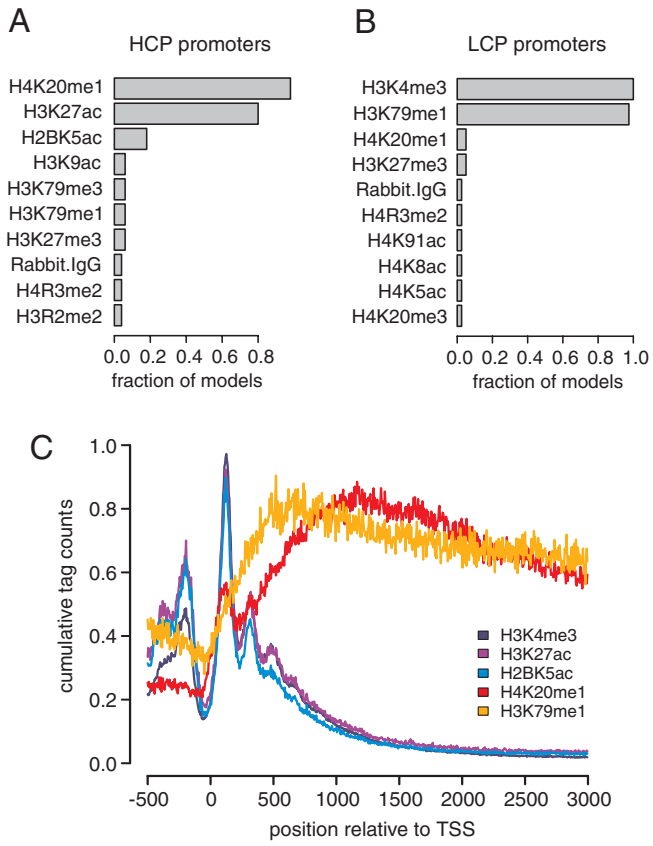


Fig. 3. Differences in transcriptional regulation between HCPs and LCPs. (*A*, *B*) Bar plots showing the frequency of appearance of different histone modifications in best scoring three-modifications models for HCPs (*B*) (50 models) and LCPs (*C*) (40 models) in CD4+ T-cells. Best scoring models are defined as reaching at least 95% of prediction accuracy of the full model trained on HCPs and LCPs, respectively. Only the top ten modifications are depicted. (*C*) Normalized cumulative tag counts in the region of −500 base pairs to 3,000 base pairs surrounding the transcription start site of RefSeq genes in CD4+ T-cells for the five important modifications identified by our analysis.

