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

Institute for Computational Genomics
01011011010
10100100101
RWTH AACHEN UNIVERSITY

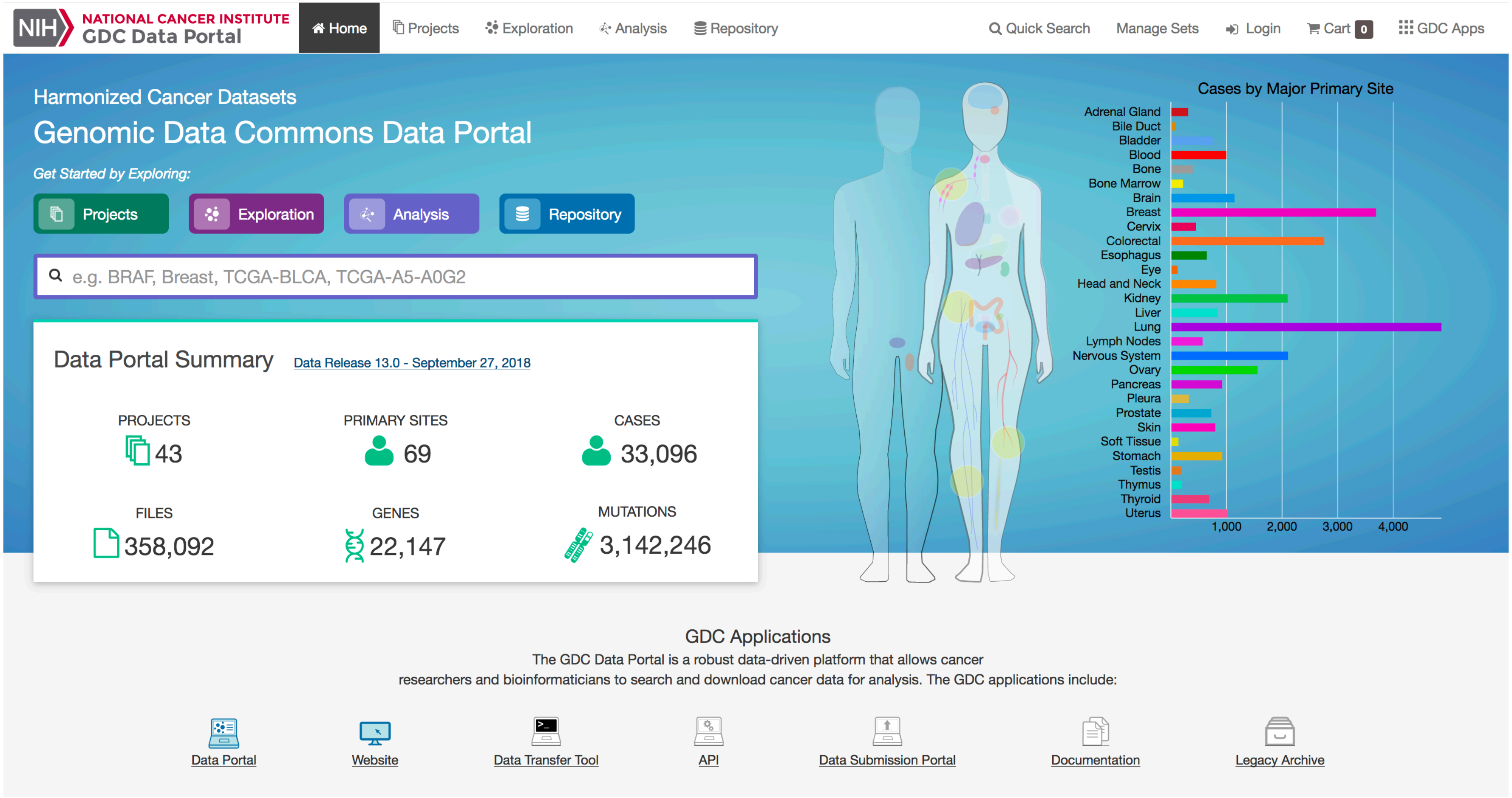
Summary

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



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

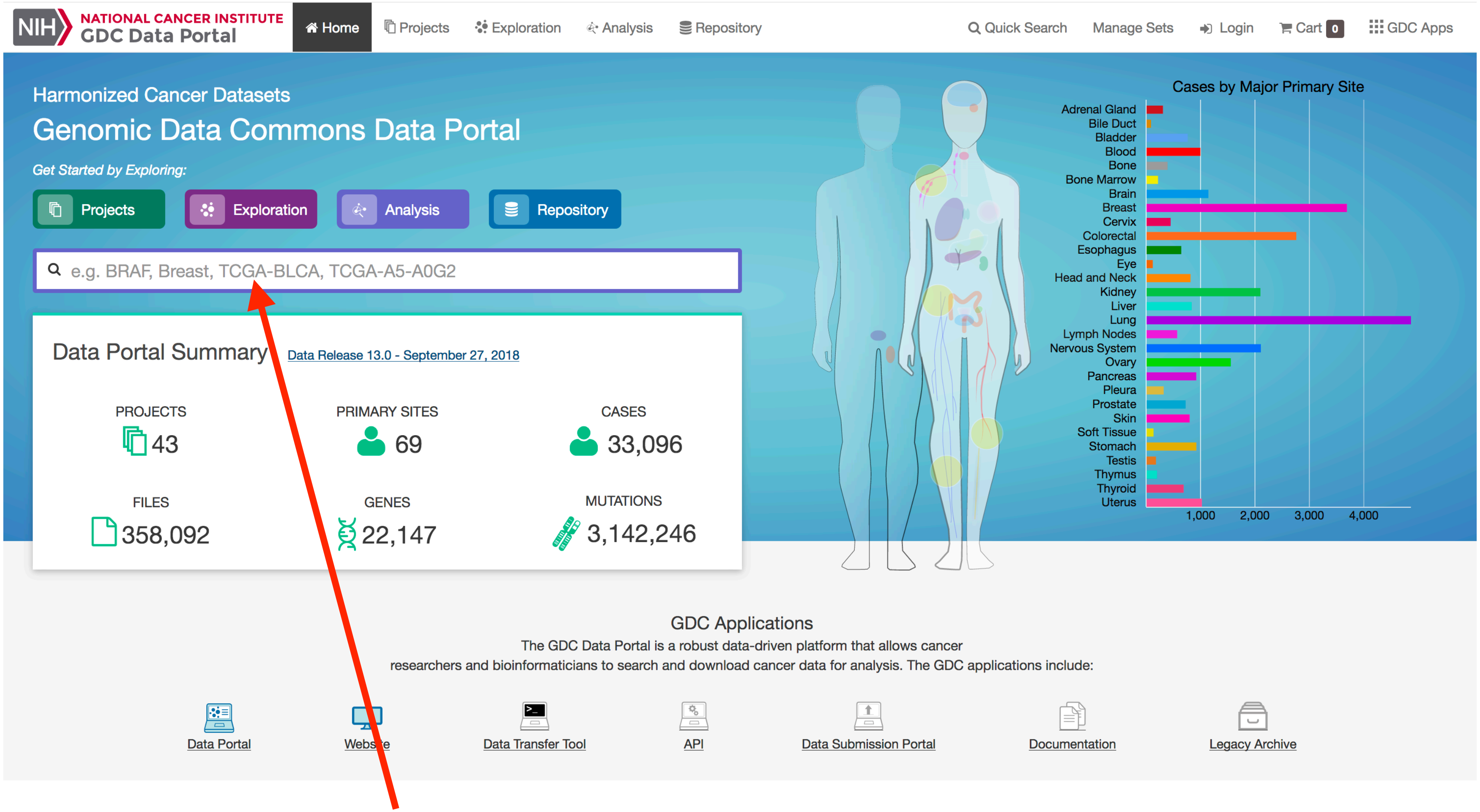
PROJECTS	PRIMARY SITES	CASES
43	69	33,096
FILES	**GENES**	**MUTATIONS**
358,092	22,147	3,142,246

