Machine Learning predictions of CRISPR/Cas9 on-target efficiency in *Drosophila melanogaster*

Pierre Merckaert

Defense 30/8/2018

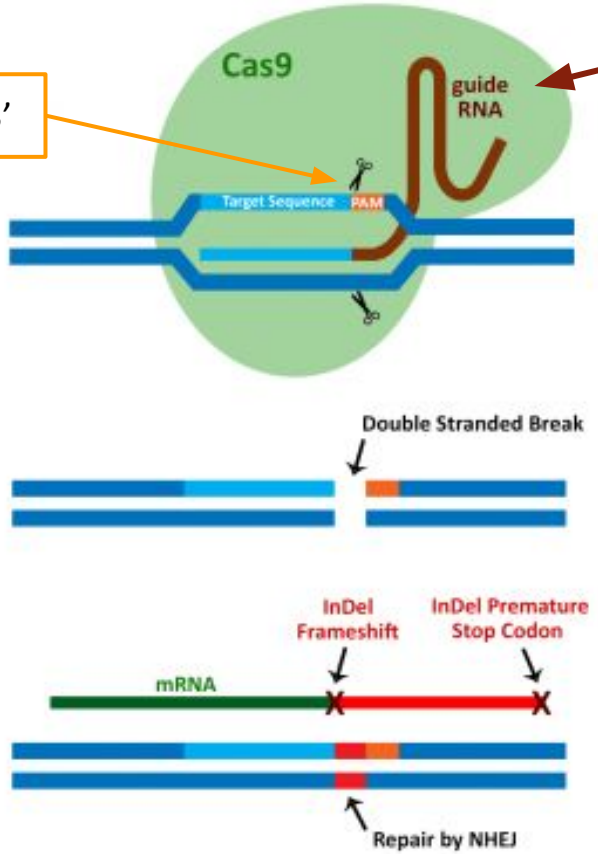
Overview

- Introduction CRISPR screens and Machine Learning
- Identification of sgRNAs targeting essential genes
- Machine Learning modelization of CRISPR sgRNA on-target tefficiency
- Performance comparison of prediction programs
- Discussion, conclusion and future directions

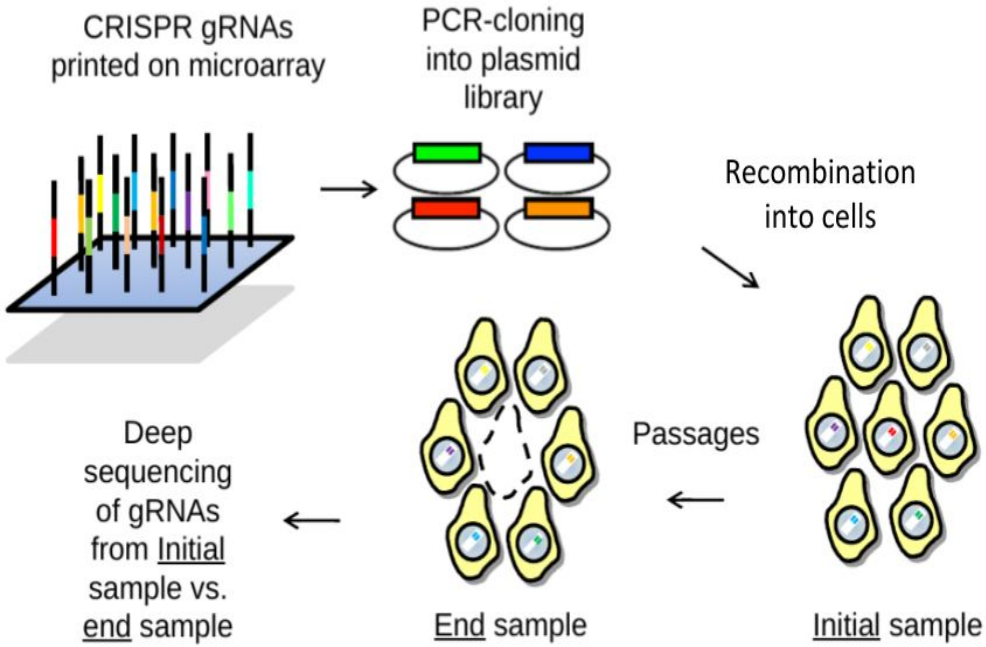
CRISPR/Cas9

PAM = 5'-**NGG**-3'

single guide RNA = **sgRNA**



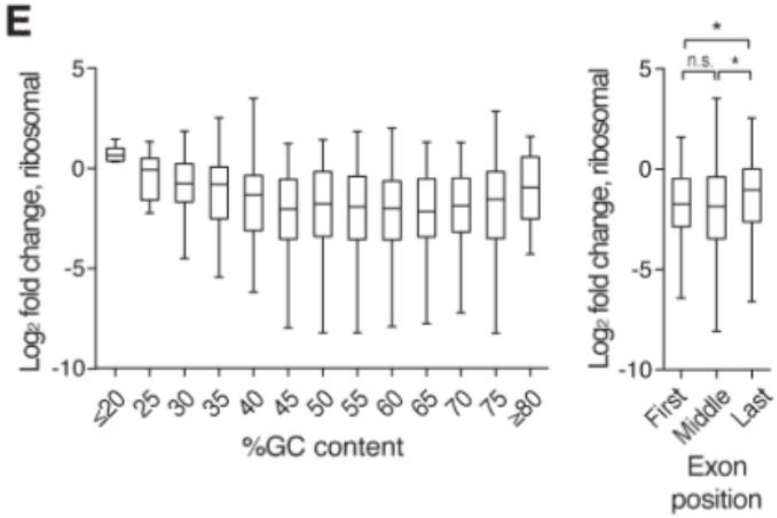
CRISPR/Cas9 genetic screens



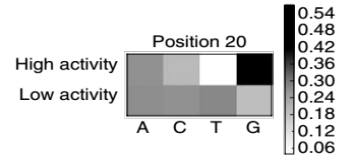
(Viswanatha et al., 2018)

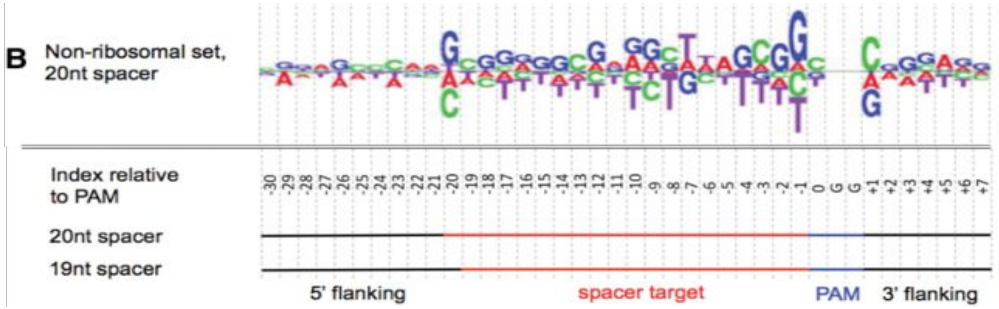
sgRNA design rules in mammals

- sgRNAs with a high or low GC% are less effective
  (Wang et al,. 2014)
- sgRNAs targeting the last exon are less effective.

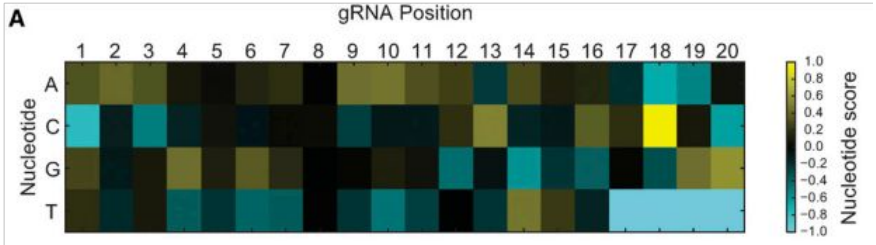


- Strong preference for G and low preference for T at position 20
  (Chari et al., 2015) (Xu et al., 2015)





- Strong preference for C at position 18
  (Hart et al., 2017)



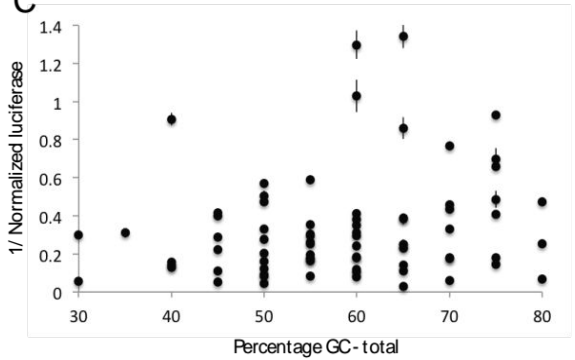
sgRNA design rules in *D.mel*

(Housden et al., 2015)

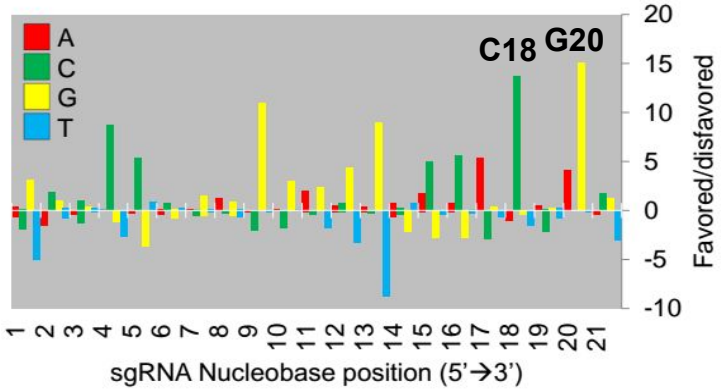
	1	2	3	4	5	6	7	8	9	10
A	0.4979	0.7959	0.7553	0.6569	**0.9481**	0.7147	0.4370	0.6212	**0.9077**	**1.0000**
T	0.6699	0.5485	0.2750	0.5972	0.6212	0.7555	**1.0000**	0.5131	**0.8608**	0.7553
C	0.4979	0.6869	0.8528	0.7643	0.5325	0.3417	0.3417	0.7643	0.6434	**0.0092**
G	0.7918	0.4461	0.4851	0.4461	0.3417	0.6869	0.2417	0.5485	**0.0947**	**0.9256**

	11	12	13	14	15	16	17	18	19	20
A	**0.1957**	0.7959	0.6212	0.8912	**1.0000**	0.5485	**0.9942**	0.5485	0.4550	**1.0000**
T	0.6569	0.3417	**1.0000**	**0.0160**	**0.9146**	0.7555	0.2906	0.4979	0.5485	0.5131
C	**0.9331**	0.5325	0.7272	0.9708	0.2905	0.7272	0.2957	0.7918	0.6434	0.5062
G	0.5325	0.8308	**0.1255**	0.7918	0.2544	0.4461	0.4979	0.6212	0.7918	0.4461

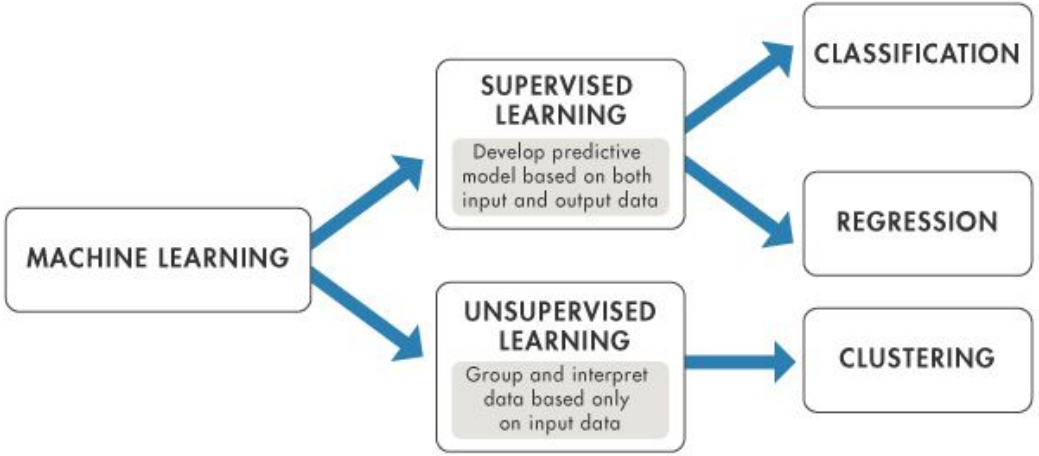
C



(Viswanatha et al., 2018)

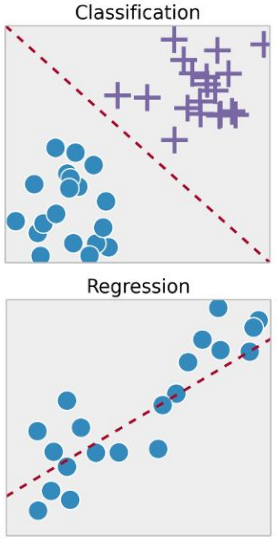


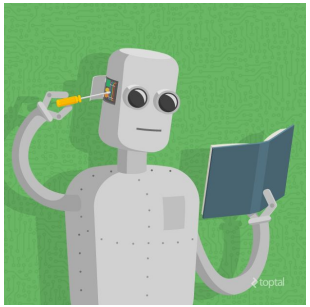
What/How to predict ?



