Bioinformatics Analysis in R

Advanced Gene Expression: Analysis of Cancer Genome Atlas

Ivan G. Costa, Martin Grasshoff

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

Institute for Computational Genomics
0101101101010
1010010010101

RWTH AACHEN UNIVERSITY

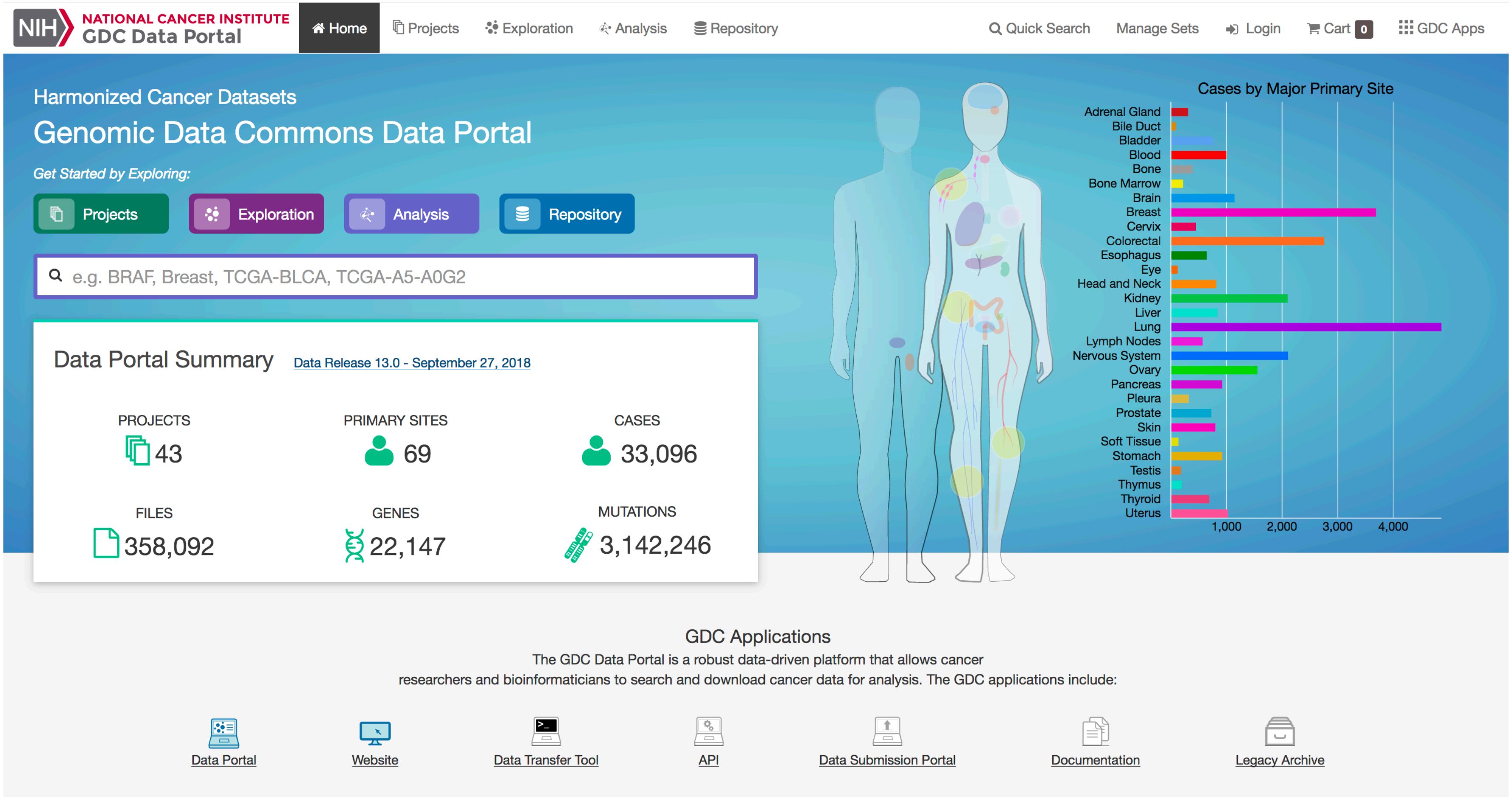
Summary

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



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

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

Data Portal Summary Data Release 13.0 - September 27, 2018

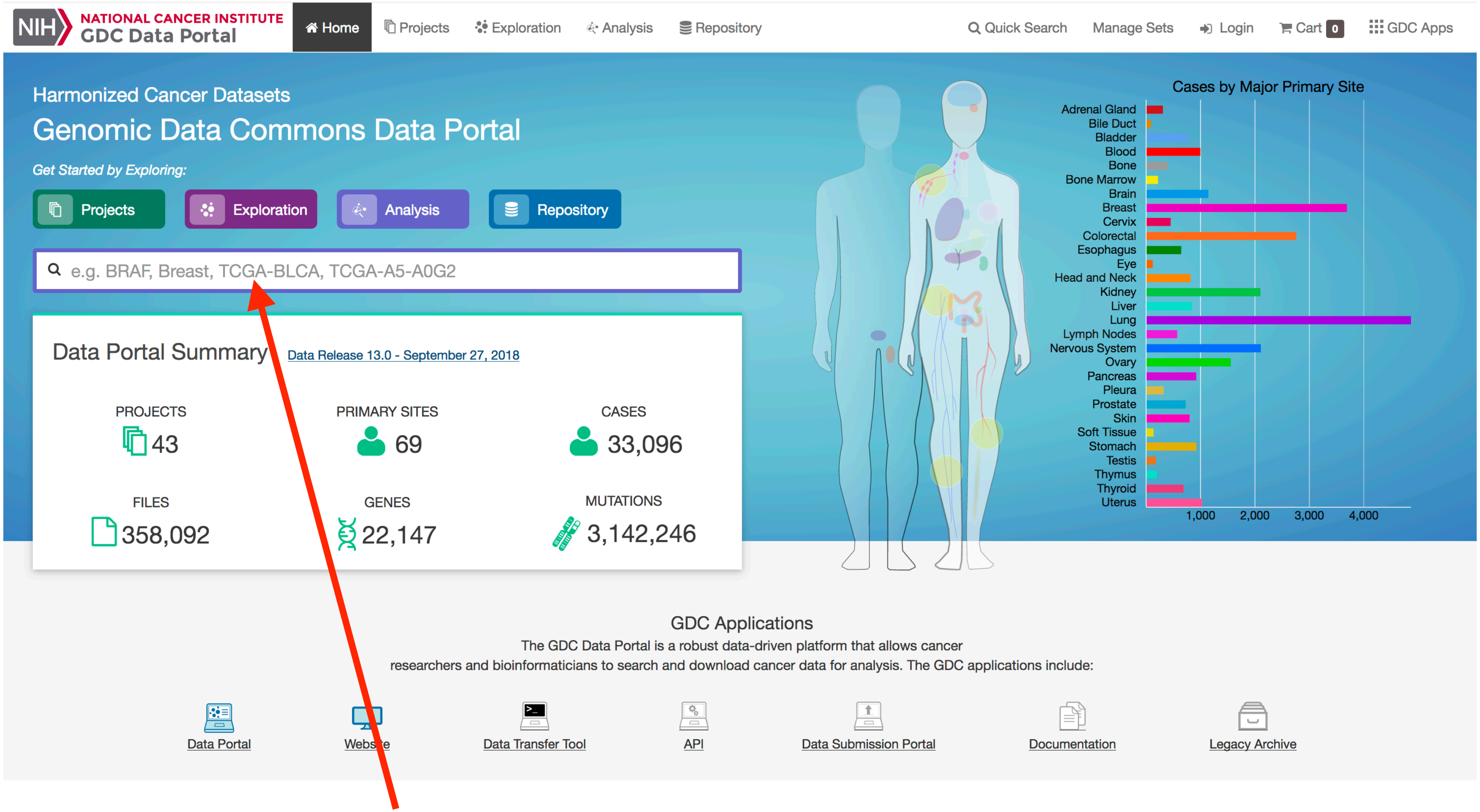
PROJECTS	PRIMARY SITES	CASES
43	69	33,096
FILES	**GENES**	**MUTATIONS**
358,092	22,147	3,142,246

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



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

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

Data Portal Summary Data Release 13.0 - September 27, 2018

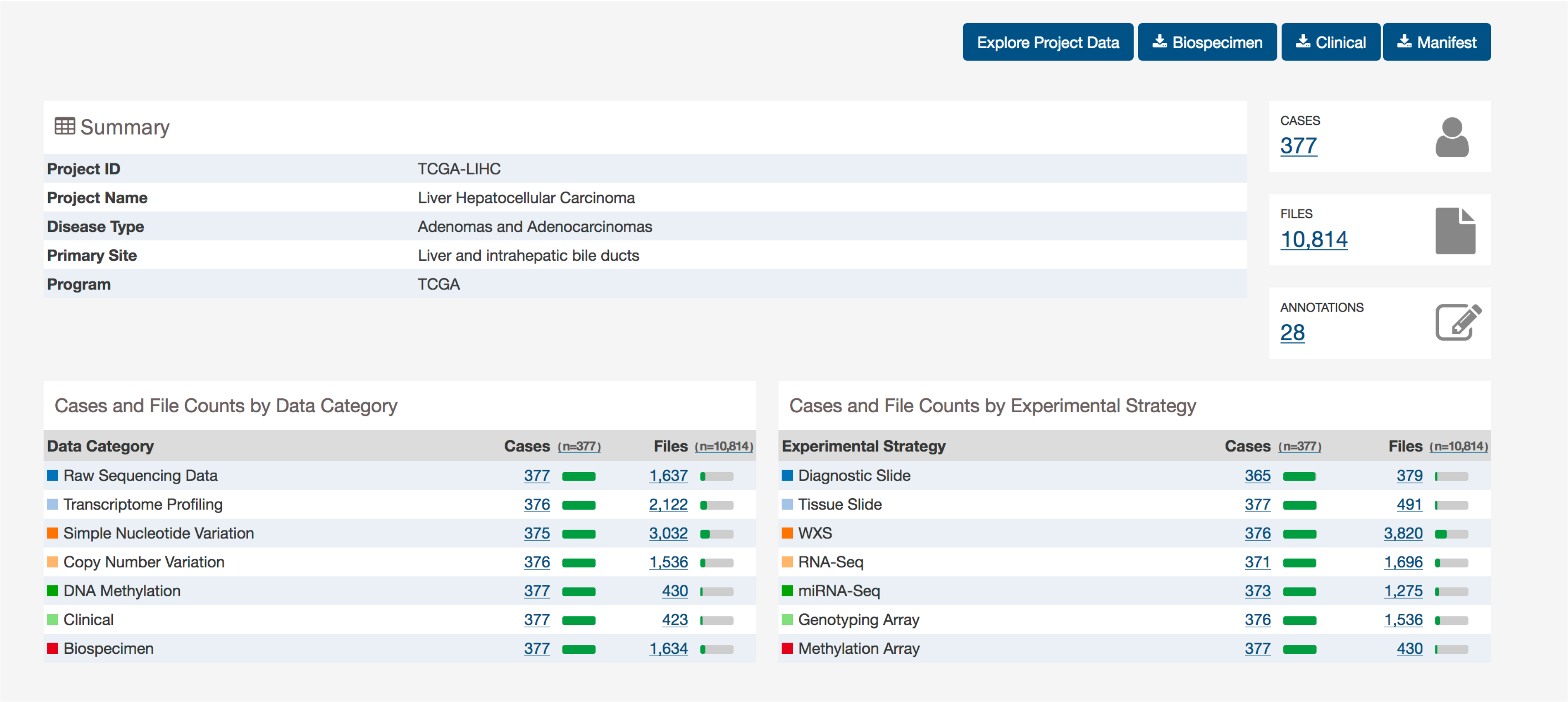
PROJECTS	PRIMARY SITES	CASES
43	69	33,096
FILES	**GENES**	**MUTATIONS**
358,092	22,147	3,142,246

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:

Check a gene or cancer type!
I will try liver ….

