BMEG-3105 Data analytics for personalized genomics and precision medicine

Lecture 15: Model Overfitting and Multi-Omics

Name: Lee KinHei SID : 1155158901

1. Model Overfitting

1.1 Definition

Overfitting occurs when the model cannot generalize, and it fits too closely to the training dataset instead.

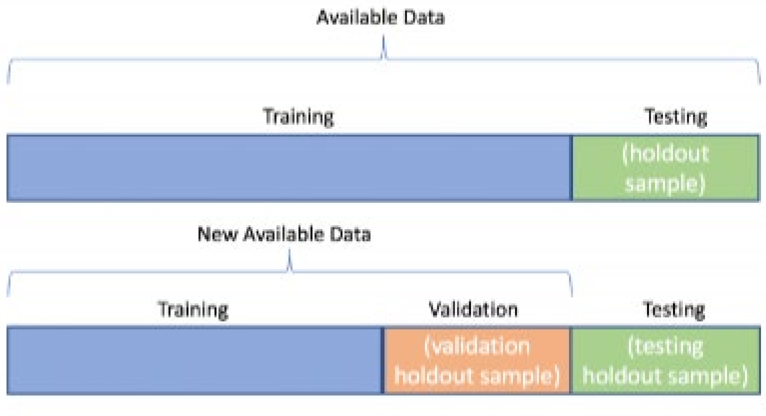
1.2 Effect

The model will fail to fit additional data or predict future observations reliably.

1.3 Evaluation (How we know)

1.3.1 Train-validation-test split

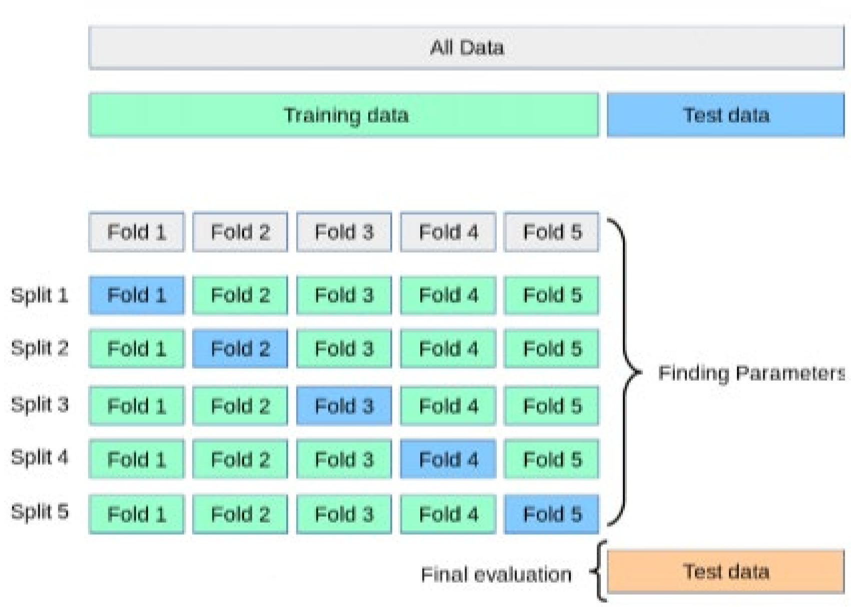
For all available data, we can separate the data like 70% for training and validation and 30% for testing. For the testing data, we will only use it to test our model after training to find out if there are any overfitting issues occurring.



For example, if we got a very high accuracy in training dataset but comparatively low accuracy in testing dataset, we know the overfitting problem may occur.

