

Bioinformatics Analysis in R

Advanced Gene Expression: Analysis of Cancer Genome Atlas

Ivan G. Costa, Tiago Maie

Institute for Computational Genomics
RWTH University Hospital
www.costalab.org


Summary

1. Obtain data from cancer patients from TCGA
2. Pre-process and analysis of RNA-seq data
3. Use machine learning to build a classifier for personalised medicine
4. Use interesting markers for survival analysis

The Cancer Genome Atlas

- TCGA is a NCI (US) funded project to generate cohorts of cancers:
 - Currently 33 cancers with 80-780 patients
- Comprehensive data from tissues:
 - Histology, clinical, gene expression profiling, copy number variation, DNA methylation using arrays or sequencing
- Data is publicly available upon generation and deposited in a portal (portal.gdc.cancer.gov)

The Cancer Genome Atlas - Portal

 **NATIONAL CANCER INSTITUTE**
GDC Data Portal

Home Projects Exploration Analysis Repository

Quick Search Manage Sets Login Cart 0 GDC Apps

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects Exploration Analysis Repository

Data Portal Summary

[Data Release 13.0 - September 27, 2018](#)

PROJECTS

43

FILES

358,092

PRIMARY SITES

69

GENES

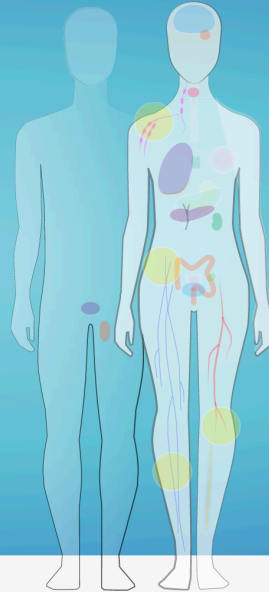
22,147

CASES

33,096

MUTATIONS

3,142,246





Cases by Major Primary Site


Adrenal Gland	100
Bile Duct	50
Bladder	100
Blood	1000
Bone	500
Bone Marrow	100
Brain	1000
Breast	3500
Cervix	500
Colorectal	2500
Esophagus	500
Eye	100
Head and Neck	1000
Kidney	2000
Liver	1000
Lung	4500
Lymph Nodes	500
Nervous System	2000
Ovary	1500
Pancreas	1000
Pleura	500
Prostate	1000
Skin	1000
Soft Tissue	100
Stomach	1000
Testis	100
Thymus	100
Thyroid	1000
Uterus	1000


GDC Applications


The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:


 Data Portal


 Website

 Data Transfer Tool

 API

 Data Submission Portal

 Documentation

 Legacy Archive

The Cancer Genome Atlas - Portal

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home Projects Exploration Analysis Repository

Quick Search Manage Sets Login Cart 0 GDC Apps

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects Exploration Analysis Repository

Q e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Data Portal Summary

Data Release 13.0 - September 27, 2018

Category	Count
PROJECTS	43
PRIMARY SITES	69
CASES	33,096
FILES	358,092
GENES	22,147
MUTATIONS	3,142,246

Cases by Major Primary Site

Primary Site	Cases
Adrenal Gland	100
Bile Duct	100
Bladder	100
Blood	100
Bone	100
Bone Marrow	100
Brain	100
Breast	100
Cervix	100
Colorectal	100
Esophagus	100
Eye	100
Head and Neck	100
Kidney	100
Liver	100
Lung	100
Lymph Nodes	100
Nervous System	100
Ovary	100
Pancreas	100
Pleura	100
Prostate	100
Skin	100
Soft Tissue	100
Stomach	100
Testis	100
Thymus	100
Thyroid	100
Uterus	100

GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

Data Portal Website Data Transfer Tool API Data Submission Portal Documentation Legacy Archive

Check a gene or cancer type!
I will try liver

LIHC - Liver Hepatocellular Carcinoma

Explore Project Data

 Biospecimen Clinical

 Manifest

Summary

Project ID	TCGA-LIHC
Project Name	Liver Hepatocellular Carcinoma
Disease Type	Adenomas and Adenocarcinomas
Primary Site	Liver and intrahepatic bile ducts
Program	TCGA

CASES

377



FILES

10,814



ANNOTATIONS

28



Cases and File Counts by Data Category

Data Category	Cases (n=377)	Files (n=10,814)
Raw Sequencing Data	377 <div><div></div></div>	1,637 <div><div></div></div>
Transcriptome Profiling	376 <div><div></div></div>	2,122 <div><div></div></div>
Simple Nucleotide Variation	375 <div><div></div></div>	3,032 <div><div></div></div>
Copy Number Variation	376 <div><div></div></div>	1,536 <div><div></div></div>
DNA Methylation	377 <div><div></div></div>	430 <div><div></div></div>
Clinical	377 <div><div></div></div>	423 <div><div></div></div>
Biospecimen	377 <div><div></div></div>	1,634 <div><div></div></div>

Cases and File Counts by Experimental Strategy

Experimental Strategy	Cases (n=377)	Files (n=10,814)
Diagnostic Slide	365 <div><div></div></div>	379 <div><div></div></div>
Tissue Slide	377 <div><div></div></div>	491 <div><div></div></div>
WXS	376 <div><div></div></div>	3,820 <div><div></div></div>
RNA-Seq	371 <div><div></div></div>	1,696 <div><div></div></div>
miRNA-Seq	373 <div><div></div></div>	1,275 <div><div></div></div>
Genotyping Array	376 <div><div></div></div>	1,536 <div><div></div></div>
Methylation Array	377 <div><div></div></div>	430 <div><div></div></div>

LIHC - Liver Hepatocellular Carcinoma

[Explore Project Data](#)[Biospecimen](#)[Clinical](#)[Manifest](#)

Summary

Project ID	TCGA-LIHC
Project Name	Liver Hepatocellular Carcinoma
Disease Type	Adenomas and Adenocarcinomas
Primary Site	Liver and intrahepatic bile ducts
Program	TCGA

CASES

[377](#)

FILES

[10,814](#)

ANNOTATIONS

[28](#)

Cases and File Counts by Data Category

Data Category	Cases (n=377)	Files (n=10,814)
Raw Sequencing Data	377	1,637
Transcriptome Profiling	376	2,122
Simple Nucleotide Variation	375	3,032
Copy Number Variation	376	1,536
DNA Methylation	377	430
Clinical	377	423
Biospecimen	377	1,634

Cases and File Counts by Experimental Strategy

Experimental Strategy	Cases (n=377)	Files (n=10,814)
Diagnostic Slide	365	379
Tissue Slide	377	491
WXS	376	3,820
RNA-Seq	371	1,696
miRNA-Seq	373	1,275
Genotyping Array	376	1,536
Methylation Array	377	430

Gene expression data!

Files

Cases

Add a File Filter

File

Q

e.g. 142682.bam, 4f6e2e7a-b...

Data Category

Simple Nucleotide Variation

3,032

Transcriptome Profiling

2,122

Raw Sequencing Data

1,637

Biospecimen

1,634

Copy Number Variation

1,536

2 More...

Data Type

Gene Expression Quantification

1,272

Isoform Expression Quantification

425

miRNA Expression Quantification

425

Experimental Strategy

RNA-Seq

1,272

miRNA-Seq

850

Workflow Type

BCGSC miRNA Profiling

850

HTSeq - Counts

424

HTSeq - FPKM

424

HTSeq - FPKM-UQ

424

Data Format

TXT

2,122

Platform

No data for this field

Access

open

2,122

Clear

Project Id

IS

TCGA-LIHC

AND

Data Category

IS

Transcriptome Profiling

Add All Files to Cart

Manifest

View 376 Cases in Exploration

View Images

Files (2,122)

Cases (376)

Primary Site

Project

Data Category

Show More

Showing 1 - 20 of 2,122 files

	Access	File Name	Cases	Project
	open	7085ee3a-b604-4a12-a877-63eef2d905e8.htseq.counts.gz	1	TCGA-LIHC
	open	acf3d05a-0ca4-4fee-8f07-44b93017b5fd.mirbase21.isoforms.quantification.txt	1	TCGA-LIHC
	open	13240f8b-ae36-4f5f-8e95-2c9d0c83e58c.FPKM-UQ.txt.gz	1	TCGA-LIHC
	open	77e29a20-68d3-4881-a3ac-a564359bbc05.FPKM-UQ.txt.gz	1	TCGA-LIHC
	open	103b1320-8c4e-44ea-9449-fdc6b6b405f94.htseq.counts.gz	1	TCGA-LIHC
	open	466776cb-6906-4da2-b788-a05a154decf3.mirbase21.mirnas.quantification.txt	1	TCGA-LIHC
	open	e4c90512-0e06-4517-95fe-c10b999f5f81.mirbase21.mirnas.quantification.txt	1	TCGA-LIHC
	open	5f94c33f-588b-4b6a-9c13-4505b0f94403.htseq.counts.gz	1	TCGA-LIHC
	open	6ce06871-a6a4-4a4a-bd08-0c448914dfcf.FPKM.txt.gz	1	TCGA-LIHC
	open	a762a98f-9041-47e2-8561-46fa6e396f12.htseq.counts.gz	1	TCGA-LIHC
	open	61ec8919-8b12-43d7-b127-8b68a66bd033.mirbase21.mirnas.quantification.txt	1	TCGA-LIHC
	open	f3e152ef-5048-4157-a195-d13ed8851170.htseq.counts.gz	1	TCGA-LIHC
	open	ca28f377-d686-41f9-90fb-9da55fec40cb.mirbase21.isoforms.quantification.txt	1	TCGA-LIHC
	open	13240f8b-ae36-4f5f-8e95-2c9d0c83e58c.FPKM.txt.gz	1	TCGA-LIHC
	open	e035a46e-6114-4a64-b5ae-9e6209223493.FPKM.txt.gz	1	TCGA-LIHC
	open	a96f2f6c-38e0-453c-961d-aa83b92652da.mirbase21.mirnas.quantification.txt	1	TCGA-LIHC
	open	a0c56eec-568a-46b0-88db-f14d64a3942b.FPKM.txt.gz	1	TCGA-LIHC
	open	9c644f65-0ebb-4862-98a9-308b81c8fb26.mirbase21.mirnas.quantification.txt	1	TCGA-LIHC
	open	ad114591-0409-4bc5-8f0b-dbb44a5ad0eb.mirbase21.isoforms.quantification.txt	1	TCGA-LIHC
	open	3edd413e-831d-442a-be8d-70b2f49e9d67.FPKM.txt.gz	1	TCGA-LIHC

Show

20

entries

LIHC - Liver Hepatocellular Carcinoma

Files Cases

Add a File Filter

File

Q e.g. 142682.bam, 4f6e2e7a-b...