LIHC - Liver Hepatocellular Carcinoma



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

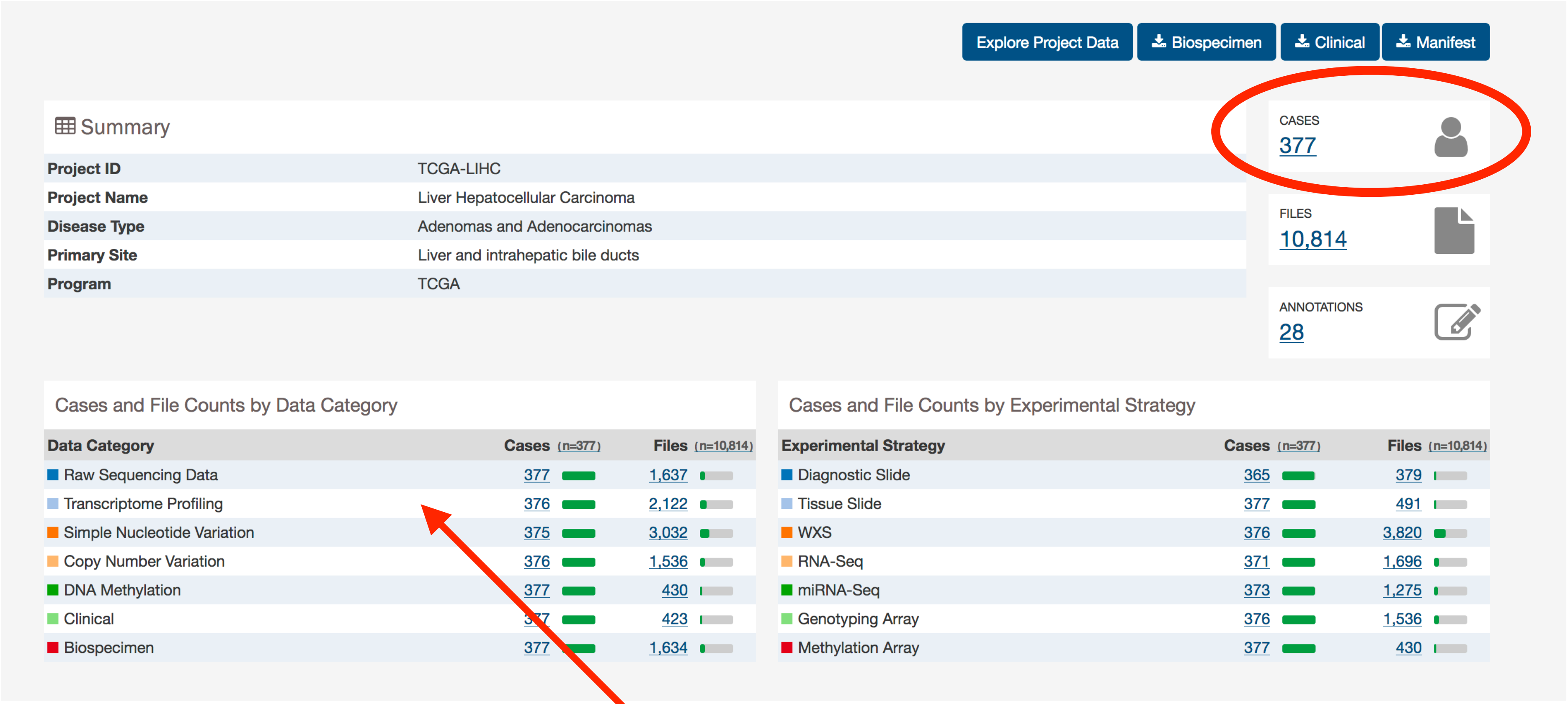
Cases and File Counts by Data Category

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

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

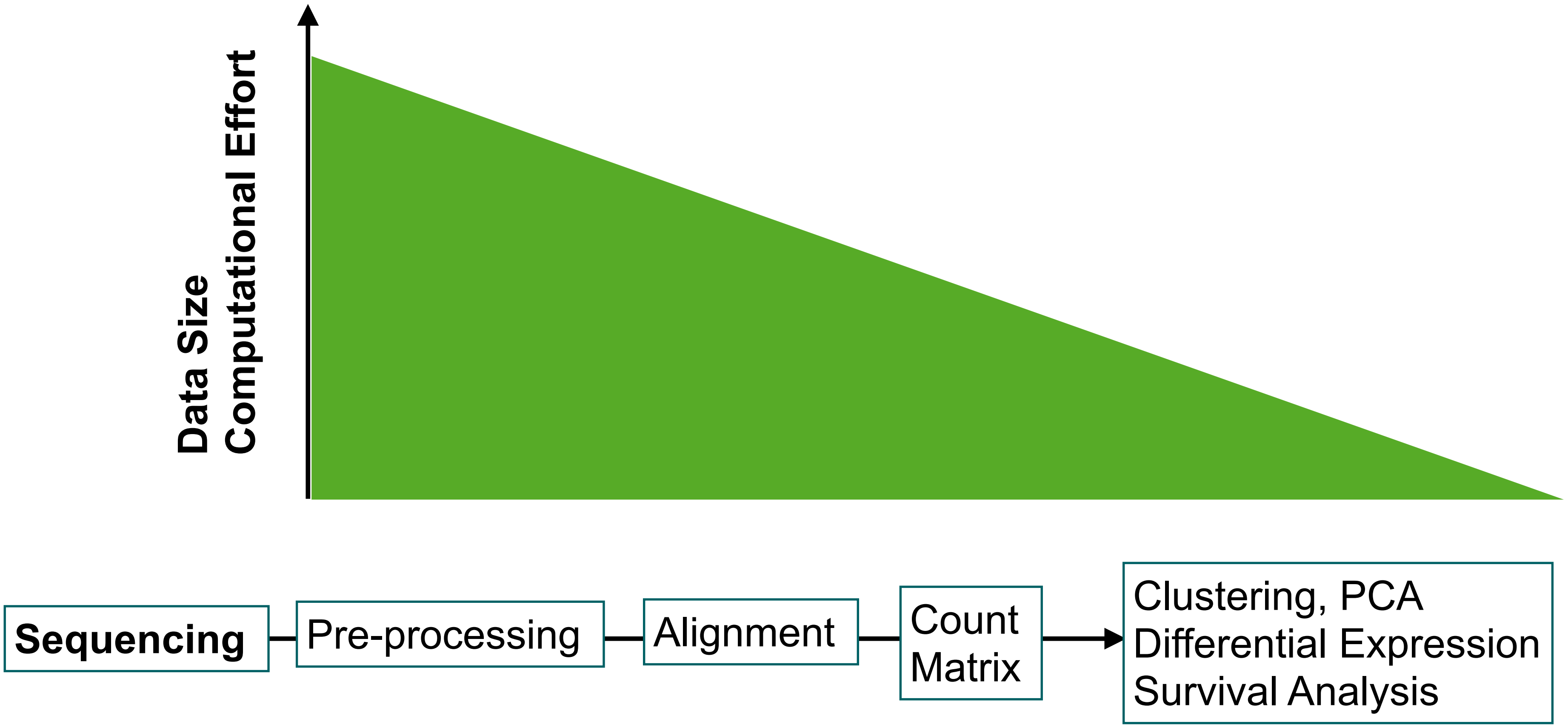
Data Category	Cases (n=377)	Files (n=10,814)
🟦 Raw Sequencing Data	377	1,637
🟦 Transcriptome Profiling	376	2,122
🟧 Simple Nucleotide Variation	375	3,032
🟧 Copy Number Variation	376	1,536
🟩 DNA Methylation	377	430
🟩 Clinical	377	423
🟥 Biospecimen	377	1,634

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!

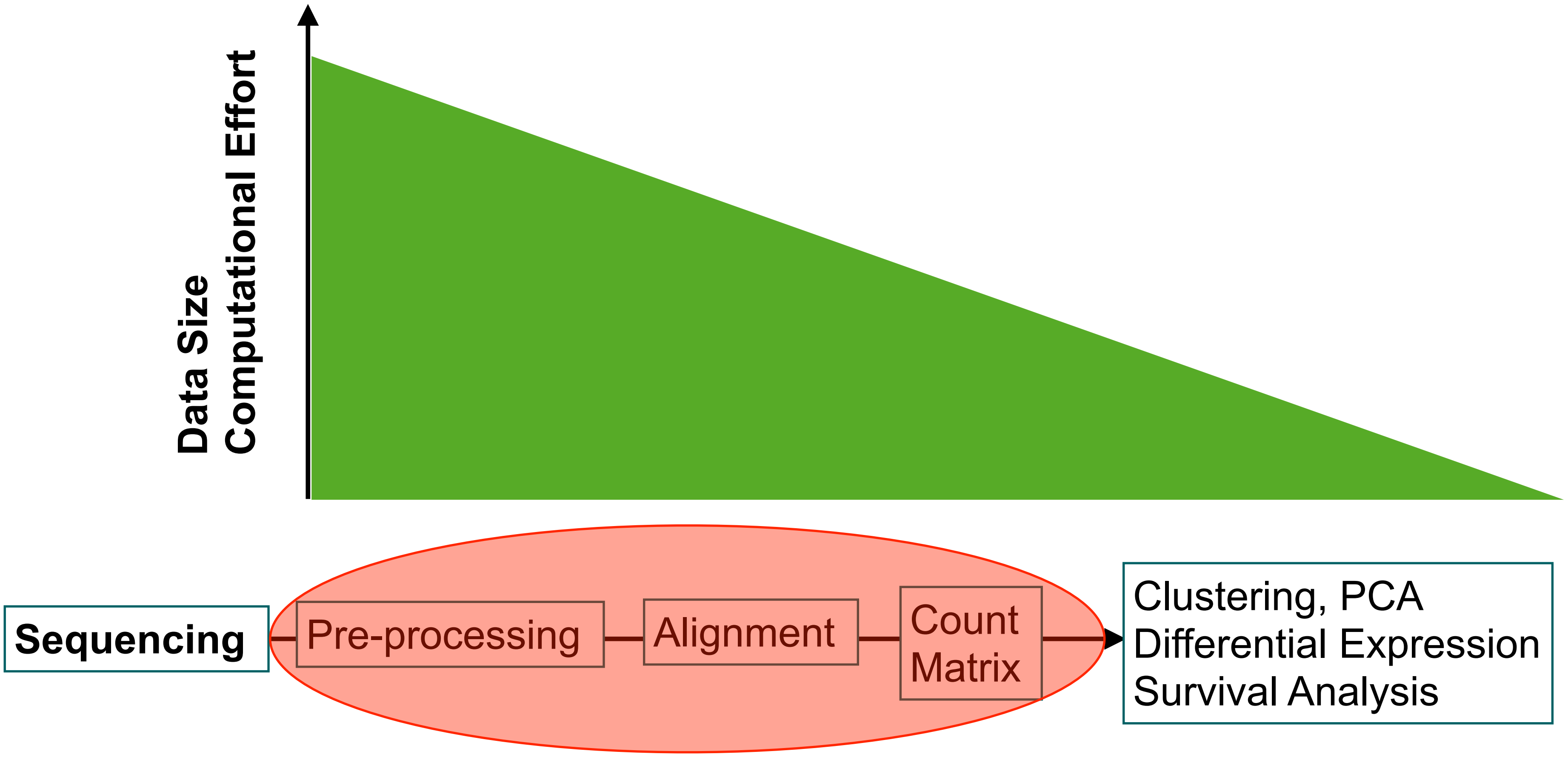
Bioinformatics Pipeline / RNA-seq



Data Size
Computational Effort

Sequencing — Pre-processing — Alignment — Count Matrix → Clustering, PCA
Differential Expression
Survival Analysis

Bioinformatics Pipeline / RNA-seq

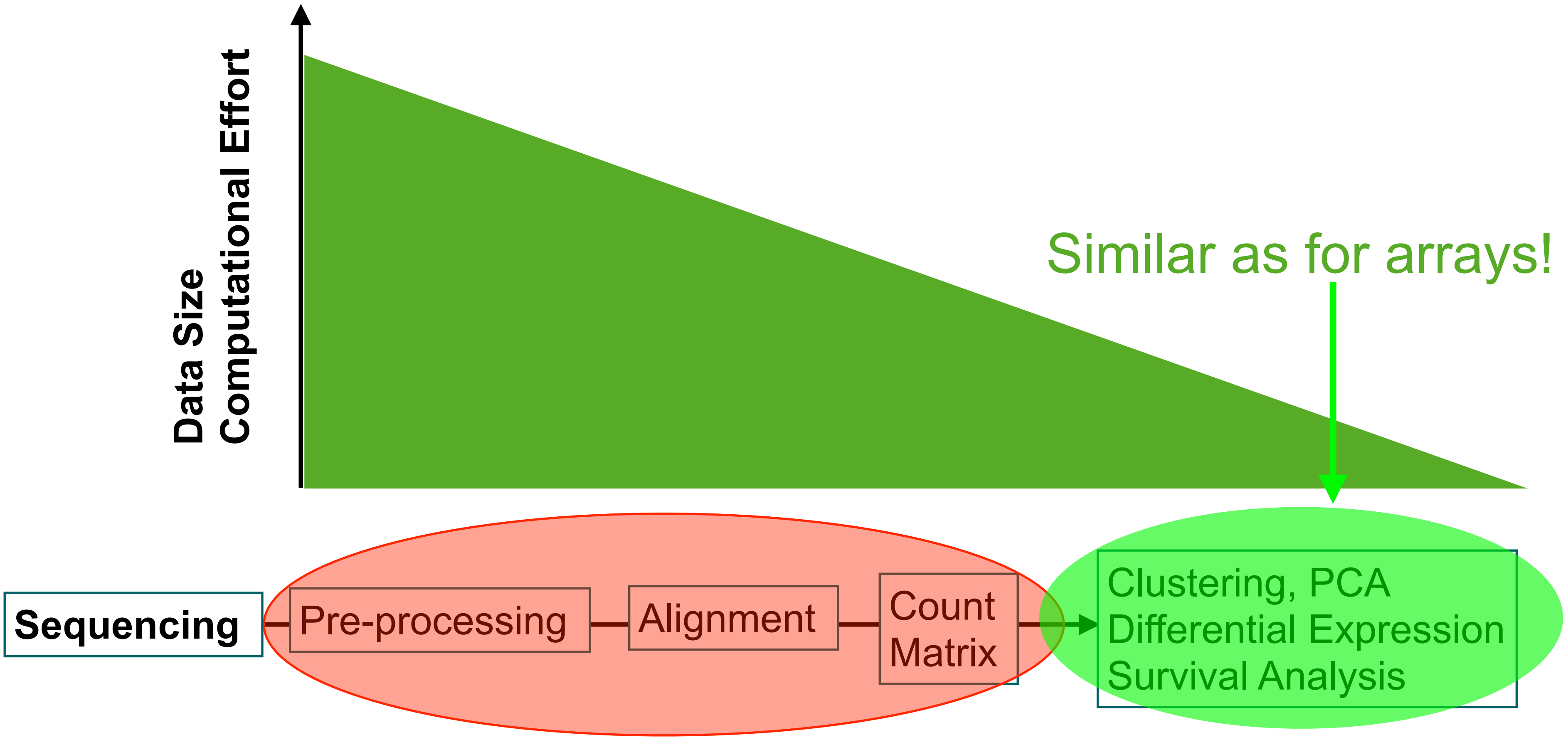


Practical part not covered!