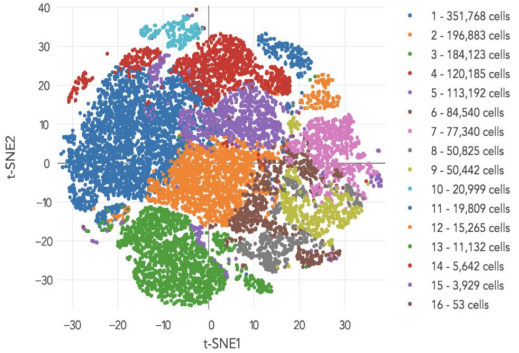
sgRNA efficiency (**binary**)	
0	1

sgRNA efficiency score (**continuous**)
[0 , 0.1 , … , 0.5 , … , 0.9 , 1]

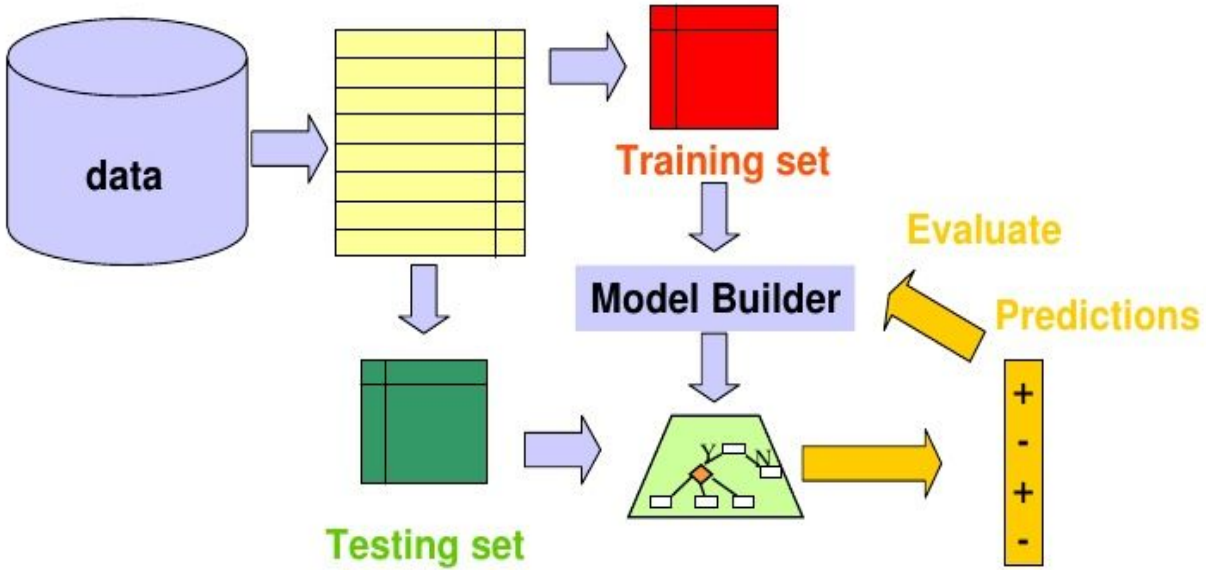




Le **Machine Learning** uses statistical algorithm to “learn” from data (e.g., progressively improve a tasks’s performance), without being explicitly programmed.



Machine Learning basics

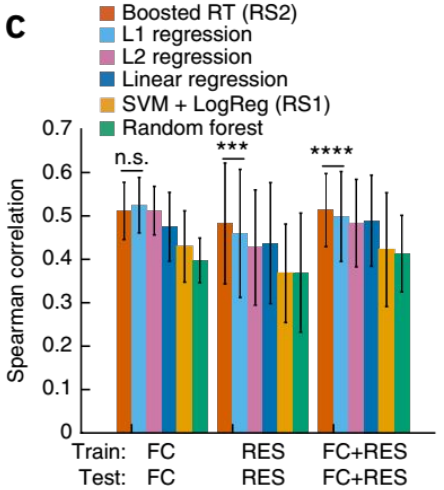


Azimuth : state-of-the-art of sgRNA efficiency prediction

Doench et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nature Biotechnology

➢ CRISPR screen of 4390 sgRNAs targeting 17 genes in human / mouse cells

➢ **ML modelization of sgRNAs efficiency based on their characteristics**



crispr.ml

Microsoft

BROAD INSTITUTE | GENETIC PERTURBATION PLATFORM

DESKTOP GENETICS

SYNTHEGO Genome Engineering Solutions

List of sgRNA design softwares

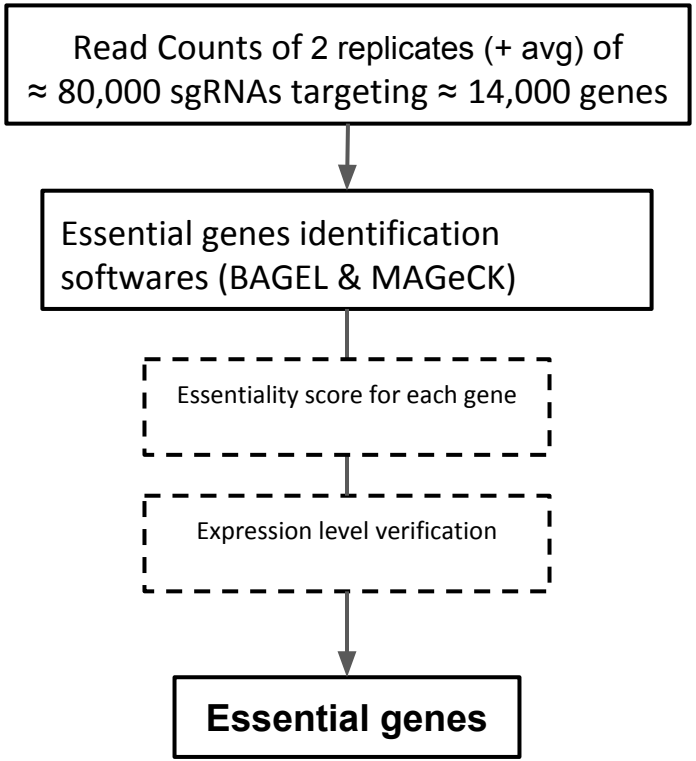
Tool Name	Provider	Searches whole genome for targets	Returns all targets of genome	Seed span and location can be defined	Maximum number of mismatches supported	Predicts gRNA activity	Available Protospacer adjacent motif (PAM) sequences
Synthego Design Tool	Synthego	Yes (over 120,000 genomes	No (Optimized for Knockout)	Yes	3	Yes	NGG
Benchling CRISPR gRNA Design	Benchling	Yes	Yes	Yes	4	Yes	User customizable
CRISPOR	University of California, Santa Cruz TEFOR	Yes (over 200 genomes)	Yes	No	4	Yes	NGG, NGA, NGCG, NNAGAA, NGGNG, NNGRRT, NNNRRT, NNNNGMTT, NNNNACA, TTTN
CHOPCHOP v2	University of Bergen	Yes	Yes	Yes	3 (0-3)	Yes	User customizable
CRISPR LifePipe	Life and Soft	Yes	Yes	Yes	0-5	yes	NGG, NGA, NGCG, TTTN, NNGRRT
DESKGEN	Desktop Genetics	Yes	Yes	Yes	Any number	Yes	Fully user customizable
Geneious CRISPR Site Finder	Geneious	Yes	Yes	Yes	Any number	Yes	User customizable
sgRNA Designer	Broad Institute	No	No	No	0	Yes	NGG
CASTING	Caagle	Yes	Yes	No	3	No	NGG and NAG
Breaking-Cas	Spanish National Center for Biotechnology	Yes (over 1000 genomes)	Yes	Yes (by weights)	4	No	User customizable
Cas-OFFinder	Seoul National University	Yes	Yes	No	0-10	No	NGG, NRG, NNAGAAW, NNNNGMTT
CCTop	University of Heidelberg	Yes	Yes	Partial	5 (0-5)	No	NGG, NRG, NNGRRT, NNNNGATT, NNAGAAW, NAAAAC
CHOPCHOP	Harvard University	Yes	Yes	Partial	0, 2	No	NGG, NNAGAA, NNNNGANN
COD	Dayong Guo	No	No	No	0, 3, 5, 8	No	NGG and NAG
CRISPR Design	Zhang Lab, MIT	Yes	No	No	4	No	NGG and NAG
CRISPRdirect	Database Center for Life Science (DBCLS)	Yes (over 200 species)	Yes	No	Any number	No	NNN
CRISPR gRNA Design Tool	DNA2.0	Yes	Yes	No	0-10	No	NGG, NAG
CRISPRseek	Bioconductor	Yes	Yes	No	Any number	No	User customizable
Genedata Selector	Genedata	Yes	Yes	Yes	Customizable	No	Customizable
GT-Scan	CSIRO & EMBL-ABR	Yes	Yes	Yes	3 (0-3)	No	User customizable
CRISPR Configurator & Specificity Tool	Dharmacon, Inc.	Yes (over 30 species)	Yes	Yes	8 (gaps or mismatches)	Internally	NGG and NAG
Off-Spotter	Thomas Jefferson University	Yes	Yes	Yes	0-5		NGG, NAG, NNNNACA, NNGRRT (R is A or G)

Internship goals

- From a CRISPR screen, identify sgRNAs targeting essential genes
- Identify the best ML models to predict sgRNA efficiency, and develop a program implementing these predictions.
- Compare my program to other sgRNA efficiency prediction tools.

Identify essential genes

Pipeline

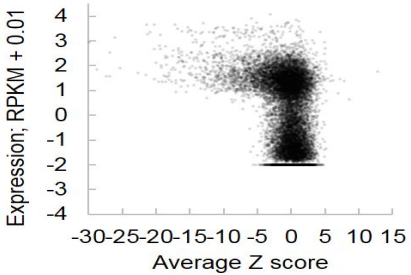


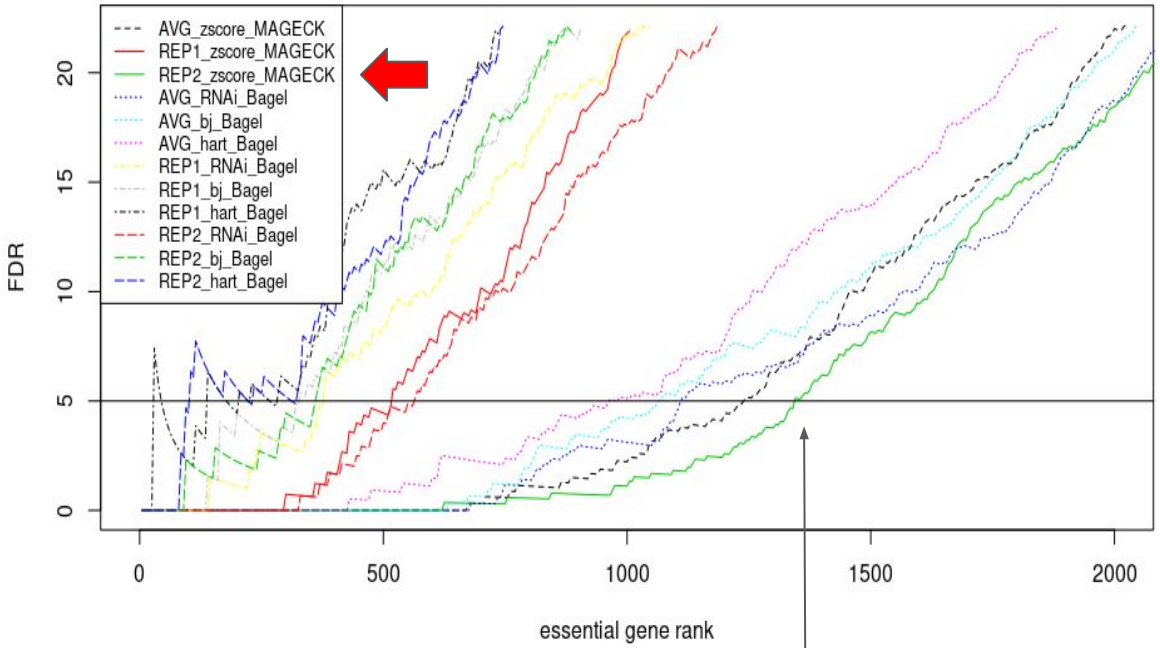
> 90% of D.mel genes

Essential genes are highly expressed

➢ **H0 : Essential genes are highly biased towards expressed genes**

(Viswanatha et al., 2018)