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

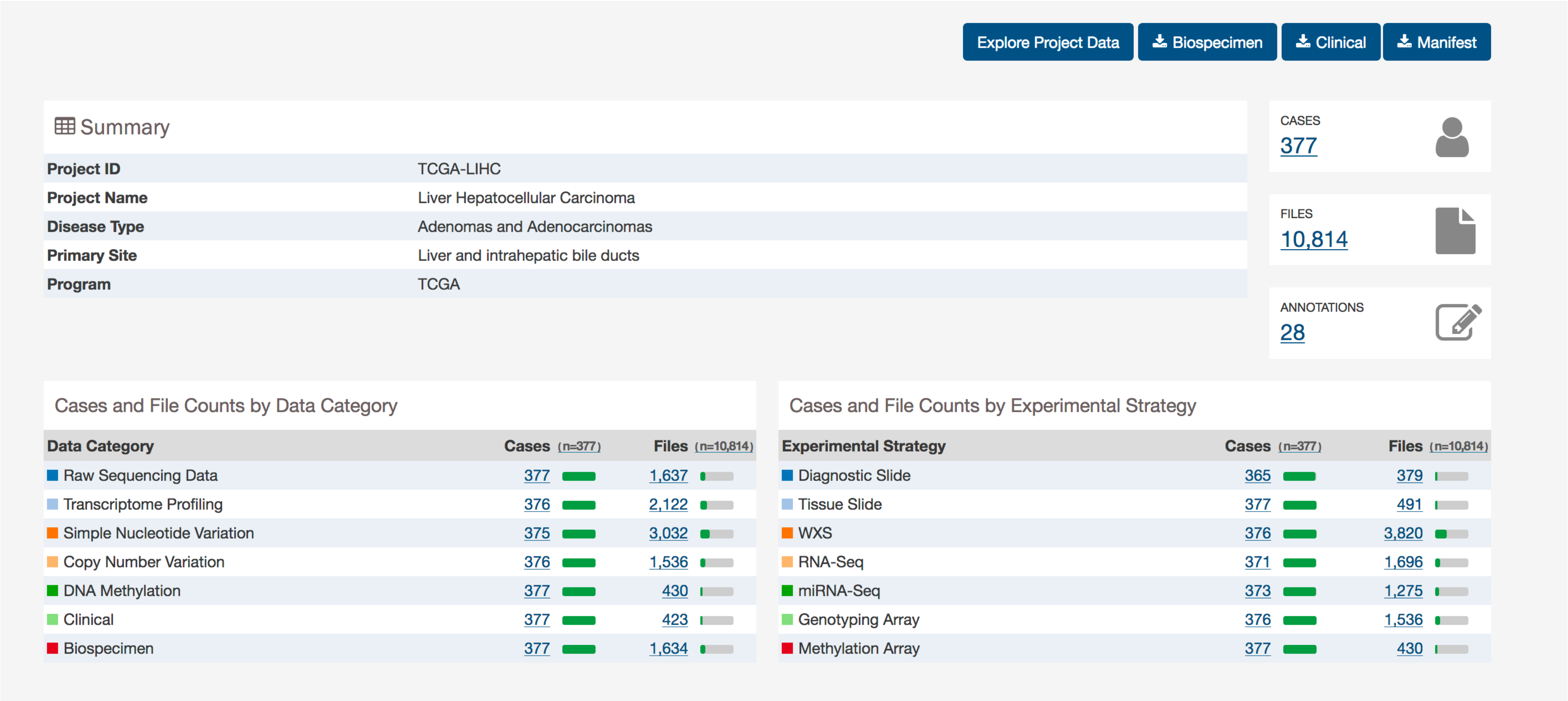
The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

Check a gene or cancer type!

I will try liver ….