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

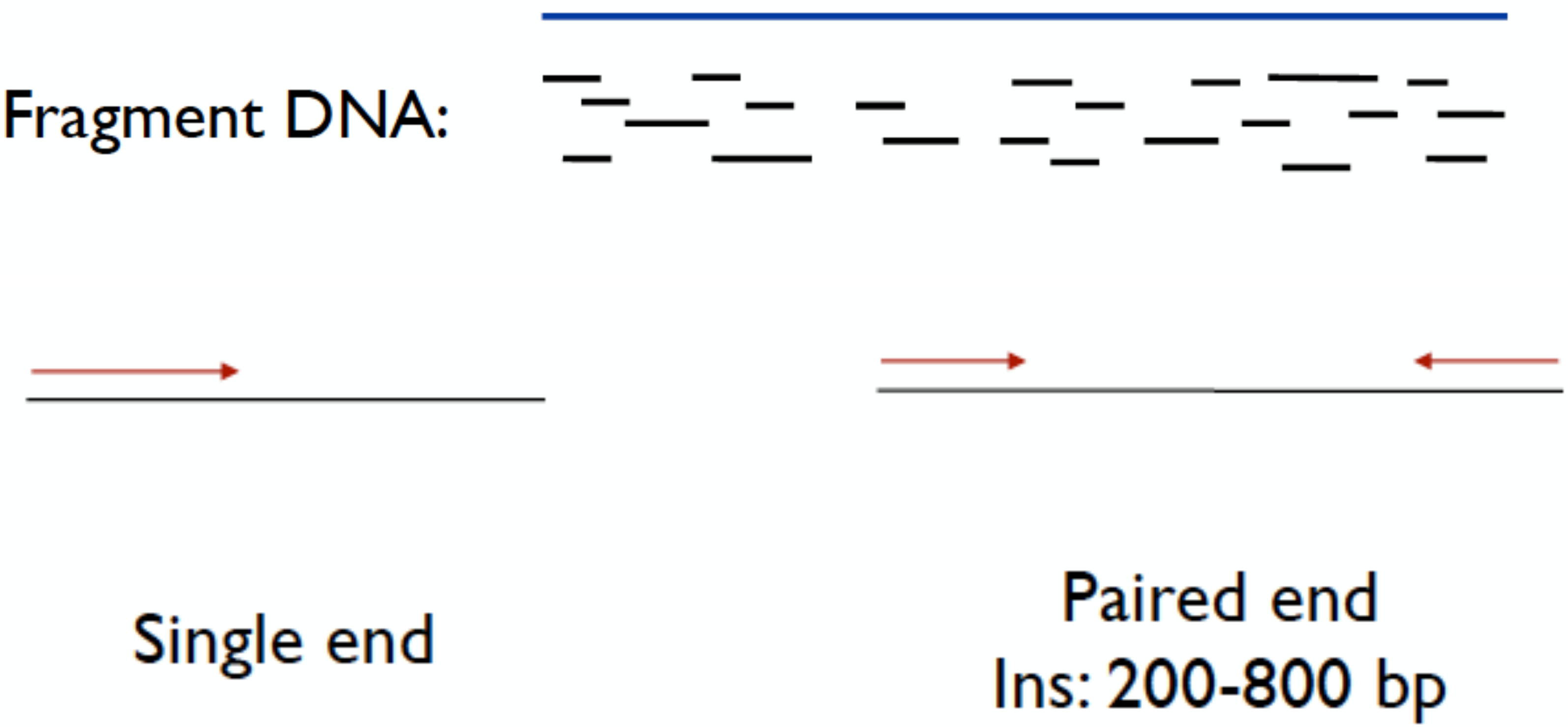
Bioinformatics Pipeline / RNA-seq



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%)

Read Types



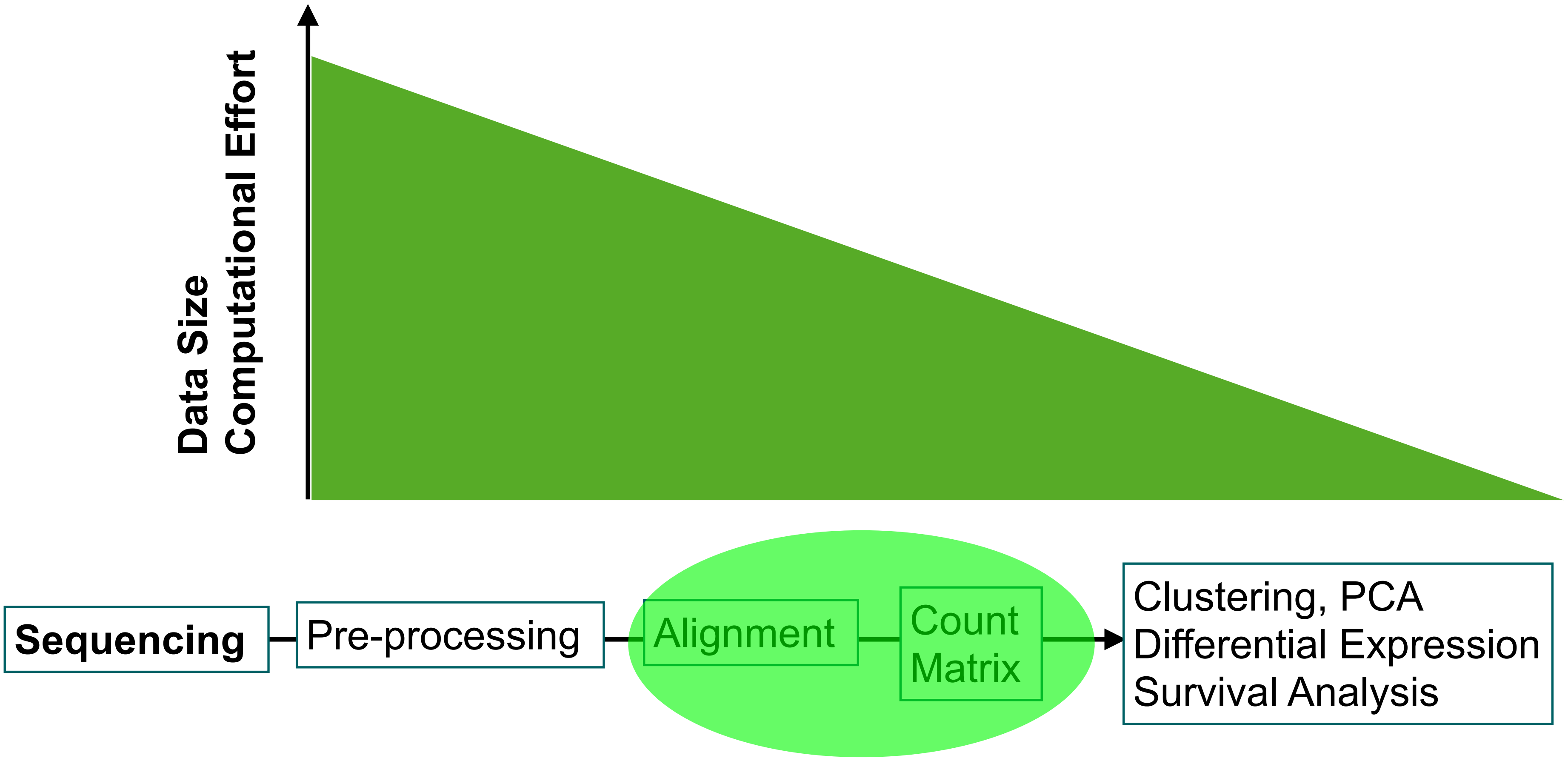
Fragment DNA:

Single end

Paired end

Ins: 200-800 bp

Bioinformatics Pipeline / RNA-seq



Sequencing

Pre-processing

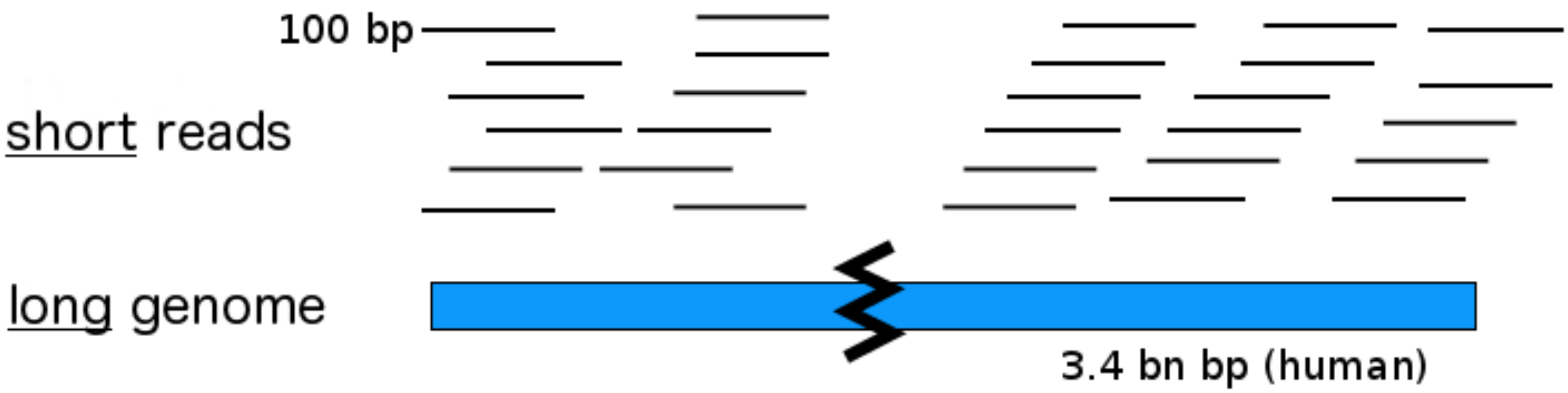
Alignment

Count Matrix

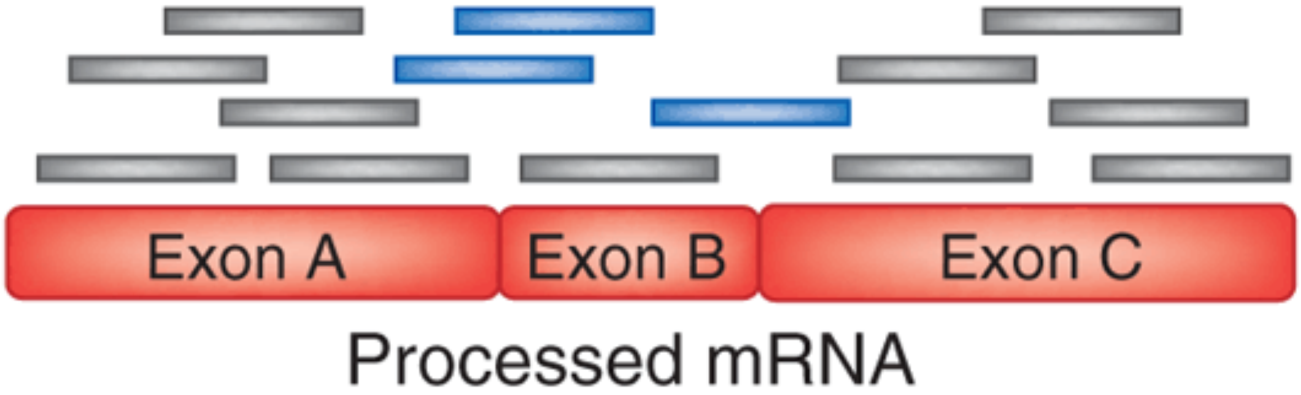
Clustering, PCA
Differential Expression
Survival Analysis

Alignment

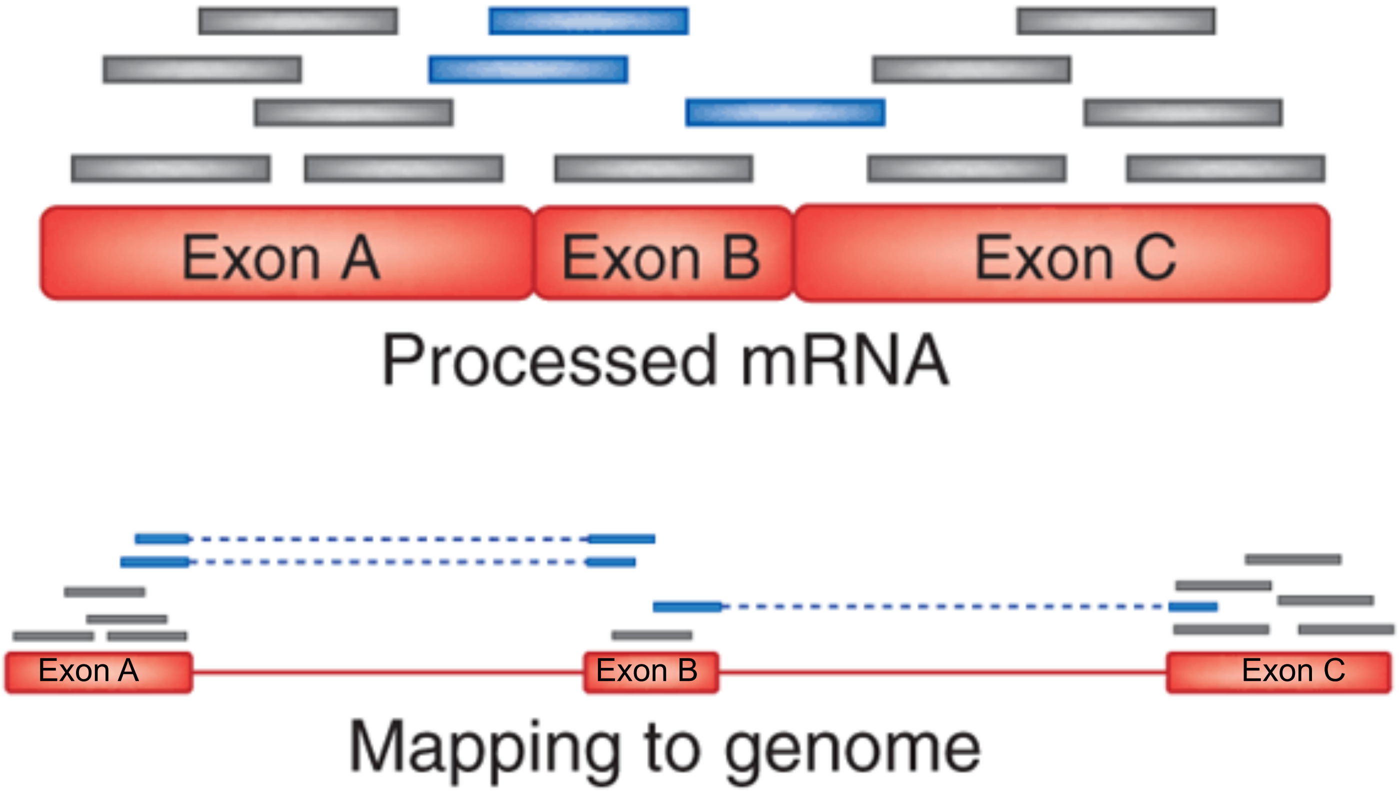
- a large reference sequence is given (genome)
    - up to billions of base pairs
- short reads (<200bps)
- find most probable position of the read in the genome (by inexact string matching)



Alignment - Split Read Mapping (RNA-Seq)



Alignment - Split Read Mapping (RNA-Seq)