Sort genes based on essentiality prediction

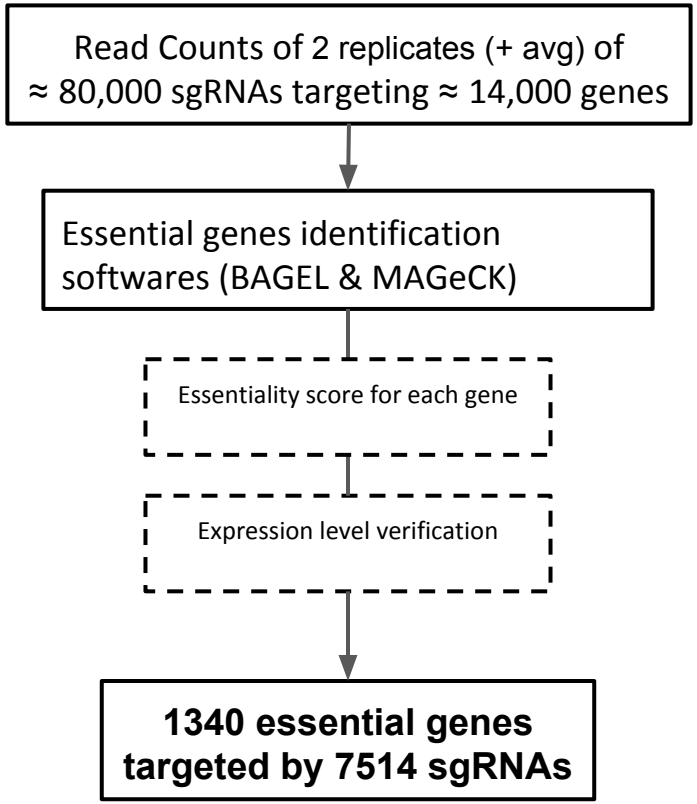
Expression level analysis

False Positive Rate computing

1340 essential genes

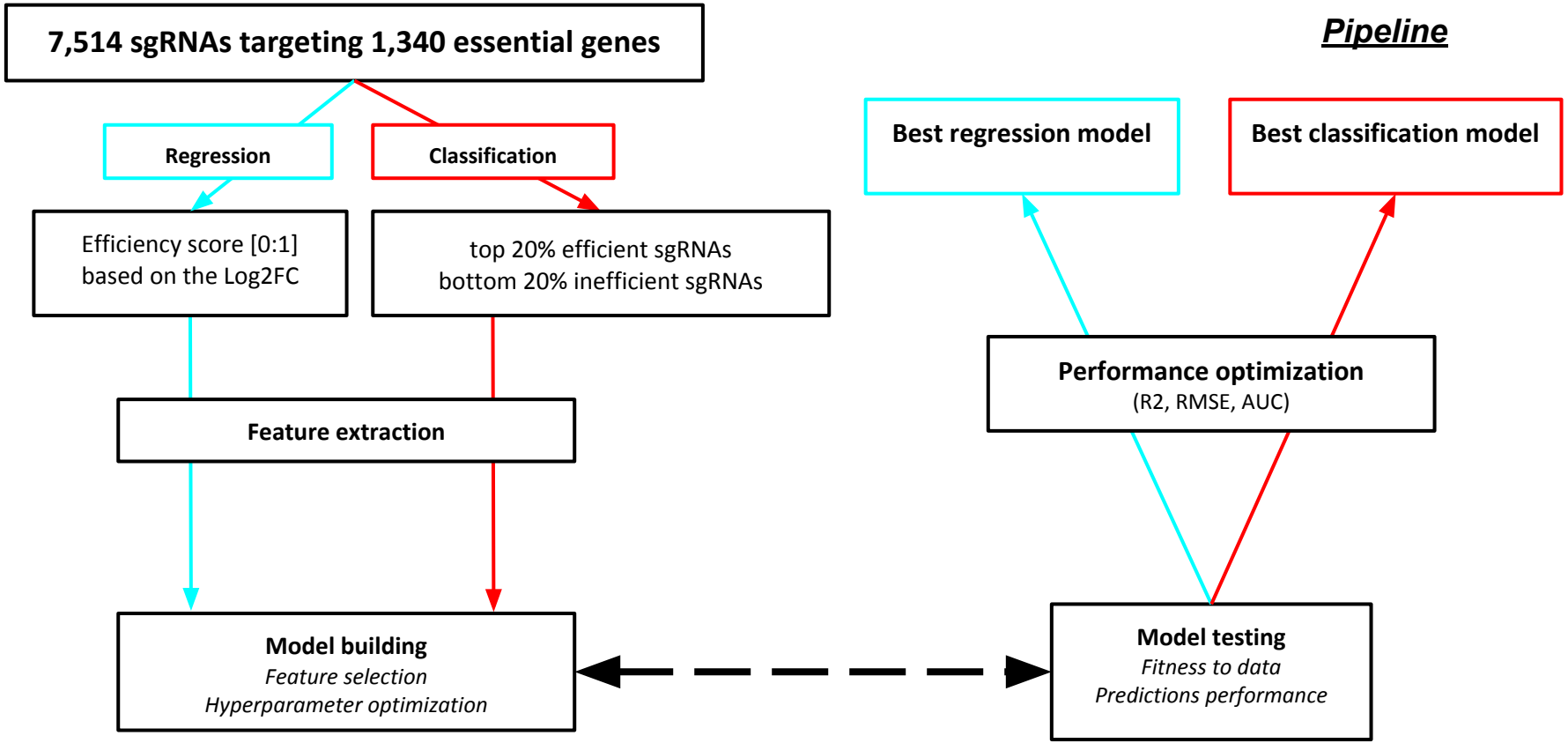
Identify essential genes

Pipeline



> 90% of D.mel genes

Find the best model to predict sgRNA efficiency



Features and performance evaluation

Feature Extraction

Extract features from 30mer sgRNA :

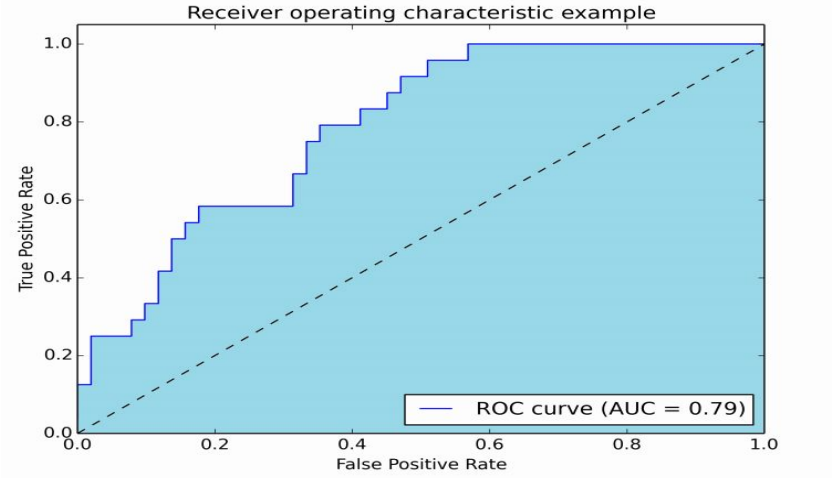
- -1st & 2nd order nucleotide positions and amount
- -GC content and Melting Temperature
- -PAM flanking nucleotides

➜ **625 features**

CATGATCATCGTACCCGAGATGACCGGCTC

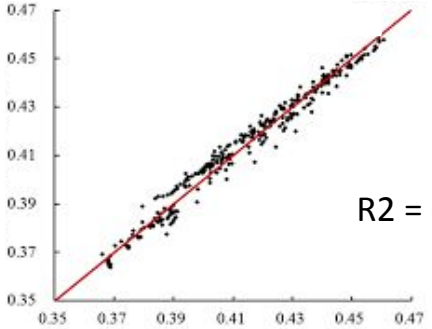
Classification

➢ Area Under the Curve (AUC)



Regression

➢ Coefficient of Determination (R2)

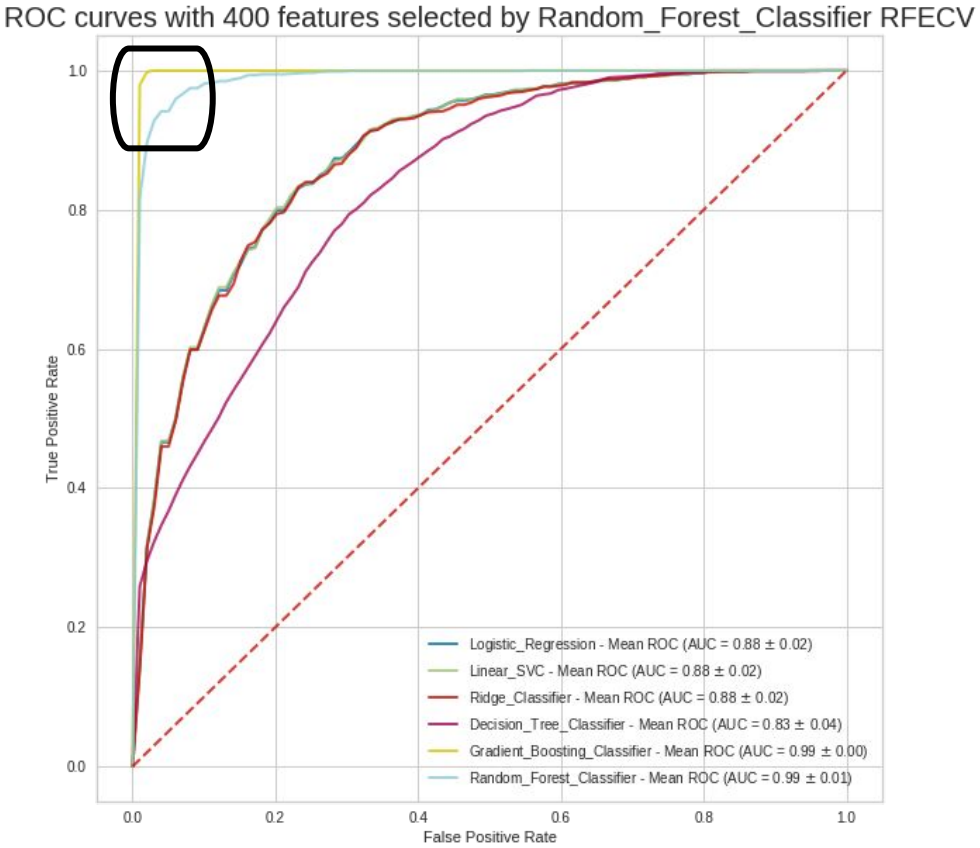


R2 = 0.97

➢ Root Mean Square Error (RMSE)

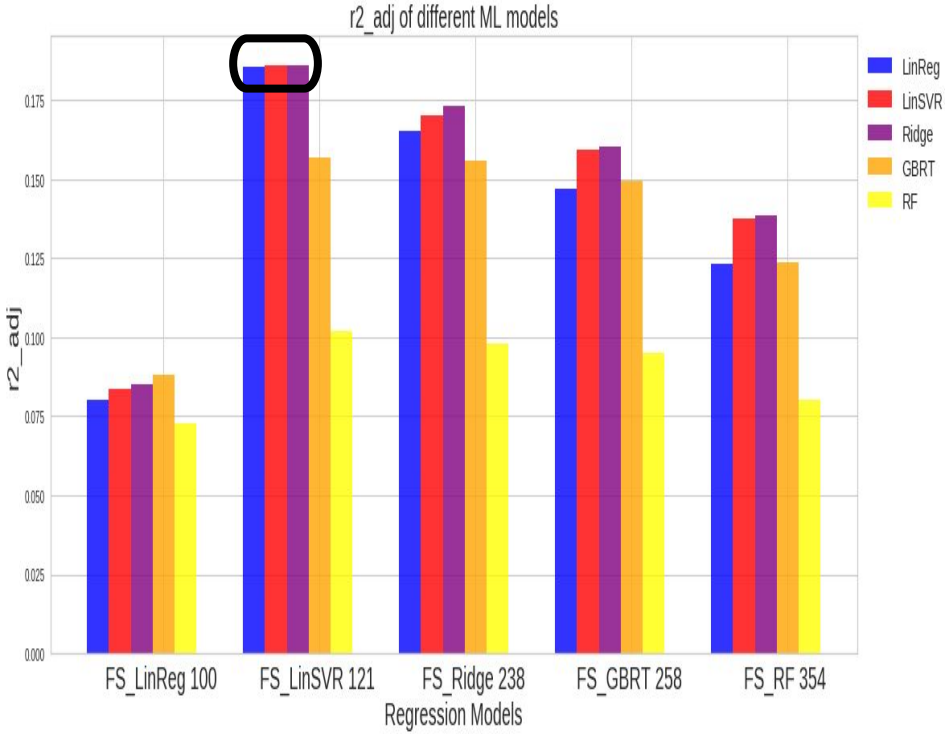
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

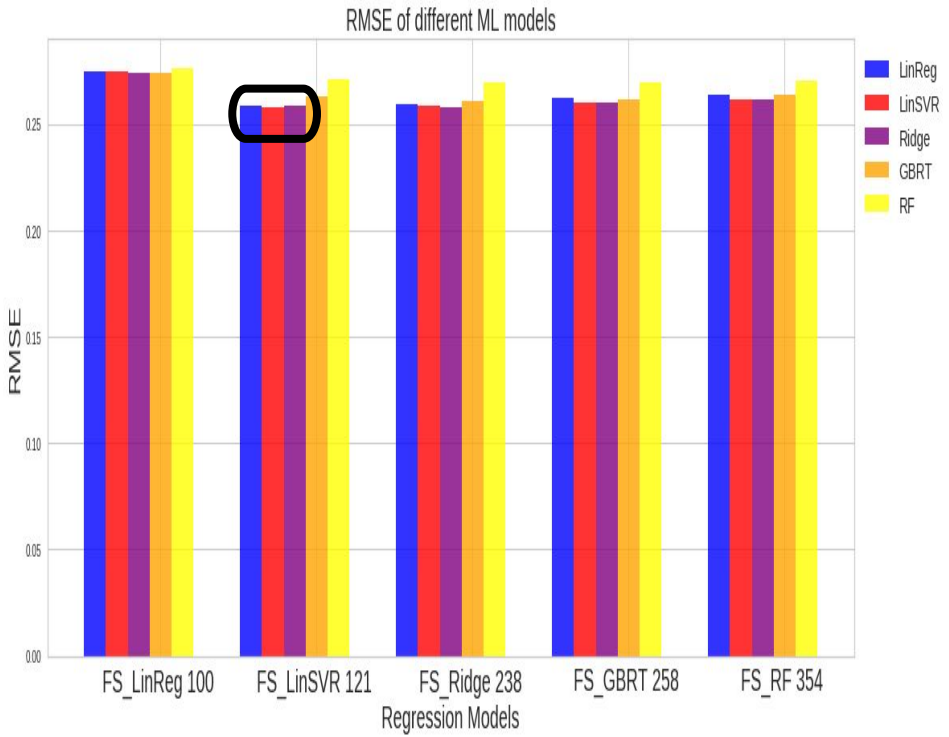
Finding the best ML model : Comparison of classification models