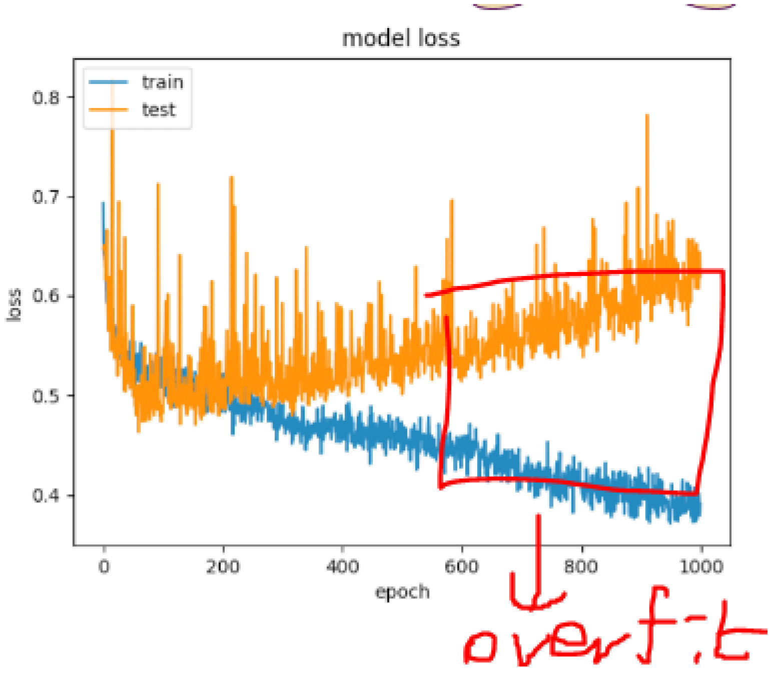
1.3.2 Cross validation

It mitigates the problem of using the same validation dataset in train-validation-test split. Instead, it splits the data into multiple folds and each time, it uses one fold to evaluate the model which makes the validation process more fair and general. But for each split, we need to have a new model training on that, so it is computationally expensive to do cross validation.



1.3.3 Model Loss

By observing the graph recording the training loss and testing loss in each epoch, we may observe the problem of overfitting by discerning the trends of two loss. For example, overfitting may happen when the testing loss increases while the training loss keeps decreasing. Also, knowing the model loss, we may also apply early stop to stop the model training before it overfit.



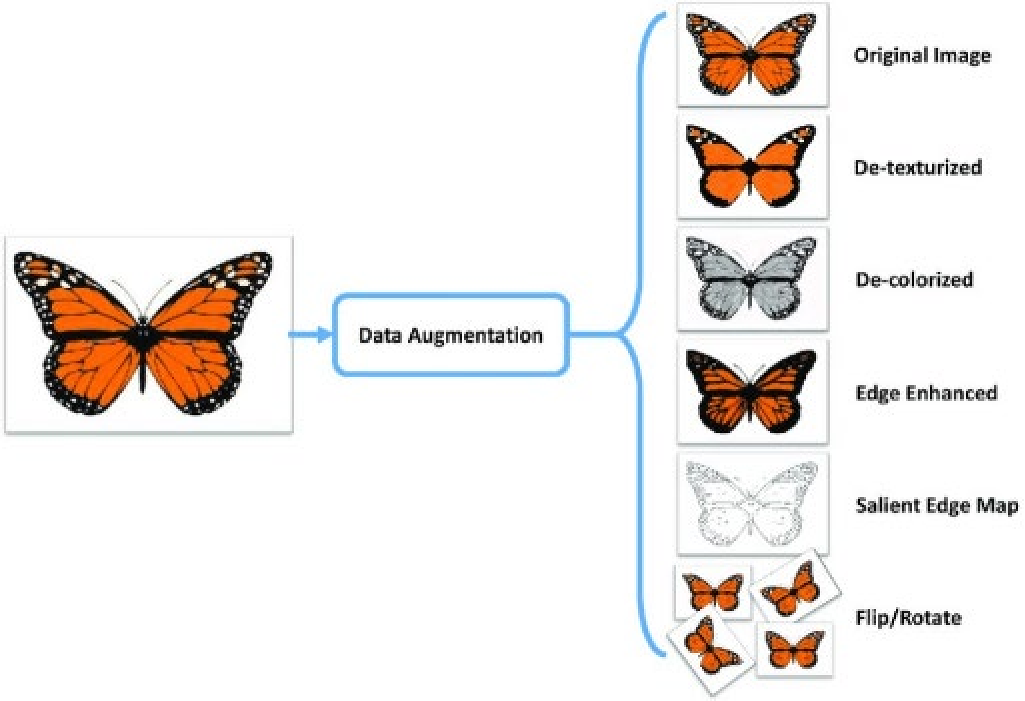
1.3.4 Performance Evaluation

By evaluating the model through different performance evaluation metrics like precision and recall, we can discern the problem of overfitting. For example, in an imbalanced dataset, the model may correctly classify most of the data which is the majority of the dataset but very poorly on the minority. In such cases, indeed, the training loss and testing loss may also be low but there is a problem that the model overfit the majority. Thus, using different performance evaluation metrics can help discover the problem.