Institute for Computational Genomics
RWTH AACHEN UNIVERSITY

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

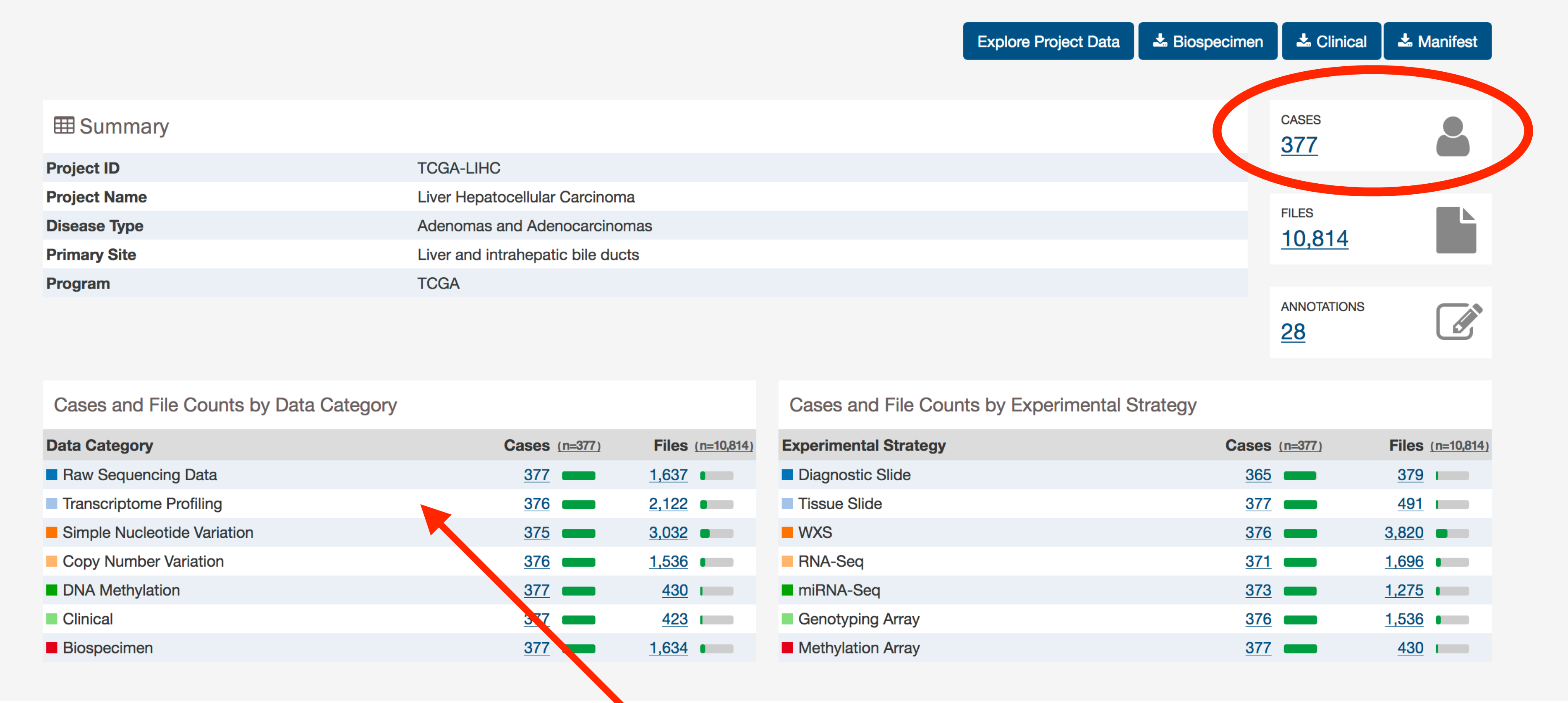
Cases and File Counts by Data Category

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



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

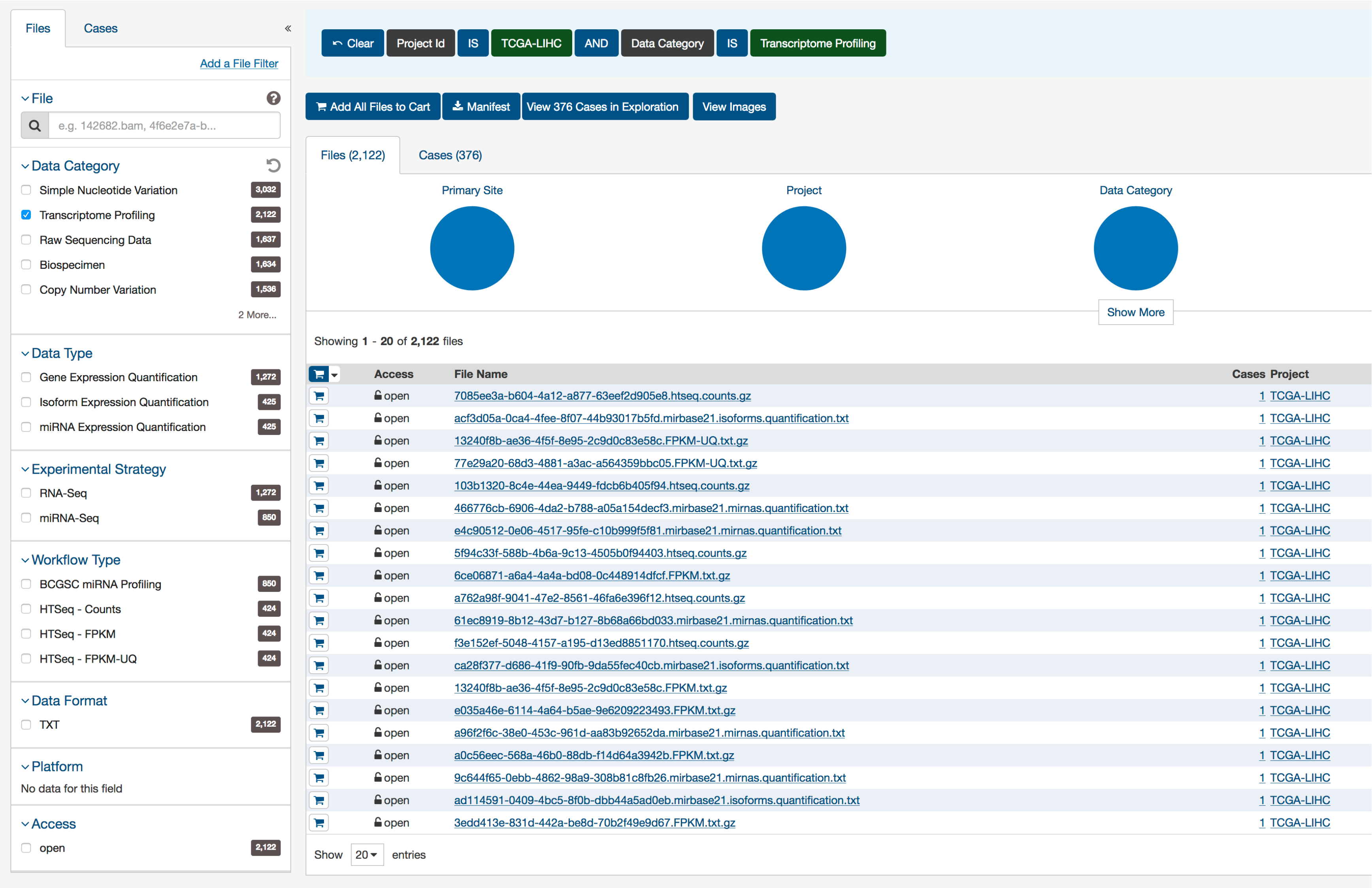
Data Category	Cases (n=377)	Files (n=10,814)
🟦 Raw Sequencing Data	377	1,637
🟦 Transcriptome Profiling	376	2,122
🟧 Simple Nucleotide Variation	375	3,032
🟧 Copy Number Variation	376	1,536
🟩 DNA Methylation	377	430
🟩 Clinical	377	423
🟥 Biospecimen	377	1,634

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!