- The best classification model for 30mer sgRNA efficiency prediction is Gradient Boosting Classifier

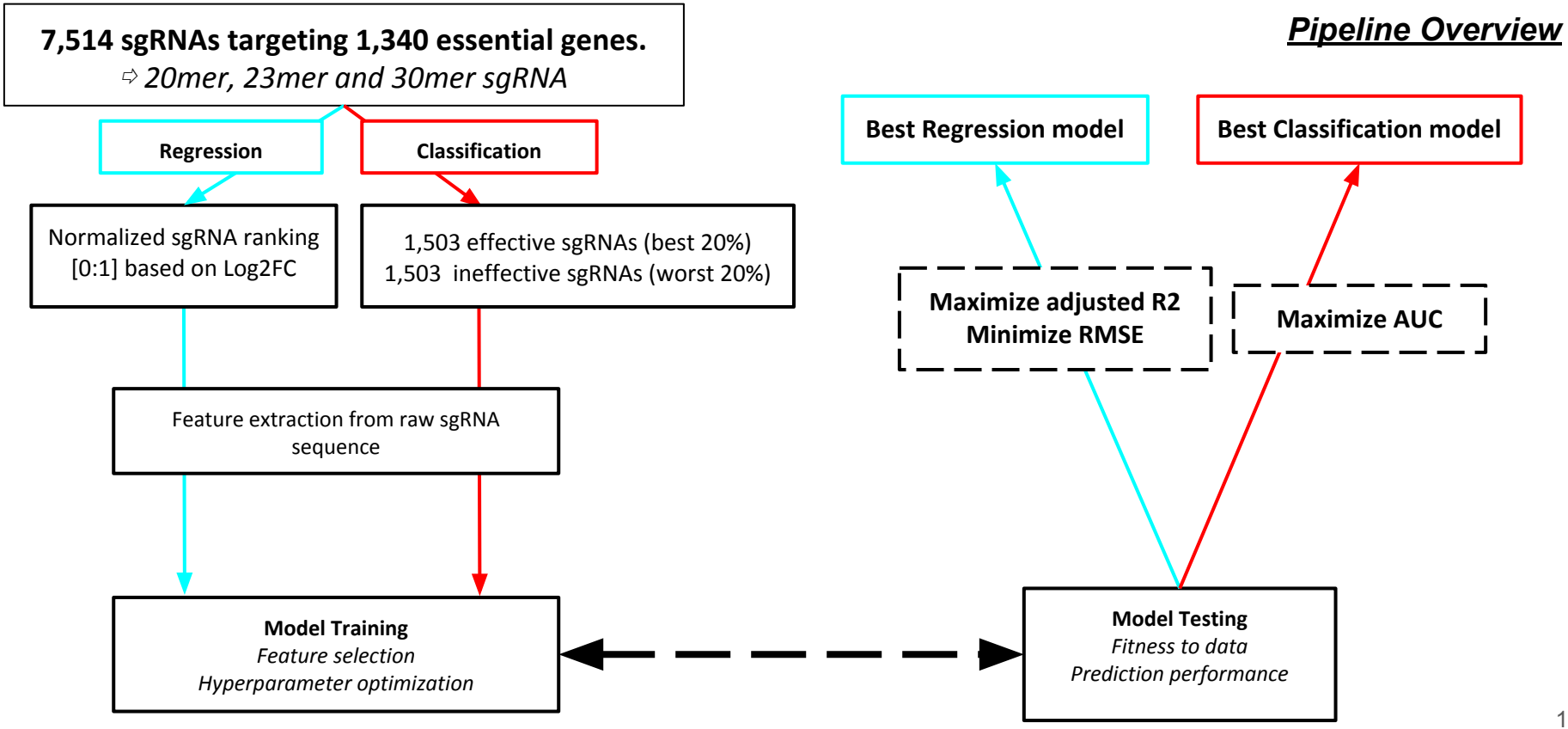
Comparison of regression models





➢ The best regression model for 30mer sgRNA efficiency prediction is Linear Support Vector Regression.

Finding the best ML model for sgRNA efficiency prediction



Tool to predict sgRNA efficiency in *D.mel*

Input
20, 23 or 30mer
sgRNA sequence

Output
Prediction scores

sgRNA 30mer
CATGATCATCGTACCCGAGATGACCGGCTC
GGCAGGCGGCAGTCCCAGAGGATTGGGTTG
TGAGACTGCAAACGGGATGCACATGGGCCG
AAGGTATATTGCGCGCTTGACACTTGGCCA

sgRNA 30mer	Efficiency prediction	Binary score
CATGATCATCGTACCCGAGATGACCGGCTC	0.512	1
GGCAGGCGGCAGTCCCAGAGGATTGGGTTG	0.485	0
TGAGACTGCAAACGGGATGCACATGGGCCG	0.738	1
AAGGTATATTGCGCGCTTGACACTTGGCCA	0.384	0

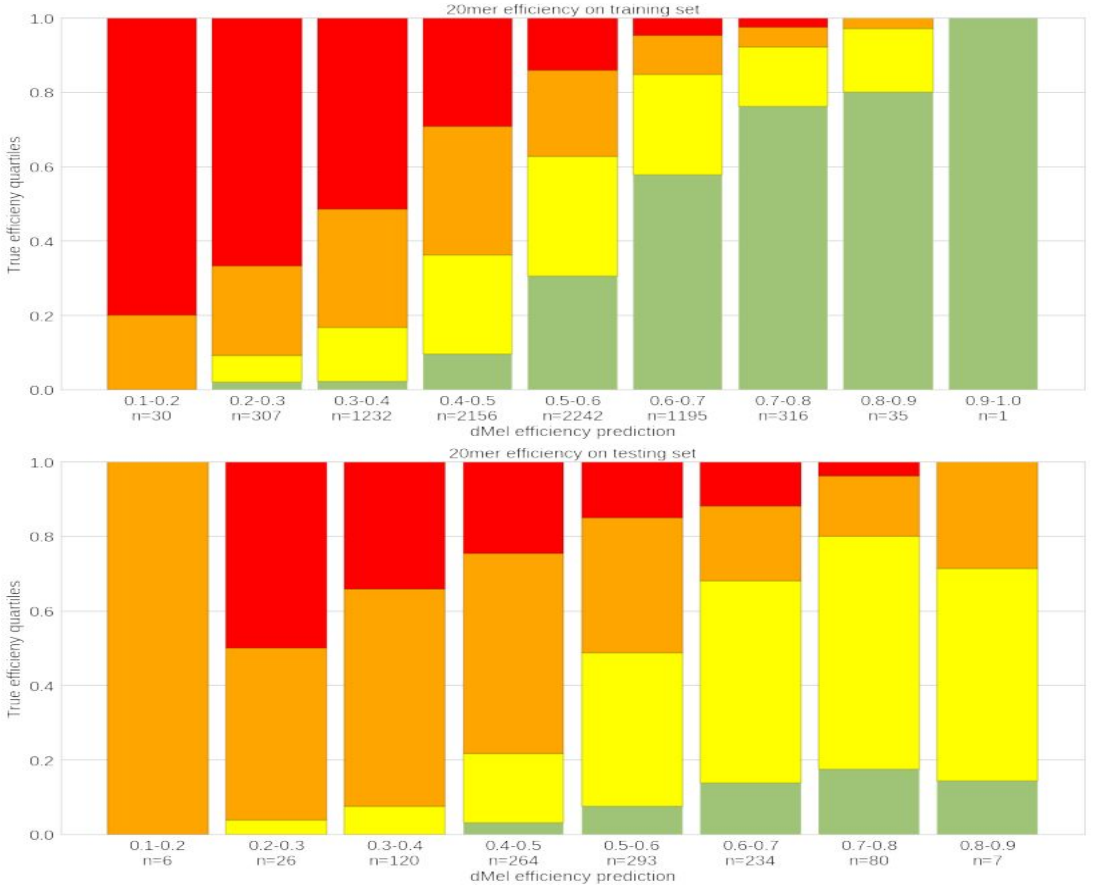
https://github.com/PierreMkt/Dmel-sgRNA-Efficiency-Prediction

dMel : regression performance

Train set : 7514 sgRNAs used to build the model

Test set : 2600 *random* sgRNAs from the CRISPR screen, not used to build the model

Red = 0% - 25% true efficiency
Orange = 25% - 50% true efficiency
Yellow = 50% - 75% true efficiency
Green = 75% - 100% true efficiency

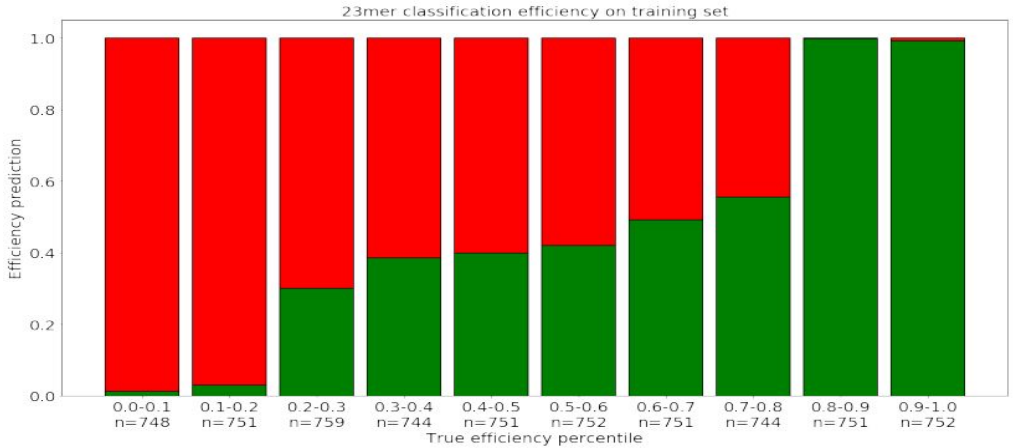


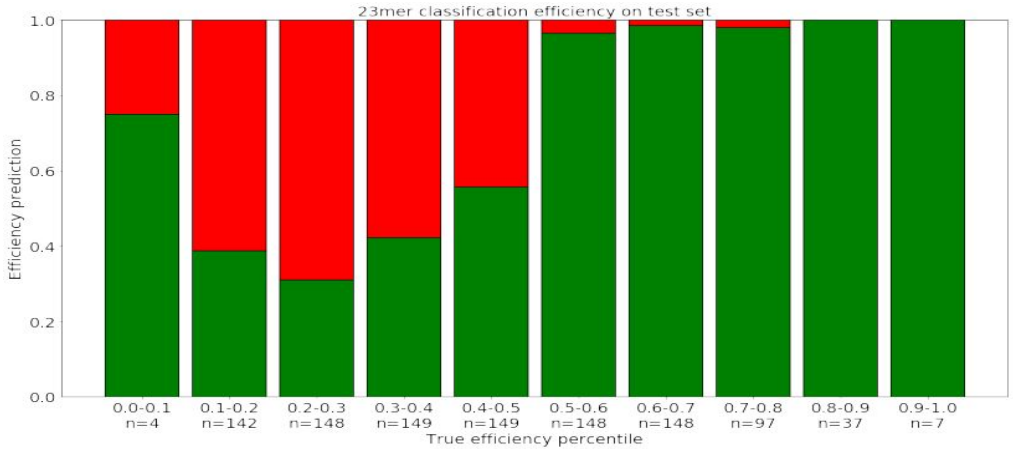
dMel classification performance

Train set : 7514 sgRNAs used to build the model

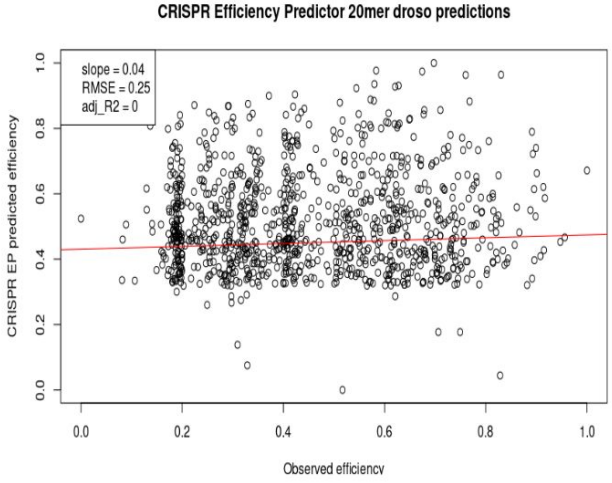
Test set : 2600 *random* sgRNAs from the CRISPR screen, not used to build the model