Data Category

- ☐ Simple Nucleotide Variation 3,032
- ☒ Transcriptome Profiling 2,122
- ☐ Raw Sequencing Data 1,637
- ☐ Biospecimen 1,634
- ☐ Copy Number Variation 1,536

2 More

Data Type

- ☐ Gene Expression Quantification 1,272
- ☐ Isoform Expression Quantification 425
- ☐ miRNA Expression Quantification 425

Experimental Strategy

- ☐ RNA-Seq 1,272
- ☐ miRNA-Seq 860

Workflow Type

- ☐ BCGSC miRNA Profiling 860
- ☐ HTSeq - Counts 424
- ☐ HTSeq - FPKM 424
- ☐ HTSeq - FPKM-UQ 424

Data Format

- ☐ TXT 2,122

Platform

No data for this field

Access

- ☐ open 2,122

Clear Project Id IS TCGA-LIHC AND Data Category IS Transcriptome Profiling

Add All Files to Cart Manifest View 376 Cases in Exploration View Images

Files (2,122) Cases (376)

Primary Site Project Data Category

Showing 1 - 20 of 2,122 files

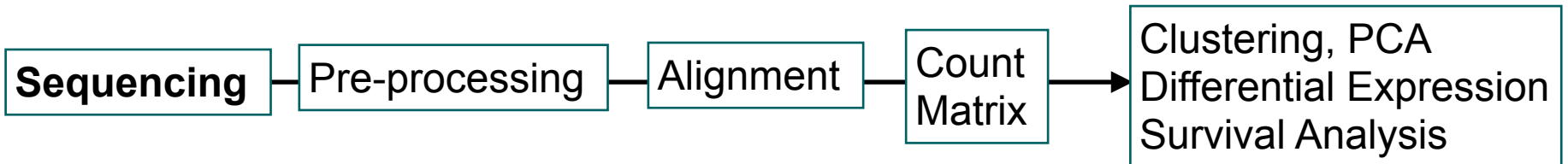
Access	File Name	Cases	Project
	7086ee31-f844-4422-a112-678e2c90508.htseq.counts.gz	1	TCGA-LIHC
	acf0003a-0ca7-4f6e-bf07-44b9d01b5fd.mirbase21.isoforms.quantification.txt	1	TCGA-LIHC
	13240f8b-ae36-4f5f-8e95-2c9d0c83e58c.FPKM-UQ.txt.gz	1	TCGA-LIHC
	77e29a20-68d3-4881-a3ac-a564359bbc05.FPKM-UQ.txt.gz	1	TCGA-LIHC
	103b1320-8c4e-44ea-9449-fdc6b6b405f94.htseq.counts.gz	1	TCGA-LIHC
	466776cb-6906-4da2-b788-a05a154decf3.mirbase21.mirnas.quantification.txt	1	TCGA-LIHC
	e4c90512-0e06-4517-95fe-c10b999f5f81.mirbase21.mirnas.quantification.txt	1	TCGA-LIHC
	5f94c33f-588b-4b6a-9c13-4505b0f94403.htseq.counts.gz	1	TCGA-LIHC
	6ce06871-a6a4-4a4a-bd08-0c448914dfcf.FPKM.txt.gz	1	TCGA-LIHC
	a762a98f-9041-47e2-8561-46fae396f12.htseq.counts.gz	1	TCGA-LIHC
	61ec819-b12-43d7-b127-8b68a661d033.mirbase21.mirnas.quantification.txt	1	TCGA-LIHC
	f3e15caf-544c-41b1-195d-11e18851170.htseq.counts.gz	1	TCGA-LIHC
	ca28f37f-d686-41f9-90fb-9da55fec40cb.mirbase21.isoforms.quantification.txt	1	TCGA-LIHC
	13240f8b-ae36-4f5f-8e95-2c9d0c83e58c.FPKM.txt.gz	1	TCGA-LIHC
	e035a46e-6114-4a64-b5ae-9e6209223493.FPKM.txt.gz	1	TCGA-LIHC
	a96f2f6c-38e0-453c-961d-aa83b92652da.mirbase21.mirnas.quantification.txt	1	TCGA-LIHC
	a0c56eec-568a-46b0-88db-f14d64a3942b.FPKM.txt.gz	1	TCGA-LIHC
	9c644f65-0ebb-4862-98a9-308b81c8fb26.mirbase21.mirnas.quantification.txt	1	TCGA-LIHC
	ad114591-0409-4bc5-8f0b-dbb44a5ad0eb.mirbase21.isoforms.quantification.txt	1	TCGA-LIHC
	3edd413e-831d-442a-be8d-70b2f49e9d67.FPKM.txt.gz	1	TCGA-LIHC

Show 20 entries

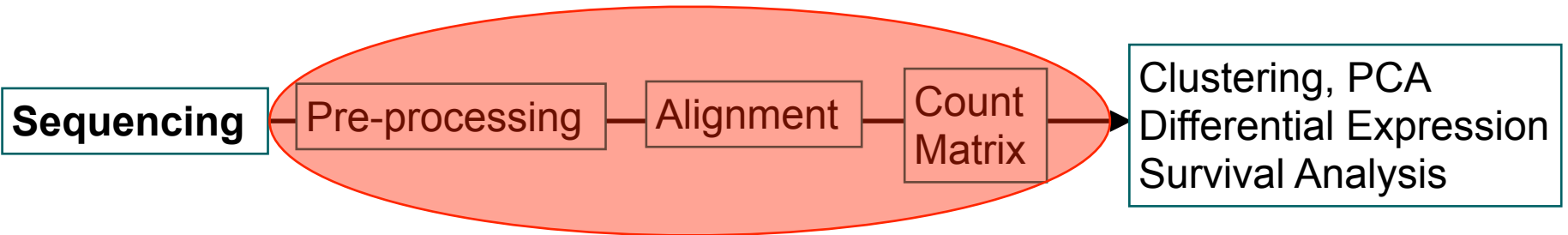
Distinct ways to represent transcripts

Distinct ways to count gene expression.

Bioinformatics Pipeline / RNA-seq

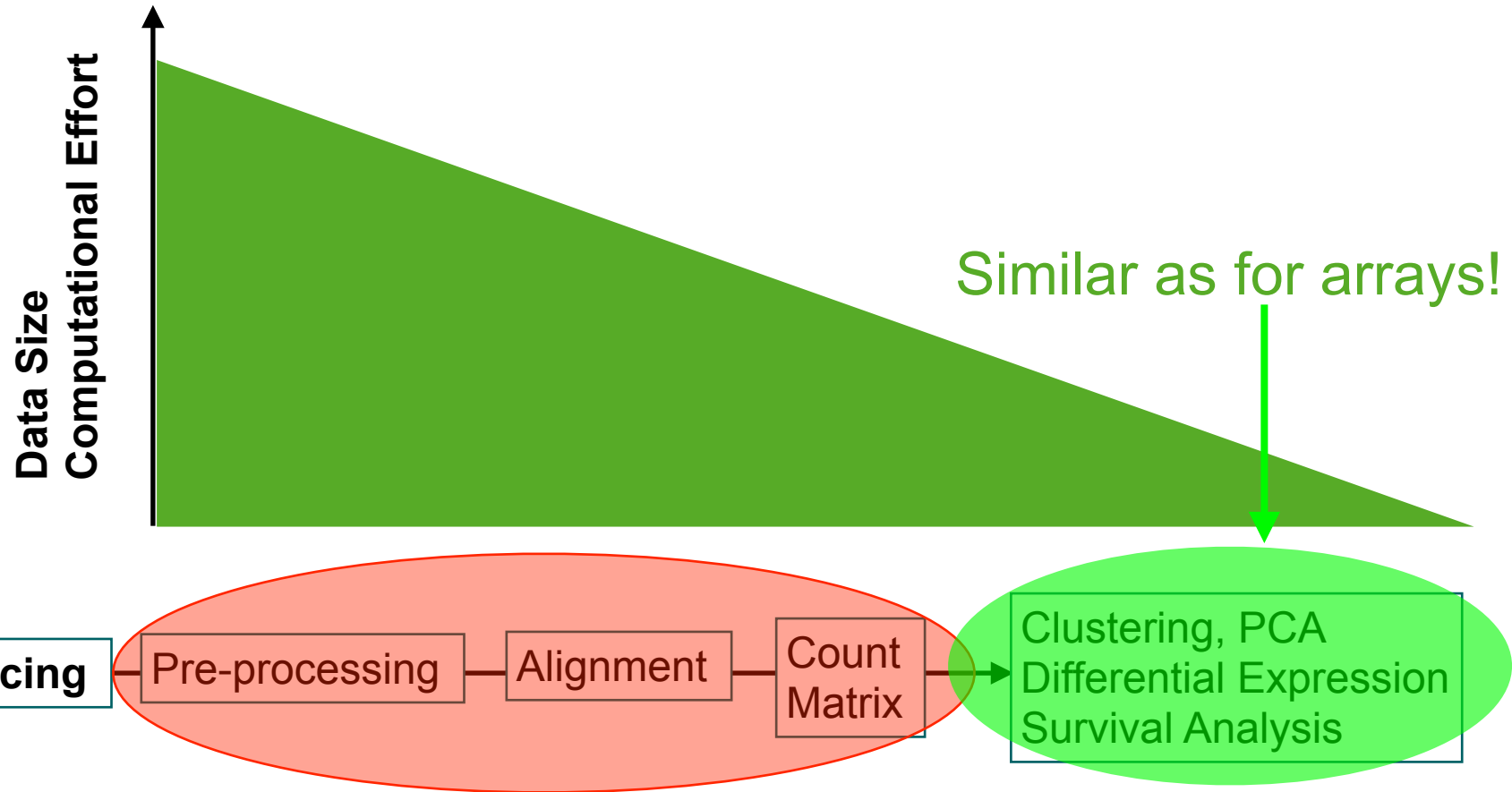


Bioinformatics Pipeline / RNA-seq



Practical part not covered!

Bioinformatics Pipeline / RNA-seq



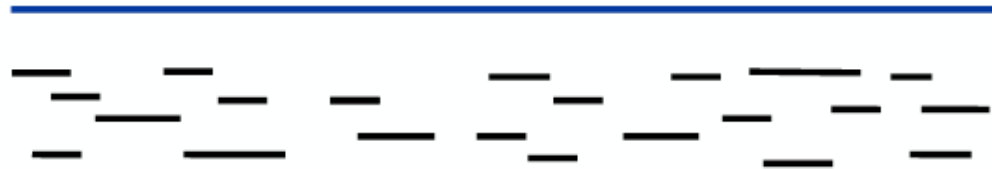
Next Generation Sequencing

- ▶ NGS take advantage of **parallelization**
 - ▶ reads millions/billions of reads per run
 - ▶ short reads (50-100 bps)
 - ▶ error rates (0.1-1%)
- ▶ commercial products:
 - ▶ 454
 - ▶ SOLiD
 - ▶ **Solexa (Illumina)**



Read Types

Fragment DNA:

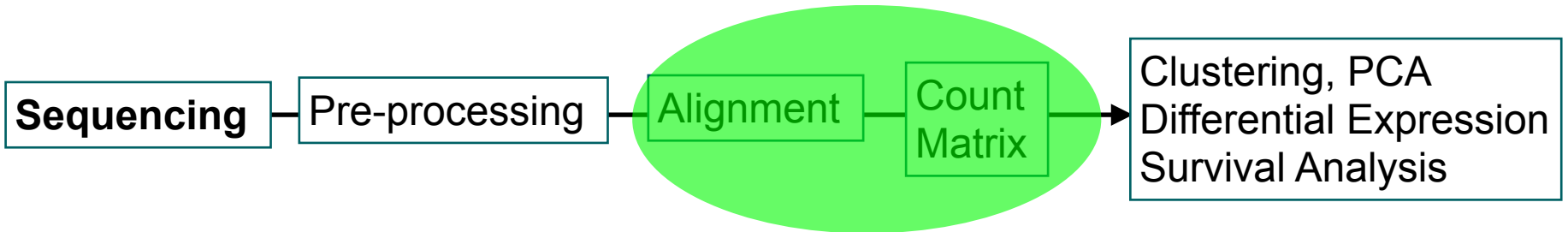


Single end



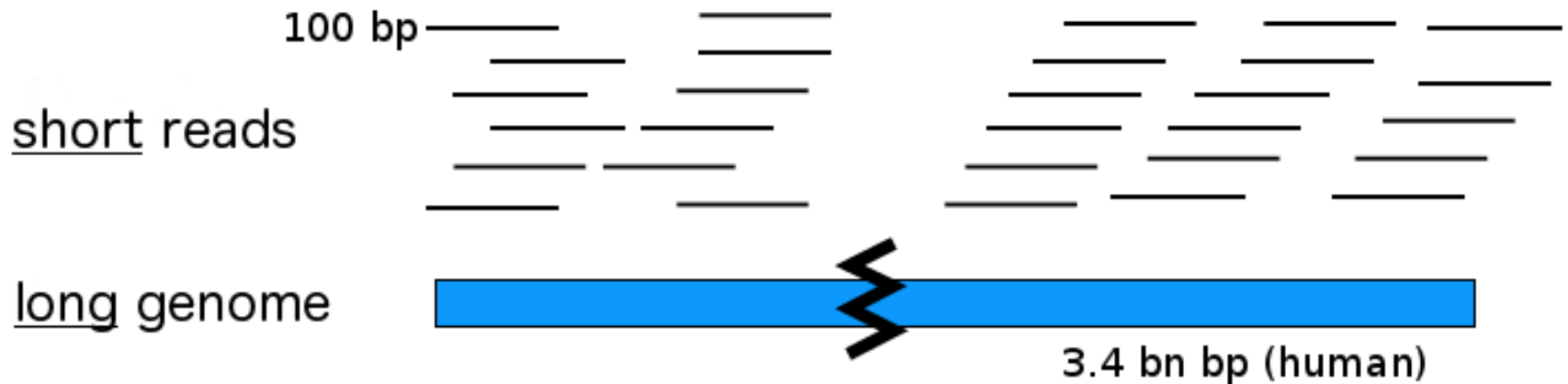
Paired end
Ins: 200-800 bp

Bioinformatics Pipeline / RNA-seq

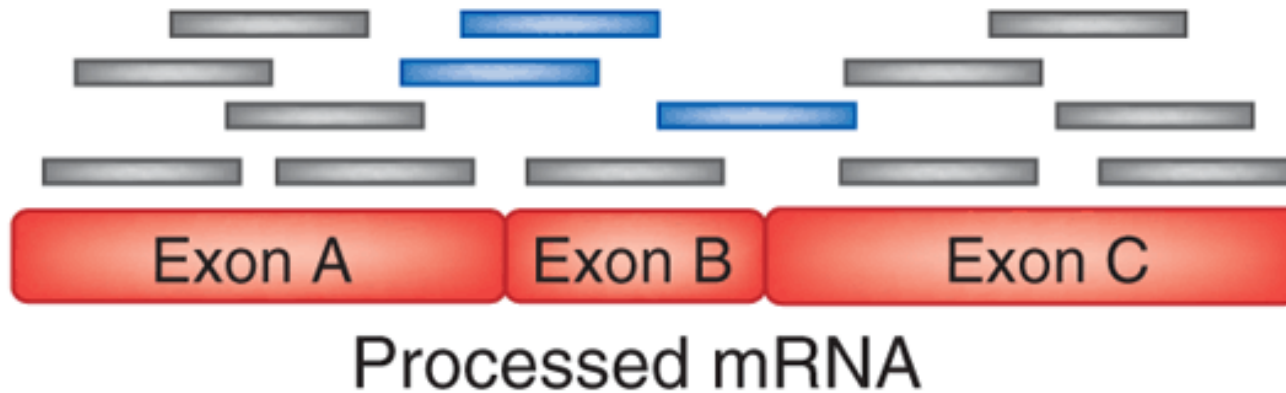


Alignment

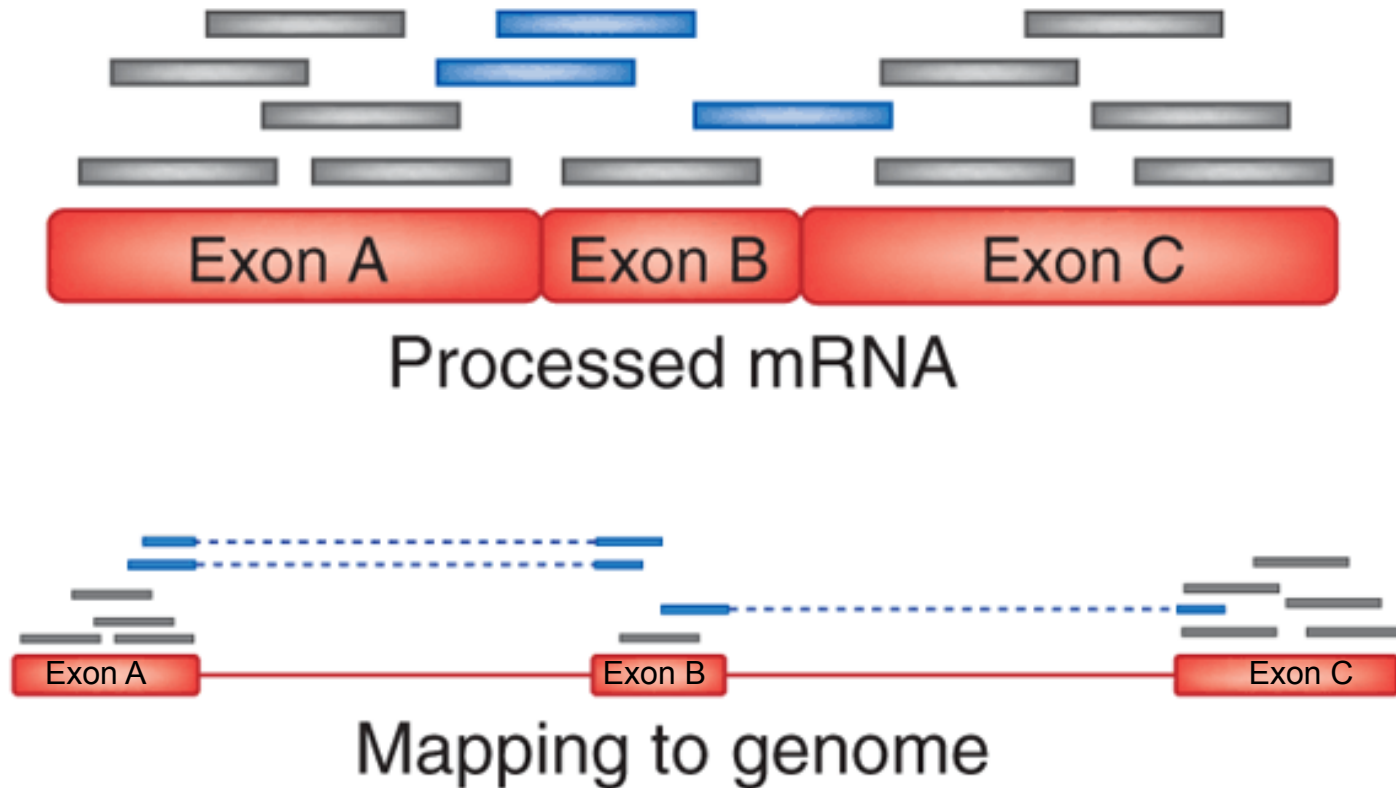
- a large reference sequence is given (genome)
 - up to billions of base pairs
- short reads (<200bps)
- find most probable position of the read in the genome (by inexact string matching)



Alignment - Split Read Mapping (RNA-Seq)



Alignment - Split Read Mapping (RNA-Seq)



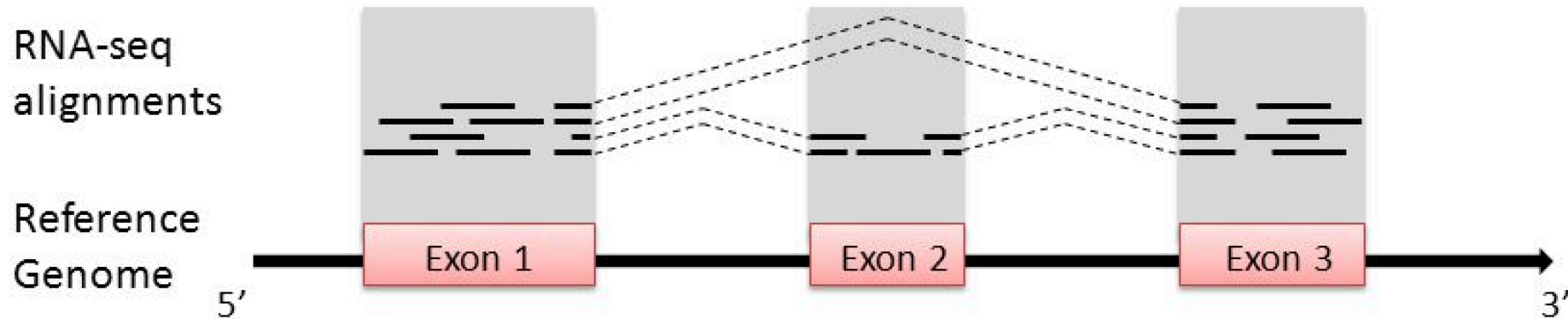
- reads are split within introns when mapped to genome
- aligners use transcript information or try to find splice events (STAR & TOPHAT)

Reference based aligners - Overview

	<i>Time</i>	<i>Precision</i>	<i>Pairs</i>	<i>GAPs</i>	<i>Phred</i>	<i>Memory</i>	<i>Application (Comments)</i>
BOWTIE	+		+	-	-	5GB	General (max. 3 mismatches)
BWA	+		+	+	+	8GB	General (max of 200bps reads)
NOVOALIGN		+	+	+	+	8GB	General (commercial license)
STAR	+		+	-	+	32GB	RNA-Seq (allow split-maps)
BISMARK	+		+	+	+	10GB	Bisulfite/reduced sequencing

Computers need large memory and a few hours of computation per experiment!

Quantification (Count Matrix)



Simple Counting Approaches

Gene Level - 17 reads

Exon level - exon 1 (8 reads), exon 2 (3 reads), exon 3 (6 reads)

Transcript Level - Exons 1,2 & 3 (10 reads) and exon 1 & 3 (7 reads) *

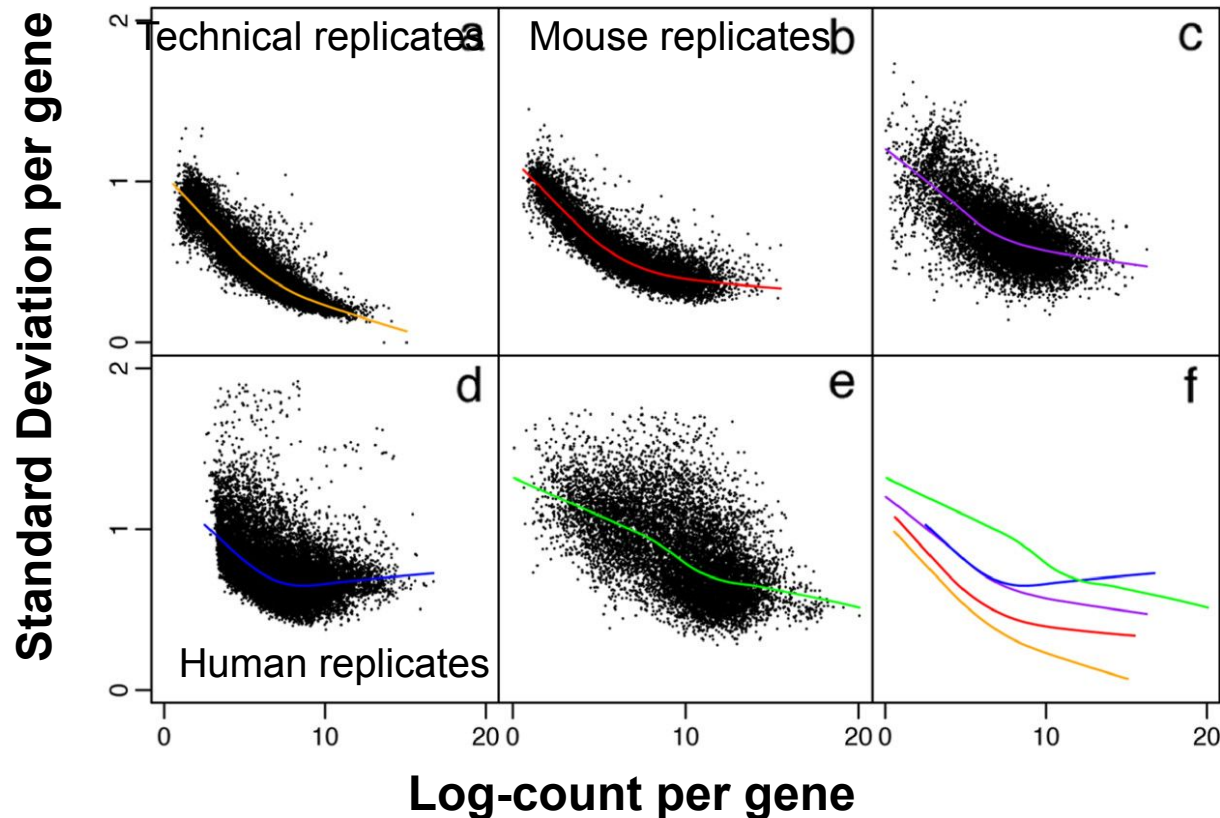
* complex computational methods required (RSe, or TopHAT needed for this)

Fragments per Kilobase (FPKM)

- normalize counts by read size (kb) and RNA-seq library size (mb)

RNA-seq and Differential Analysis

Arrays and RNA-seq have distinct distributions

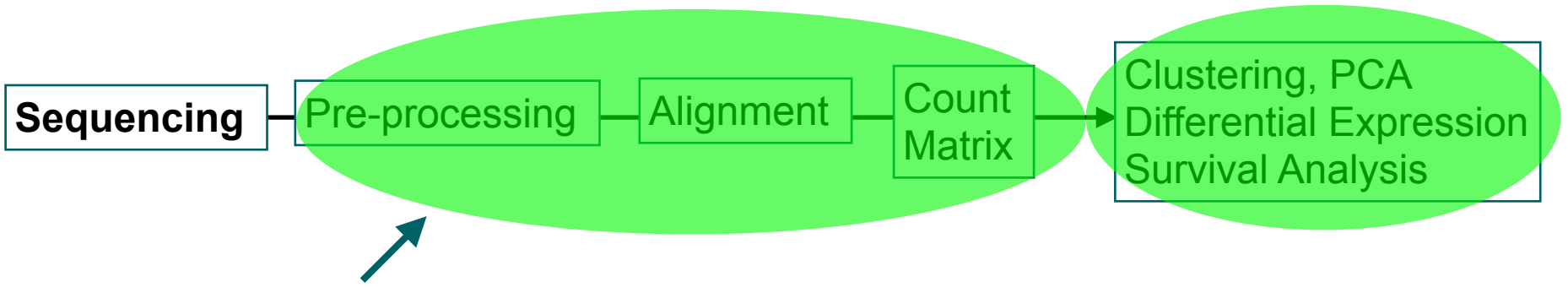


VOOM analysis is necessary to make variance similar to arrays.

