3000788 Intro to Comp Molec Biol

Lecture 28: Biomarker discovery

Fall 2025

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Today's agenda

- What are biomarkers?
- Designing a biomarker study
- Biomarker discovery
- Biomarker validation and interpretation
- Translating biomarkers into assays

What is a biomarker?

- A **measurable** characteristic of a biological system
  - Molecular
  - Physical
  - Behavioral
- **Indicative** of a normal/abnormal biological process
  - Sensitivity vs specificity
- **Predictive** of future response to treatment or exposure
- May be (or may not be) **explainable**

Biomarker in health checkup

Urinalysis

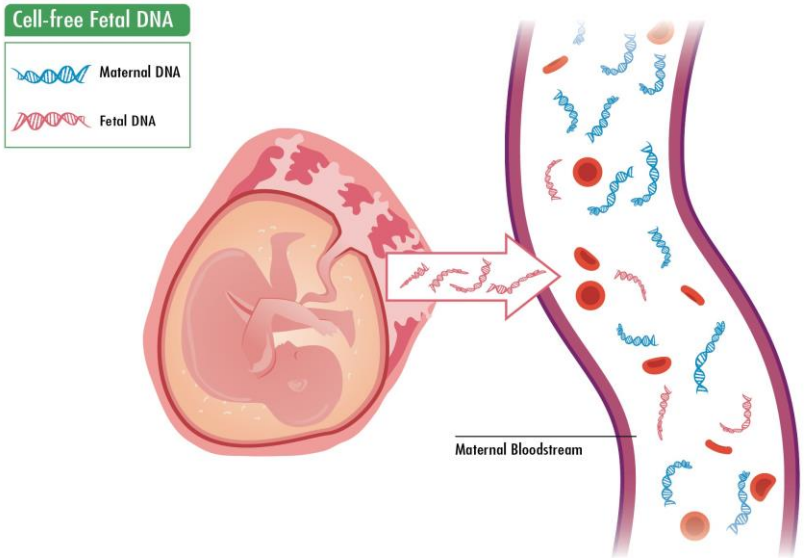
Blood Testing

Blood Pressure Screening

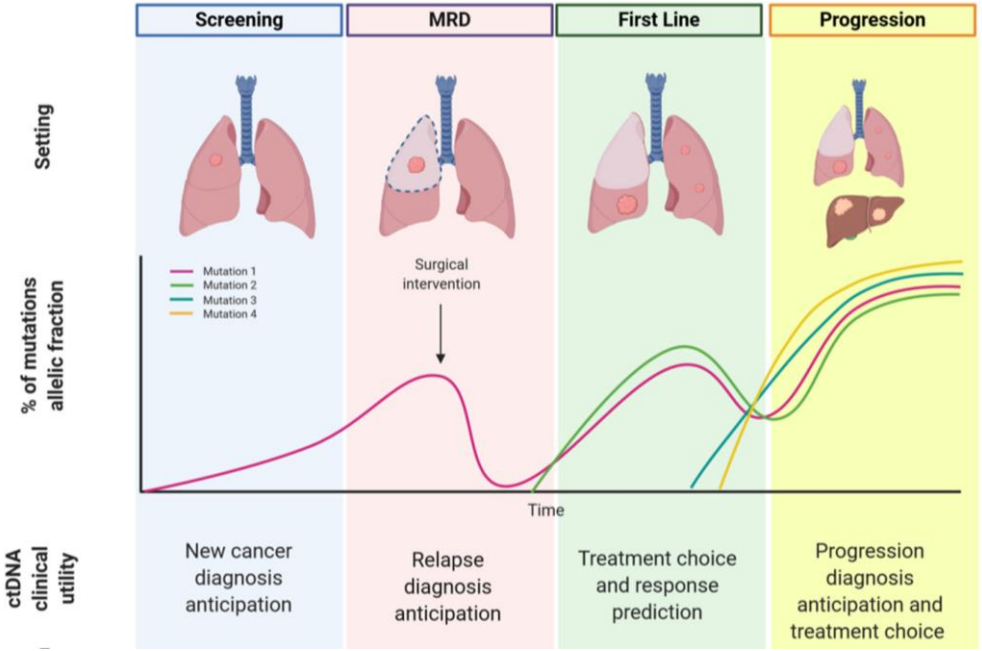
https://www.hudsonalpha.org/biomarkers-the-human-bodys-early-warning-system/

- High cholesterol and triglycerides → CVD risk
- ALT, AST, bilirubin → liver function
- Protein in urine → kidney function

Cell free DNA monitoring



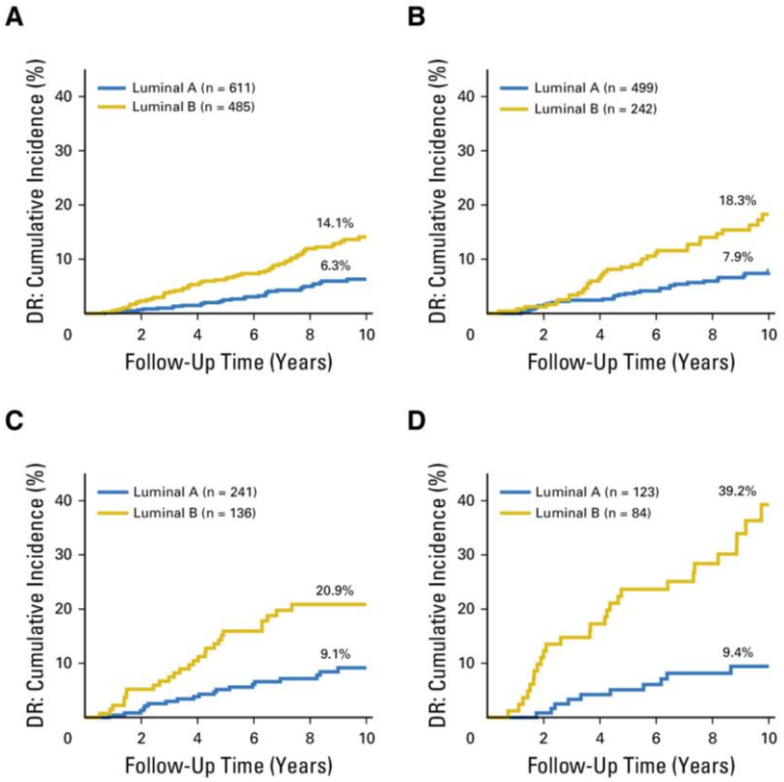
Yu, T. et al. Biomedical Research and Therapy 12:7418-7423 (2025)