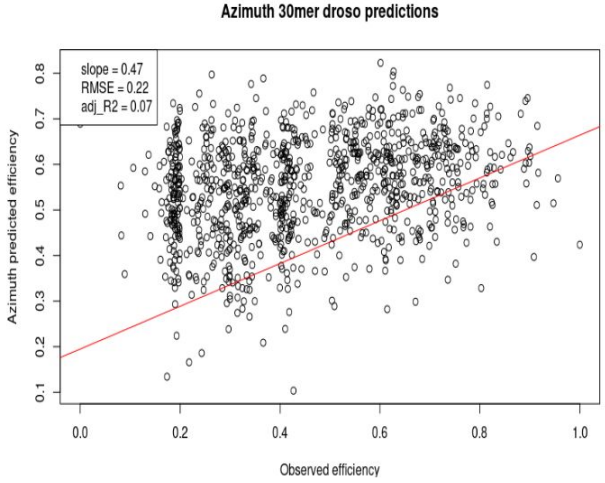
Vert = sgRNA is predicted efficient
Rouge = sgRNA is predicted inefficient

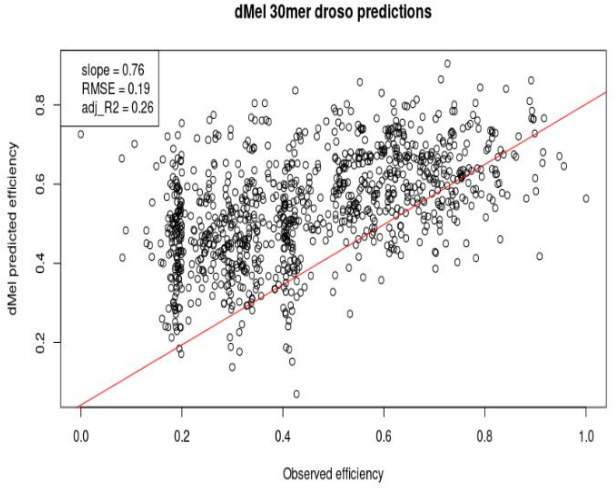


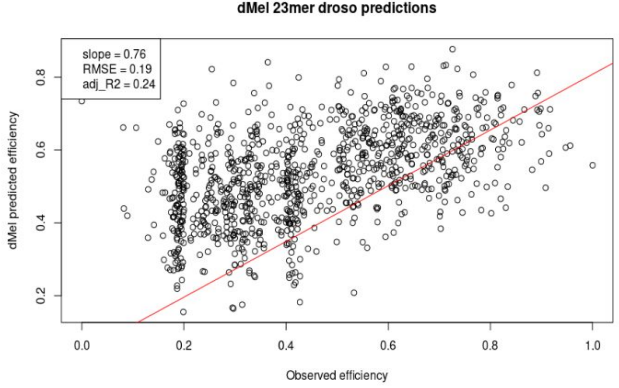


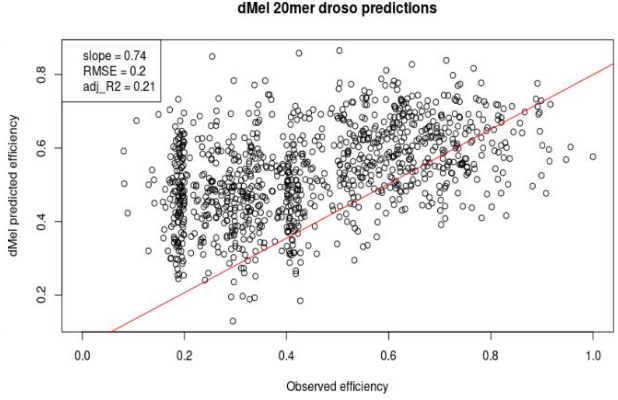
Comparison of performance in *D.mel*



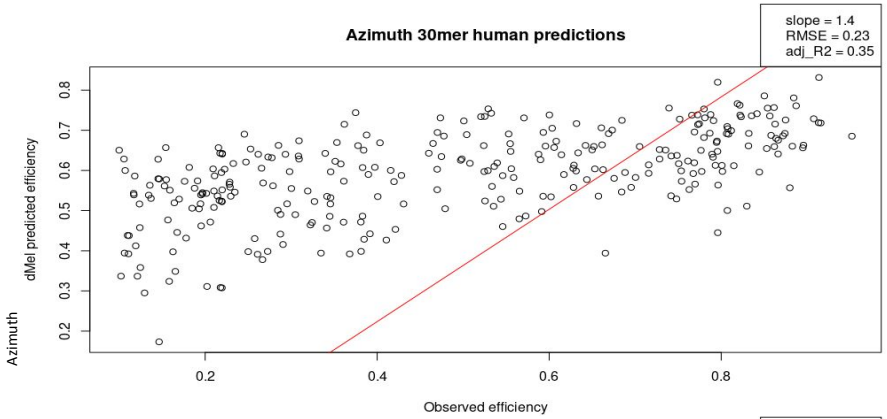


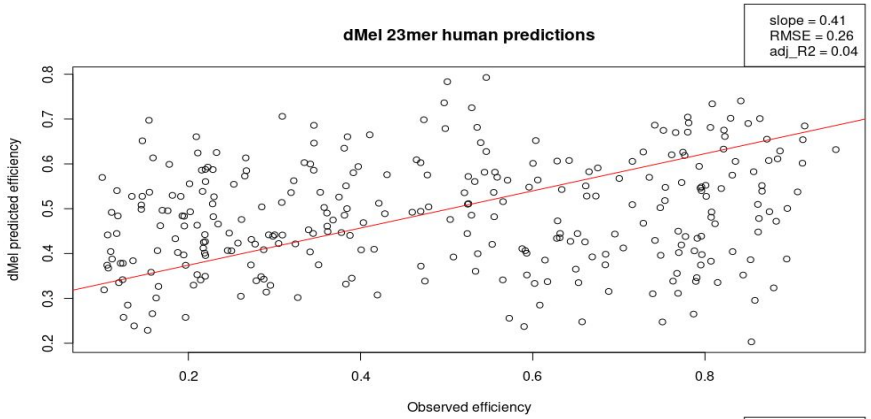


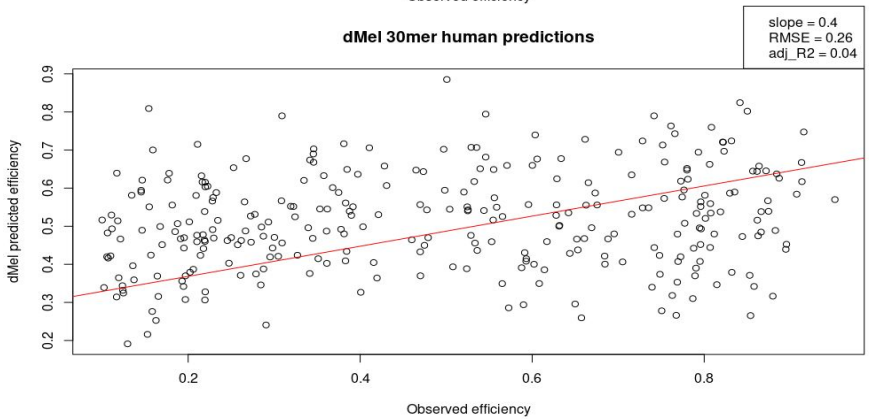


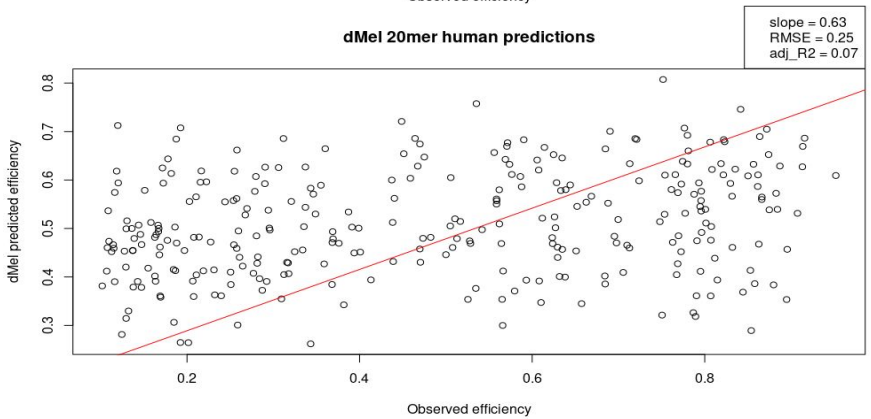


Comparison of performance in *Human*









Conclusion & Future directions

- Model predictions depend on the species from which the input data are retrieved.
- The raw sgRNA sequence is sufficient to build accurate predictive models.
- My program will help my thesis lab to develop more efficient sgRNAs for CRISPR screens in *D.mel.*

- **Implement my program** in the sgRNA design pipeline of my thesis lab
- Make new CRISPR screens in *D.mel* with a sgRNA library based on my program's predictions, and composed of lots of sgRNAs targeting a few non-essential genes.
  - Rebuild the predictive models based on the data from this new screen.
- Models based on **Deep Learning** may have better prediction accuracy.
  Kim et al. (2018). Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. Nature Biotechnology, 36(3), 239–241.

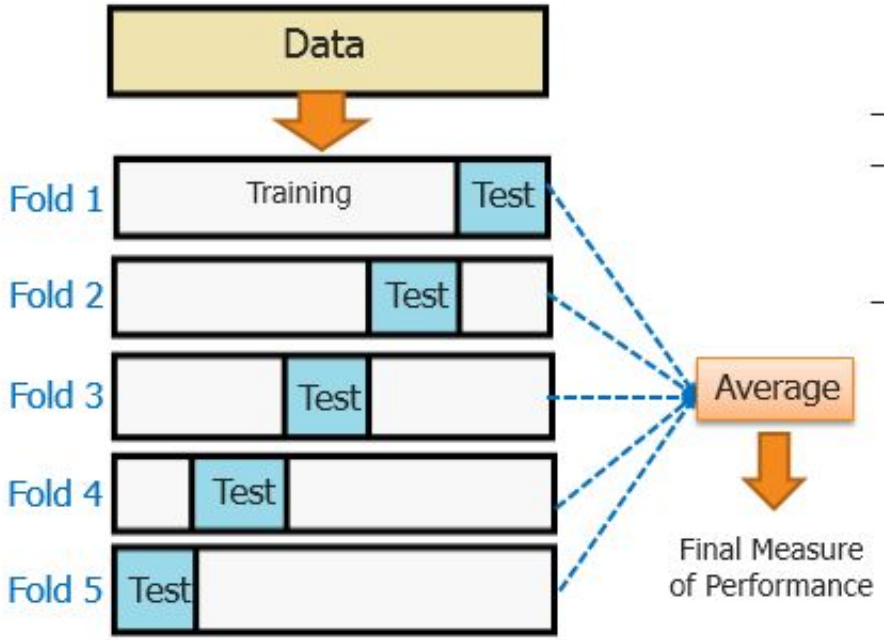
HARVARD MEDICAL SCHOOL

THANK YOU !

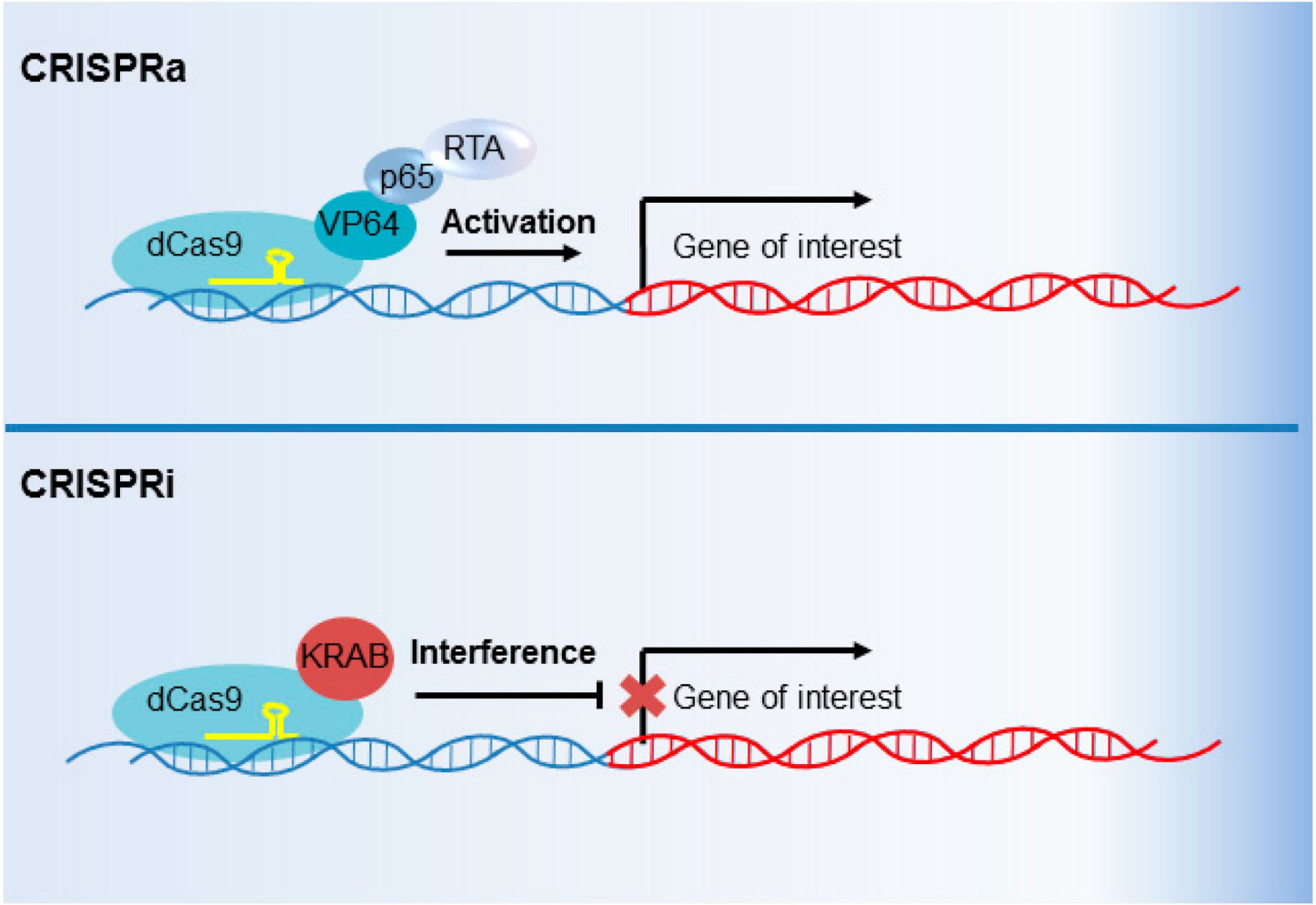
I spent way too much time on this...