Bioinformatics Pipeline / RNA-seq



We will see this today!



Provided by TGCA or your Core Facility!

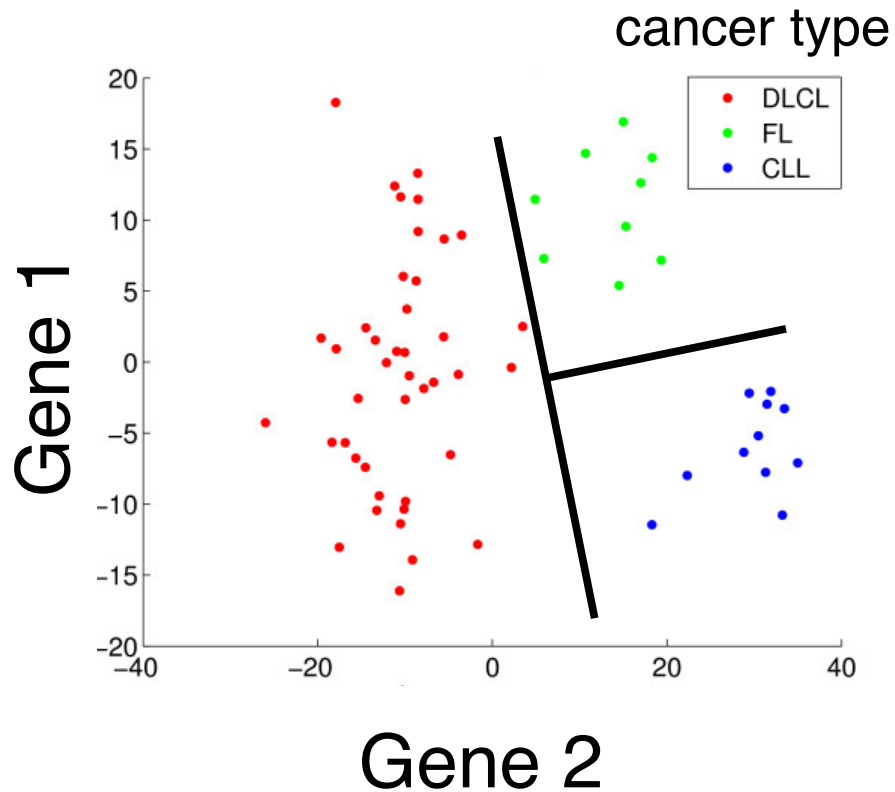
Personalized Medicine

Diagnosis and treatment choices is mostly carried on macromolecular features:

- morphology of tumours (image), symptoms, blood levels

Challenges: use molecular markers (expression or genetics) for diagnosis or treatment selection.

Machine Learning - Classifier



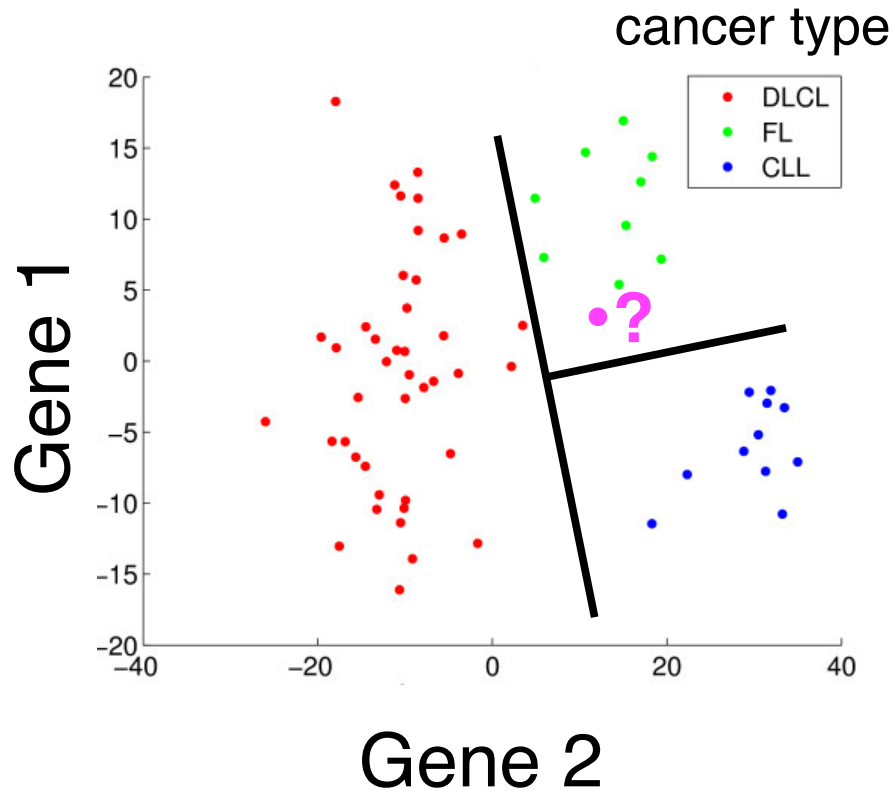
Data

Expression matrix X
(genes vs samples)
classification vector Y
(diagnosis)

Find a function:

$$f(x) \rightarrow y$$

Machine Learning - Classifier



Data

Expression matrix X
(genes vs samples)
classification vector Y
(diagnosis)

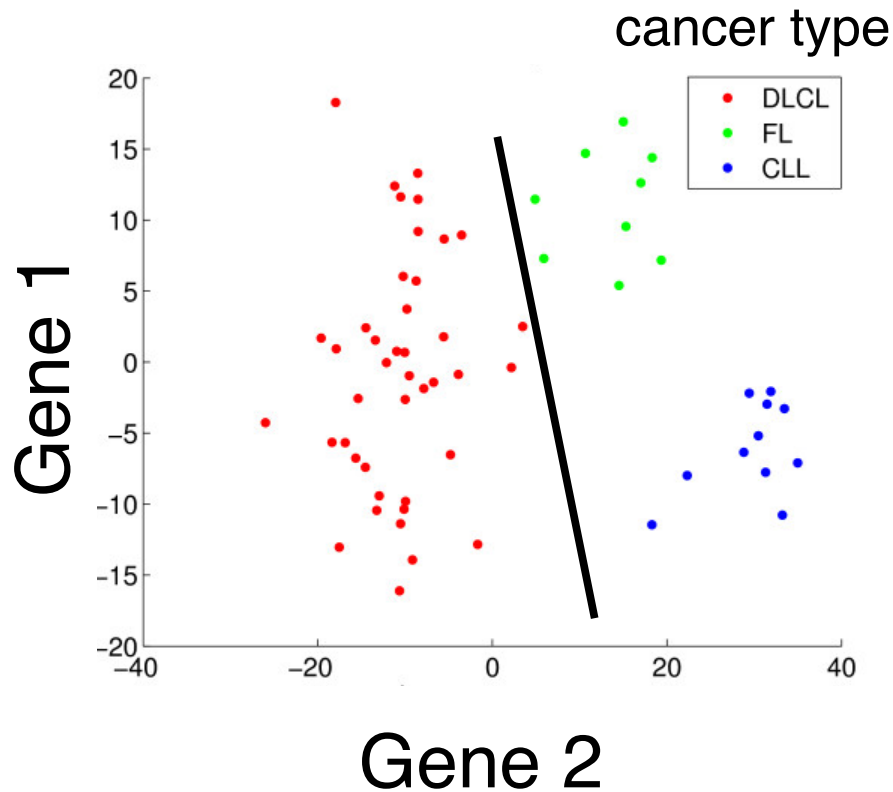
Find a function:

$$f(x) \rightarrow y$$

For new patients X' :

$$f(x') \rightarrow y'$$

Linear Classifier



Linear Function:

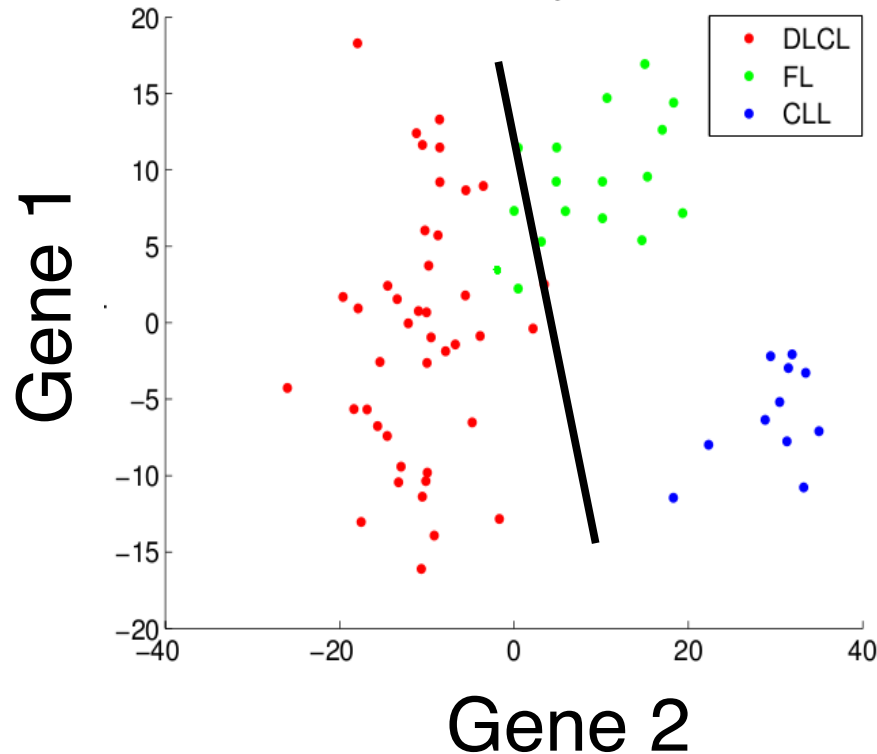
$$f(x, A) = a_0 + a_1x_1 + \dots + a_Lx_L$$

$$f(x, A) > 0 \Rightarrow \text{classe A}$$

$$f(x, A) \leq 0 \Rightarrow \text{classe B}$$

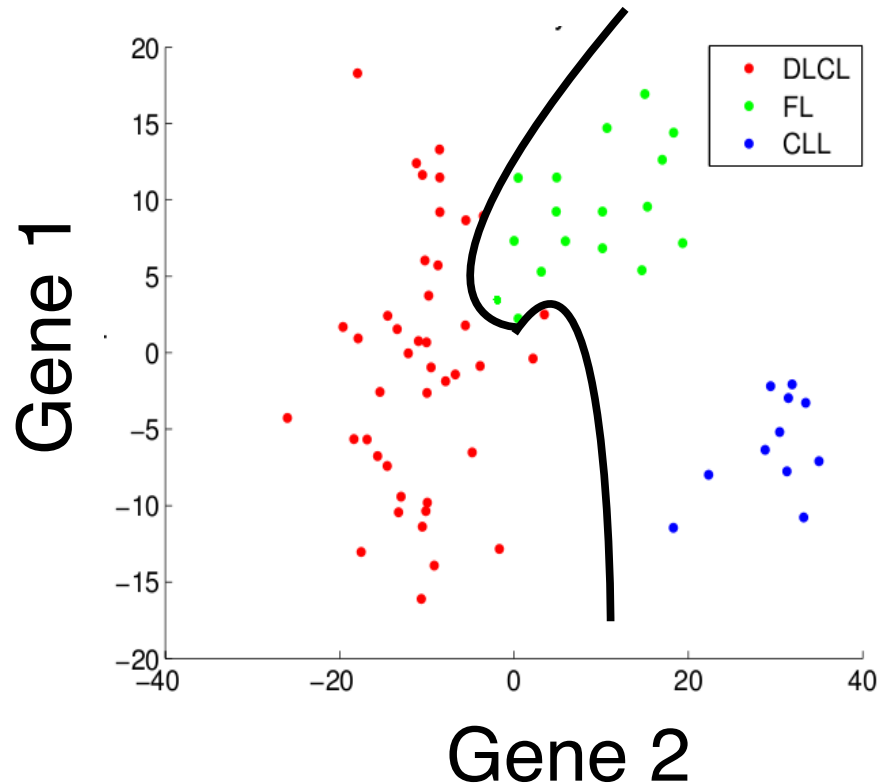
- Works for 2 classes only
 - Train a function for each cancer type
- Find coefficients A
 - estimated with neural networks or support vector machines