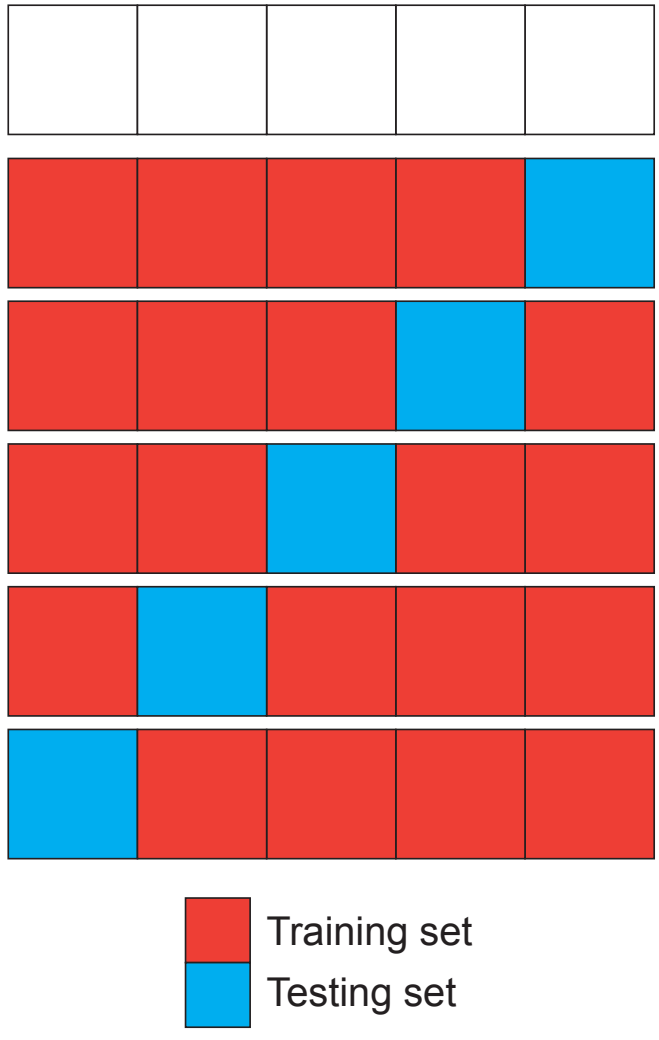
Overfitting

We want to get good generalisation performance and avoid over-fitting based on the data

Does it work well on something it hasn't seen? Is it generalisable?