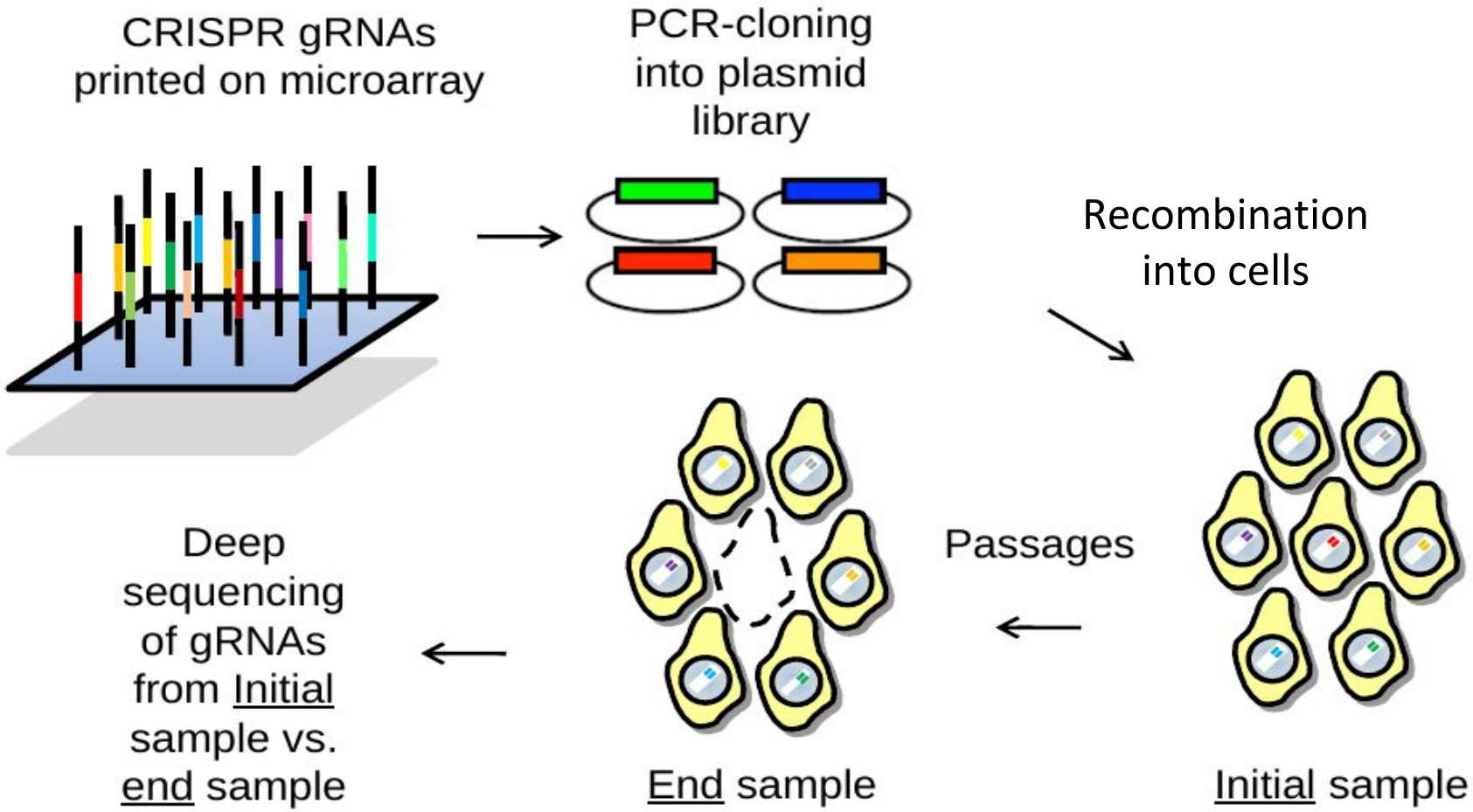
Cross-Validation (CV)

edureka!



- → Technique to validate models/classifiers
- → Method to estimate how accurately the model generalizes to unseen data i.e., how well it performs/predicts
- → K-fold CV
  - » Most popular
  - » k is typically set to 10
  - » Every sample/record is used both in training and test sets


CRISPRa
RTA
p65
VP64
dCas9
Activation
Gene of interest
CRISPRi
KRAB
dCas9
Interference
Gene of interest


CRISPR gRNAs printed on microarray
PCR-cloning into plasmid library
Recombination into cells
Initial sample
Passages
End sample
Deep sequencing of gRNAs from Initial sample vs. end sample

Identifying essential genes from pooled CRISPR KnockOut screen

MAGeCK
(Li et al., 2014)

Read counts mean and variance modeling
(Maximum Likelihood Estimation)

↓

Z-score

↓

Negative value = fitness / essential gene
Positive value = growth / suppressive gene

BAGEL
(Hart et al., 2016)

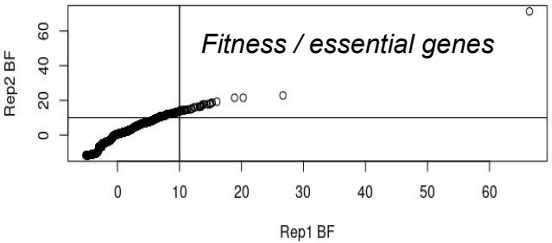
Log2FC probability density function modeling
(Kernel Density Estimation)

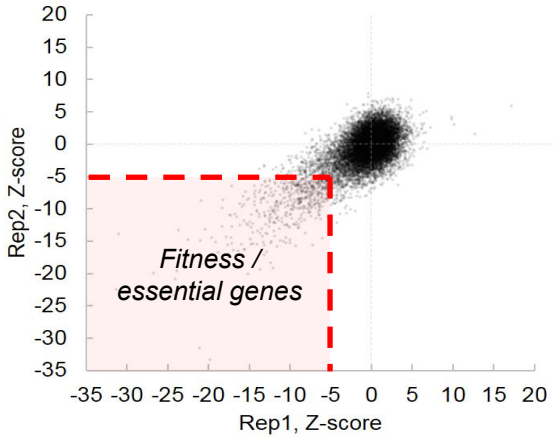
↓

Bayesian Factor

↓

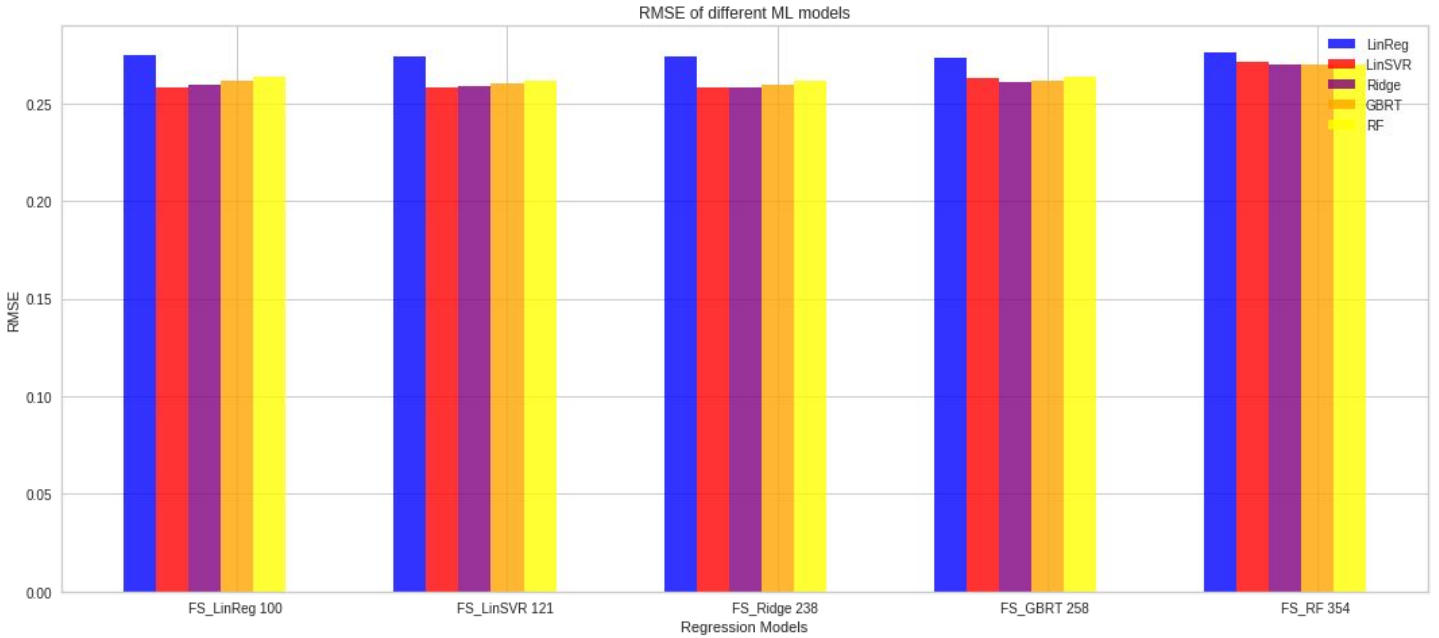
BF > 10 ⇨ fitness / essential gene





Ram Viswanatha

Finding the best ML model : Comparison of regression models



Finding the best ML models : Feature Extraction & Selection

-RNAi, bj, Hart datasets
-BAGEL MAGeCK

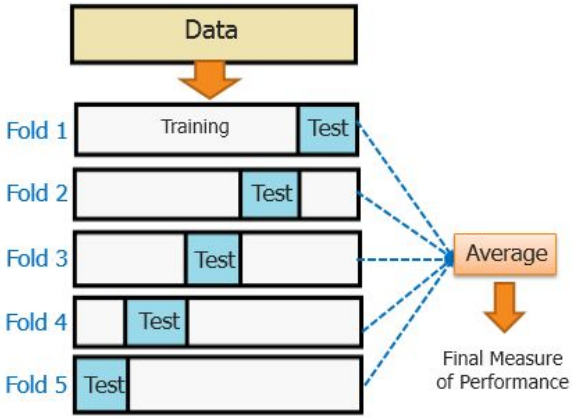
-Describe each ML Model
-Hyperparameters