Linear Classifier - Problems



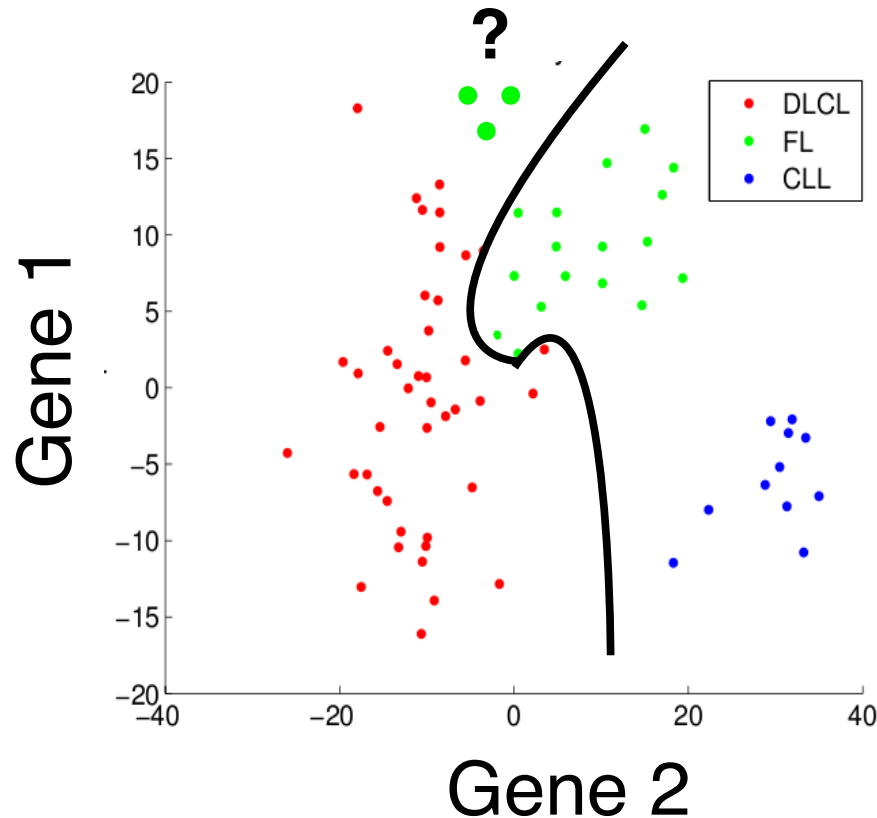
- Most real word problems are not linearly separable!
- There will be always some error!
- Solution: non-linear functions

Nonlinear Classifier - Problems



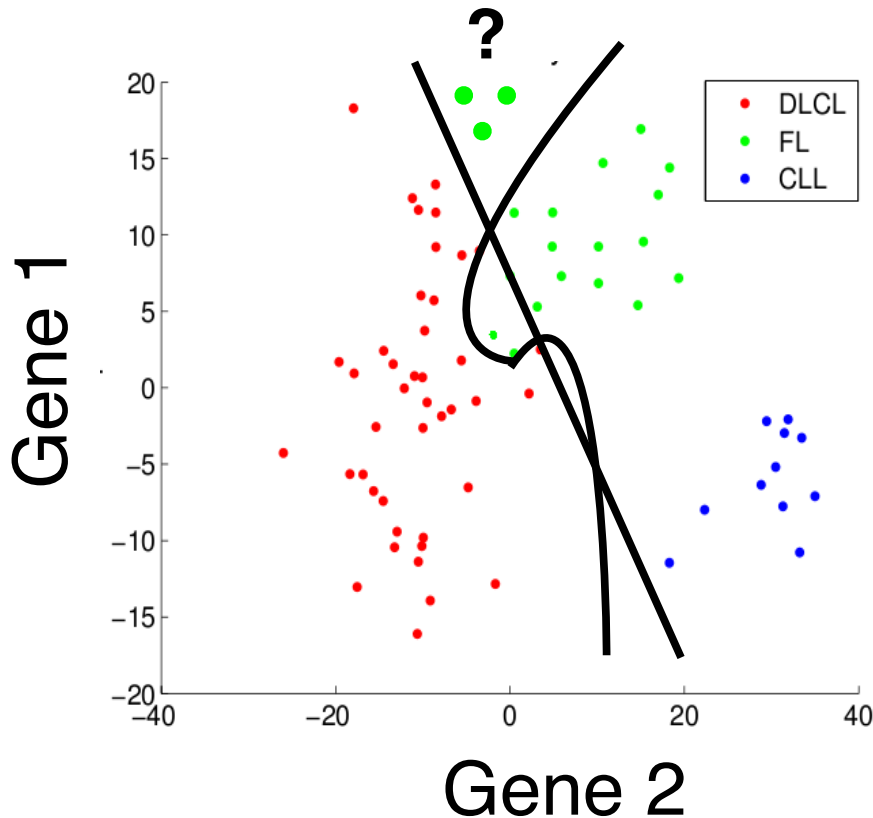
- Polynomial Function
- $$f(x, A) = a_0 + a_{11}x_1^3 + \dots + a_{L1}x_L^3 + a_{12}x_1^2 + \dots + a_{L2}x_L^2 + a_{12}x_1 + \dots + a_{L2}x_L$$
- Third order polynomial
- Problem: overfitting

Nonlinear Classifier - Problems



- Polynomial Function
- $$f(x, A) = a_0 + a_{11}x^3_1 + \dots + a_{L1}x^3_L$$
$$a_{12}x^2_1 + \dots + a_{L2}x^2_L$$
$$a_{12}x_1 + \dots + a_{L2}x_L$$
- Third order polynomial
- Problem: overfitting

Nonlinear Classifier - Problems



- Polynomial Function
- $$f(x, A) = a_0 + a_{11}x_1^3 + \dots + a_{L1}x_L^3 + a_{12}x_1^2 + \dots + a_{L2}x_L^2 + a_{12}x_1 + \dots + a_{L2}x_L$$
- Third order polynomial
- Problem: overfitting

Curse of Dimensionality

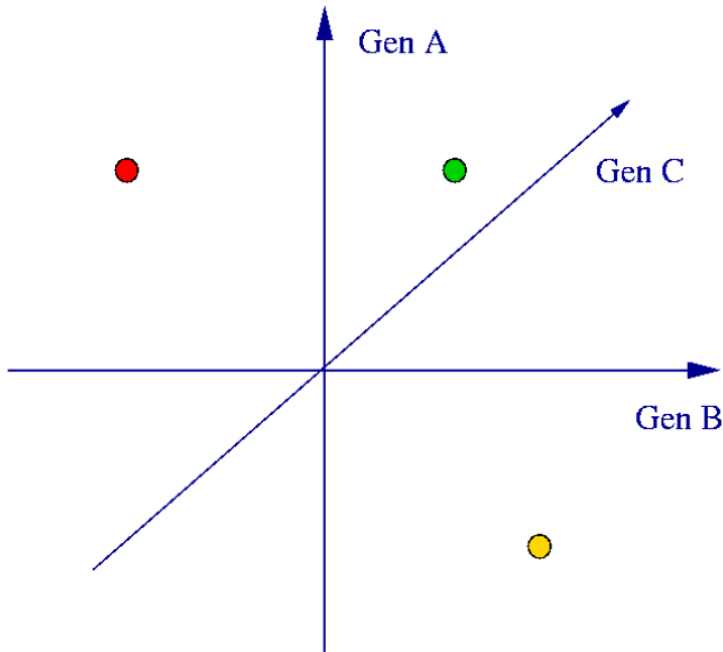
Size of a Euclidean space grows
with dimension (number of genes)

Dots (patients) are sparsely
distributed in space

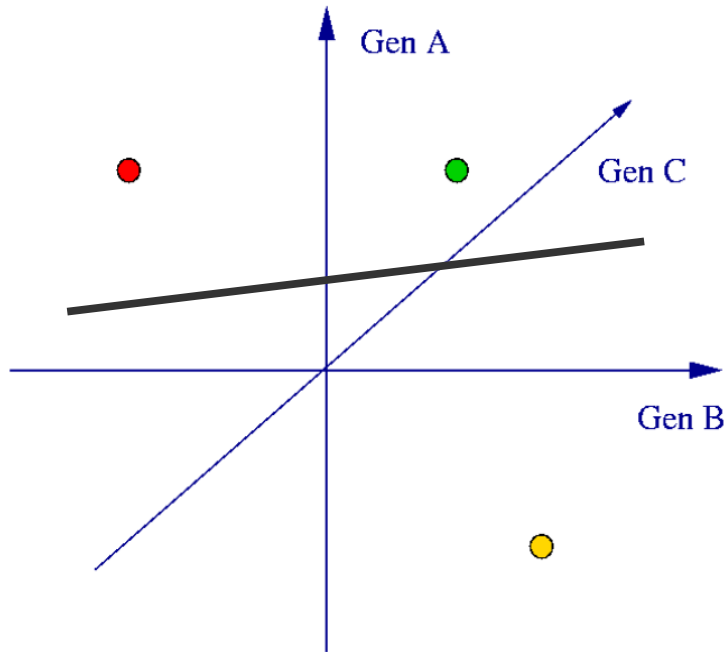
Curse of Dimensionality : Example

Sparse data

- three genes
- 2 patients with known cancer (red vs yellow)
- 1 unknown (green)



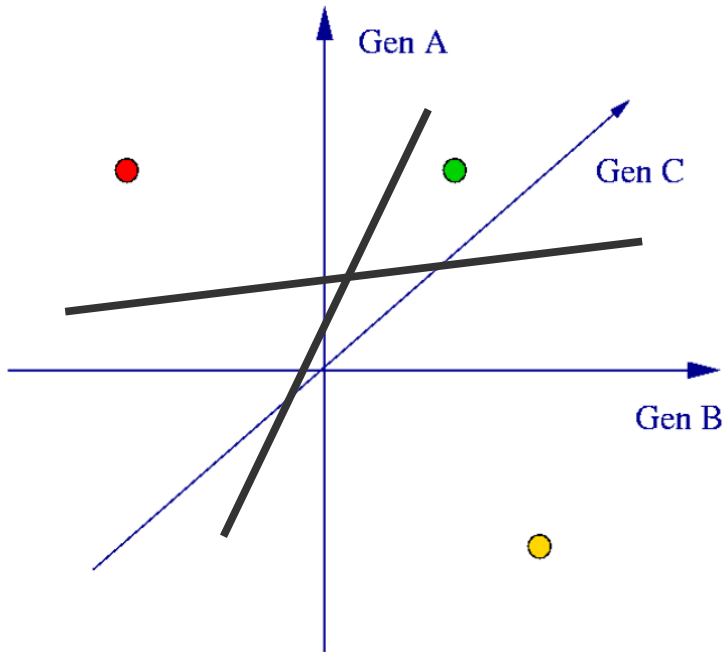
Curse of Dimensionality : Example



- Sparse data
 - three genes
 - 2 patients with known cancer (red vs yellow)
 - 1 unknown (green)

Perfect classifier (on training)

Curse of Dimensionality : Example

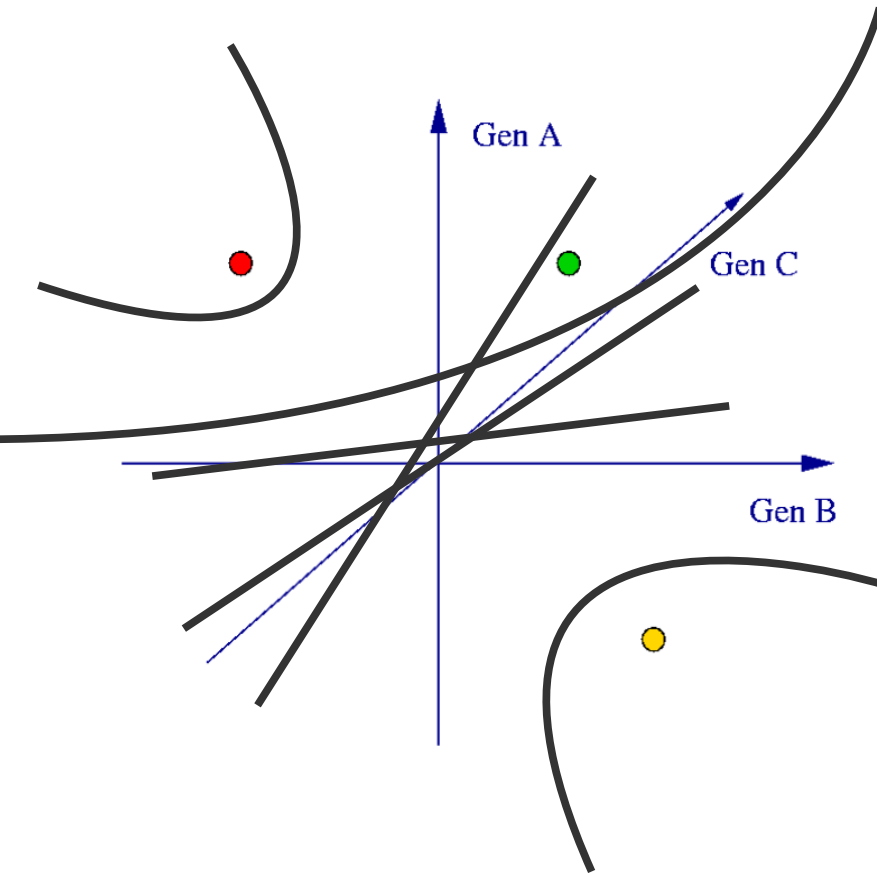


- Sparse data
 - three genes
 - 2 patients with known cancer (red vs yellow)
 - 1 unknown (green)

Both are perfect classifiers
(on training)

Hard to generalise!