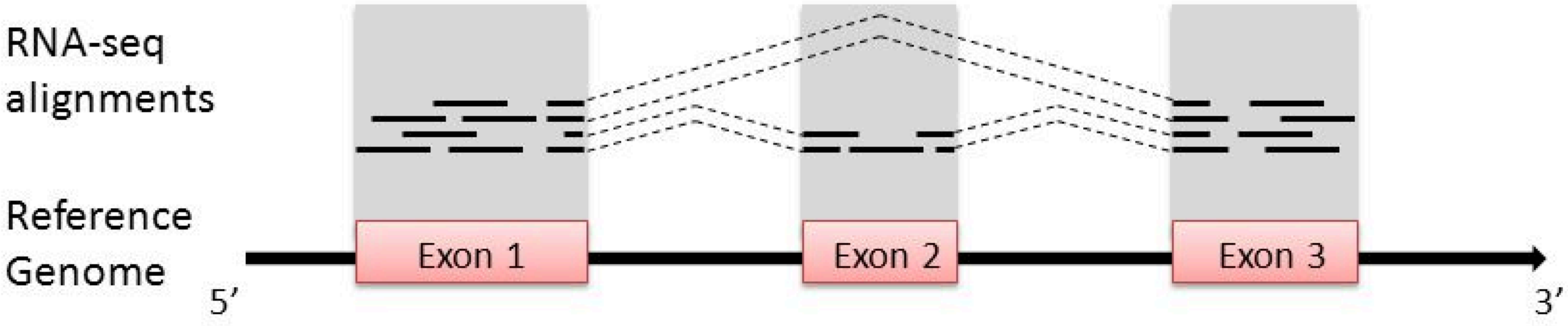
- reads are split between exons when mapped to genome
- aligners use transcript information or try to find splice events (STAR & TOPHAT)

Reference based aligners - Overview

	Time	*Precision*	*Pairs*	*GAPs*	*Phred*	*Memory*	*Application (Comments)*
BOWTIE	+		+	-	-	*5GB*	*General* *(max. 3 missmatches)*
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

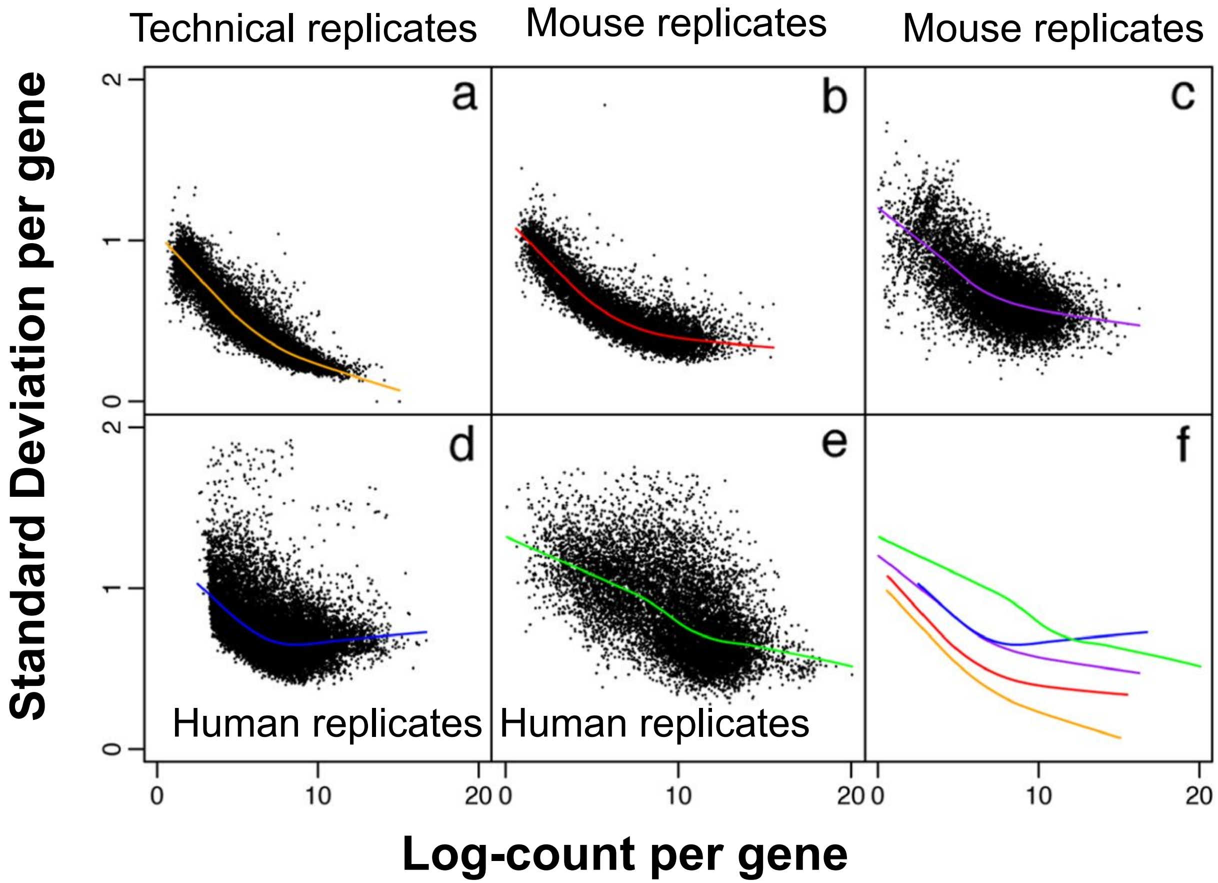
Gene Level - 17 reads
Exon level - exon 1 (8 reads), exon 2 (3 reads), exon 3 (6 reads)
Transcript Level - Exons 1,2 & 3 (10 reads) and exon 1 & 3 (7 reads) *
* complex computational methods required (RSe, or TopHAT needed for this)

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

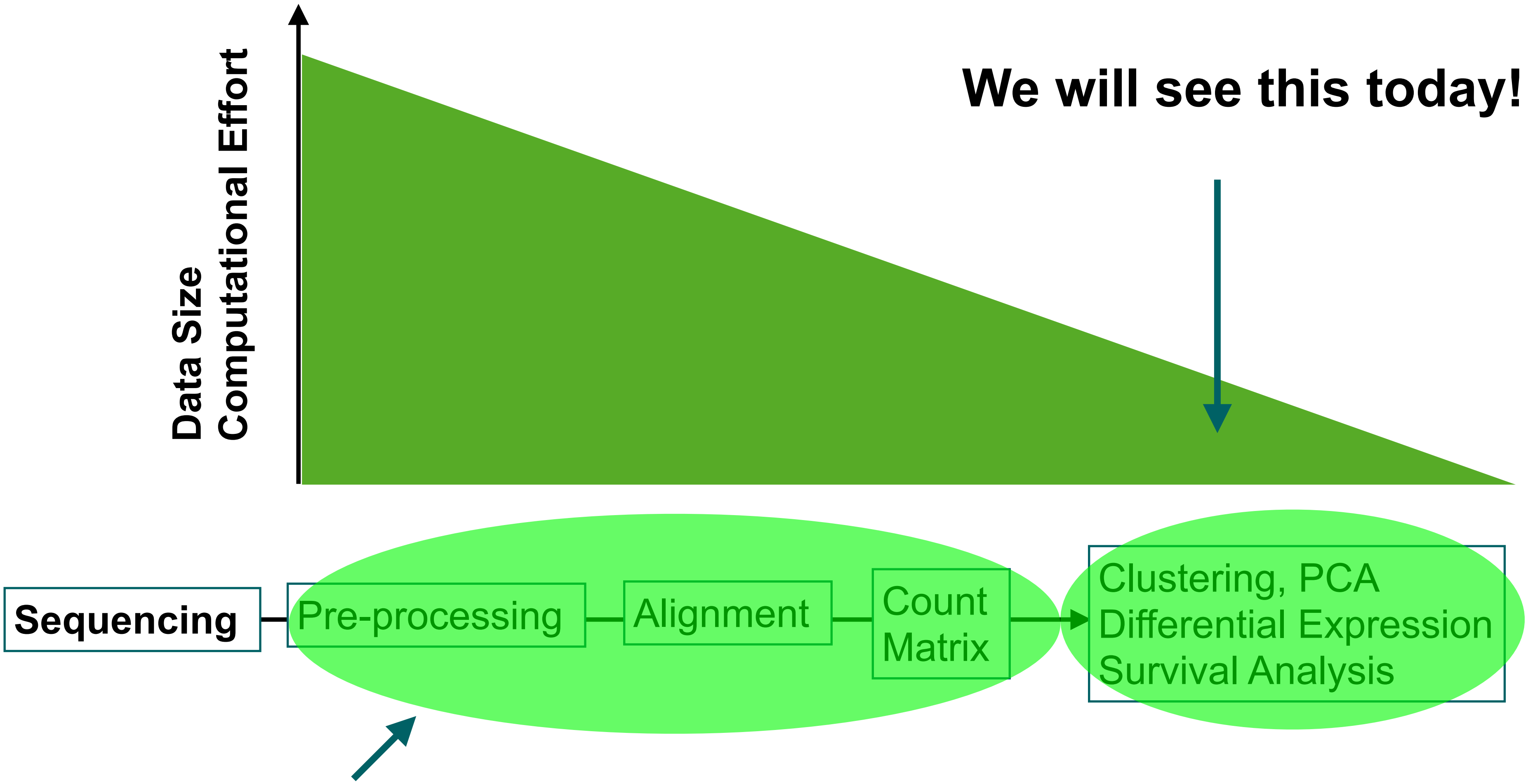


VOOM analysis is necessary to make variance similar to arrays.

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

Bioinformatics Pipeline / RNA-seq



We will see this today!

Provided by TGCA or your Core Facility!

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

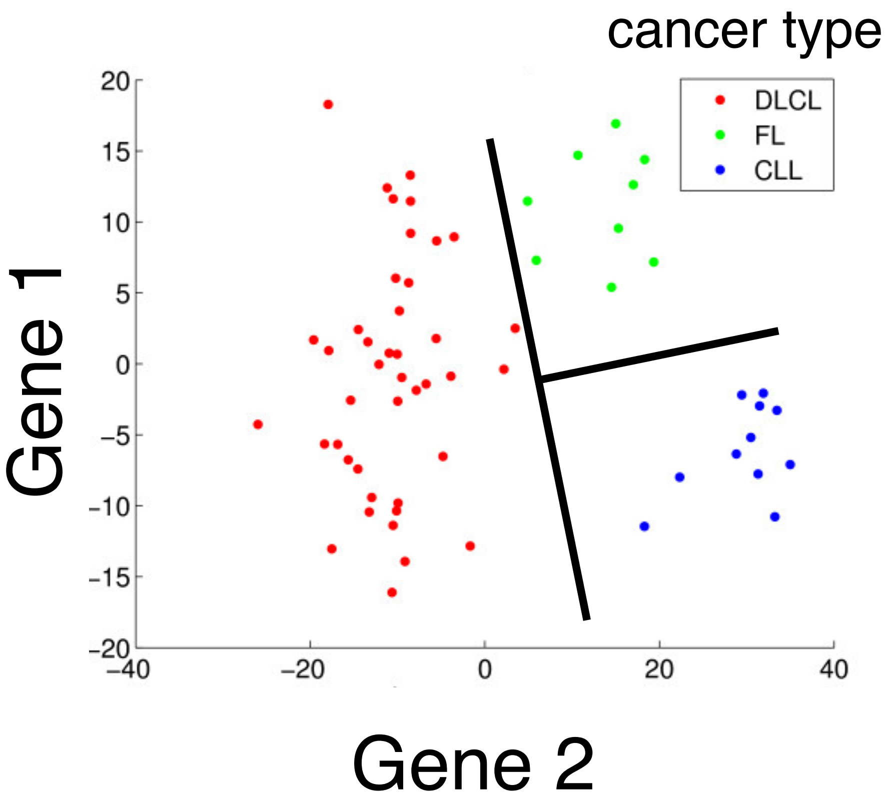
Personalized Medicine

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

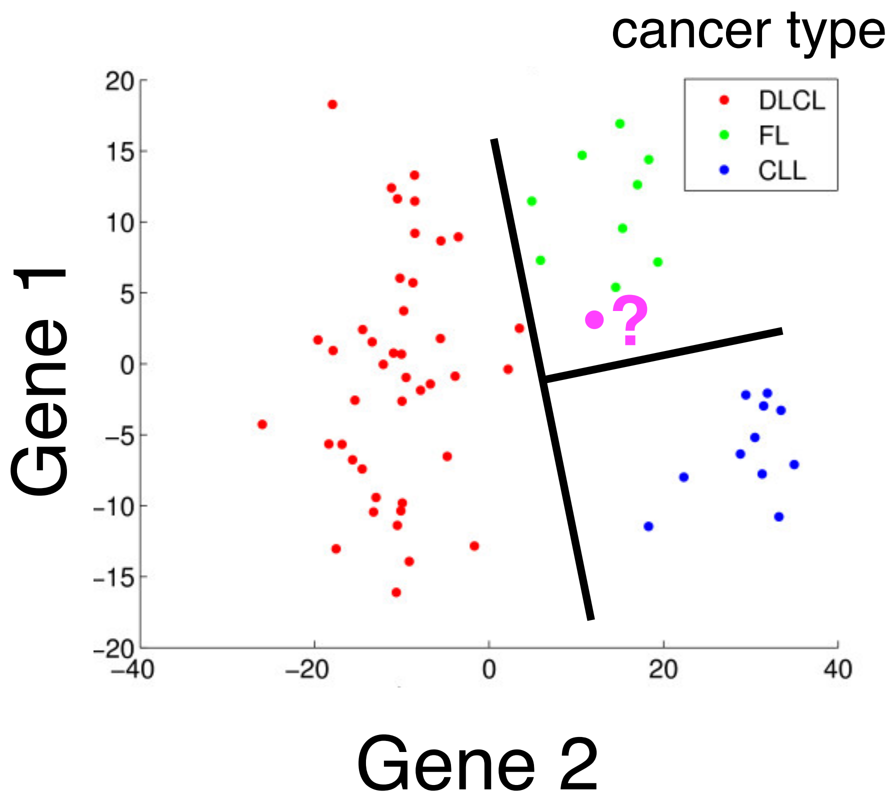
Expression matrix X (genes vs samples)

classification vector *Y* (diagnosis)

Find a function:

f(*x*) → *y*

Machine Learning - Classifier



Data

Expression matrix X (genes vs samples)

classification vector *Y* (diagnosis)