-RFECV

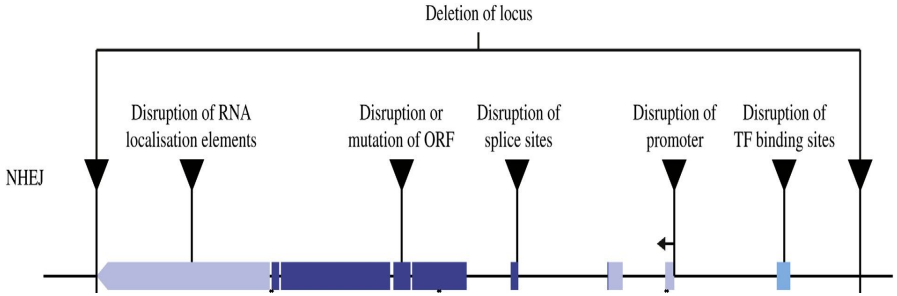
-tilted residuals

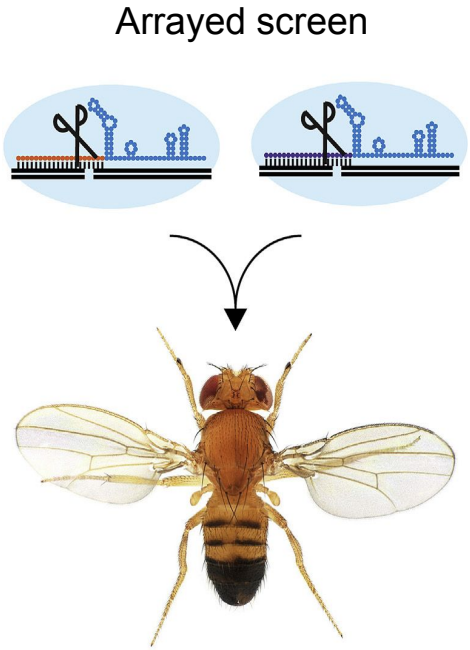
-feature importance

Cross-Validation

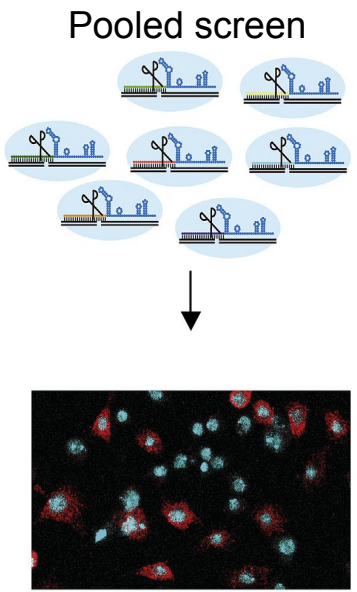


CRISPR/Cas9 for *Drosophila melanogaster* genome editing





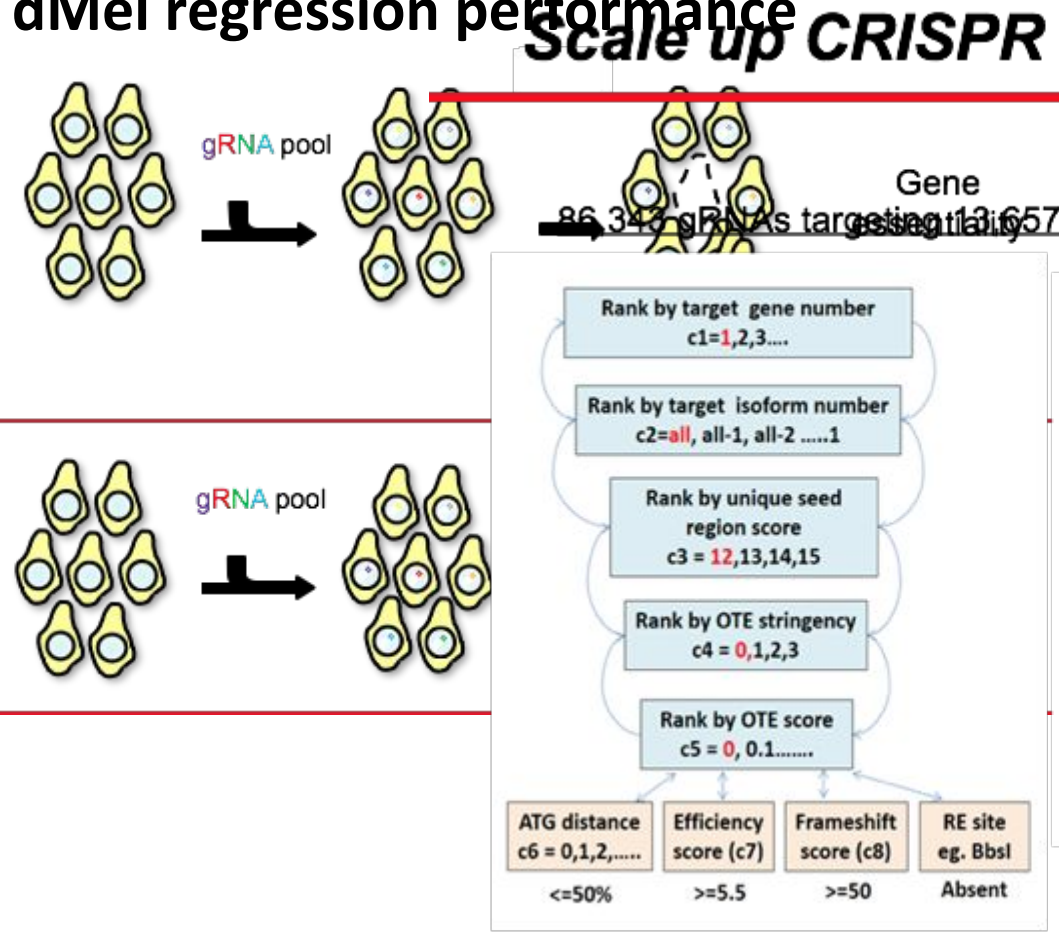
Screen for phenotype (cells or flies)
Compounding of mutations



Select for phenotype (cells or flies)
Identify mutations by targeted sequencing

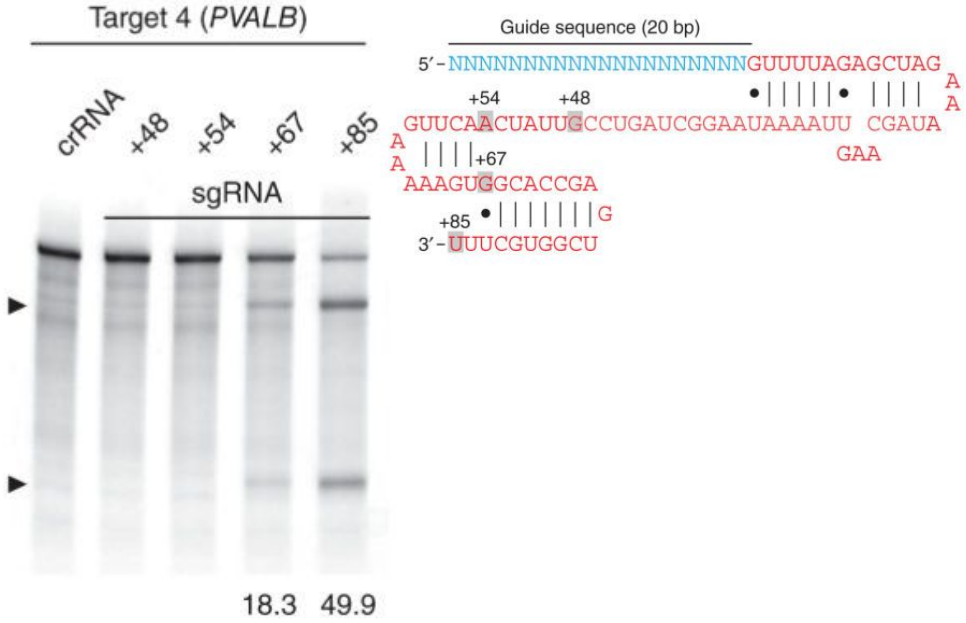
A. Bassett, et al (2014). Journal of Genetics and Genomics, 7-19

dMel regression performance

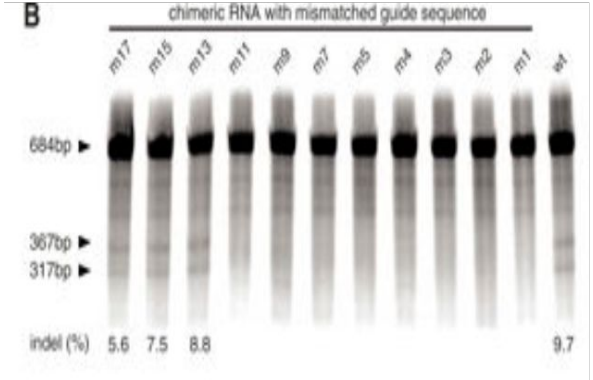


Règles du design des sgRNA chez les mammifères (1)

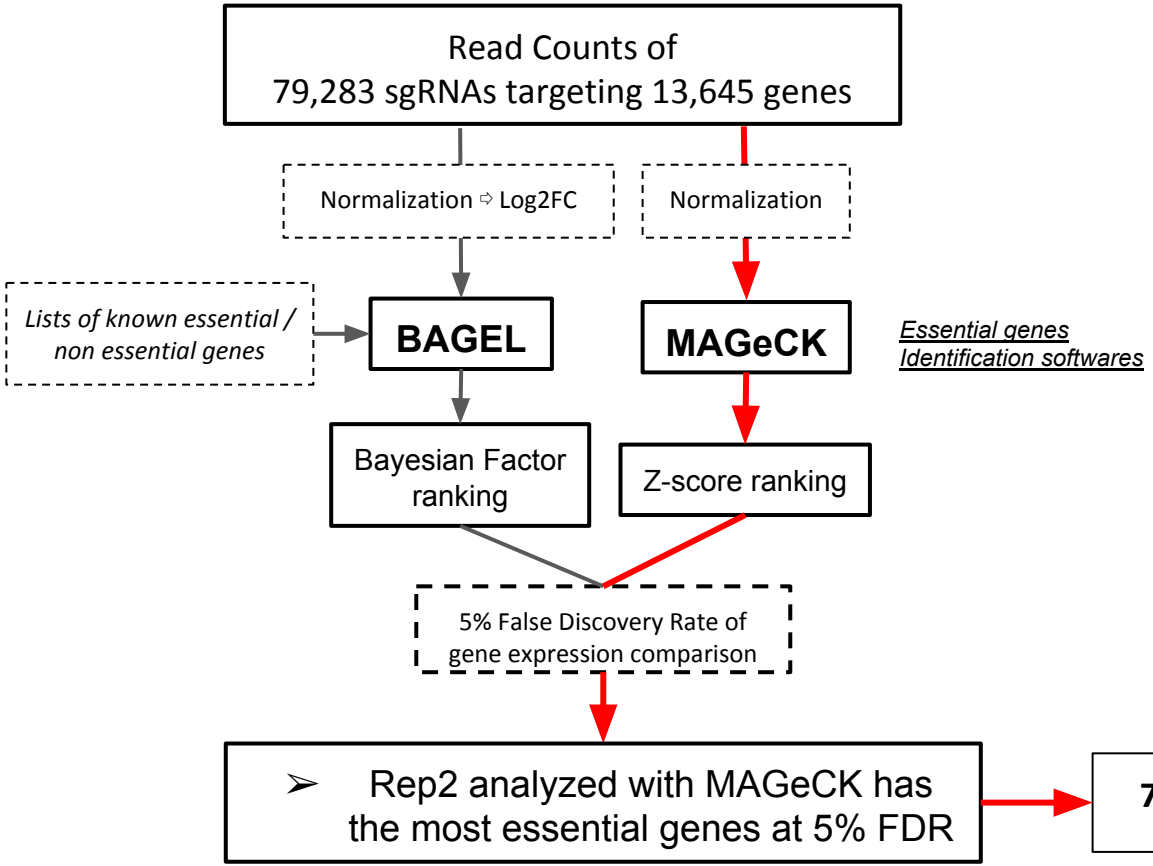
- Augmenting tracrRNA length improves sgRNA efficiency (Hsu et al., 2013)



- Single-base mismatches in the seed region abolishes cleavage by Cas9. (Cong et al., 2013)

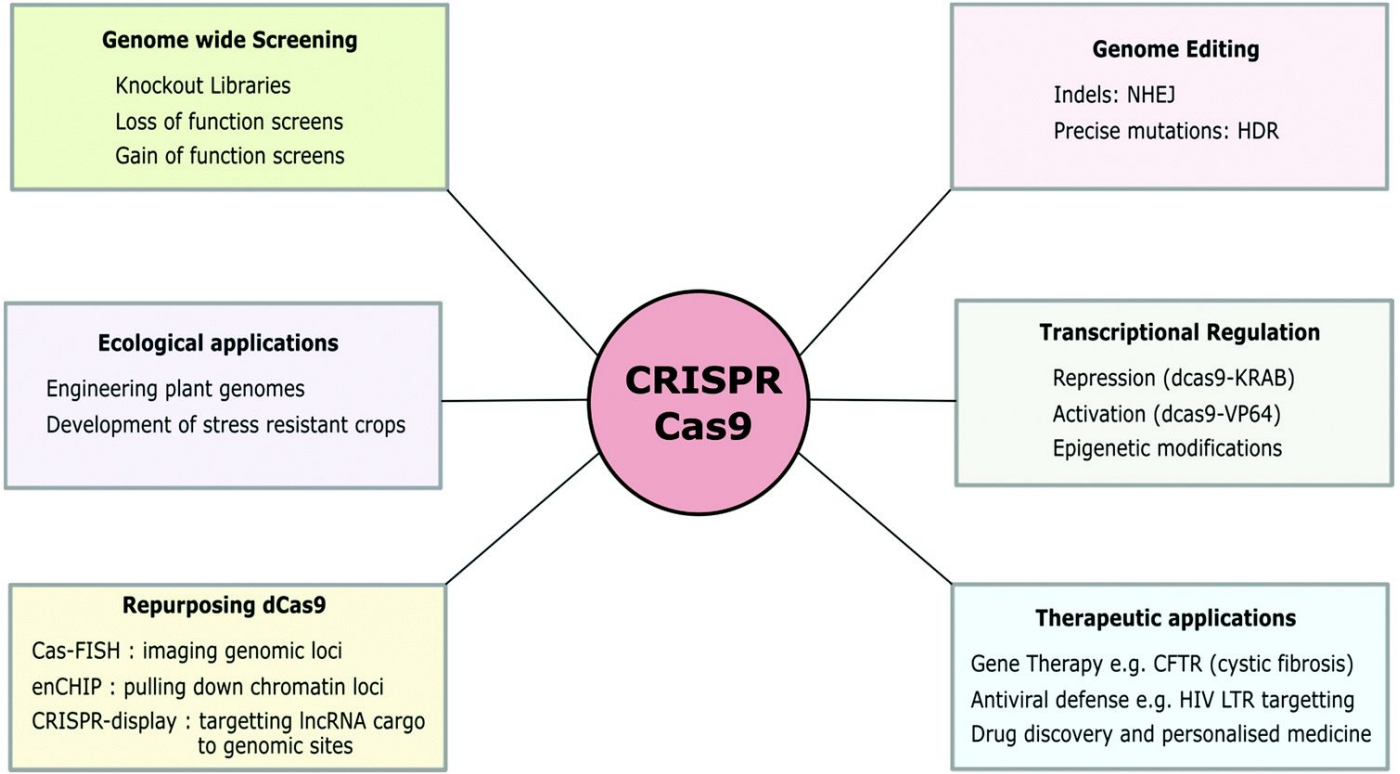


Identifier les sgRNAs visant les gènes essentiels



Pipeline

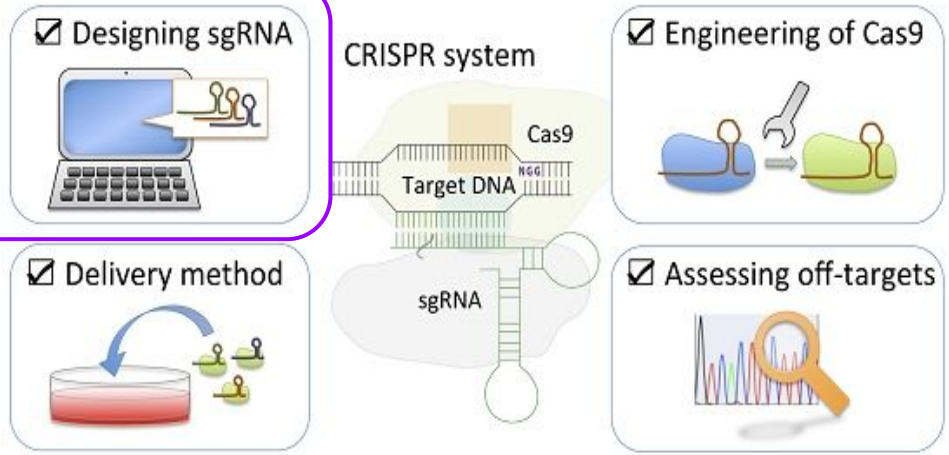
Applications et potentiel de CRISPR/Cas9