Curse of Dimensionality : Example



- There are millions of perfect linear classifiers
- And even more non-linear classifiers!

Dealing with Curse of Dimensionality

- Have a proper training / test evaluation procedure
- Use classifiers which are as simple as possible
- Reduce the dimension of your data (feature selection or PCA)

Classifier Evaluation

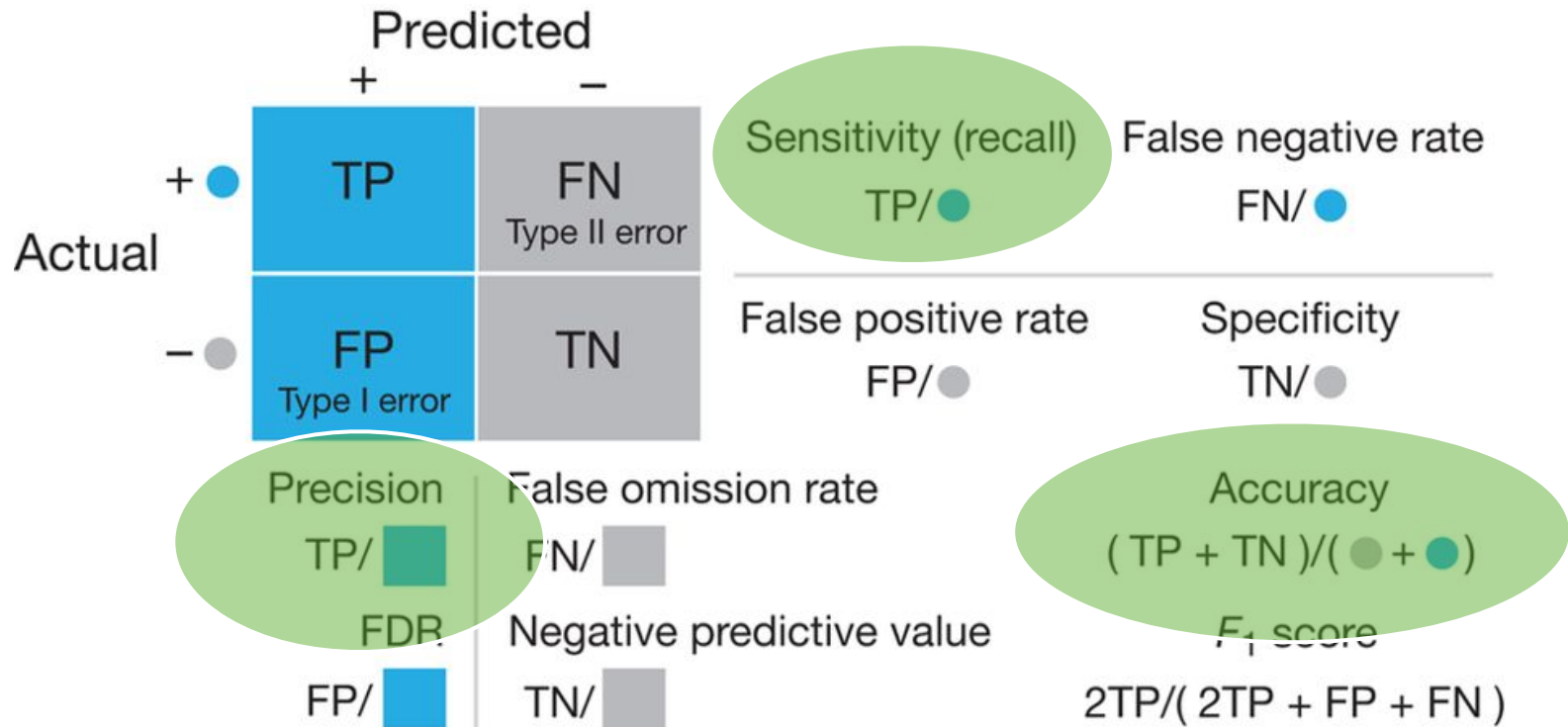
Measures for a two class problem (cancer + vs. non-cancer -)

		Predicted	
		+	-
Actual	+ ●	TP	FN Type II error
	- ●	FP Type I error	TN

Source: Lever et al., Nat. Methods (2016)

Classifier Evaluation

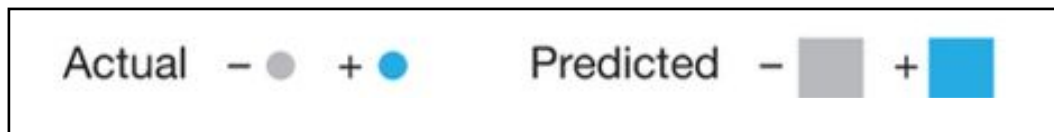
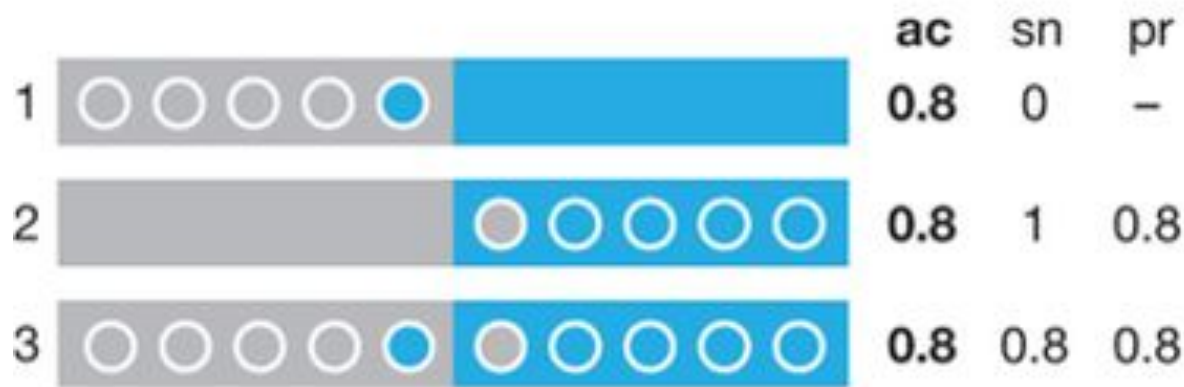
Measures for a two class problem (cancer + vs. non-cancer -)



Source: Lever et al., Nat. Methods (2016)

Classifier Evaluation

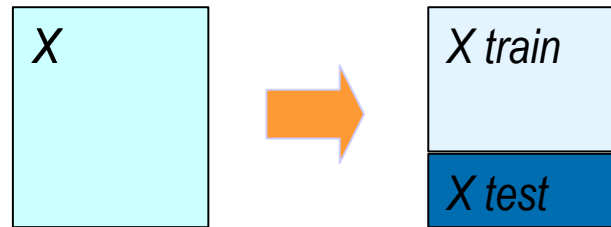
Measures for a two class problem (cancer + vs. non-cancer -)



Source: Lever et al., Nat. Methods (2016)

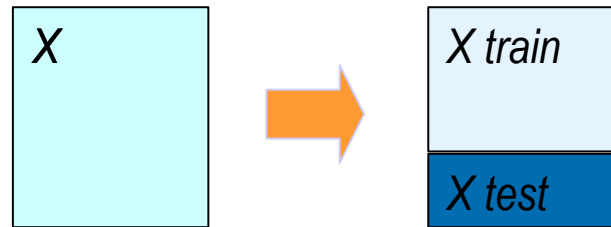
Classifier Evaluation

- The performance of your classifier needs to be evaluated at your test data:
 - an independent "validation cohort"
 - or a large (1/3 of samples) and have similar distribution of classes as train data



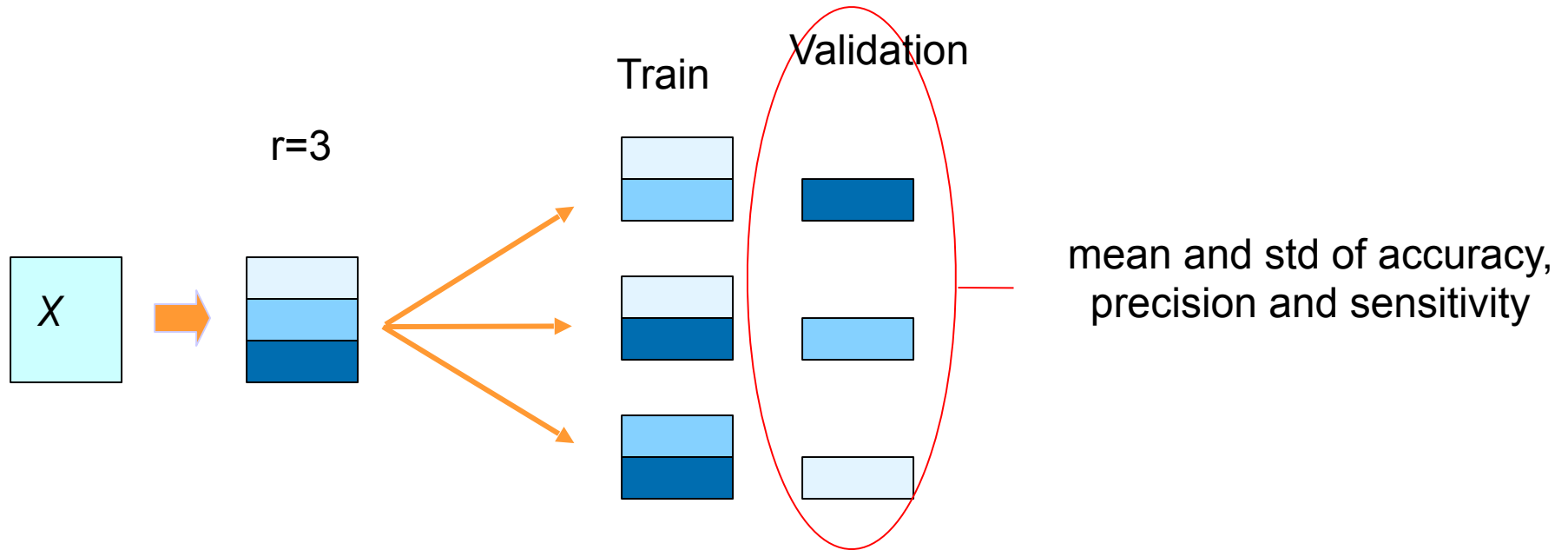
Classifier Evaluation

- The performance of your classifier needs to be evaluated at your test data:
 - an independent "validation cohort"
 - or a large (1/3 of samples) and have similar distribution of classes as train data



- Never use test data to improve classification (choose a better classifier or marker gene)
 - For this you need to establish validation data (or cross validation)

Cross-validation



Elastic Net

Is based on a linear function:

$$f(x, A) = a_0 + a_1x_1 + \dots + a_Lx_L$$

$$f(x, A) > 0 \Rightarrow \text{classe A}$$

$$f(x, A) \leq 0 \Rightarrow \text{classe B}$$

- Find coefficients A , *while most of them have 0*.
 - A shrinkage factor (λ) controls the number of genes selected.
 - Shrinkage factor can be automatically identified with cross-validation.