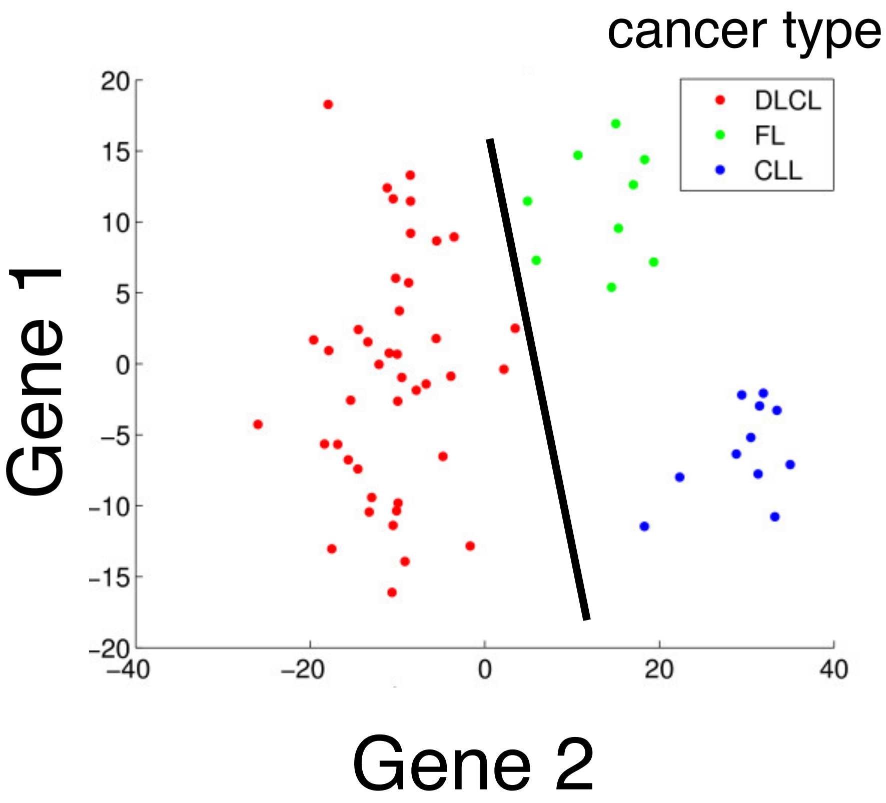
Find a function:

f(*x*) → *y*

For new patients X':

f(*x*') → *y*'

Linear Classifier



Linear Function:

$f\ (x\ ,\ A) = a_0 + a_1 x_1 + \ldots + a_L x_L$

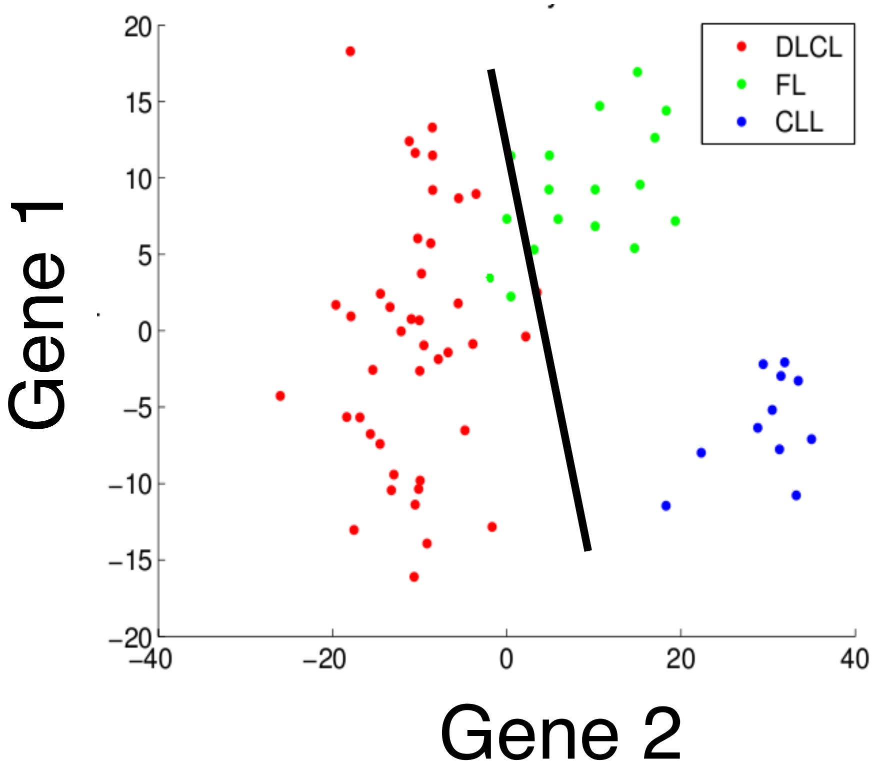
$f\ (\ x,\ A) > 0 \Rightarrow$ class A

$f\ (\ x,\ A) \leq 0 \Rightarrow$ class B

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

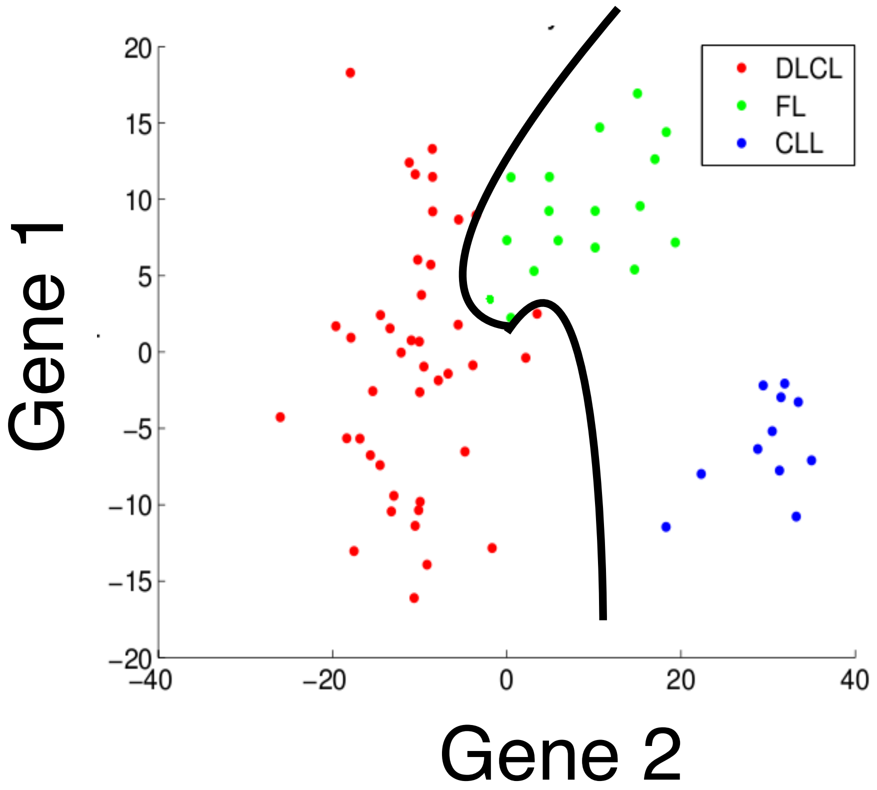
Institute for Computational Genomics
RWTH AACHEN UNIVERSITY

Linear Classifier - Problems



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

Nonlinear Classifier - Problems



- Polinomial Function
- $f\,(x\,,A) = a_0 + a_{11}x^3{}_1 + \ldots + a_{L1}x^3{}_L$

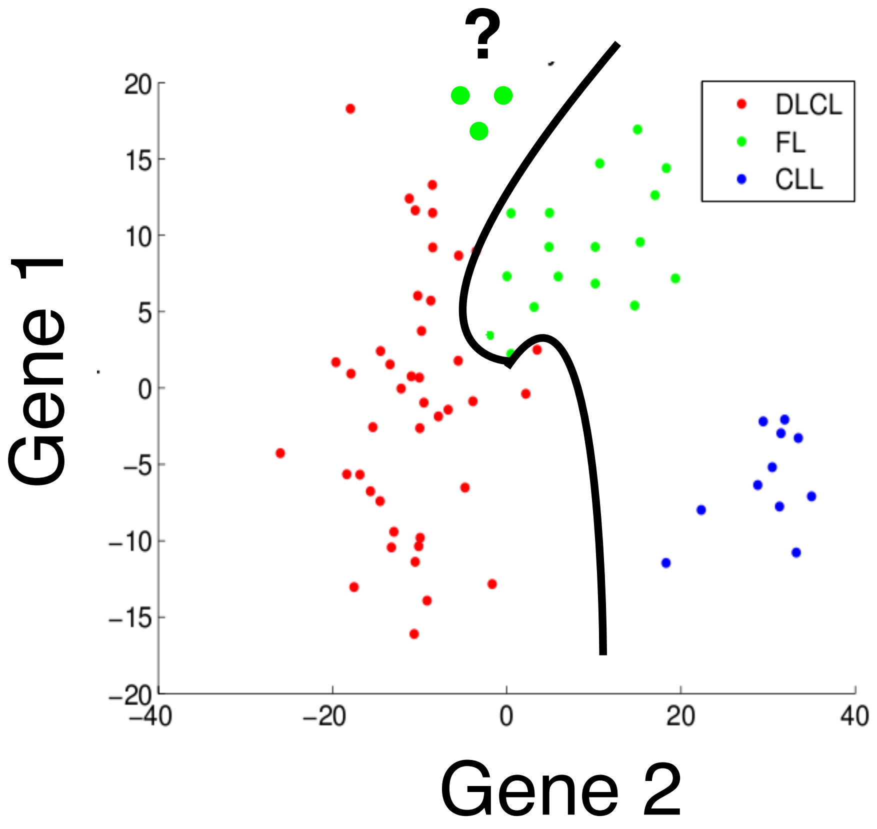
  $a_{12}x^2{}_1 + \ldots + a_{L2}x^2{}_L$

  $a_{12}x_1 + \ldots + a_{L2}x_L$
- Third order polynomial
- Problem: overfitting

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

Nonlinear Classifier - Problems

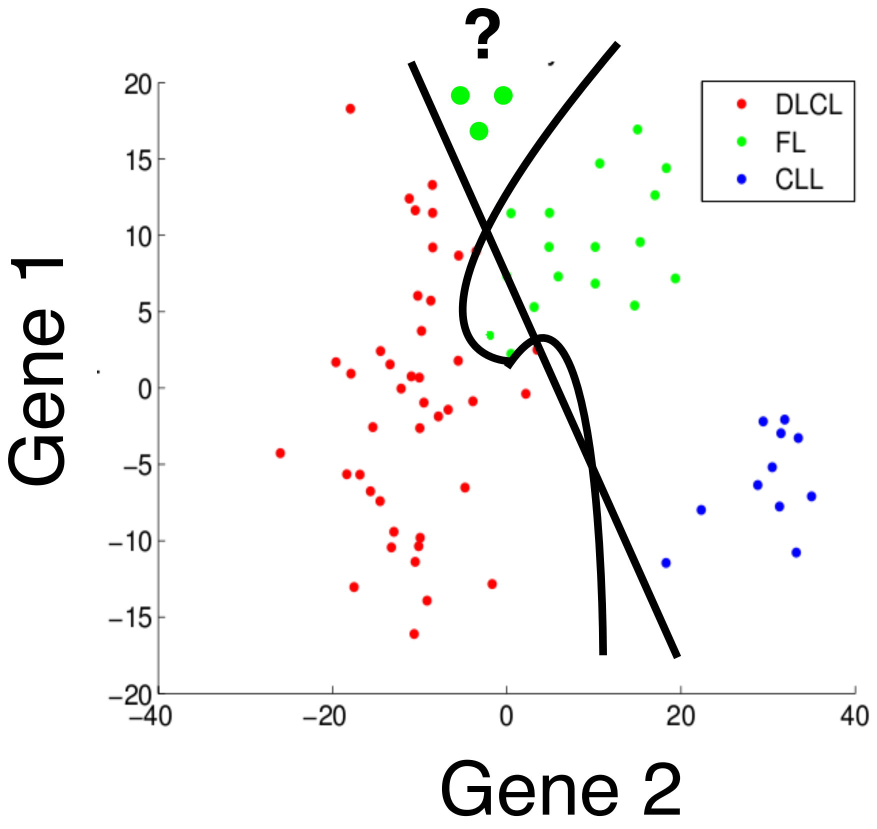


- Polinomial Function
- $f\ (x\ ,\ A) = a_0 + a_{11}x^3{}_1 + \ldots + a_{L1}x^3{}_L$

  $a_{12}x^2{}_1 + \ldots + a_{L2}x^2{}_L$

  $a_{12}x_1 + \ldots + a_{L2}x_L$
- Third order polynomial
- Problem: overfitting

Nonlinear Classifier - Problems



- Polinomial Function
- $f\,(x\,,\,A) = a_0 + a_{11}x^3{}_1 + \ldots + a_{L1}x^3{}_L$

  $a_{12}x^2{}_1 + \ldots + a_{L2}x^2{}_L$

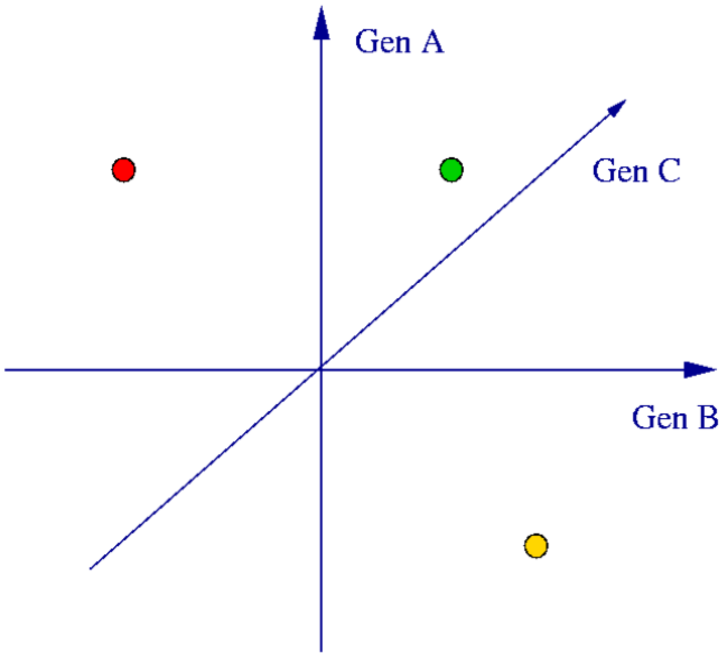
  $a_{12}x_1 + \ldots + a_{L2}x_L$
- Third order polynomial
- Problem: overfitting