Gobbini, E. et al. Cancers 12:3112 (2020)

- Fetal abnormal DNA or circulating tumor DNA

Cancer subtyping and prognostic biomarkers



- Use omics data to group patients
- Some subtypes exhibit high expressions or mutation rates of cancer-related genes
- Different survival and response → optimize treatment

Laenkholm, A.-V. et al. Journal of Clinical Oncology 36:725-740 (2018)

Preferred characteristics of a biomarker

- **Specific** to the biological condition of interest
  - How to rule out related conditions with shared biomarkers?
- Easy to **detect**
  - Aware of detection limit
- **Reproducible**
  - Include sample preparation, analysis, and interpretation
- **Non-invasive**, **scalable**, and **affordable** assay

Biomarker discovery-development workflow

- **Project design**: pick target application and biomarker type
- **Biomarker discovery**: assay selection, data handling, data analysis
- **Biomarker validation & interpretation**: data splitting, performance measurement, biological interpretation
- **Assay development**: calibration, reproducibility & scalability test

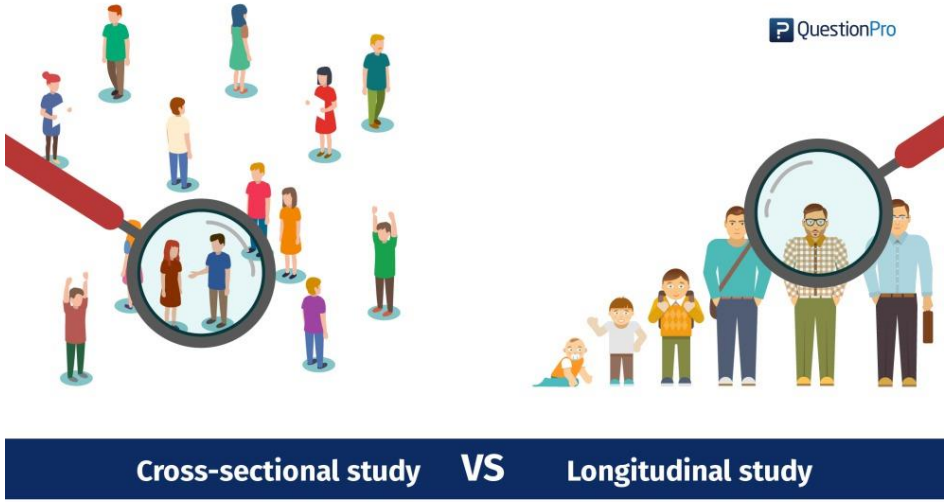
Designing a biomarker study

Pinpoint biological hypothesis / application

- What do you need the biomarker for?
  - Condition of interest, related conditions
- Which type of biomarker is feasible?
  - Molecular, physical, or behavioral
  - Longitudinal data
- What would the final deployment assay look like?
  - sample acquisition
  - assay cost and scalability
  - human interpretation of the test result

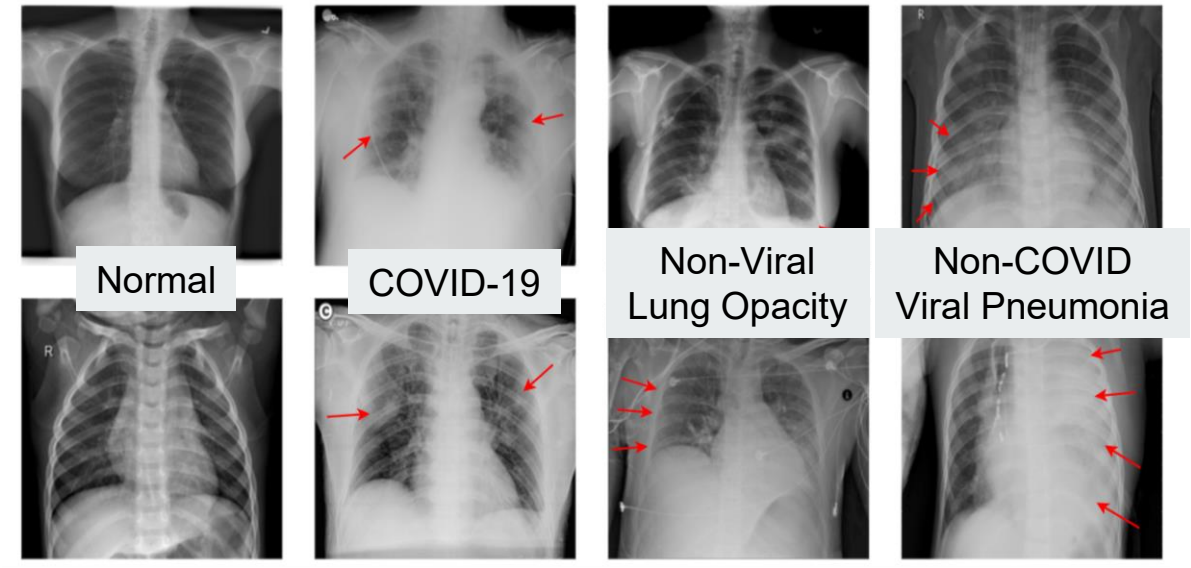
Cross-sectional vs longitudinal

https://www.questionpro.com/blog/cross-sectional-study-vs-longitudinal-study/



- Longitudinal data may be needed (but takes time to collect):
  - High biological variation
  - Multiple causes of disease
  - Biomarker is the change in signal

Confounding factors and covariates



Islam, M.N. et al. Healthcare 11:410 (2023)

- Be wary of **related conditions** that can exhibit the same biomarkers
- Be wary of **confounding factors** that can influence the level of biomarkers

Strategies for handling covariates

- Tighten inclusion/exclusion criteria
  - **Pros**: Minimize sample size, simplify analysis
  - **Cons**: Limited usability of identified biomarkers

- Collect more data and include them in the analysis
  - Stratified splitting or as model variables
  - **Pros**: Biomarkers can be applied broadly
  - **Cons**: More sample size, careful design & analysis, data input burden

What determines the sample size?