1.4 Potential Solution

1.4.1 Data

Small data size may not reflect the true distribution which may mislead the model to learn a wrong distribution and reduce its generalizability. Considering that, the most obvious way will be to increase the size. Other than that, sometimes we may not be able to collect more data, or it is too expensive to do so, we can have data augmentation to increase the variety of the data.



1.4.2 Model

For insufficient data in a particular task, we can actually use transfer learning to enhance the performance of the model on that task. For transfer learning, we may pretrain the model with a large dataset like ImageNet, to gain some basic knowledge and re-use the knowledge in related tasks to boost the performance. For instance, we may have a

large dataset to train the model to recognize different cars and this model can be transferred to recognize trucks as we have few images of trucks.

For a very complex model, we can also try to reduce the model complexity to see if overfitting issues still occur.

1.4.3 Connectivity

The connection between hidden layers in the model may be too strong that it may learn unimportant features. In this case, we can adopt a dropout strategy, to discard some connections between nodes.

1.4.4 Parameter value

Sometimes, some parameter values may be too large and may affect the contribution of other features. Thus, we may panelize the model for overly focusing on one feature which means that we want to have more evenly distributed weights across the features.

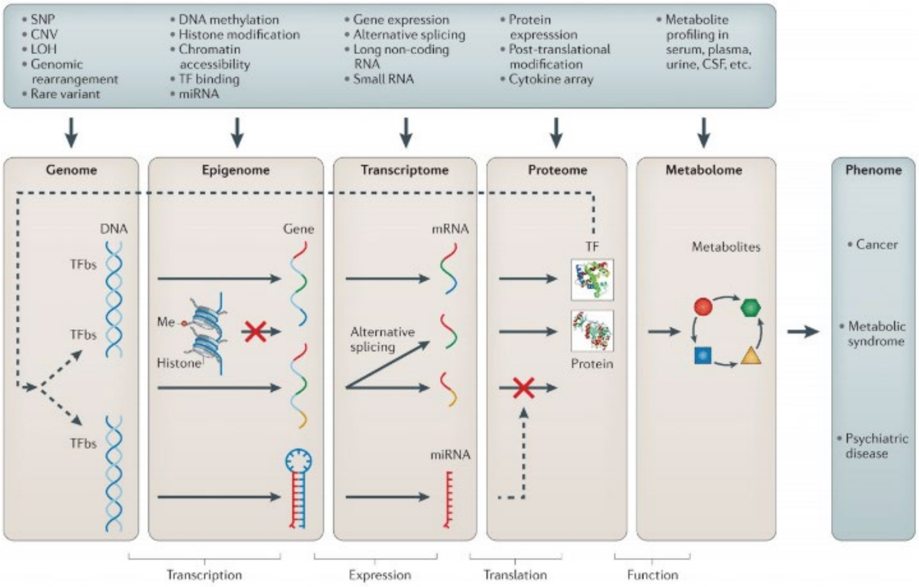
1.4.5 Training time

By observing the loss graph, if the test loss starts increasing while training loss keeps decreasing after a long time of training, we should stop the model and train to find a suitable epoch number from the graph.

2. Multi-Omics

2.1 What is omics?

Omics aims at the collective characterization and quantification of pools of biological molecules as shown below that translate into the structure, function, and dynamics of an organism or organisms. We should consider all omics together as a whole as they closely interact with each others.