Curse of Dimensionality

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

Dots (patients) are sparsely distributed in space

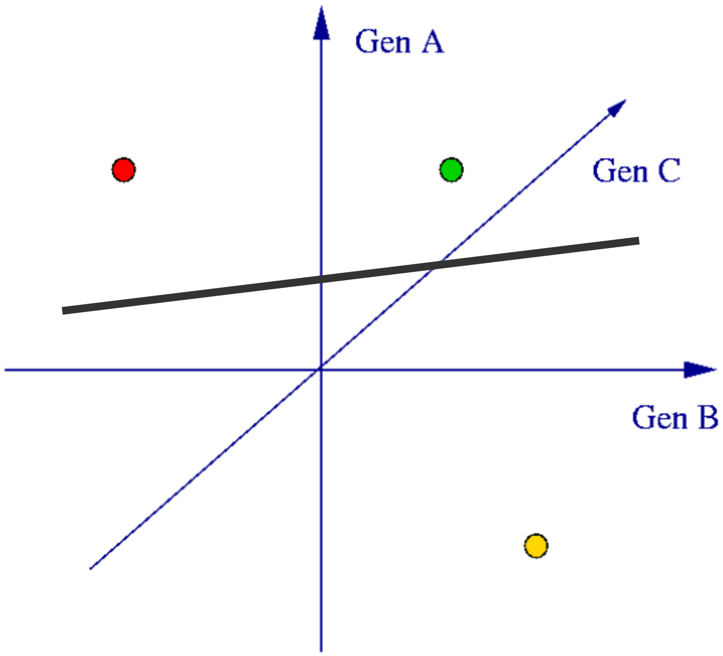
Curse of Dimensionality : Example



Sparse data

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

Curse of Dimensionality : Example



- Sparse data

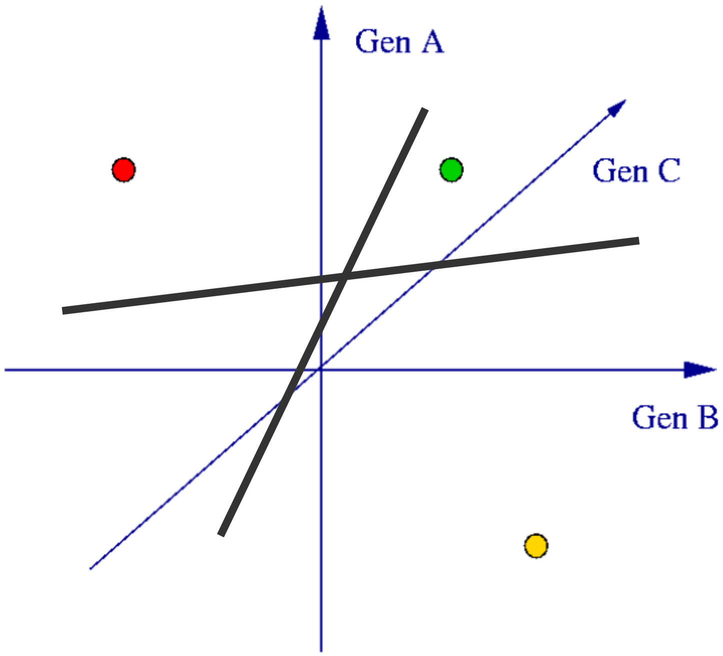
  - three genes

  - 2 patients with known cancer (red vs yellow)

  - 1 unknown (green)

Perfect classifier (on training)

Curse of Dimensionality : Example



- Sparse data

  - three genes

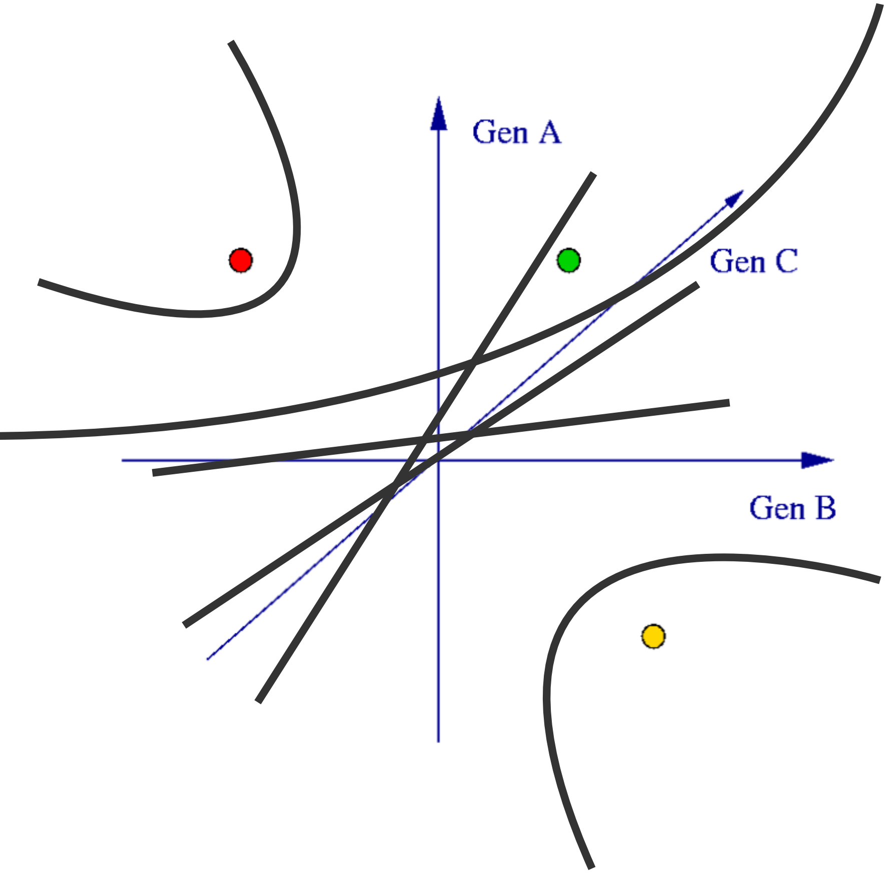
  - 2 patients with known cancer (red vs yellow)

  - 1 unknown (green)

Both are perfect classifiers (on training)

Hard to generalise!

Curse of Dimensionality : Example



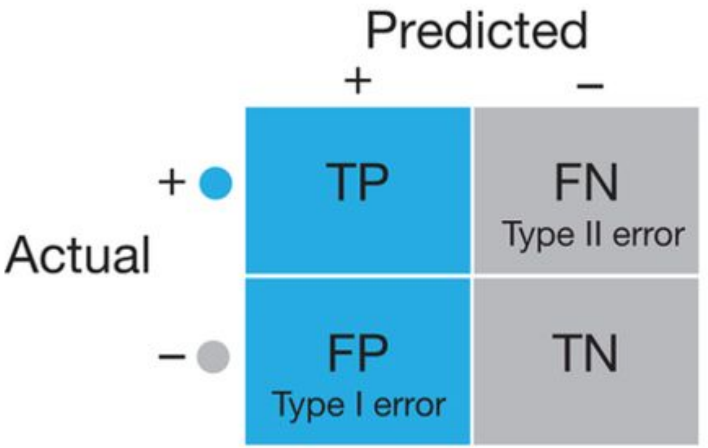
- There are millions of perfect linear classifiers
- And even more non-linear classifiers!

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 -)

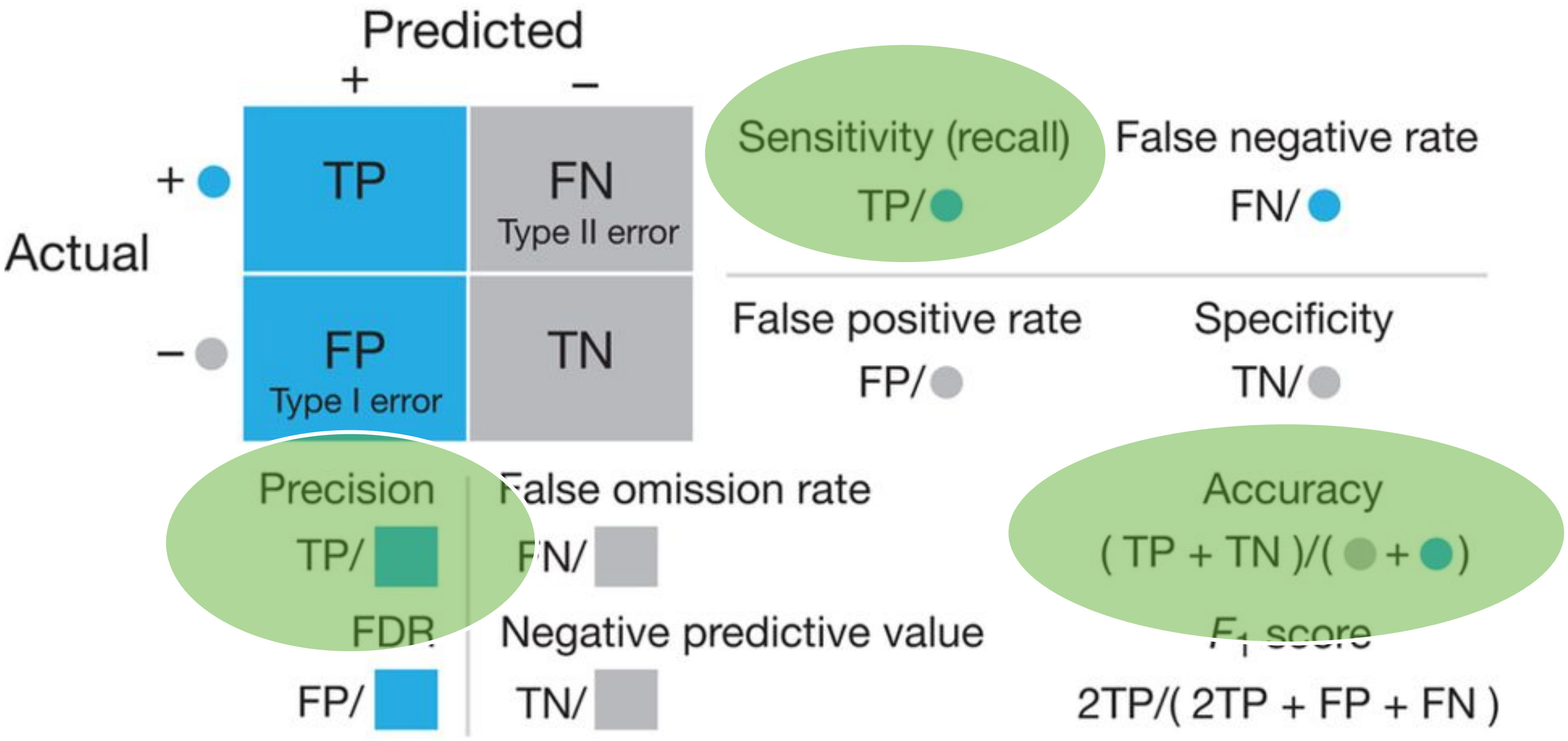


		Predicted +	Predicted –
Actual	+ ●	TP	FN Type II error
Actual	– ●	FP Type I error	TN

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

Classifier Evaluation

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



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

Sensitivity (recall)
TP/●

False negative rate
FN/●

False positive rate
FP/●

Specificity
TN/●

Precision
TP/■

FDR
FP/■

False omission rate
FN/■

Negative predictive value
TN/■

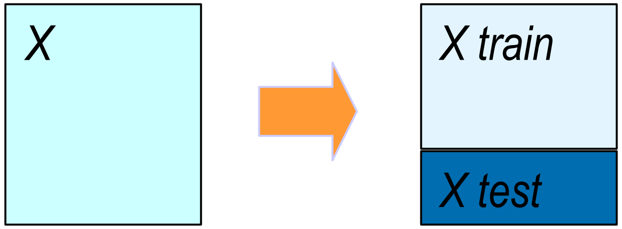
Accuracy
(TP + TN)/(● + ●)

F_1 score
2TP/(2TP + FP + FN)

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

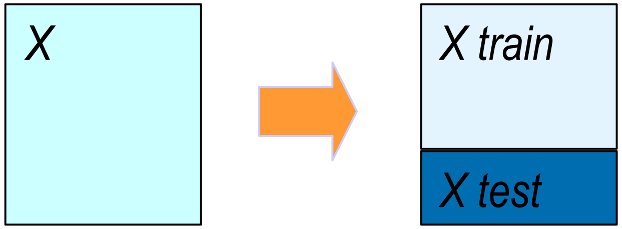
Classifier Evaluation

- The performance of your classifier needs to be evaluated on your test data:
  - an independent "validation cohort"
  - or retain a set of samples (1/3) that has similar distribution of classes of your total data



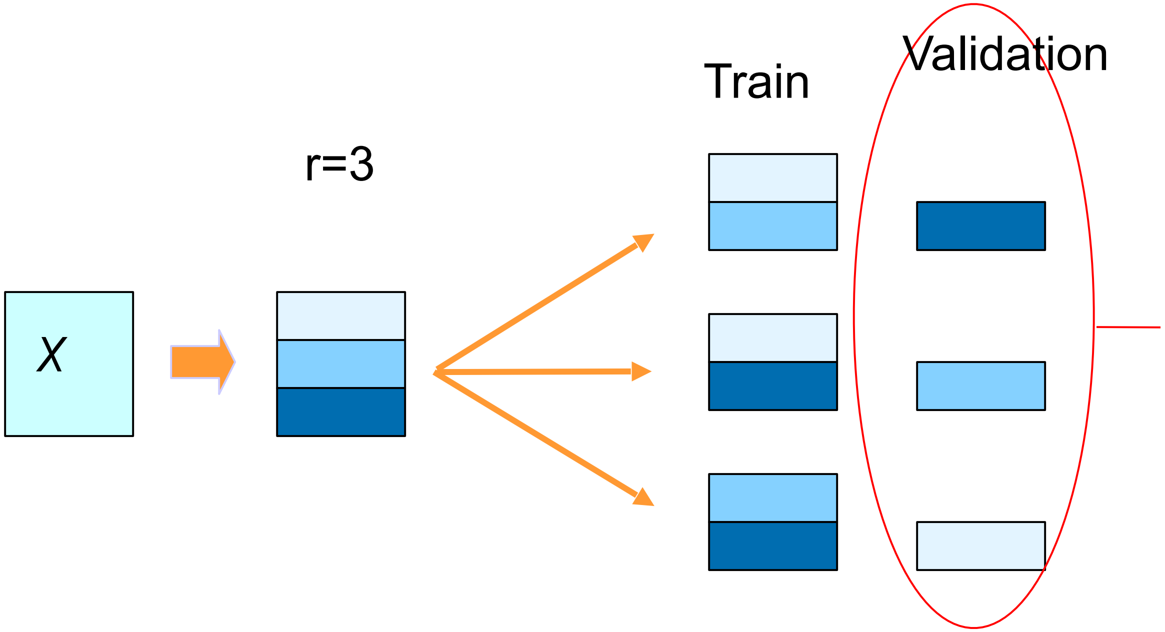
Classifier Evaluation

- The performance of your classifier needs to be evaluated on your test data:
  - an independent "validation cohort"
  - or retain a set of samples (1/3) that has similar distribution of classes of your total 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



mean and std of accuracy, precision and sensitivity

Elastic Net

Is based on a linear function:

$f\,(x\,,A) = a_0 + a_1 x_1 + \ldots + a_L x_L$

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

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

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

Hands on!