- Desired **uncertainty level** of the analysis result
- **Number of variables**
- **Biological variation**
  - Need enough data to represent the population
  - Depend on number of distinct subtypes / subpopulation / biological states

Data collection and management

- Balance budget and **potential exploration**
  - Unlikely to be able to collect more variables from the same population
- Keep track of **metadata**
  - Date / time / sample ID / etc.
  - Help tracing back when problem with data arise
- **Prevent human errors**
  - Enforce data type
  - Define common keywords: male, M, Male

Reproducibility and batch effect

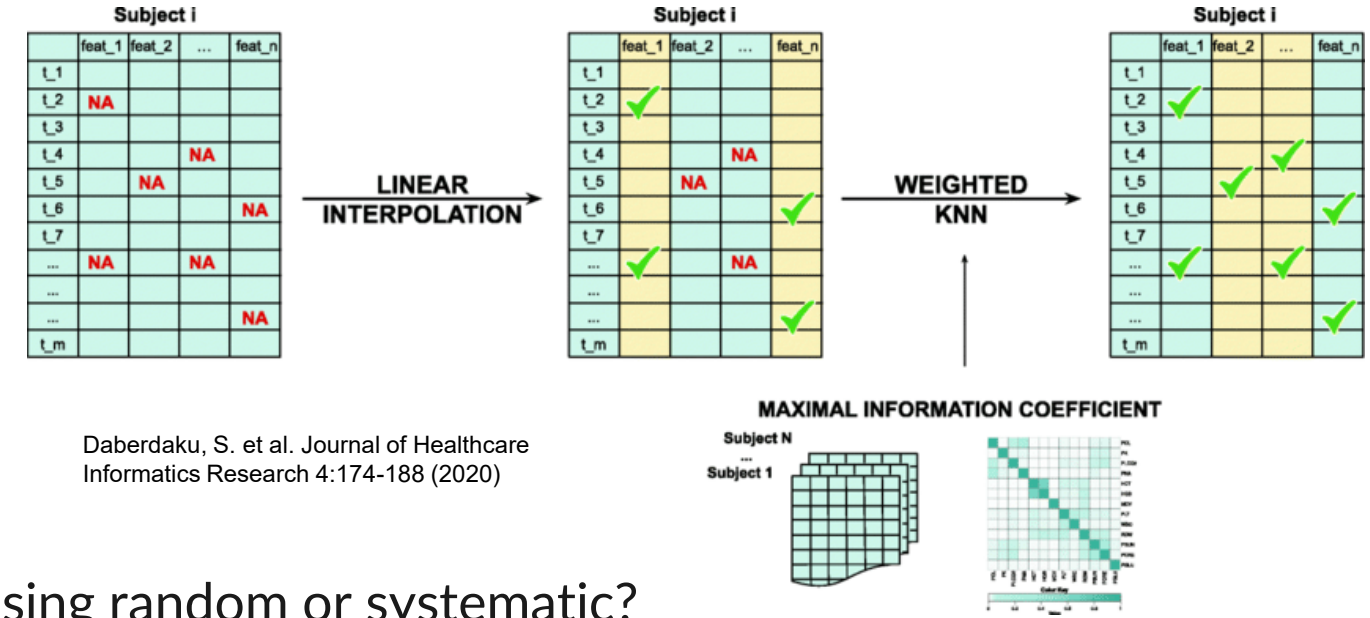
- Collecting **replicates for a subset of samples**
- Utilizing a **panel of experts to annotate** data
- If samples must be analyzed in batches, mix up the sample classes so that batches and covariates are uncorrelated
- Batch correction method typically requires all batches to contain every sample class (to use as anchors)

Data handling tips

Simple quality and error check

- View min and max values of numerical data
- Try converting (supposedly numerical) data to numbers
- View unique values for categorical data
- **Plot histograms and frequency tables**

Impute missing values

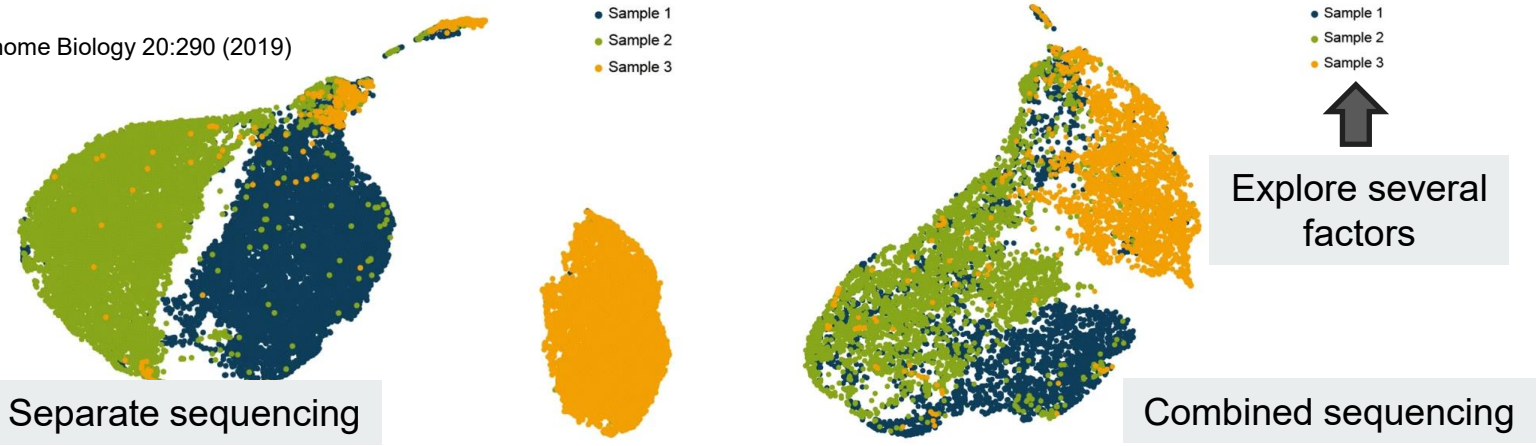


Daberdaku, S. et al. Journal of Healthcare Informatics Research 4:174-188 (2020)

- Is missing random or systematic?
- Impute by basic statistics: mean, median, min, max
- Impute by prediction: linear model, nearest neighbor

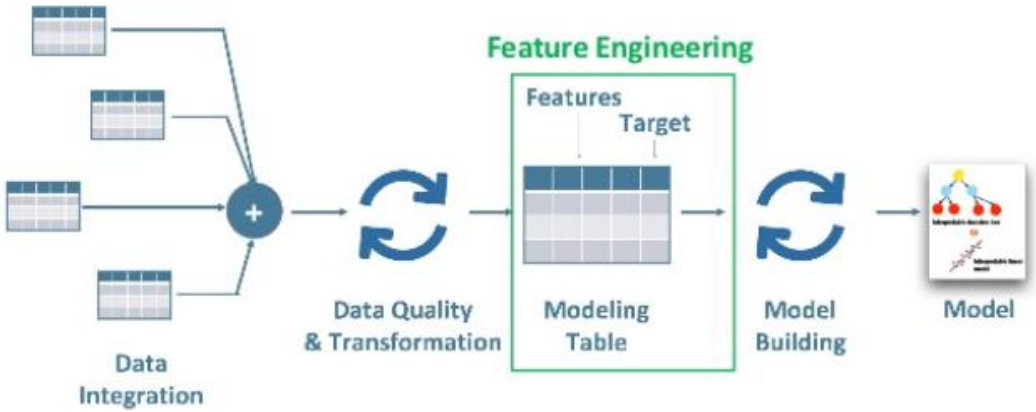
Batch effect

Xu, J. et al. Genome Biology 20:290 (2019)