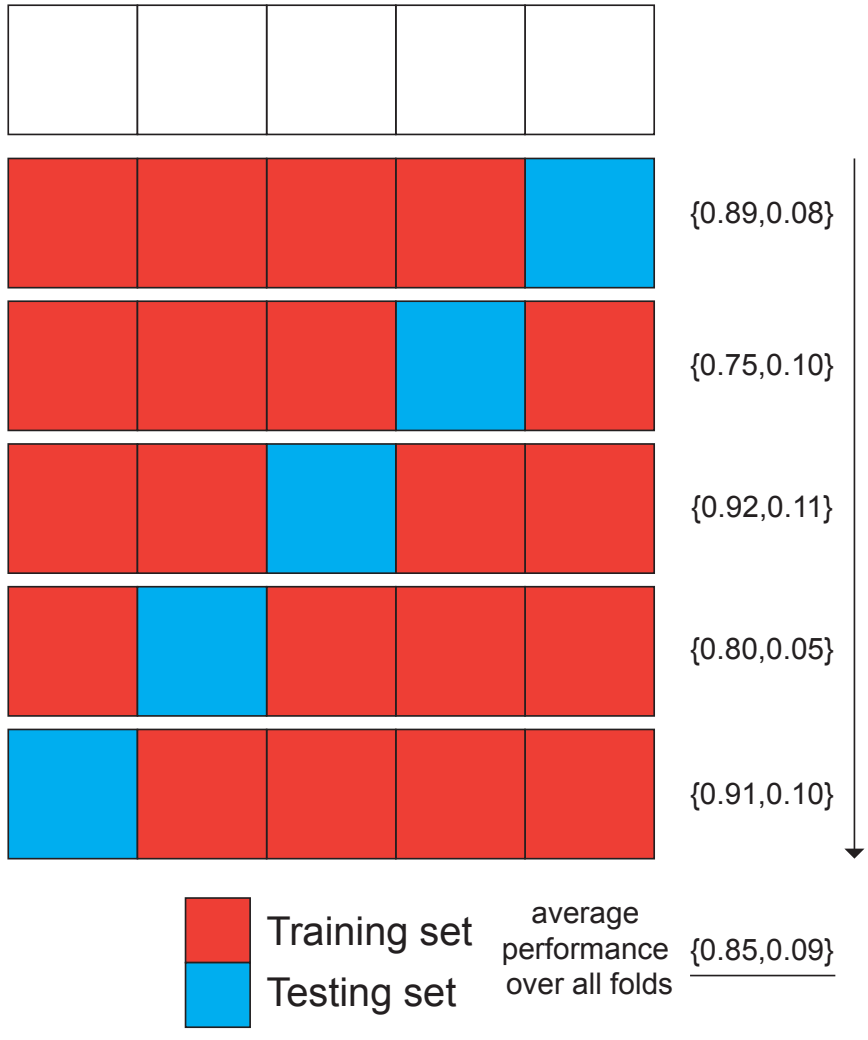
Cross validation

5 fold cross validation



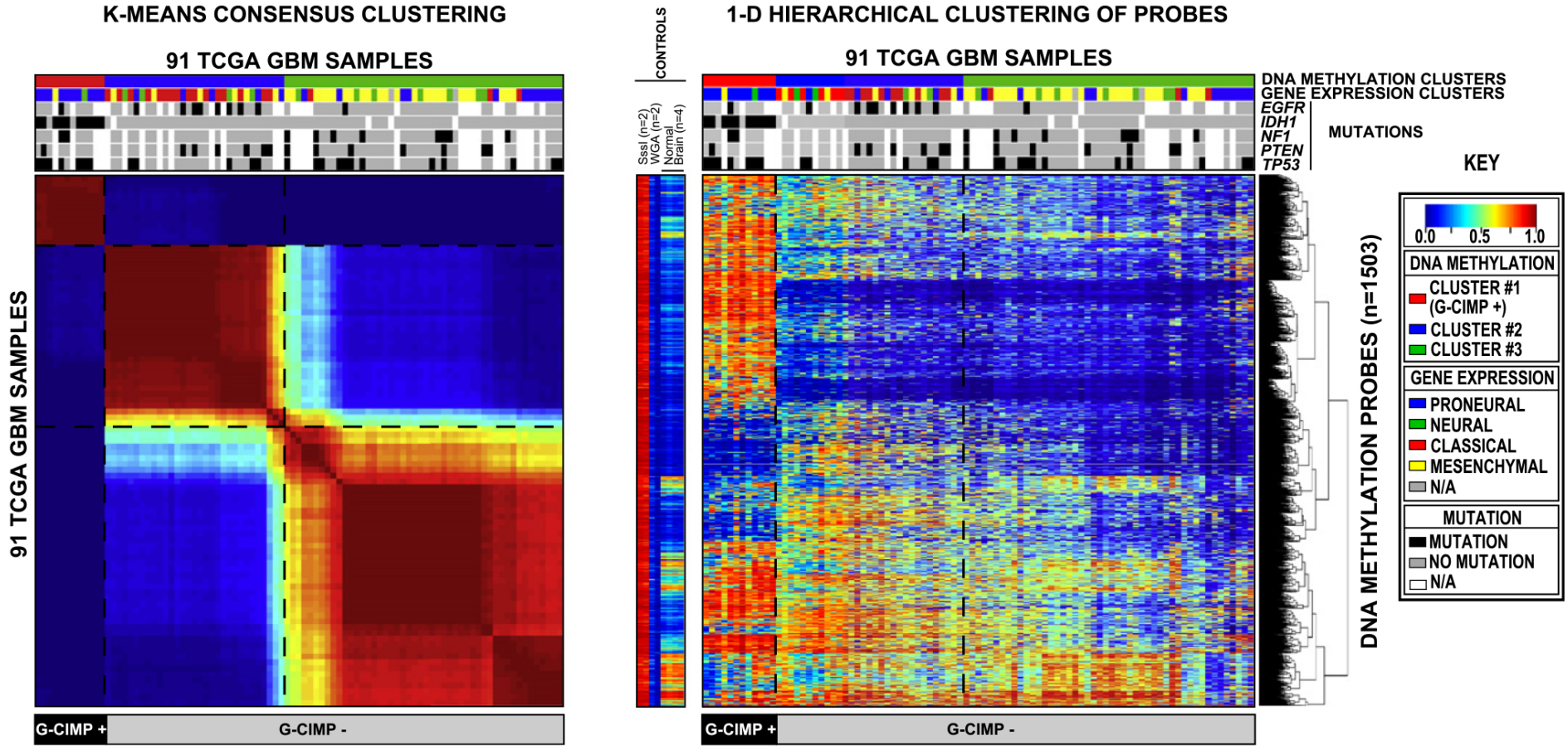
Cross validation

5 fold cross validation



CIMP vs non-CIMP

CpG island methylator phenotype (CIMP): cancer-specific CpG island hypermethylation of a subset of genes in a subset of tumours



Noushmehr *et al.* 2010

On Tuesday

- Answers to problemset 2
- Summary
- Why Most Published Research Findings Are False