Exercise (after the handout)

You should perform clustering of tissues with liver cancer. Tip: use code similar to the one seen in gene expression data (day 3). Since, we are interested in grouping patients, you can transpose the matrix with the function **t***.*

1. Can you see nice clusters in the dendrogram?

2. What about genes associated to each group? Are they associated to some particular biological function? Use differential expression analysis and GO enrichment analysis to solve this task.

www.costalab.org

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

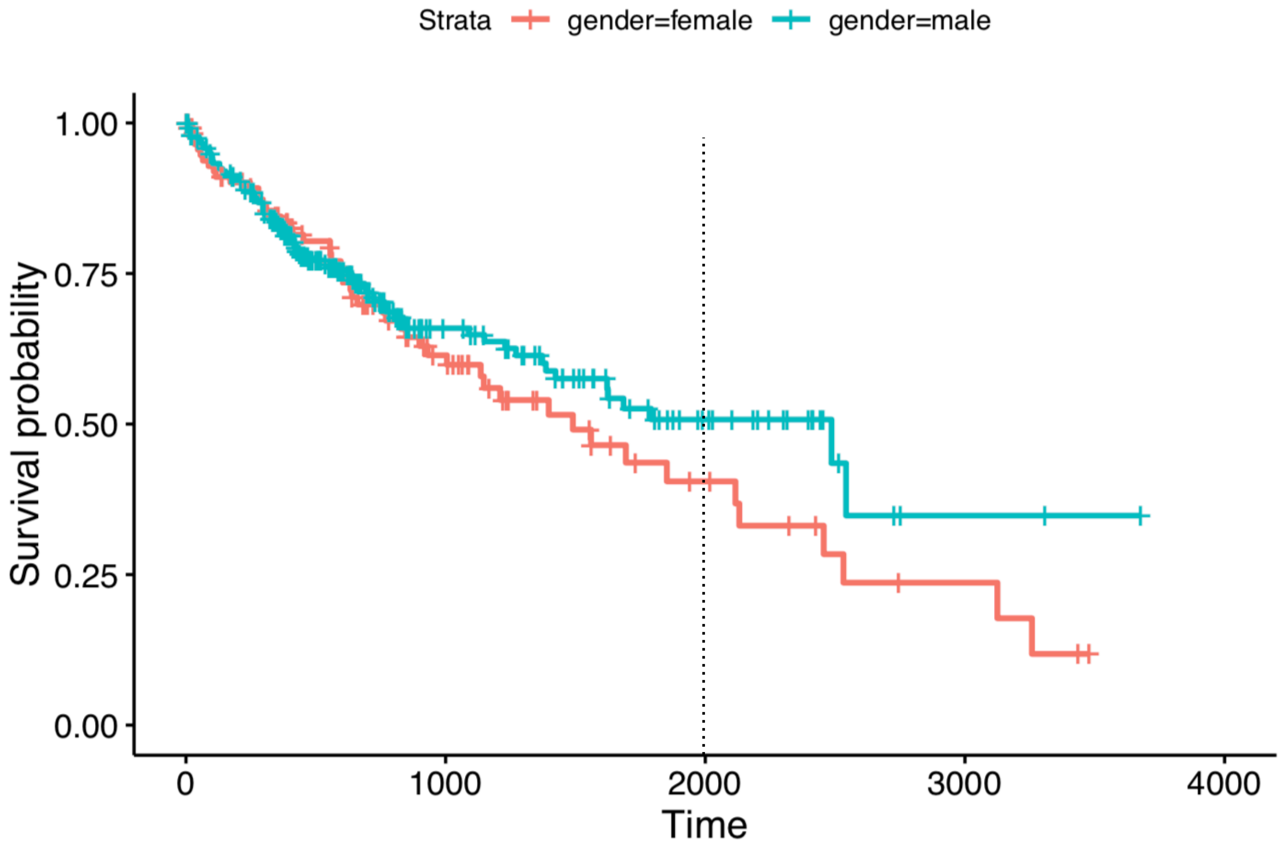
Data:

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

Patient	*Status*	*Time*	*Sex*
1	*Dead*	*343*	*Male*
2	*Alive*	*20*	*Male*
3	*Alive*	*300*	*Female*
4	*Dead*	*200*	*Male*

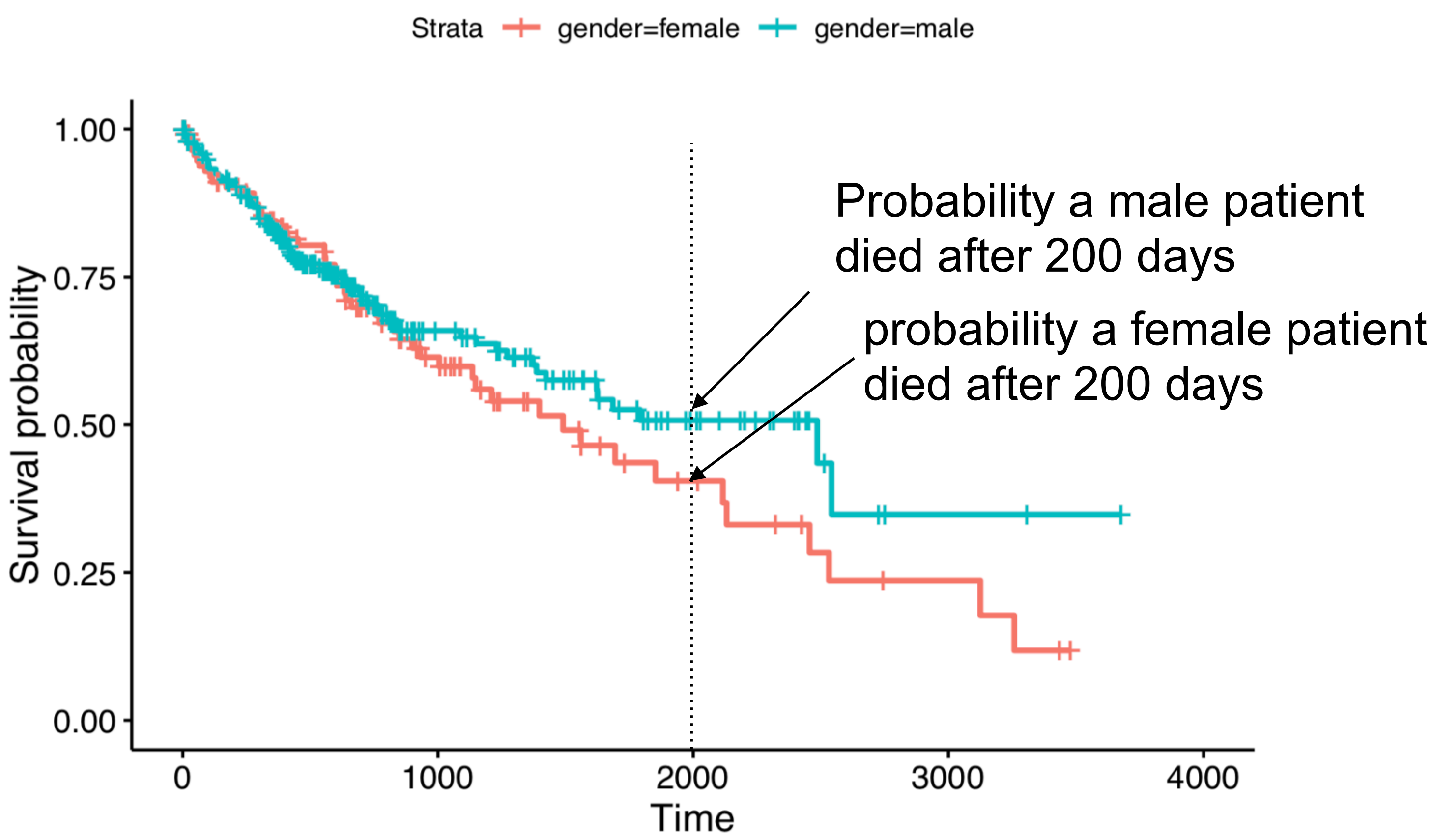
Kaplan-Meier plot

Survival of LIHC patients - male vs. Female



Kaplan-Meier plot

Survival of LIHC patients - male vs. Female



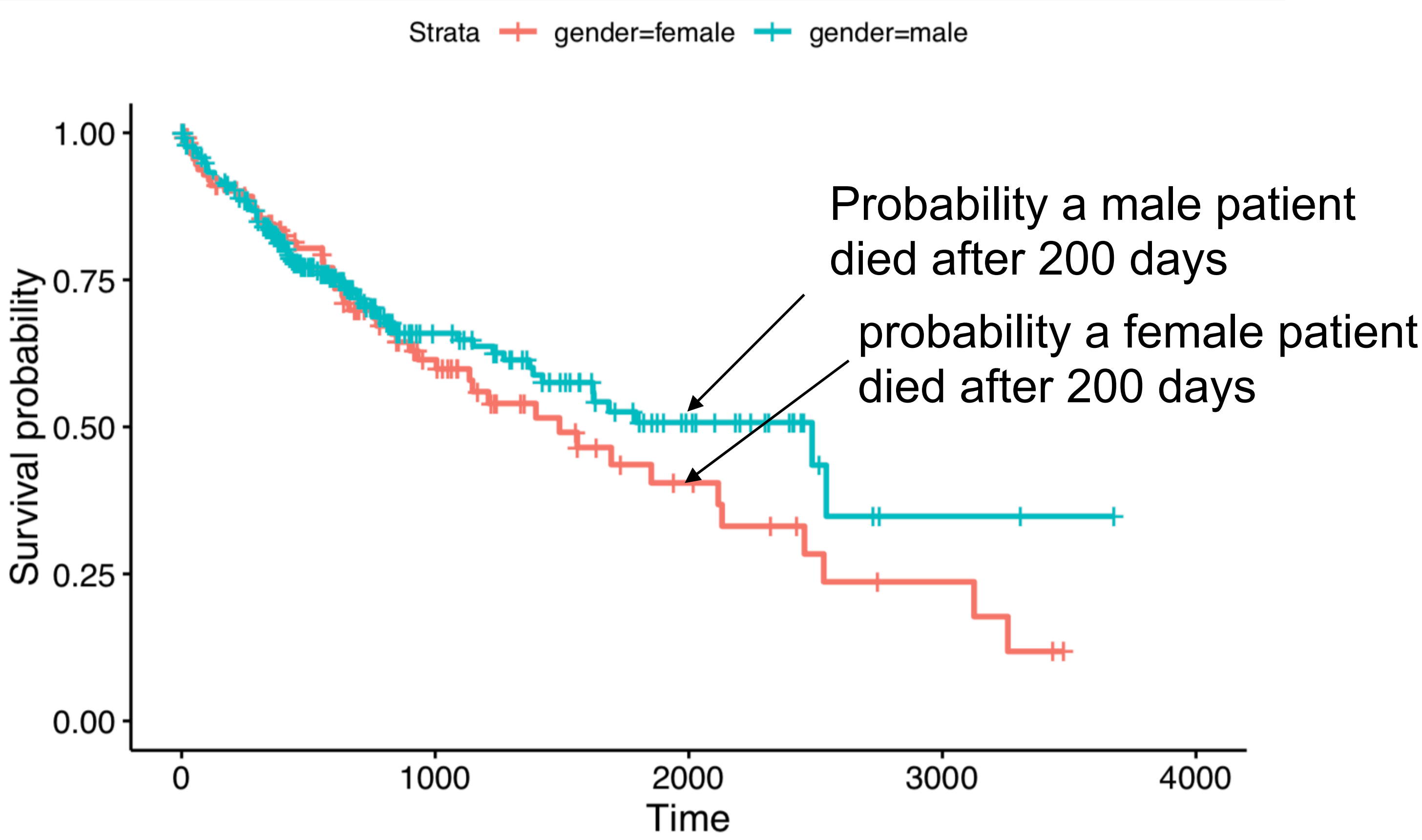
Probability a male patient died after 200 days

probability a female patient died after 200 days

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

Log-rank test

Is the survival difference significant?

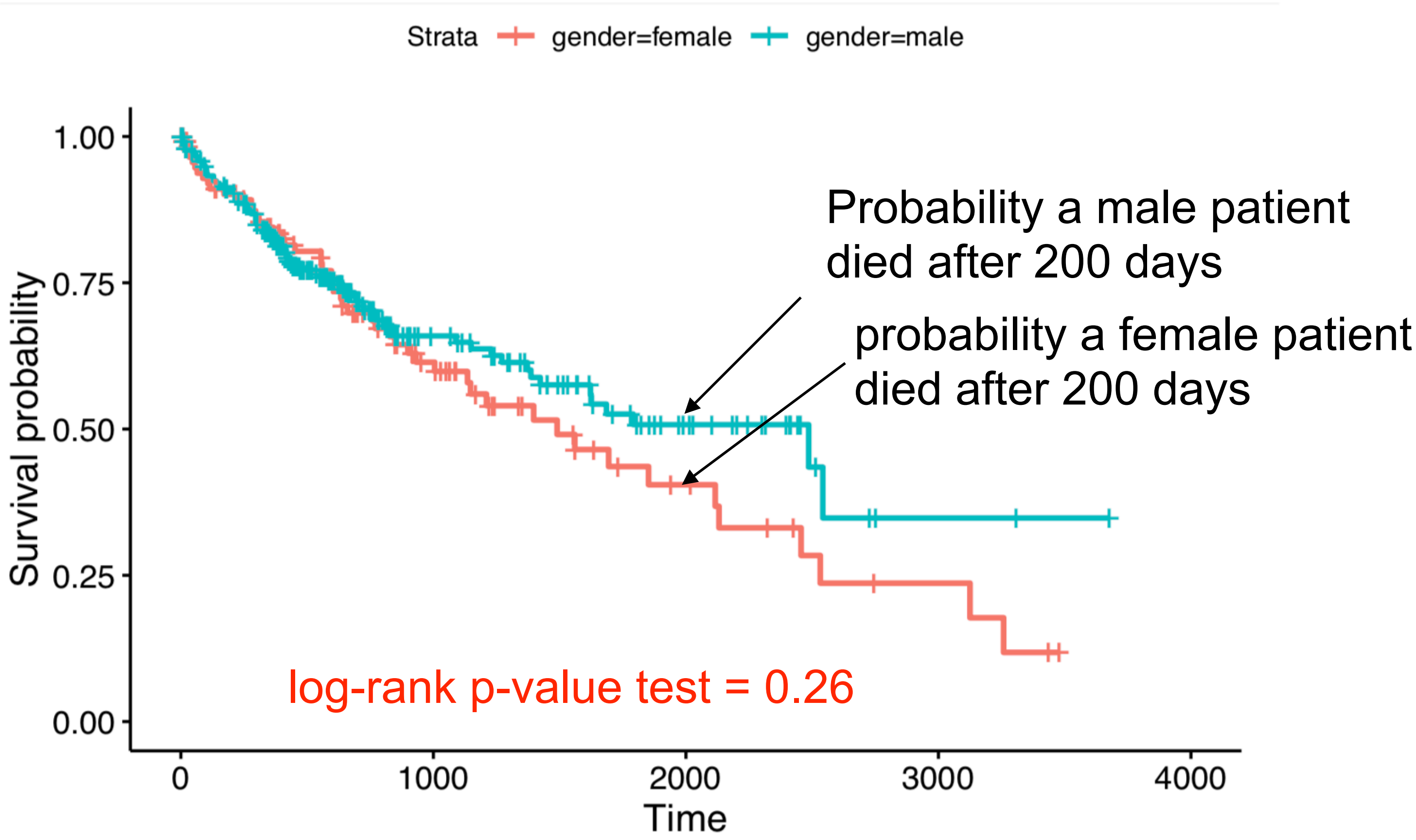


Probability a male patient died after 200 days

probability a female patient died after 200 days

Log-rank test

Is the survival difference significant?