- Visualizing using dimensionality reduction: PCA, *t*-SNE, UMAP
- Compare with biological knowledge
  - Color by biological state or feature values

Feature engineering

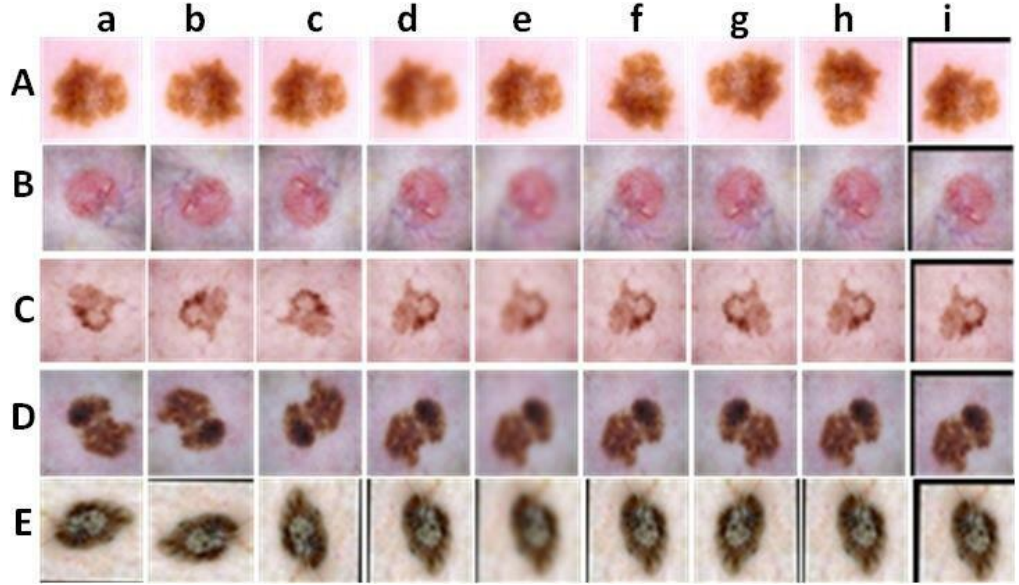


https://www.analyticsvidhya.com/blog/2020/10/getting-started-with-feature-engineering/

- Assay might not directly capture relevant biomarkers
  - **Normalization**: Food intake per body weight
  - **Nonlinear transformation**: log, power
  - **Interaction term**: product of two features

Data augmentation

Maher, H. and Kashmola, M. International Journal of Computing and Digital Systems 13:2210-2142 (2023)

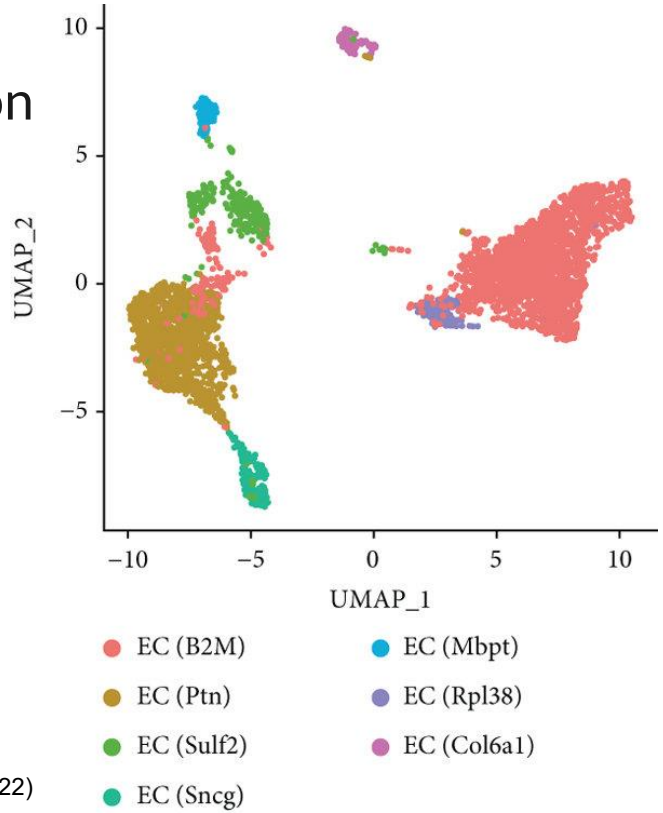


- Synthesize new data by following natural variation of the data
  - Geometric transformation of cell images
  - Adding noises to measurement / generative AI

Biomarker discovery analysis

Subtype / subpopulation inspection

- Visual inspection with dimensionality reduction
- Perform clustering
- Identify biomarker for each clusters
  - Univariate statistics
- Biological interpretation
  - Are clusters real or should be collapsed?



Huang, X. et al. BioMed Research International 2022:1-10 (2022)

Statistics for identifying differences

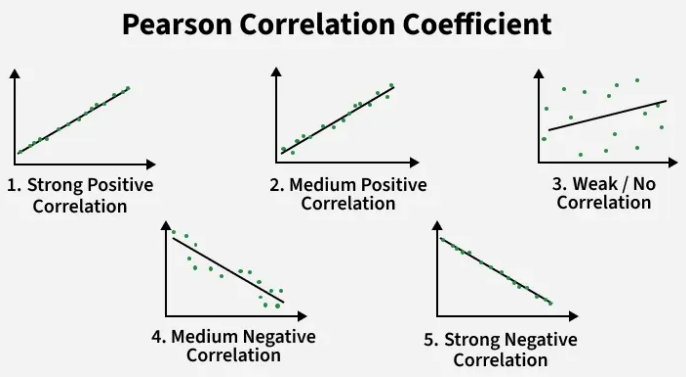
- For omics data with established model (e.g., read count), use those tools