Break time !

Survival Analysis

Can be used to evaluate if characteristics of a patients indicates an increase/decrease risk of survival

- clinical: tumour type, gender
- Molecular: expression of a gene, mutation

Common Survival Tests:

- Cox proportional hazards regression (not seen here)
 - Compares survival with a numeric variable
- Kaplan-Meier graph / Log-rank test
 - compares the survival of groups of individuals

Kaplan-Meier graph / Log-rank test

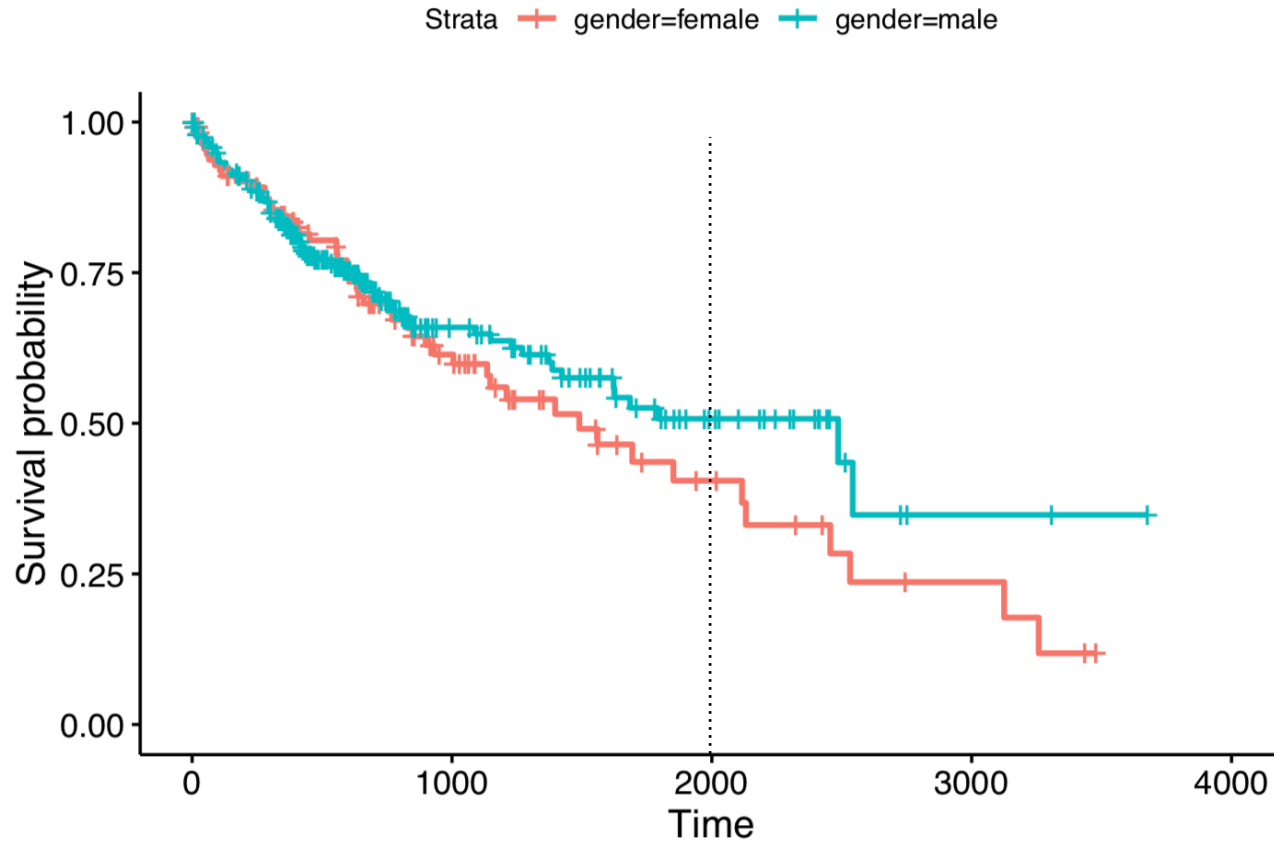
Data:

- **Event:** death / alive
- **Time:** period between first and last observation.
- **Characteristics:** sex, tumor grade

<i>Patient</i>	<i>Status</i>	<i>Time</i>	<i>Sex</i>
<i>1</i>	<i>Dead</i>	<i>343</i>	<i>Male</i>
<i>2</i>	<i>Alive</i>	<i>20</i>	<i>Male</i>
<i>3</i>	<i>Alive</i>	<i>300</i>	<i>Female</i>
<i>4</i>	<i>Dead</i>	<i>200</i>	<i>Male</i>

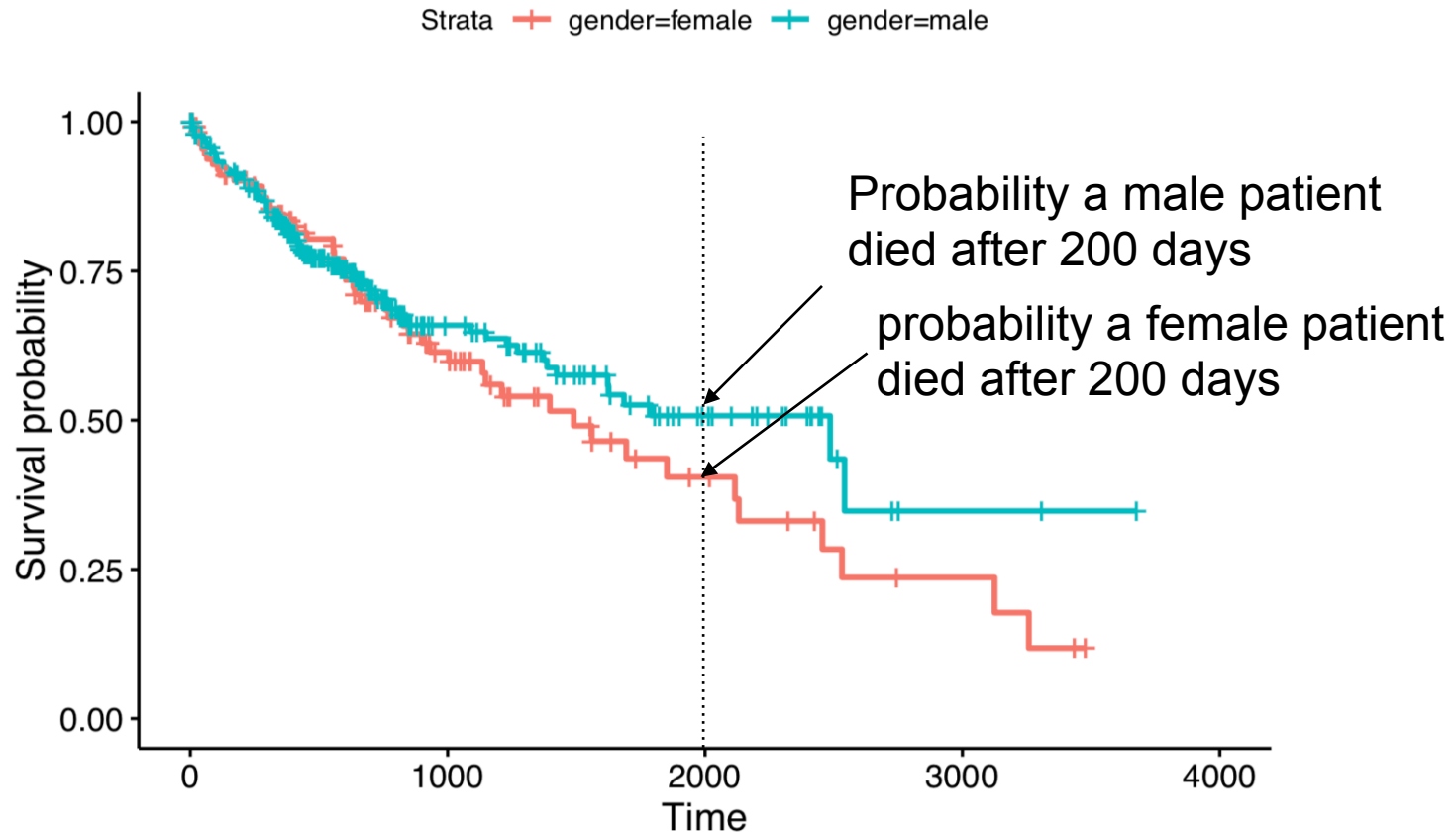
Kaplan-Meier plot

Survival of LIHC patients - male vs. Female



Kaplan-Meier plot

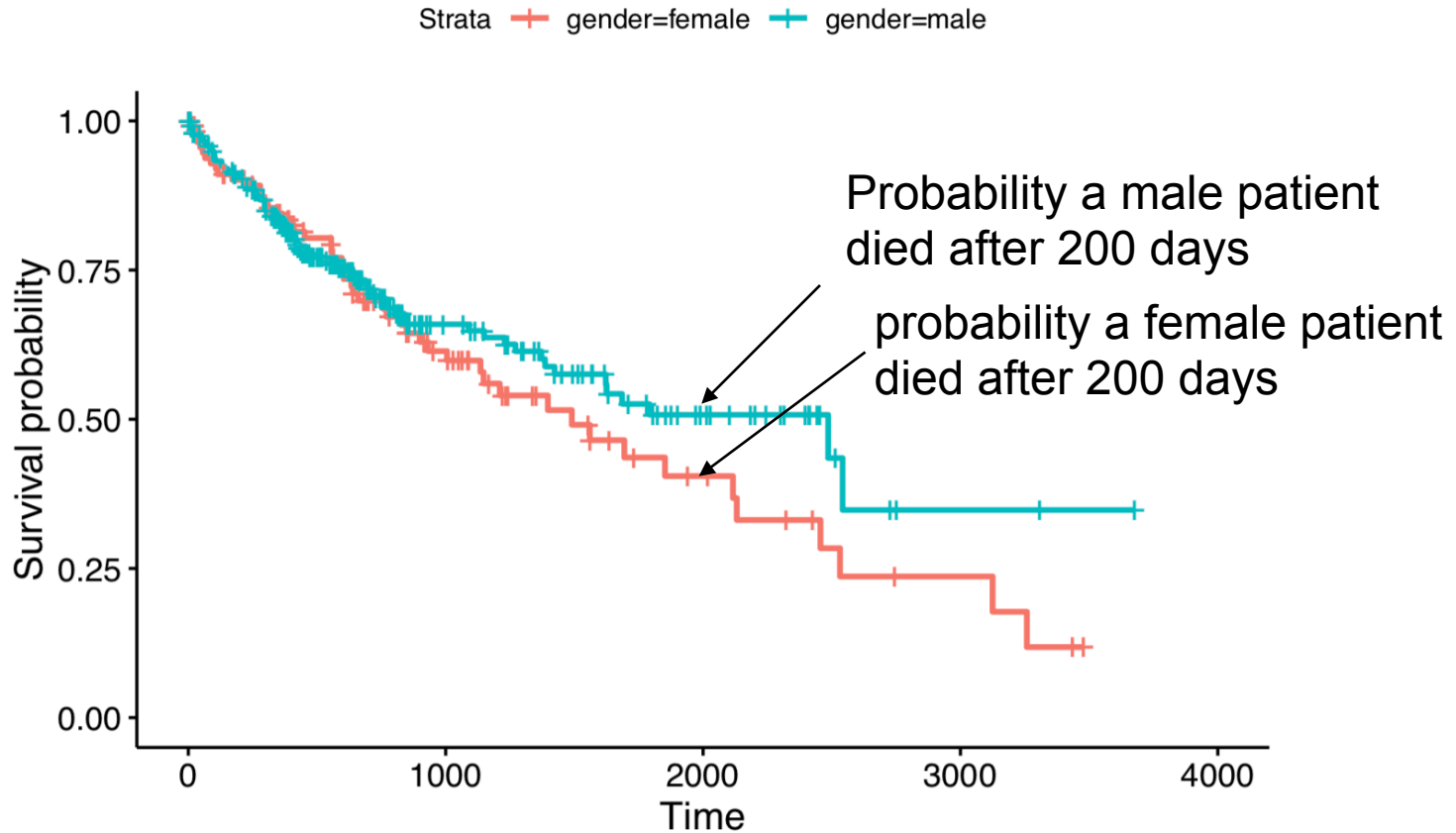
Survival of LHC patients - male vs. Female



$$\text{Probability (X days)} = \frac{\text{\# cases alive after X days}}{\text{\# cases measured after X days}}$$