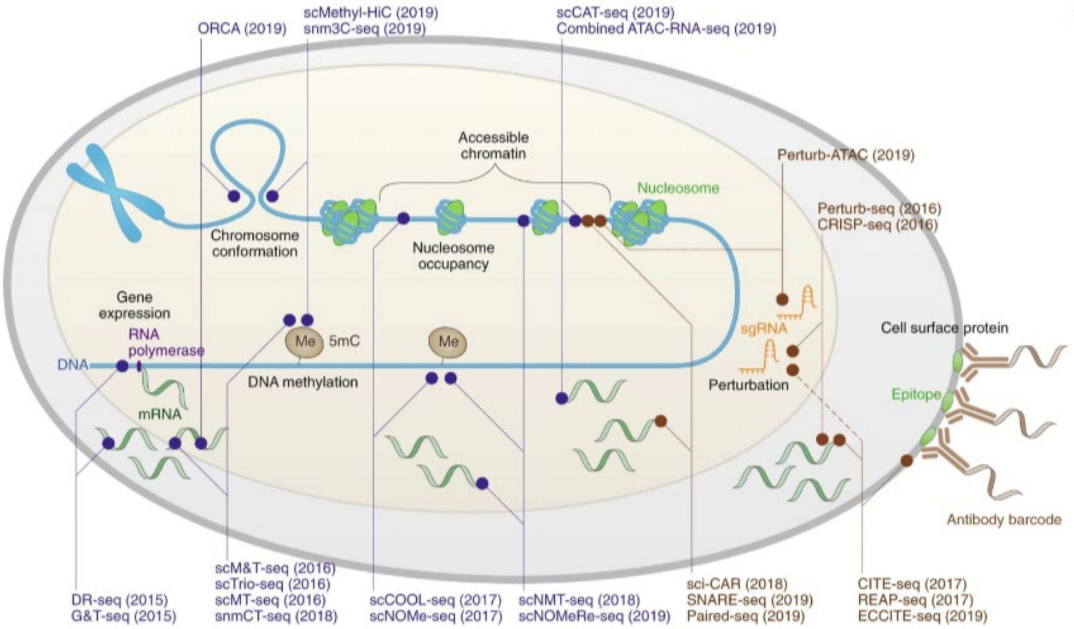
Transcriptome: Transcribe the DNA into RNA

Alternative splicing: Modify the immature mRNA into mature mRNA

Protein: Perform different functions on our body including regulate the transcription in our body

Single-cell multi-omics

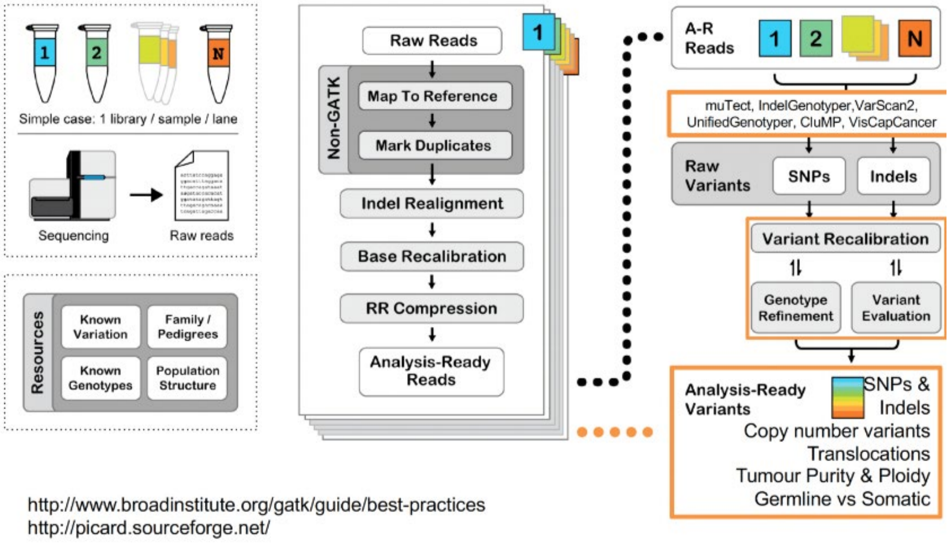
There are multiple types of cells which are very hard for us to analyze, and we want to separate each cell for exhaustive investigation.