- Otherwise:
  - **2 classes, unpaired**: *t*-test / Mann-Whitney U / Wilcoxon rank-sum
  - **2 classes, paired**: paired *t*-test / Wilcoxon signed-rank
  - **>2 classes**: ANOVA / Kruskal-Wallis
  - **Categorical**: Chi-squared / Fisher's exact test / McNemar (paired)

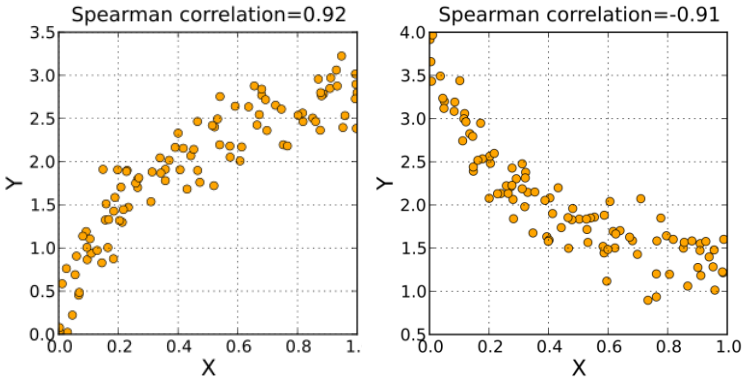
- Statistical significance is nice, but not an absolute requirement
  - Multivariate relationship can still be found
  - Good predictive performance can be still be achieved

Statistics for identifying associations

- **Pearson's** correlation (linear)
- **Spearman's** correlation (rank)
  - Difference in rank values
- **Kendall's** correlation (rank)
  - Concordance in rank across pairs of samples
  - Related to C-index in survival analysis

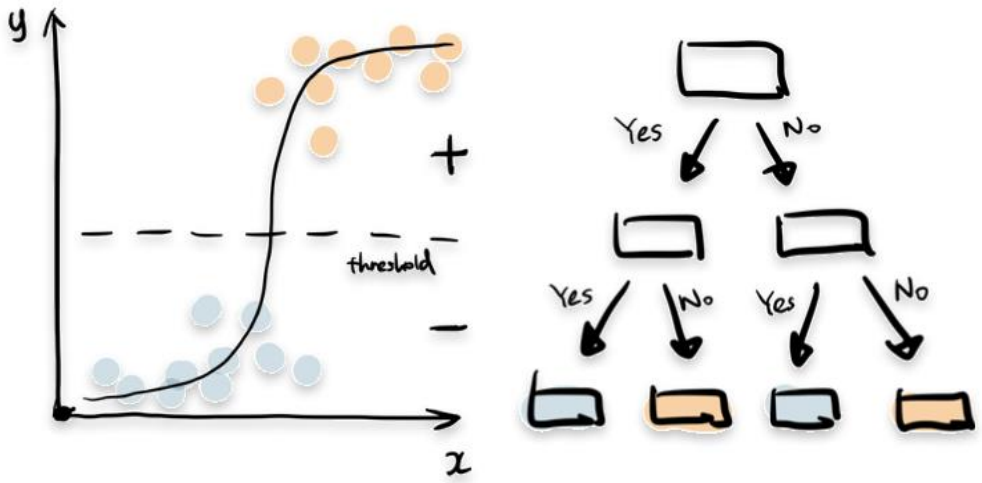


https://www.geeksforgeeks.org/maths/pearson-correlation-coefficient/



https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

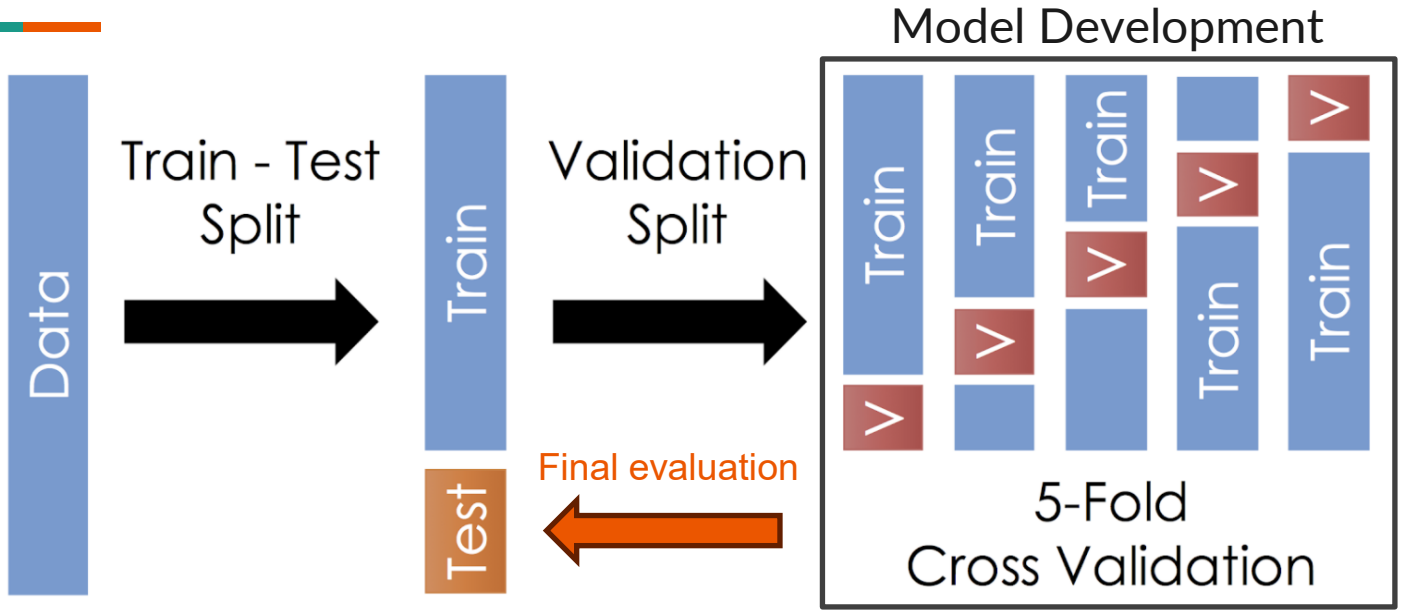
Basic statistical and ML models



https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

- Generalized linear model
  - $f(y) = a_1x_1 + \cdots + a_nx_n$
- Logistic regression
  - $\log\left(\frac{p}{1-p}\right) = a_1x_1 + \cdots + a_nx_n$
- Tree models
  - Random forest
  - Gradient boosting tree

How to identify the best models?