LIHC - Liver Hepatocellular Carcinoma



File

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

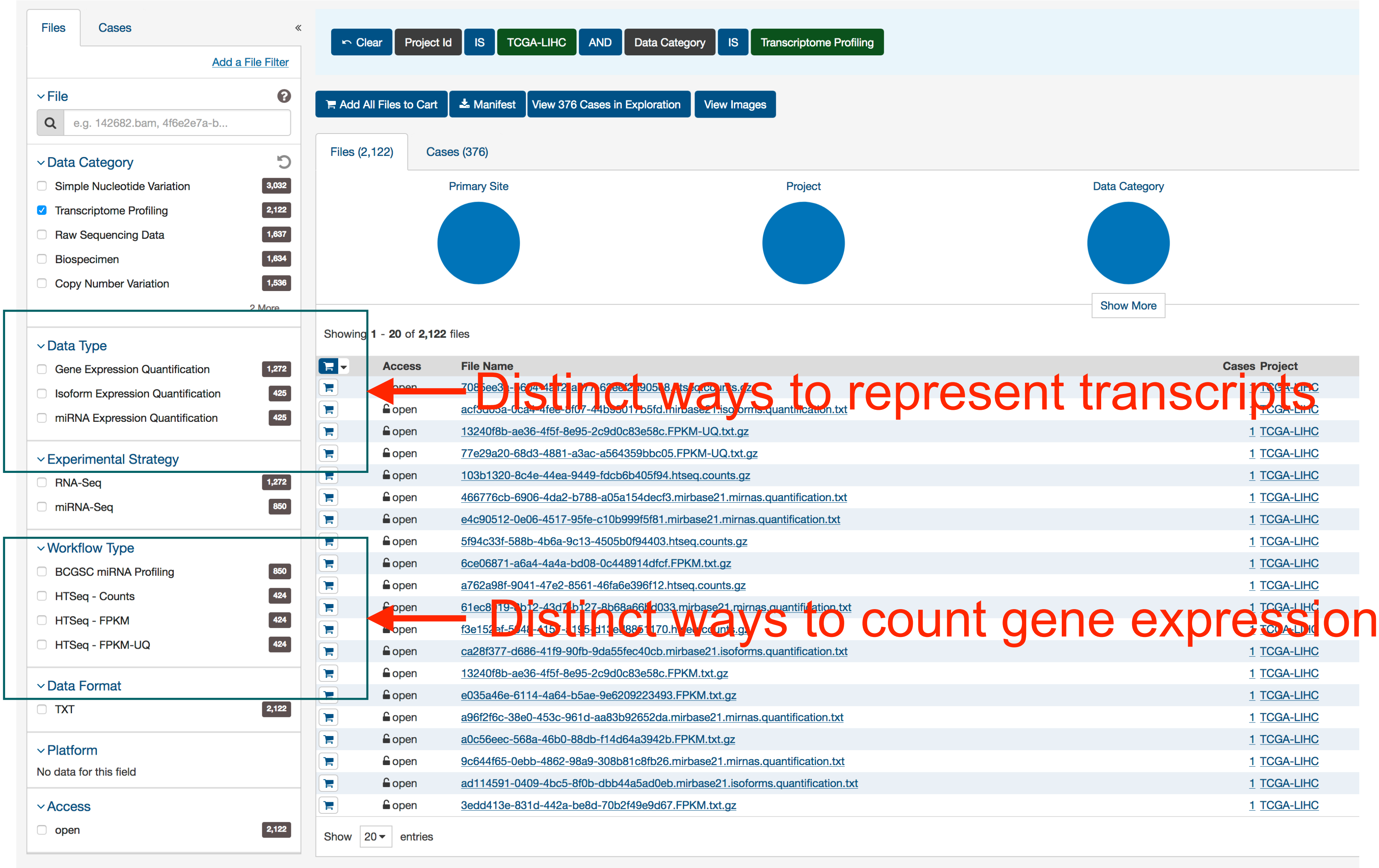
Clear | Project Id | IS | TCGA-LIHC | AND | Data Category | IS | Transcriptome Profiling

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-fdcb6b405f94.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

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)