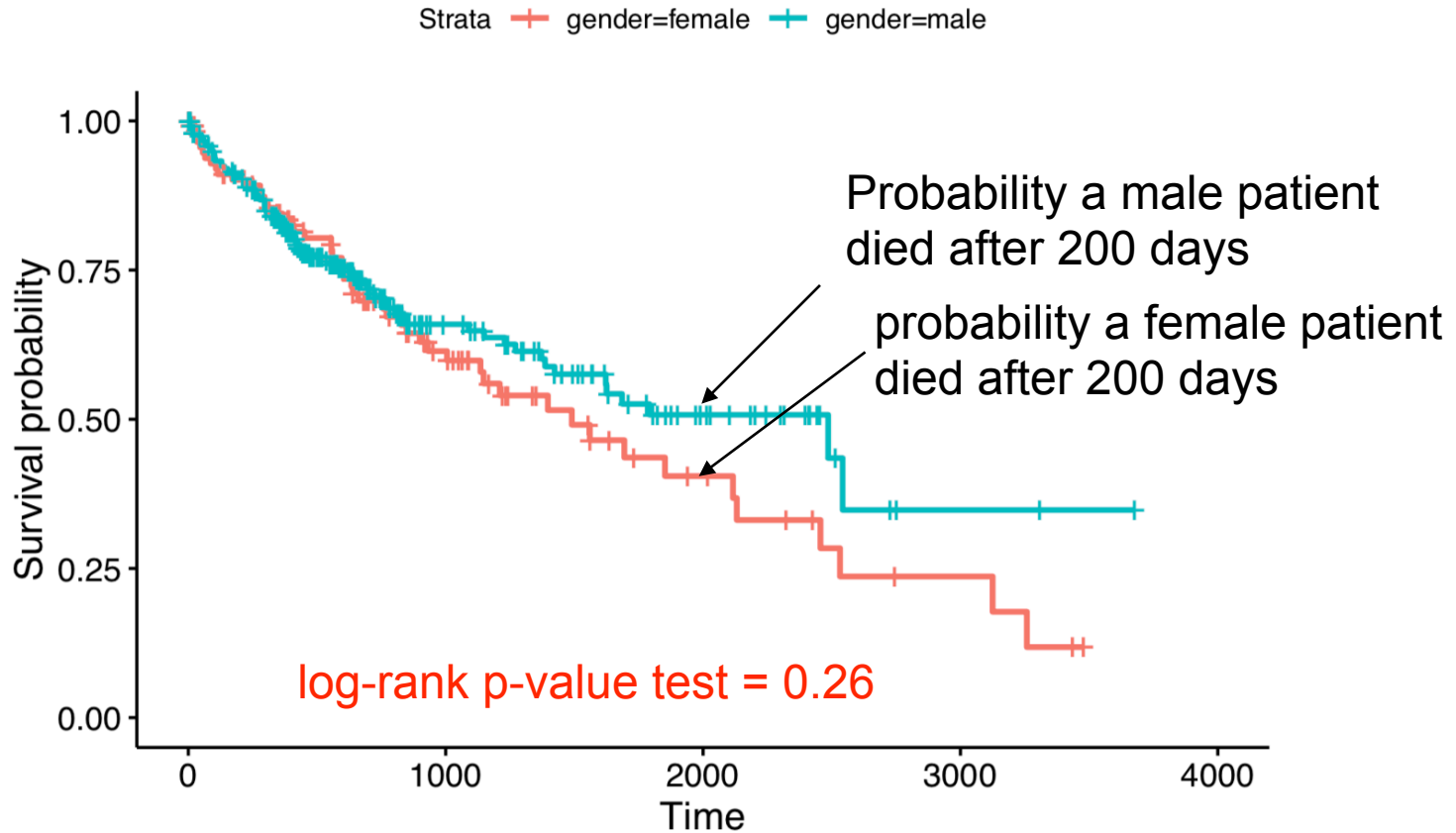
Log-rank test

Is the survival difference significant?

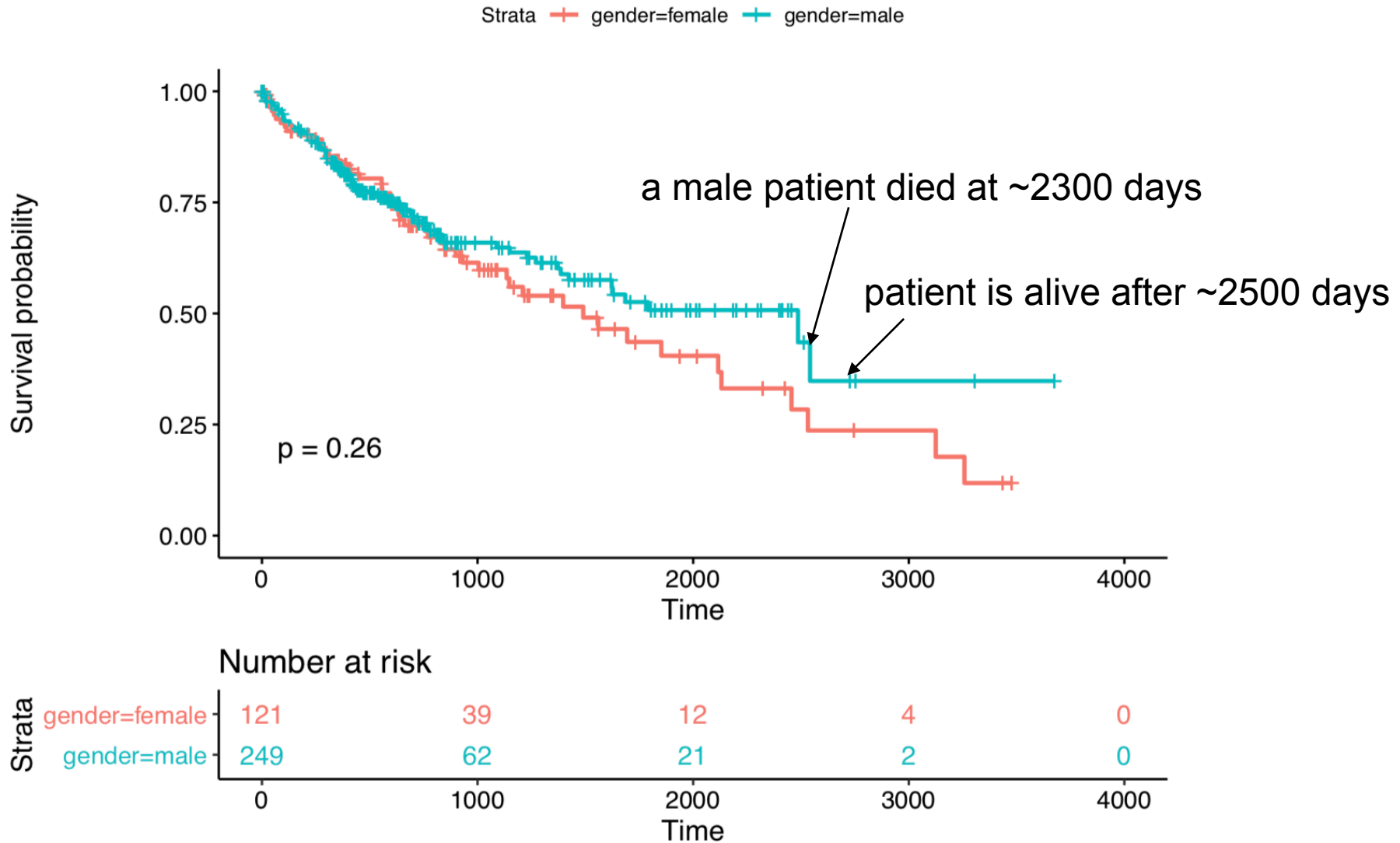


Log-rank test

Is the survival difference significant?



Kaplan-Meier plot



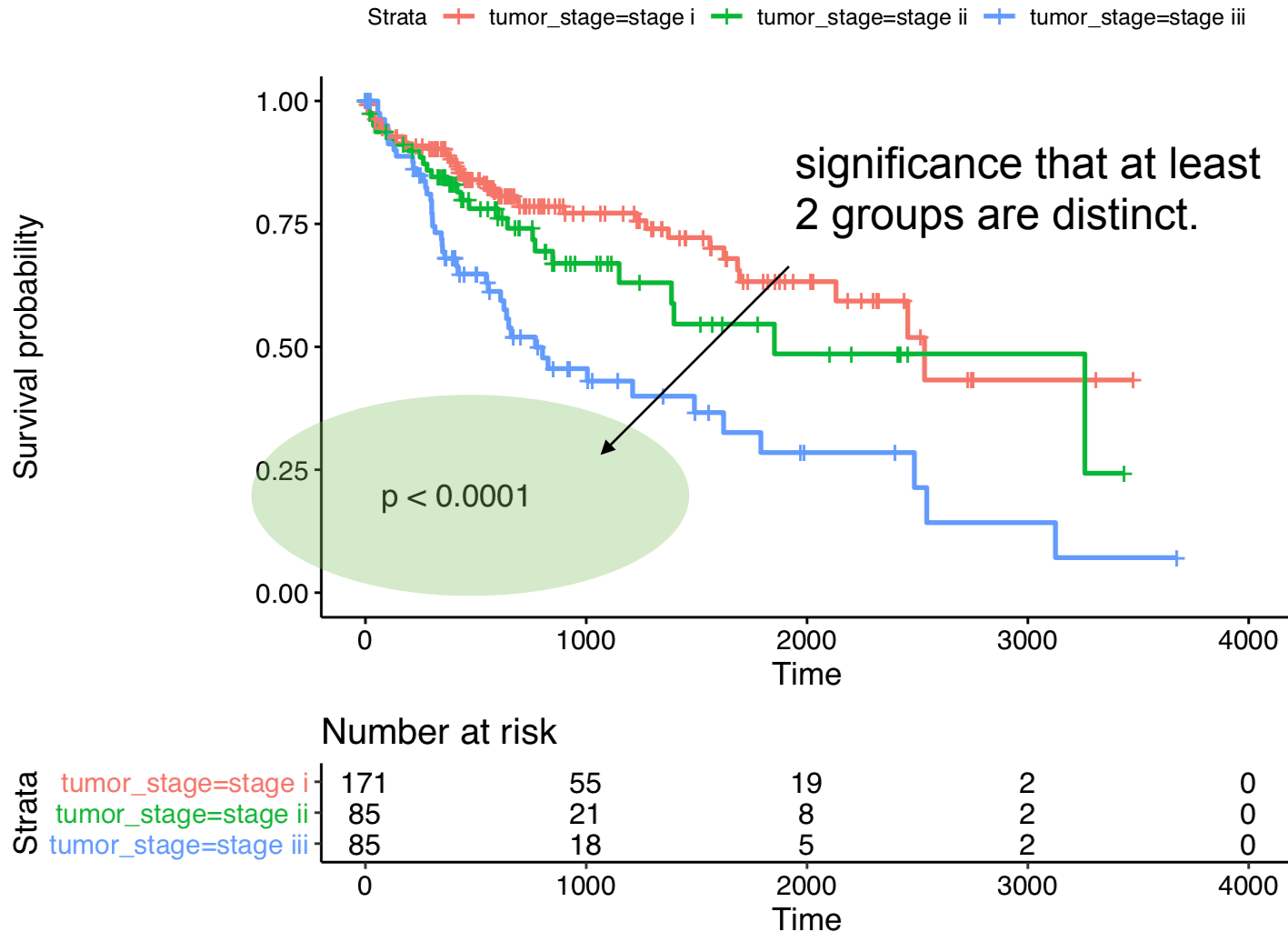
Kaplan-Meier plot



Kaplan-Meier / Log-Rank Test

KM and LRT can compare several groups at a time.

Survival vs Tumour stage at diagnosis



Survival Analysis and Biological Markers

How to perform survival analysis on biological markers?

1. Given their continuous nature of gene expression, Cox hazards test is recommended.
2. An alternative is to group patients by expression of a gene (low/high expression) and use Kaplan-Meier plots (seen in practical).

Important: if you test several markers you need to correct for multiple testing!!!