Source: medium.com

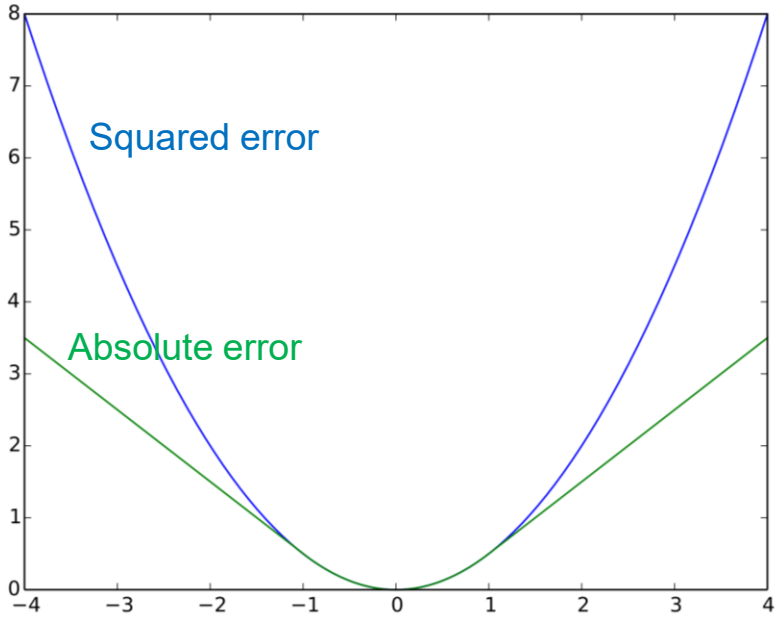
- Hold out a **Test** set for an unbiased evaluation
- During development, cycle through splitting (cross-validation) to train and evaluate the models

Performance metrics (for regression)

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Mean Absolute Percentage Error (MAPE)

- R^2 (Coefficient of Determination)

- **Measure how "off" are the predictions from observed values**



Performance metrics (for classification)

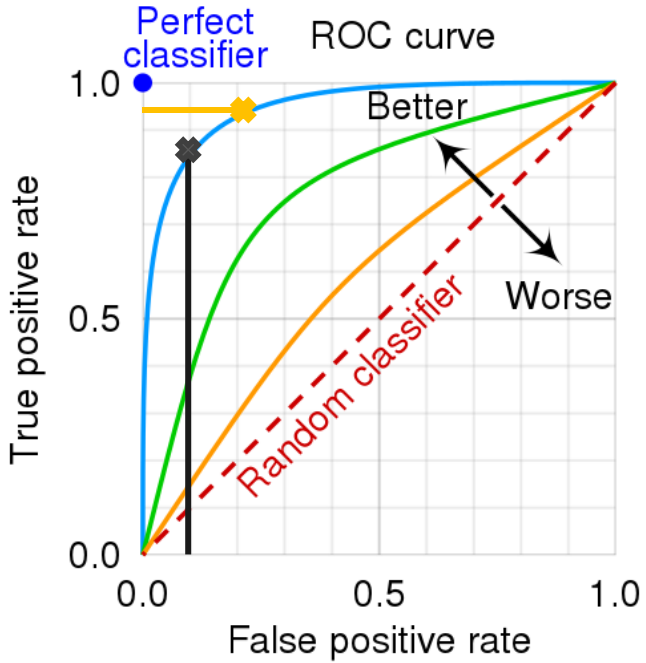
		Predicted: Negative	Predicted: Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive
		Predicted < 0.5	Predicted > 0.5

- Accuracy = (TN + TP) / total
- Precision = TP / (TP + FP) = Positive predictive value

- Recall = TP / (TP + FN) = Sensitivity
- Specificity = TN / (TN + FP)

Specific to the cutoff value

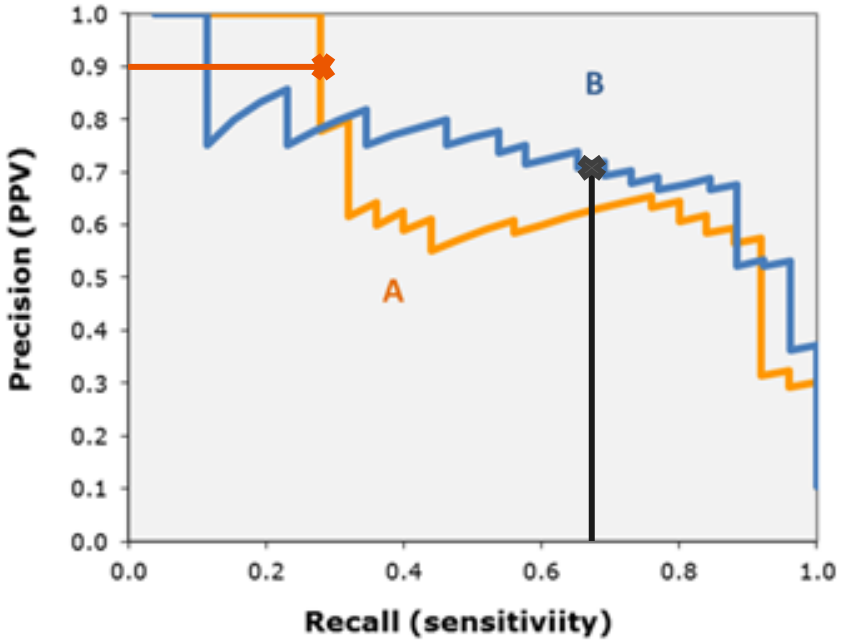
ROC: Receiver Operating Characteristic curve



- Visualize sensitivity-specificity tradeoff at every output threshold
- Area under the ROC curve (AUROC, AUC)
  - Random guess = 0.5
  - Perfect model = 1.0
- Pick threshold based on application needs
  - Specificity >0.9
  - Sensitivity >0.9