Showing **1** - **20** of **2,122** files

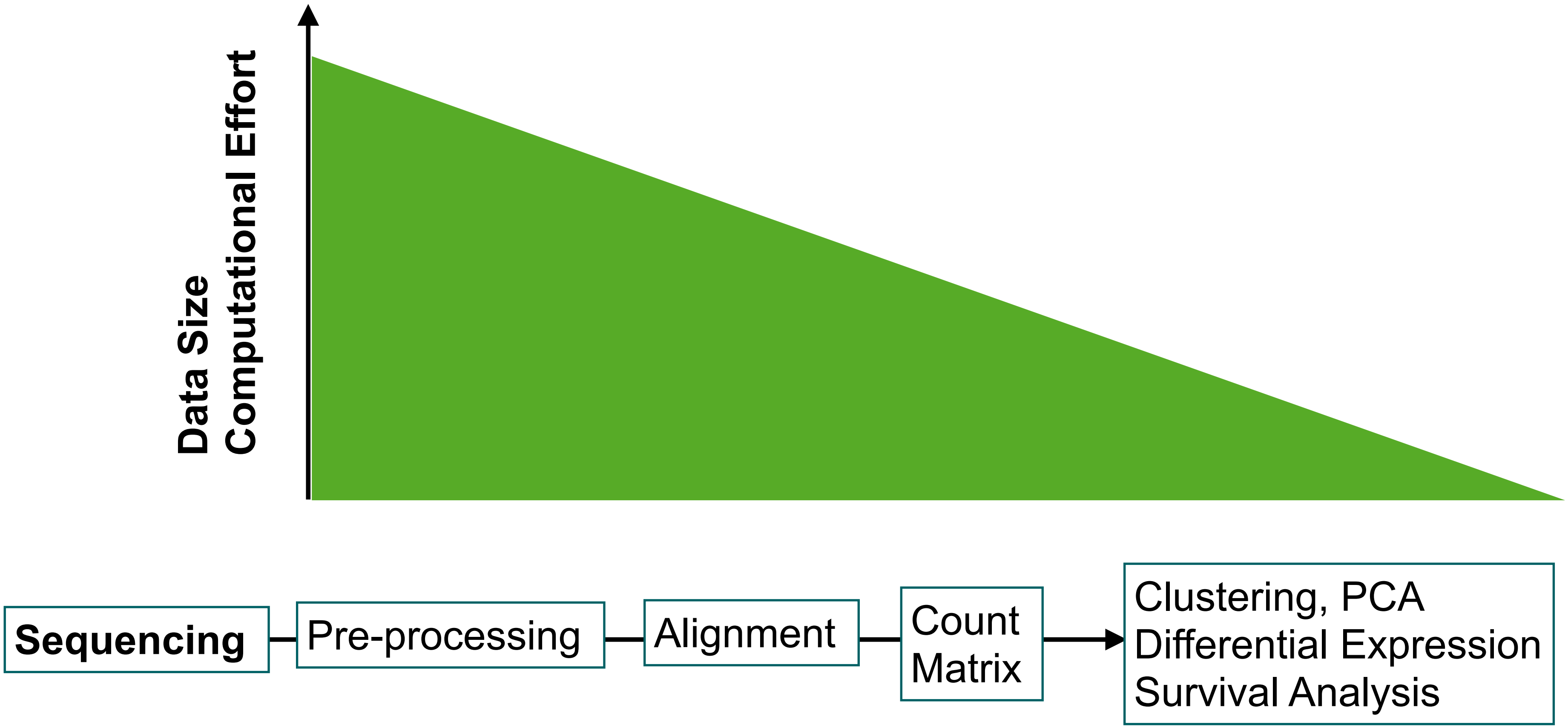
	Access	File Name	Cases	Project
	open	7085ee3[illegible].htseq.counts.gz	1	TCGA-LIHC
	open	acf3d05a-0ca4-4fee-8f07-44b93017b5fd.mirbase21.isoforms.quantification.txt	1	TCGA-LIHC
	open	13240f8b-ae36-4f5f-8e95-2c9d0c83e58c.FPKM-UQ.txt.gz	1	TCGA-LIHC
	open	77e29a20-68d3-4881-a3ac-a564359bbc05.FPKM-UQ.txt.gz	1	TCGA-LIHC
	open	103b1320-8c4e-44ea-9449-fdcb6b405f94.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	61ec8[illegible]19-[illegible]b12-43d7-b127-8b68a66[illegible]d033.mirbase21.mirnas.quantification.txt	1	TCGA-LIHC
	open	f3e152ef-5[illegible]48-415[illegible]-[illegible]195-d13e[illegible]8851170.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

Distinct ways to represent transcripts

Distinct ways to count gene expression.