Sorry I really don't have lots of knowledge in biology :/

2.2 Genome (next lecture) :>

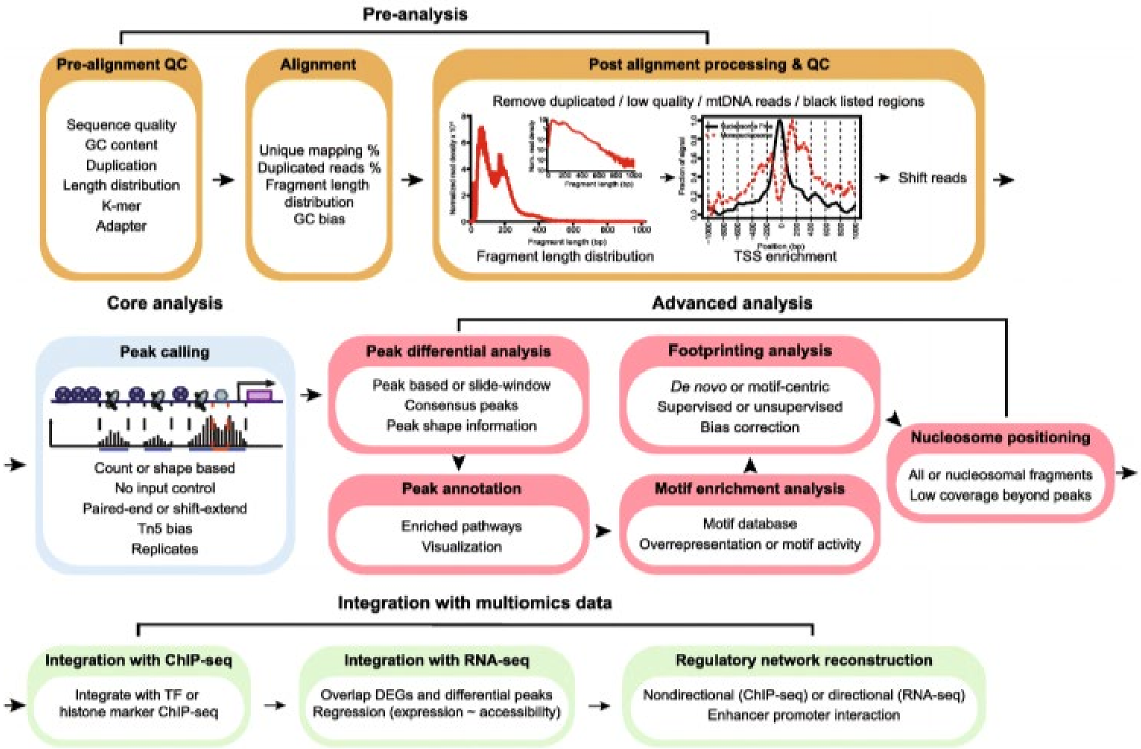
Try to find the differences and similarities between different people's genome and link the differences with the phenotype differences.



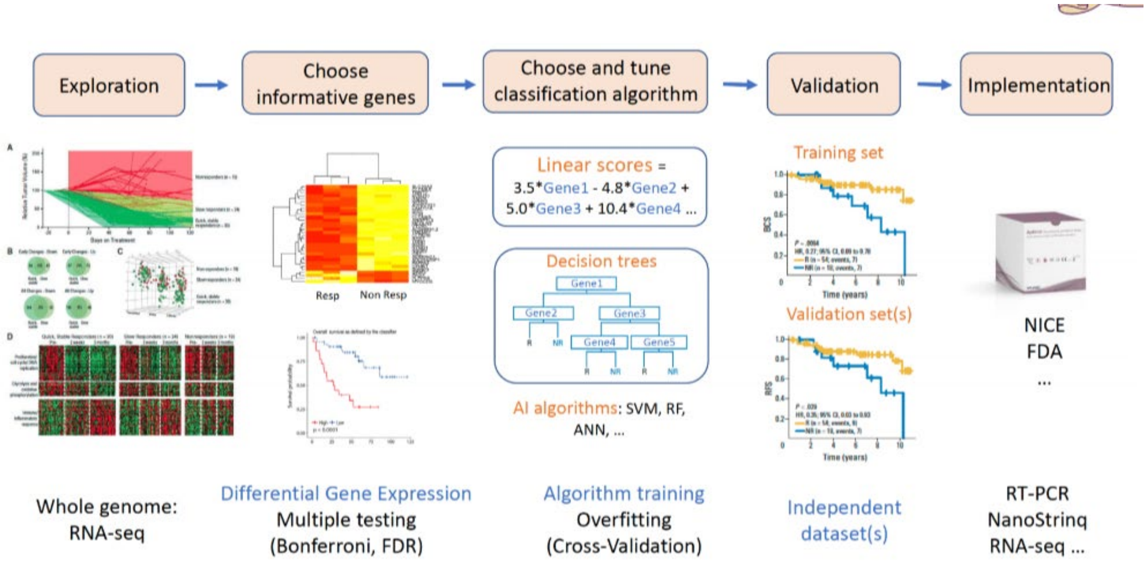
http://www.broadinstitute.org/gatk/guide/best-practices
http://picard.sourceforge.net/