https://commons.wikimedia.org/wiki/File:Roc_curve.svg

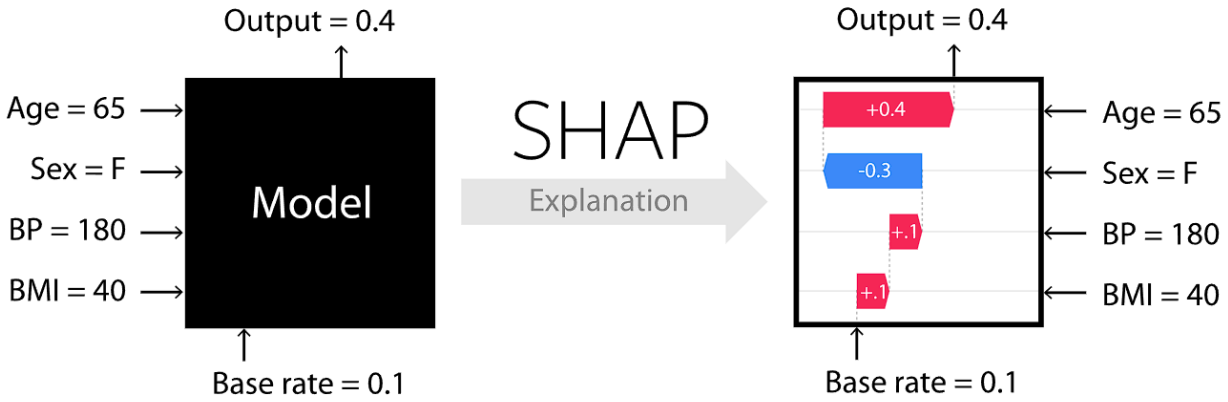
Precision-Recall curve



- Visualize sensitivity (recall)-precision tradeoff at every output threshold
- Area under the PR curve (AUPRC)
  - Average precision
- Pick threshold based on application needs

https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used

Biomarker importance and selection



- Explainable biomarkers are more trustworthy
- Sometimes not possible, e.g., image-based biomarkers
- Need replication and validation on unrelated datasets

Biomarker validation

What can affect biomarker performance?

- Completeness of the **discovery study's design**
  - Unaccounted for covariates and related conditions
  - Inclusion & exclusion criteria

- **Chance** (when sample size is low)

- **Technical variation** across sample preparation and assay instrument
  - Especially when deployed assay differs from discovery assay

Cross-platform validation

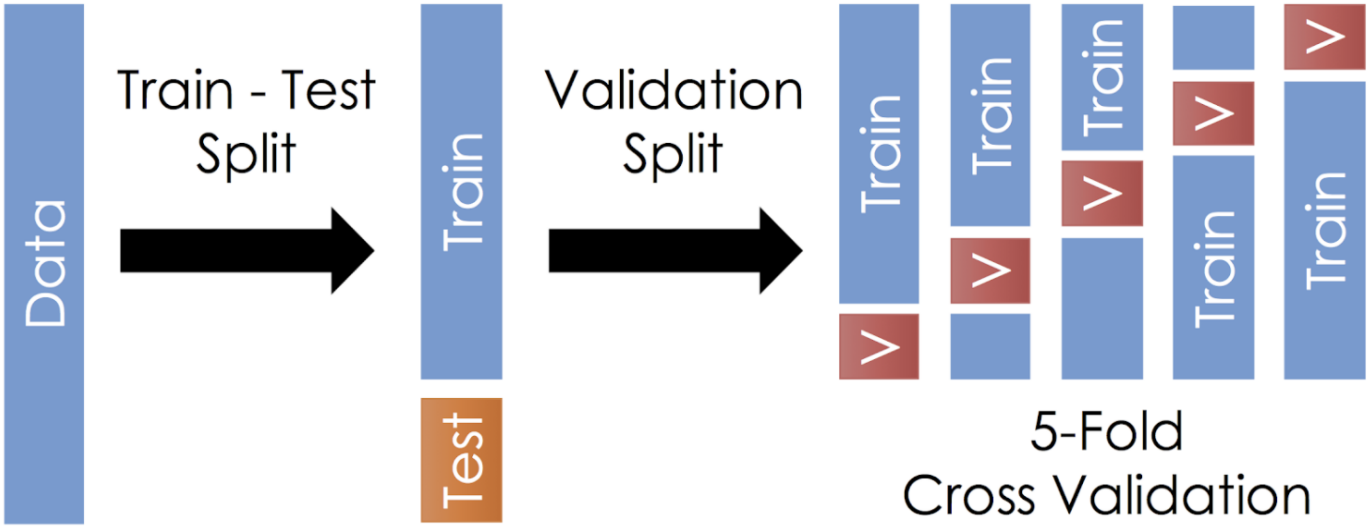
Chia, R. et al. Nature Medicine 31:3440-3450 (2025)

Table 1 | Proteins associated with ALS based on differential abundance analysis

Protein	Olink Explore 3072 (plasma)				SomaScan 7k (CSF)		
	Discovery Cohort		Replication Cohort		Cohort		
	$\log_2$ fold change	FDR-adjusted *P* value	$\log_2$ fold change	FDR-adjusted *P* value	Estimate	s.e.	*P* value
NEFL	2.34	2.22×10^{-88}	2.36	**1.39×10^{-19}**	2.02	0.11	**8.71×10^{-35}**
ALDH3A1	2.22	6.27×10^{-40}	2.64	**4.47×10^{-15}**	−0.04	0.02	0.057
MEGF10	0.88	2.14×10^{-39}	0.97	**1.22×10^{-11}**	0.12	0.02	**5.34×10^{-10}**
CORO6	1.24	1.39×10^{-32}	0.72	**0.0174**	NA	NA	NA
HS6ST2	0.77	5.81×10^{-30}	0.68	**4.61×10^{-4}**	−0.02	0.02	0.44
CSRP3	2.19	2.56×10^{-29}	1.53	**7.18×10^{-4}**	0.11	0.10	0.25
MYBPC1	1.50	5.31×10^{-28}	0.87	**0.0392**	−0.09	0.09	0.30
CA3	1.08	5.91×10^{-27}	0.59	**0.0374**	0.36	0.32	0.27
MYLPF	1.56	1.01×10^{-25}	1.26	**6.01×10^{-4}**	NA	NA	NA
MYOM3	1.32	1.58×10^{-23}	0.69	0.166	−0.03	0.02	0.10
RBFOX3	0.73	2.26×10^{-22}	0.45	0.135	NA	NA	NA