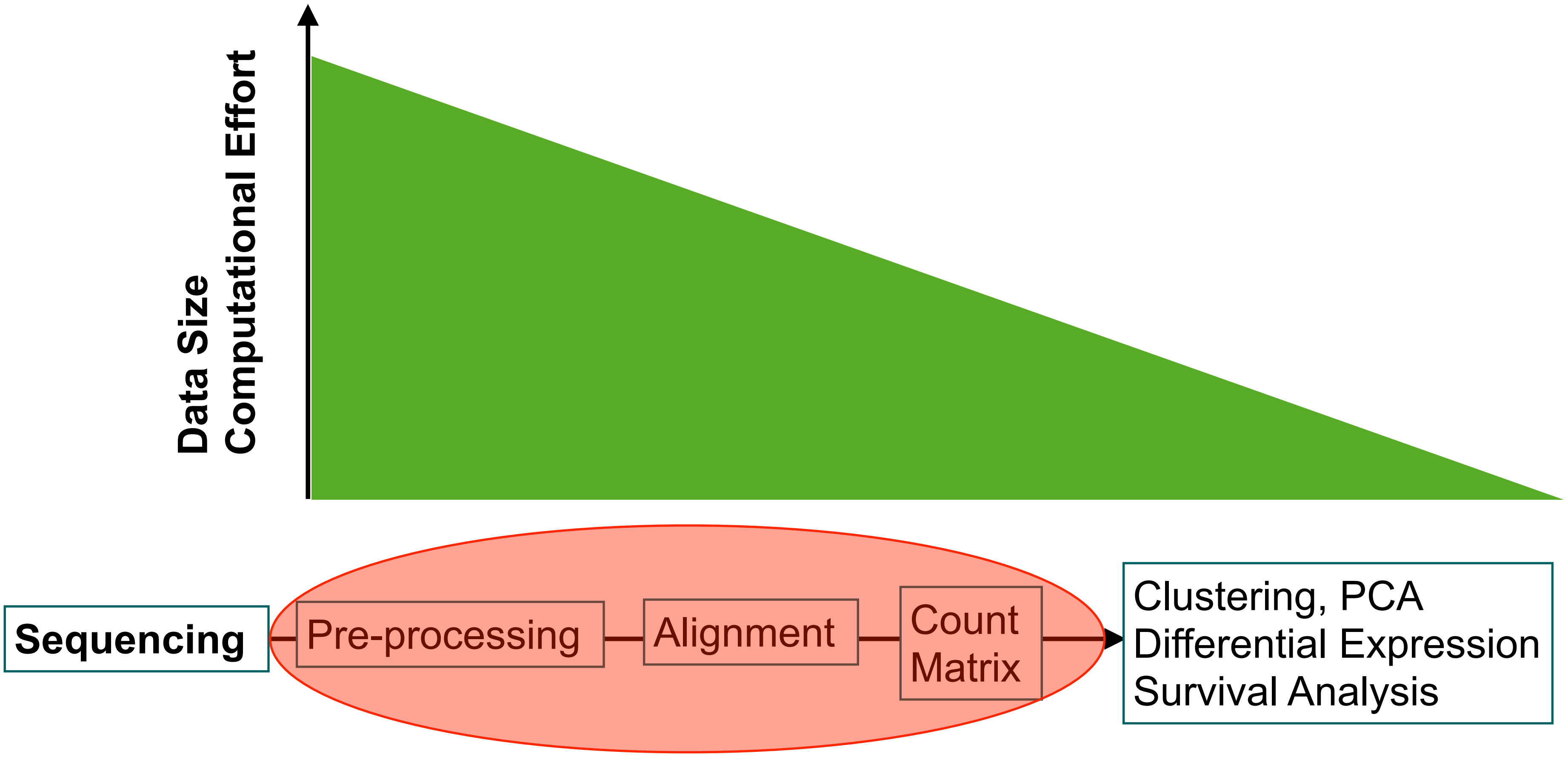
Bioinformatics Pipeline / RNA-seq



Data Size
Computational Effort

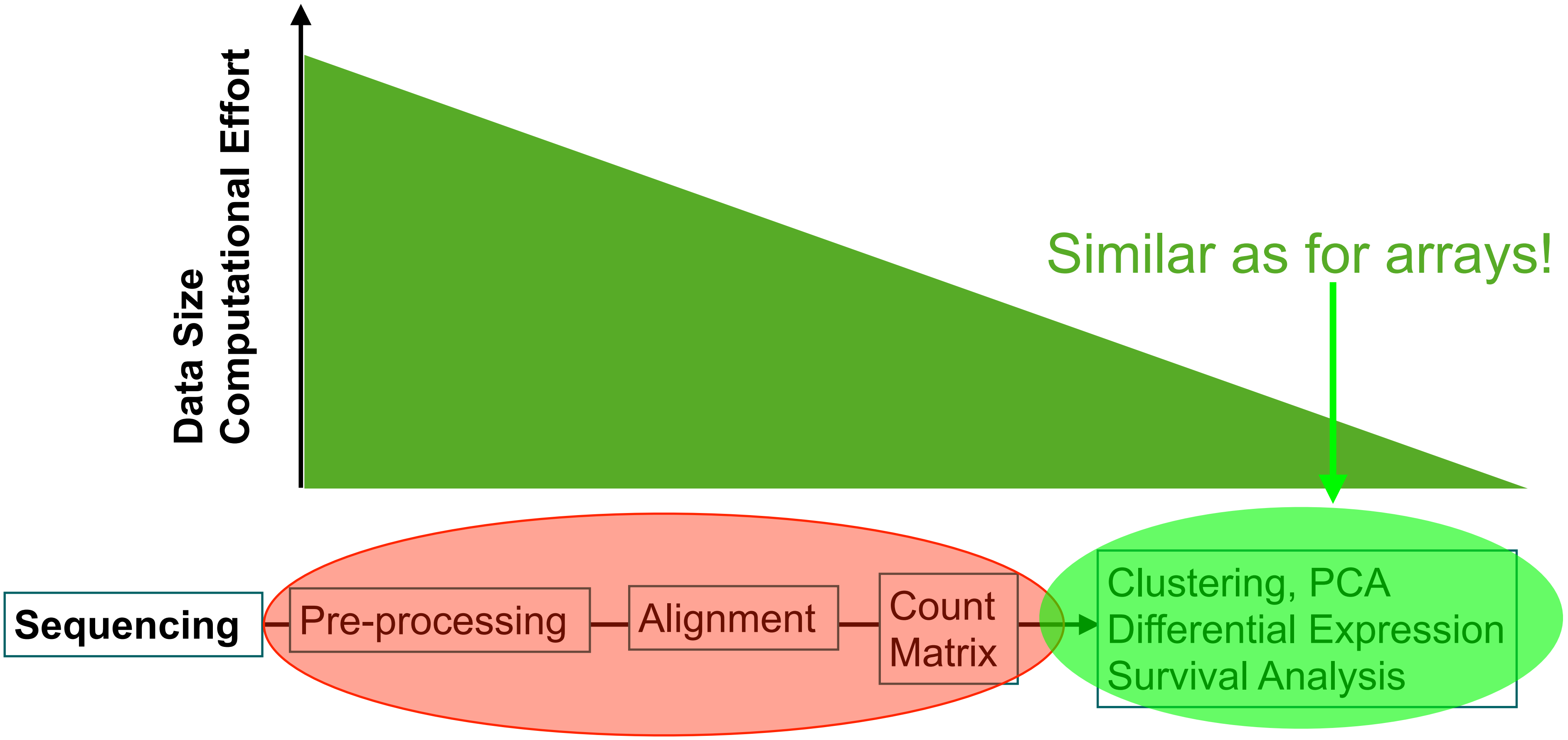
Sequencing → Pre-processing → Alignment → Count Matrix → Clustering, PCA
Differential Expression
Survival Analysis

Bioinformatics Pipeline / RNA-seq



More details on next class!
Practical part not covered!