2.3 Epigenome

We may have different phenotypes even if there are not mutations in our genes because modifications which are reversible are not going to change the DNA sequences, but this kind of chemical processes will affect the final gene expression and phenotypes. We are trying to identify this epigenome modification.


Pre-analysis
Pre-alignment QC
Sequence quality
GC content
Duplication
Length distribution
K-mer
Adapter
Alignment
Unique mapping %
Duplicated reads %
Fragment length distribution
GC bias
Post alignment processing & QC
Remove duplicated / low quality / mtDNA reads / black listed regions
Fragment length distribution
TSS enrichment
Shift reads
Core analysis
Peak calling
Count or shape based
No input control
Paired-end or shift-extend
Tn5 bias
Replicates
Advanced analysis
Peak differential analysis
Peak based or slide-window
Consensus peaks
Peak shape information
Peak annotation
Enriched pathways
Visualization
Footprinting analysis
De novo or motif-centric
Supervised or unsupervised
Bias correction
Motif enrichment analysis
Motif database
Overrepresentation or motif activity
Nucleosome positioning
All or nucleosomal fragments
Low coverage beyond peaks
Integration with multiomics data
Integration with ChIP-seq
Integrate with TF or histone marker ChIP-seq
Integration with RNA-seq
Overlap DEGs and differential peaks
Regression (expression ~ accessibility)
Regulatory network reconstruction
Nondirectional (ChIP-seq) or directional (RNA-seq)
Enhancer promoter interaction

2.4 Transcriptome



3. Statistical testing

3.1 Statistical analysis

It is used to discover quantitative changes in expression levels between experimental groups.

Mean and variance may not be most representative to compare different groups.

3.2 T-test

Check whether there are significant differences between different groups.

1. Calculate a test statistic which follows a student’s t-distribution
2. Compare it with the p-value

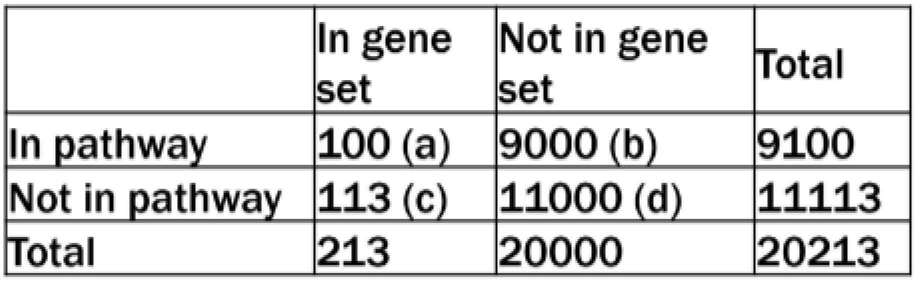
   If the test statistic is smaller than p-value(0.05), we can say we are confident about it.

We also have different t-tests like paired which we cannot shuffle the values and unpaired, one-tailed and two-tailed which we have different p-value.

3.3 Gene enrichment analysis

We want to find a biological pathway which is a series of interactions among molecules in a cell that leads to certain changes in a cell.

Given different pathways and their related and unrelated genes, we can check their relationship by creating contingency tables.



	In gene set	Not in gene set	Total
In pathway	100 (a)	9000 (b)	9100
Not in pathway	113 (c)	11000 (d)	11113
Total	213	20000	20213

They are related if a and d are large, and b and c are small.

How large is it so we can confidently say they are related?

Fisher’s exact test is a statistical significance test used in the analysis of contingency tables.