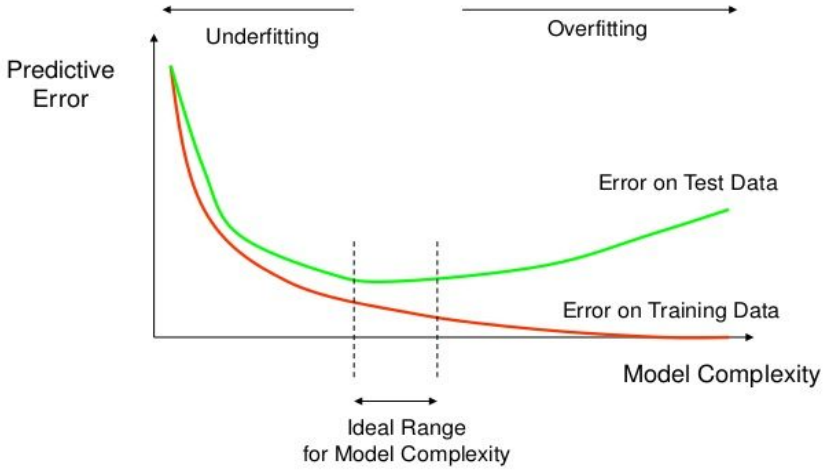
A. Singh et al, (2016). *Chem. Soc. Rev.*, 45

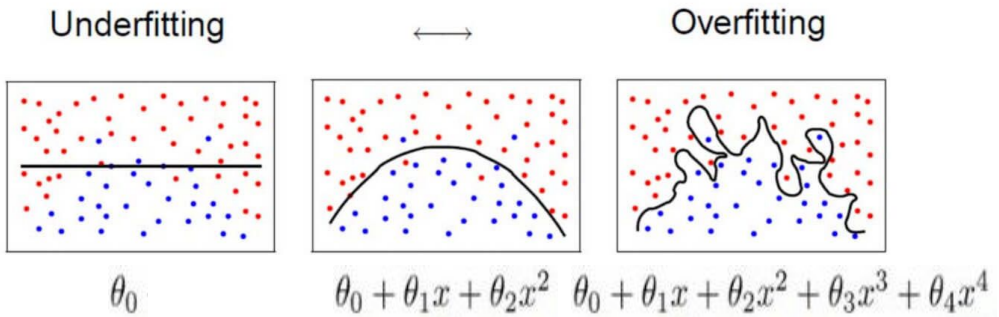
CRISPR/Cas9 challenges

- **Maximiser On-target**
- **Minimiser Off-target**



Optimisation des modèles de Machine Learning





Model Building

Feature Extraction

Extract features from 30mer sgRNA :

-1st & 2nd order nucleotide positions and amount

-GC content and Melting Temperature

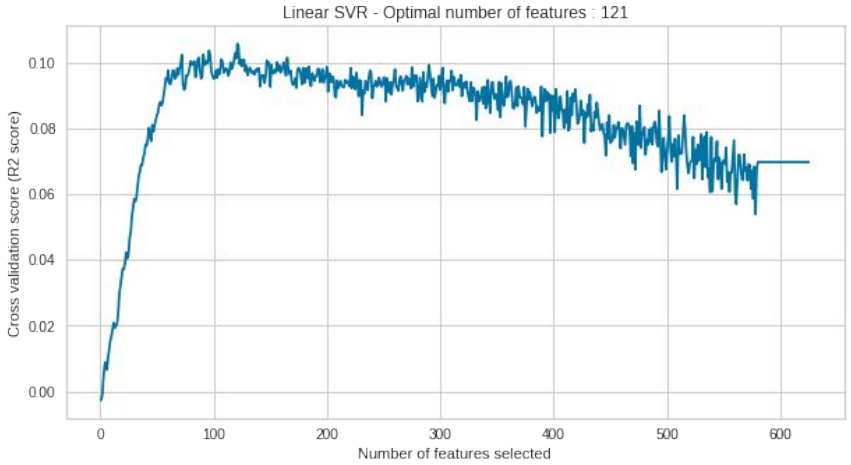
-PAM flanking nucleotides

➜ **625 features**

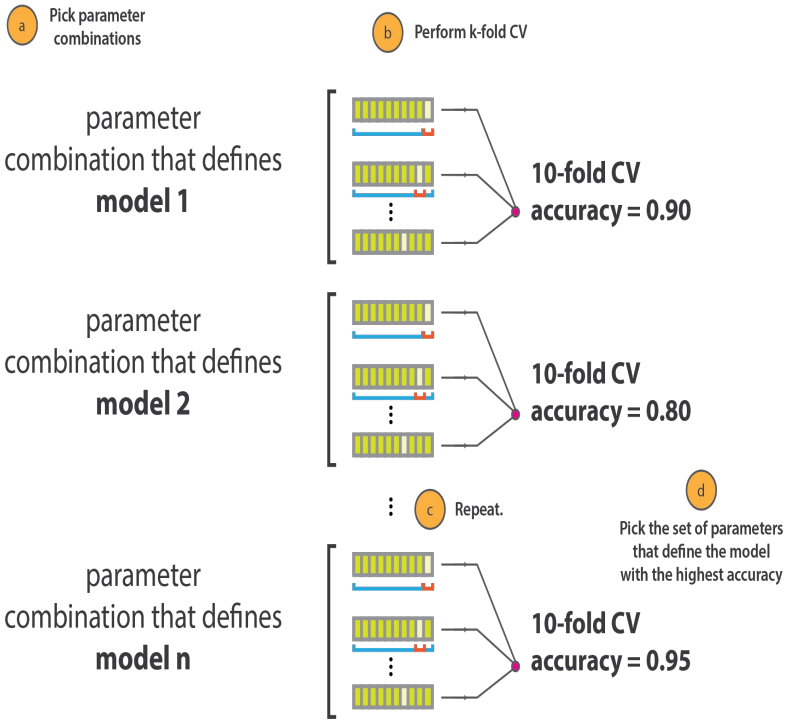
CATGATCATCGTACCCGAGATGACCGGCTC

Feature Selection

Cross-Validated Recursive Feature Elimination (RFECV)



Hyperparameter Optimization



Discussion

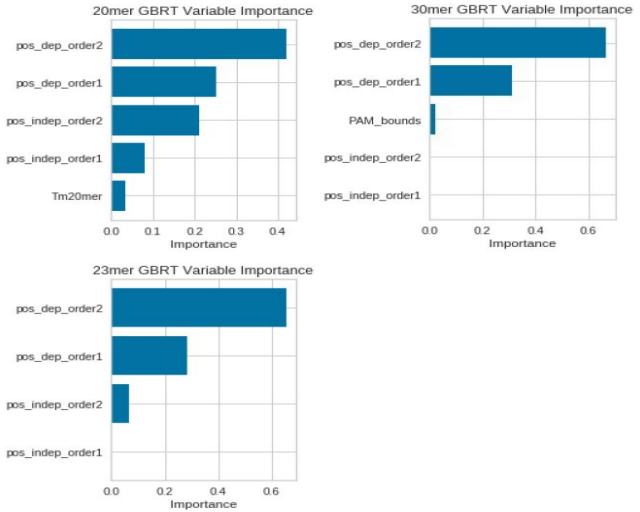
- ➢ Lots of targeted genes, but few sgRNAs per gene, may be suboptimal for on-target efficiency predictions
  - All genes are not equally essentials, nor are expressed at the same levels

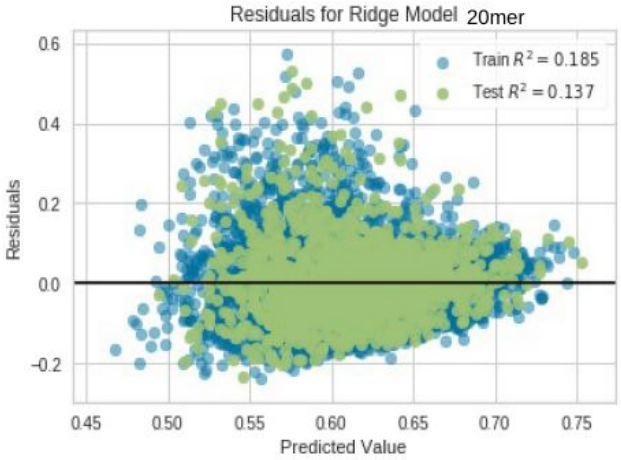
⇨ **Lots of sgRNA targeting a few genes**

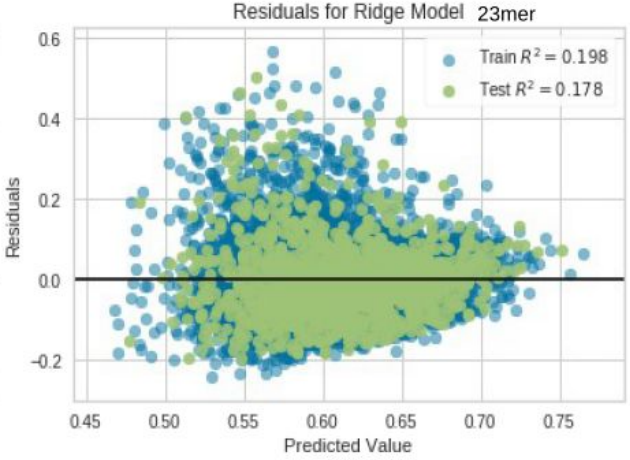
- ➢ Essentiality screen may also be suboptimal to test sgRNA efficiency
  - A “bad” guide that cuts everywhere would be falsely interpreted as a “good” guide

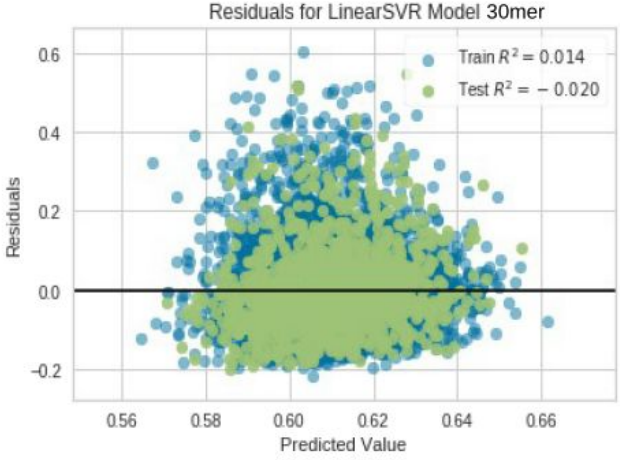
⇨ **Target genes that code for cell surface markers**

- ➢ Several potentially relevant features were excluded to promote generalization of the models
  - Chromatin state
  - Distance PAM - exon start
  - Peptide percent




Residuals for Ridge Model 20mer
Train $R^2 = 0.185$
Test $R^2 = 0.137$
Residuals
Predicted Value
0.6
0.4
0.2
0.0
−0.2
0.45
0.50
0.55
0.60
0.65
0.70
0.75


Residuals for Ridge Model 23mer
Train $R^2 = 0.198$
Test $R^2 = 0.178$
Residuals
Predicted Value
0.6
0.4
0.2
0.0
−0.2
0.45
0.50
0.55
0.60
0.65
0.70
0.75


Residuals for LinearSVR Model 30mer
Train $R^2 = 0.014$
Test $R^2 = -0.020$
Residuals
Predicted Value
0.6
0.4
0.2
0.0
−0.2
0.56
0.58
0.60
0.62
0.64
0.66