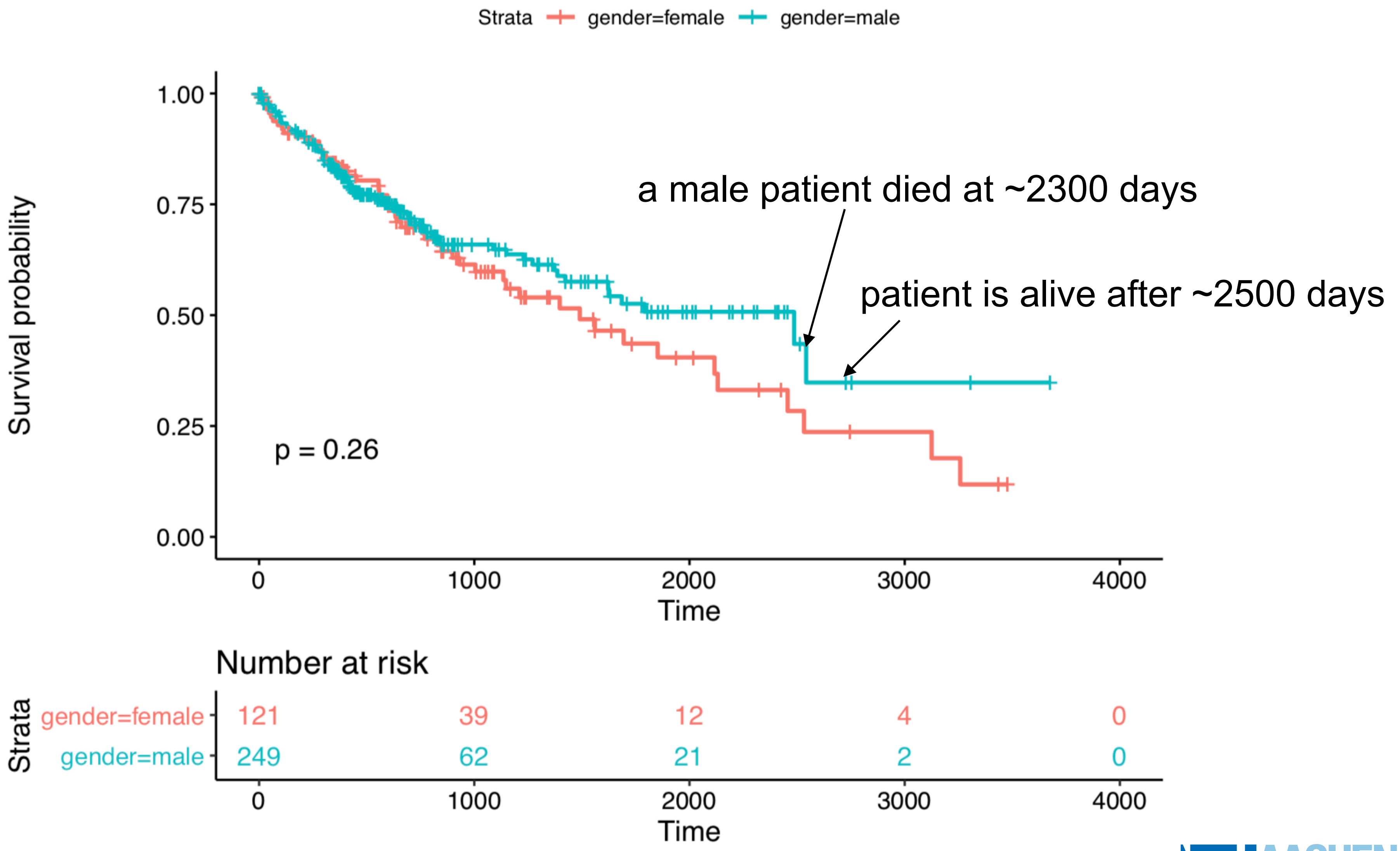


Probability a male patient died after 200 days

probability a female patient died after 200 days

log-rank p-value test = 0.26

Kaplan-Meier plot

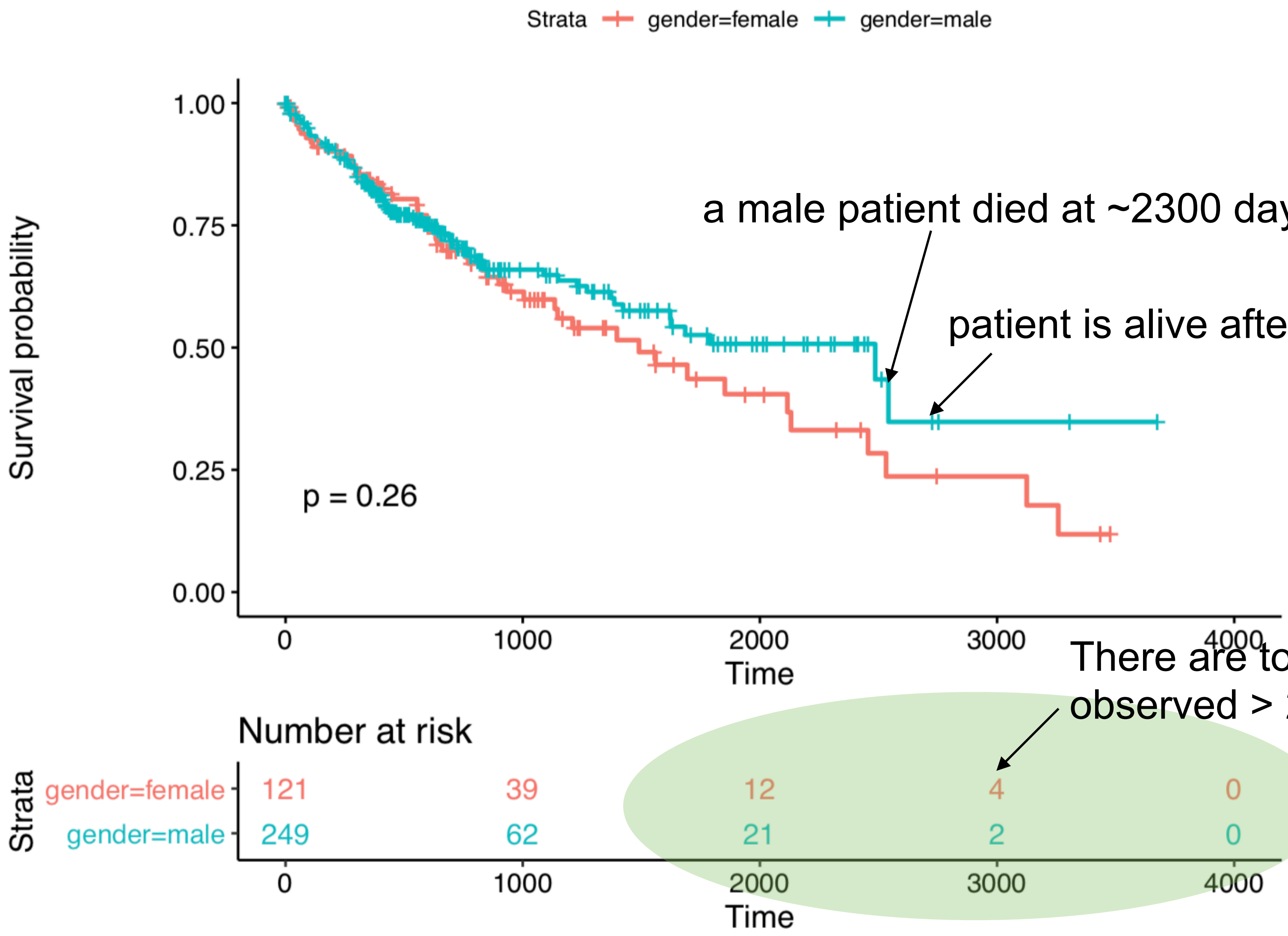


a male patient died at ~2300 days

patient is alive after ~2500 days

Strata	0	1000	2000	3000	4000
gender=female	121	39	12	4	0
gender=male	249	62	21	2	0

Kaplan-Meier plot



a male patient died at ~2300 days

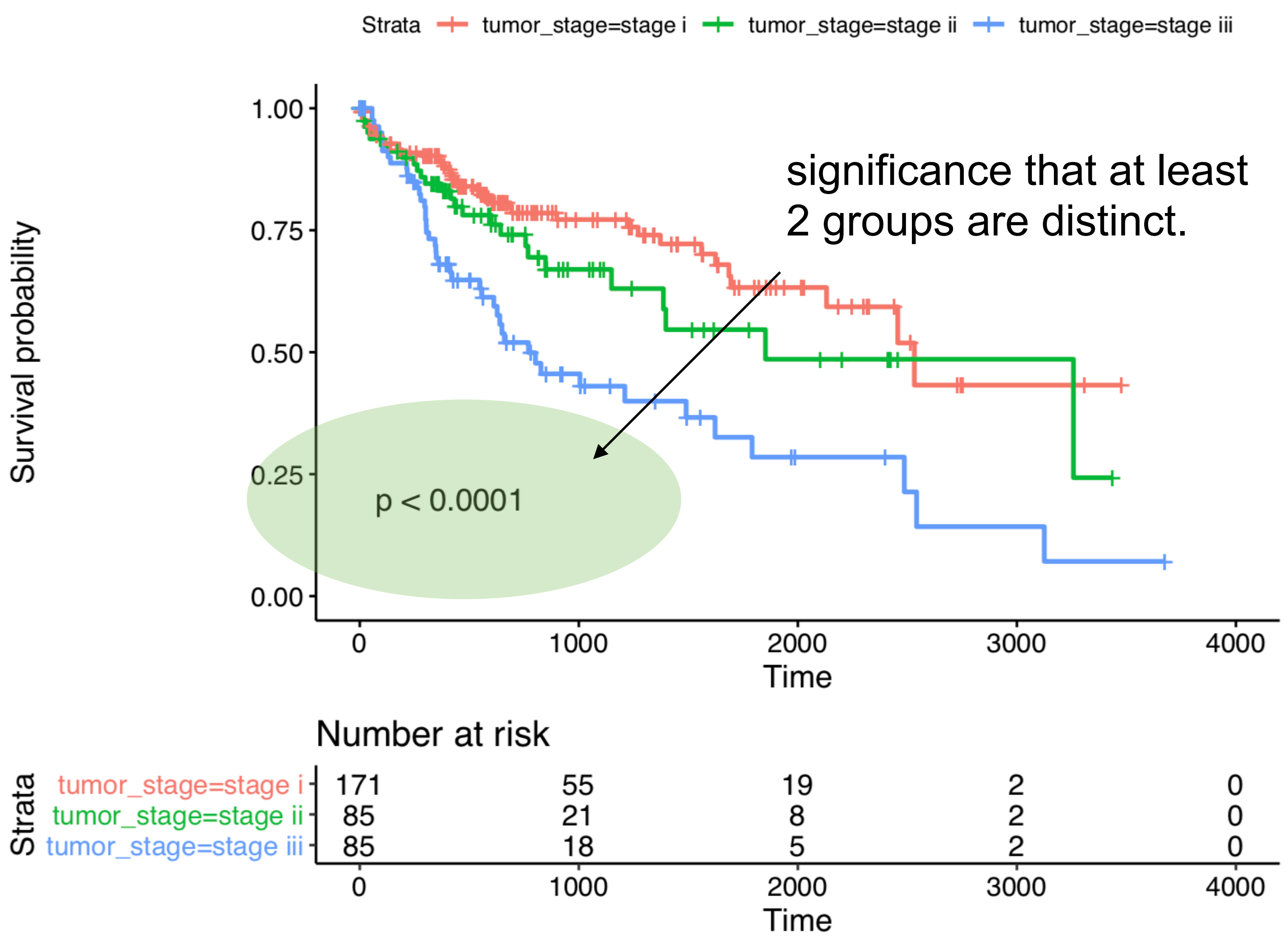
patient is alive after ~2500 days

There are too few patients observed > 2000 days

Kaplan-Meier / Log-Rank Test

KM and LRT can compare several groups at a time.

Survival vs Tumour stage at diagnosis



significance that at least 2 groups are distinct.

Number at risk

Strata	0	1000	2000	3000	4000
tumor_stage=stage i	171	55	19	2	0
tumor_stage=stage ii	85	21	8	2	0
tumor_stage=stage iii	85	18	5	2	0

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-Meyer plots (seen in practical).

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

Next week - Final Project

Ideas:

- Perform an analysis of a real gene expression data set
- Project can be developed in groups of 2-3 students
- Groups need to create an R code and a 10 minutes presentation showing the analysis

Schedule:

9:30 - Problem explanation
15:00 - Delivery of code and presentation slides
15:00 to 17:00 - Presentations

Hands on!

Hands on!