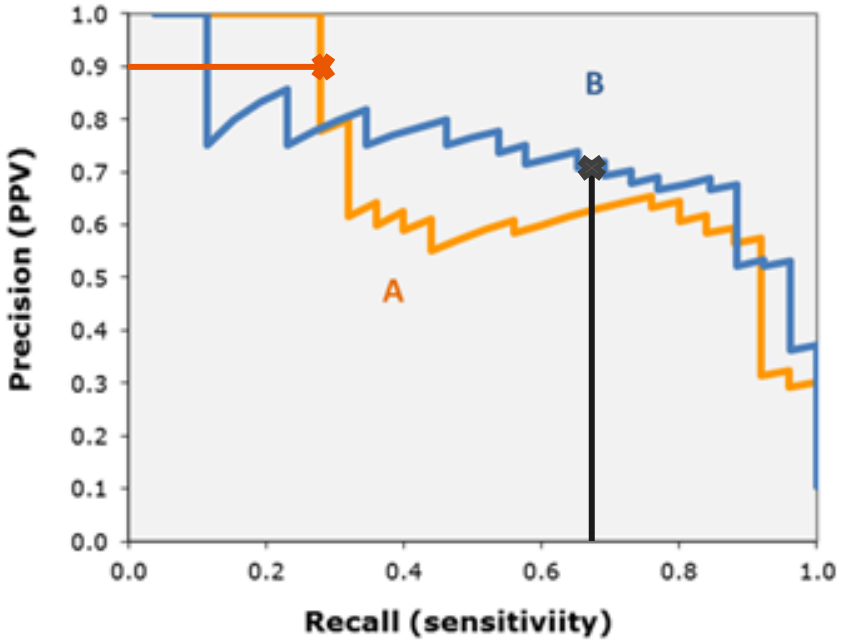
- Confirm biomarkers across sample, technology, and omics layers

The need for external validation



- Cross-validation **overestimates real-world performance**
  - Minimal technical variation
- Need unbiased validation, e.g., newer samples, different batch

Model cutoff calibration

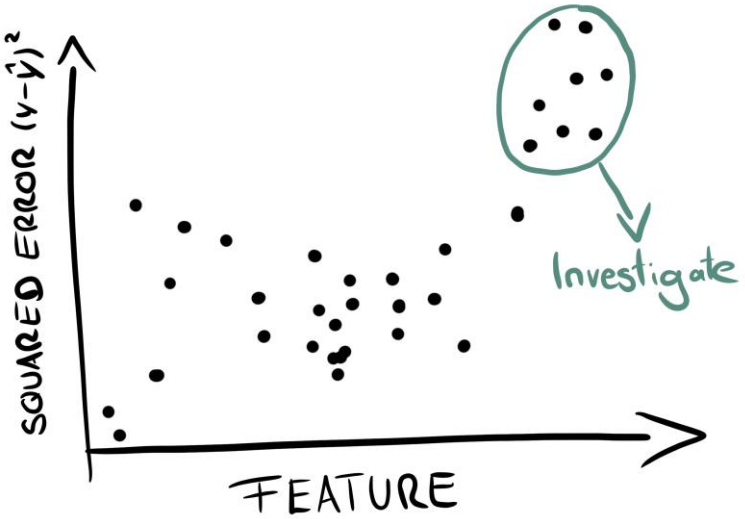


https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used

- When deploying assay, a cutoff on biomarker level or model output must be selected
- Which performance metrics to optimize? What are the requirements?
- How should “intermediate” outputs, e.g., 0.4 and 0.6 be interpreted?

Error analysis

- Not all errors are equal!
- There are **easy and hard cases**
- **Systematic errors** can reflect missing covariates, limited model capability, or bias in data collection
  - **Ex**: Less drug allergy incidence in elderly (because physicians are inclined to use only safe drugs)



https://mindfulmodeler.substack.com/p/a-simple-recipe-for-model-error-analysis

Uncertainty estimation

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.8885	0.057	-15.524	0.000	-1.013	-0.764
year	-6.9170	1.255	-5.510	0.000	-9.652	-4.182
Air pollution	1.3691	0.444	3.086	0.009	0.402	2.336
Alcohol use	-0.7543	0.156	-4.821	0.000	-1.095	-0.413
Dietary risks	-0.1104	0.126	-0.878	0.397	-0.384	0.164
High LDL cholesterol	-0.0110	0.058	-0.191	0.852	-0.137	0.115
High body-mass index	1.4454	1.461	0.989	0.342	-1.738	4.629
High fasting plasma glucose	0.0754	0.185	0.407	0.691	-0.328	0.479
High systolic blood pressure	0.1769	0.049	3.635	0.003	0.071	0.283
Kidney dysfunction	-0.3302	0.059	-5.576	0.000	-0.459	-0.201
Low physical activity	1.3026	0.181	7.189	0.000	0.908	1.697
Tobacco	-3.6491	0.700	-5.213	0.000	-5.174	-2.124

- Use **bootstrapping** to generate similar datasets
- **Estimate variance** for each parameter and model performance
- Average coefficient doesn't show consistency
  - **Ex**: -0.1, -0.01, 3.7, 5.2, 7.1
  - **Ex**: 1.2, 1.6, 1.7, 2.0, 2.2

Translating biomarker into assays

Picking the right assay to deploy

- Target specific biomarkers (molecules, imaging modalities, etc.)

- **Scalable & accessible**:
  - Easy to collect samples
  - Common instrument & cheap reagents
  - Easy to use for non-technical person

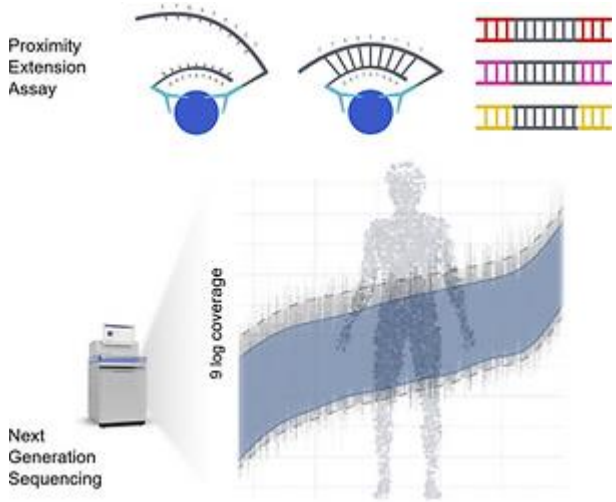
- **Reproducible**:
  - Streamline the pipeline, from sample preparation to result interpretation
  - Develop clear guideline

- Plan for troubleshooting and improvement
  - Collect extra details and feedback

Moving from discovery result to deployed assay

- During validation phase, identified biomarkers can be measured using candidate deployed assays right away
  - Some biomarkers can fail on the new assay
  - Or perform in parallel with discovery assay
  - **Ex**: From RNA-seq to PCR panel

- Look out for new targeted techniques
  - **Ex**: Detecting proteins via DNA sequencing



https://www.immunodiagnostics.com.hk/olink-3072

Any question?

- Congratulations on completing the course!

- Recommended next steps
  - Hands-on omics analysis practice on the topics of your interest
  - **MIT OCW 7.91** More in-depth bioinformatics algorithms
  - **MIT OCW 6.0001/6.0002** Computer programming and data science
  - https://www.youtube.com/@ManolisKellis1/courses Advanced topics in computational molecular biology x machine learning