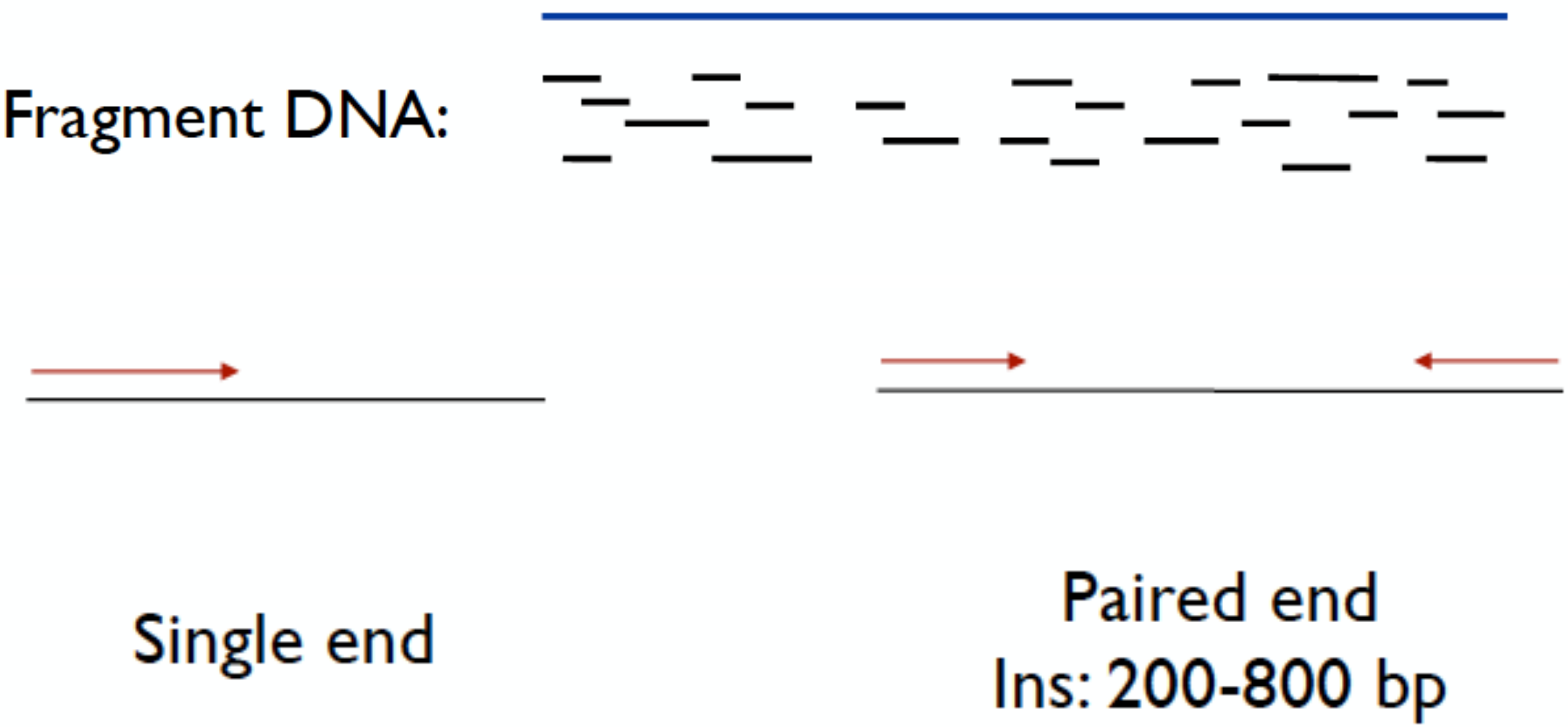
Bioinformatics Pipeline / RNA-seq



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

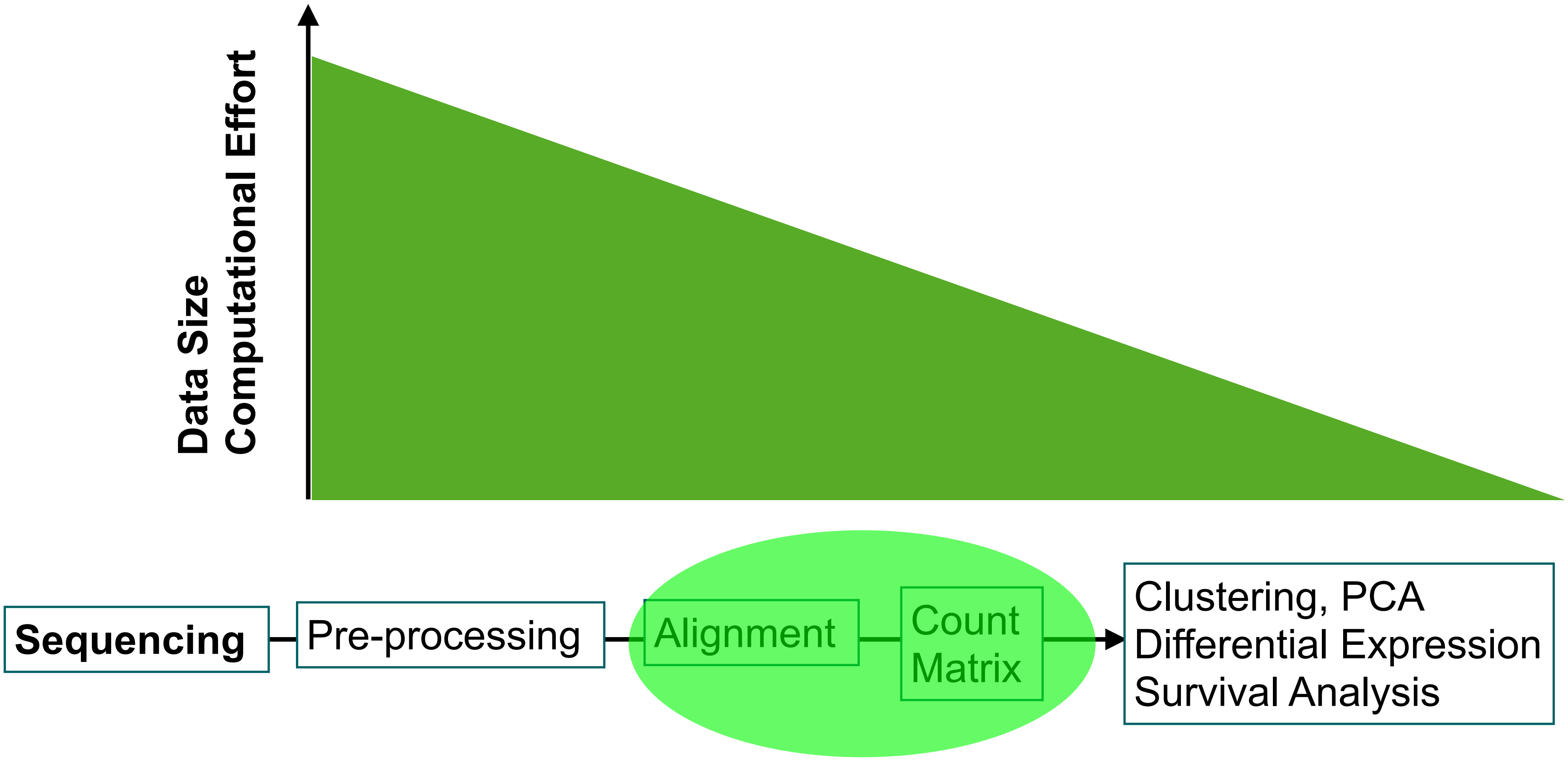


Fragment DNA:

Single end

Paired end
Ins: 200-800 bp

Bioinformatics Pipeline / RNA-seq

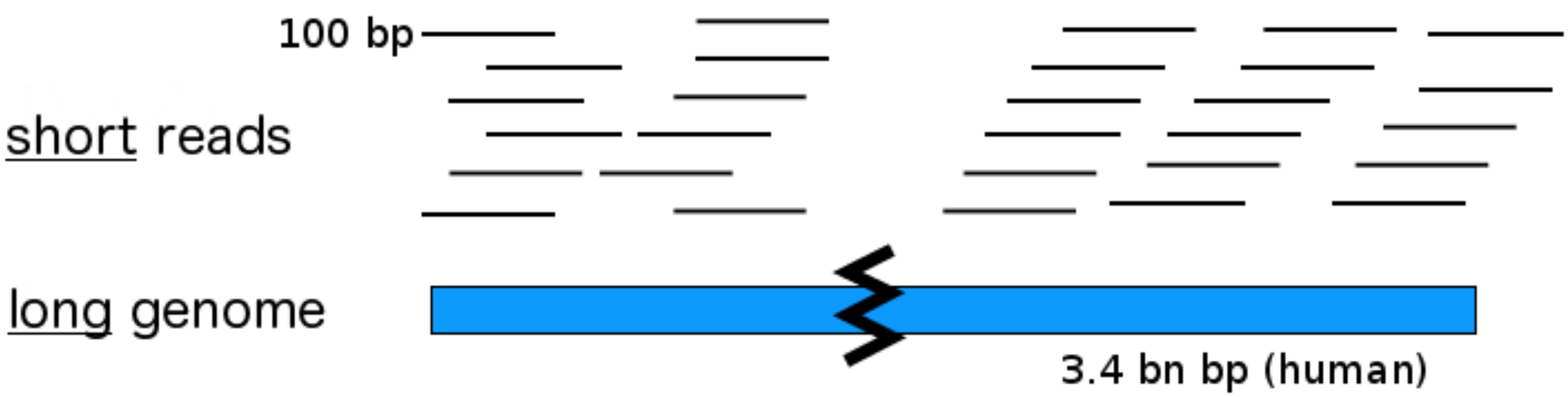


Data Size
Computational Effort

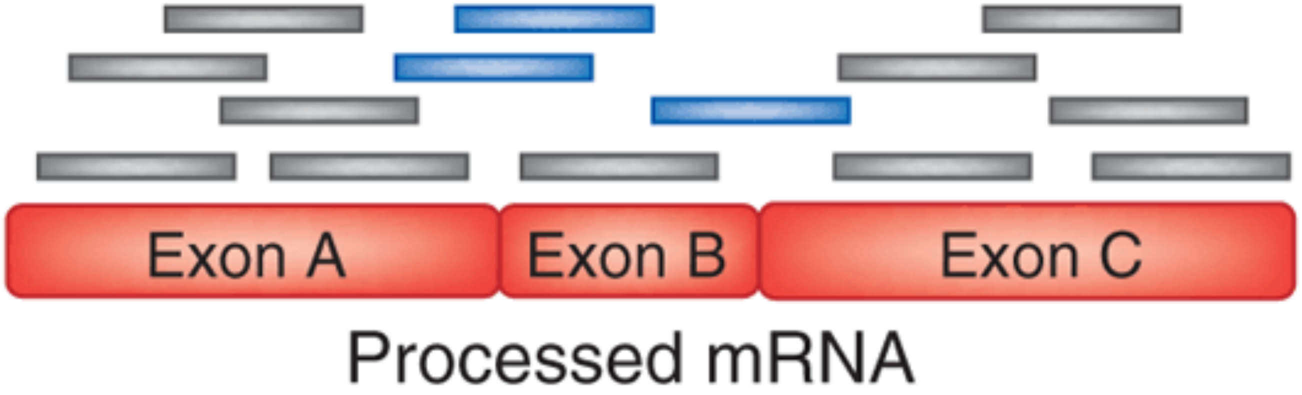
Sequencing → Pre-processing → Alignment → Count Matrix → Clustering, PCA
Differential Expression
Survival Analysis

Alignment

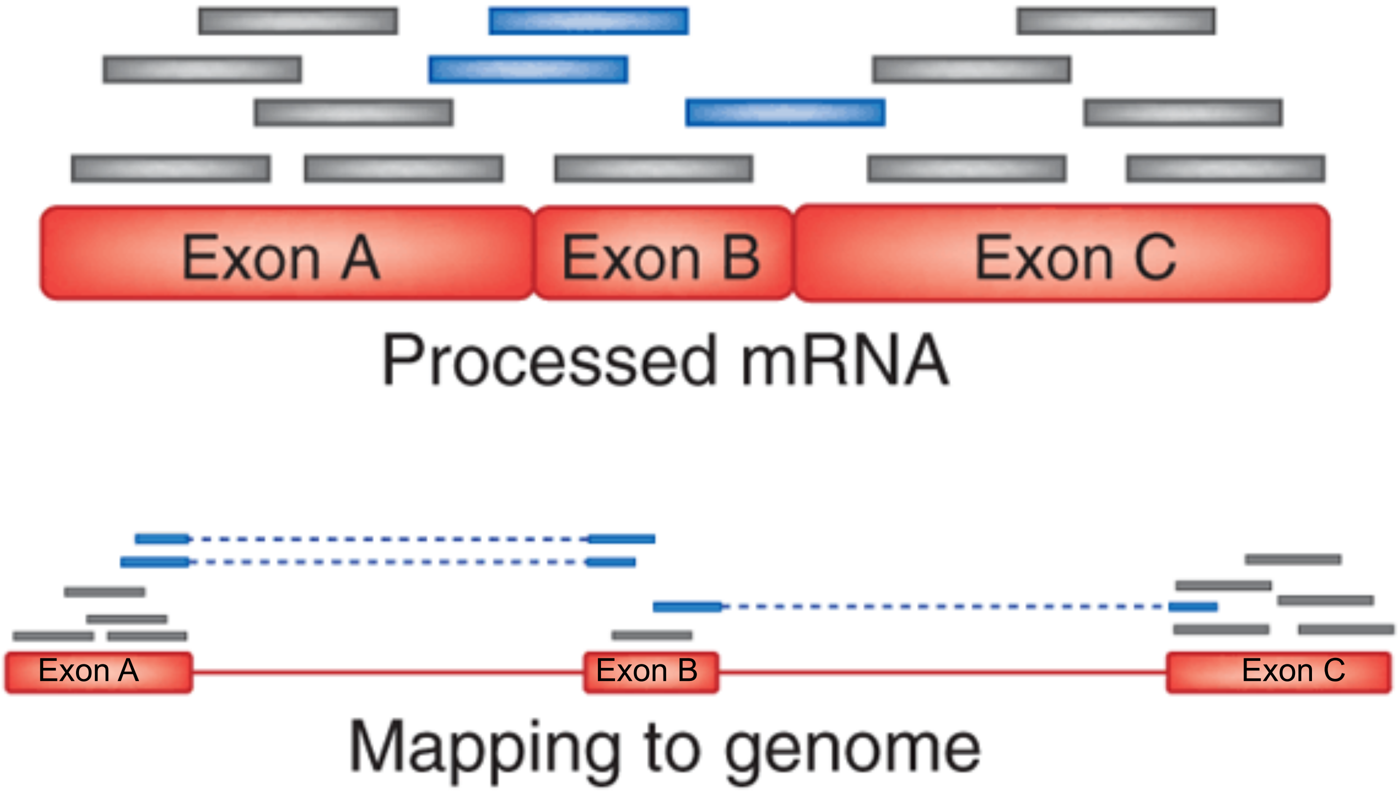
- a large reference sequence is given (genome)
  - up to billions of base pairs
- short reads (<200bps)
- find most probable position of the read in the genome (by inexact string matching)



Alignment - Split Read Mapping (RNA-Seq)



Alignment - Split Read Mapping (RNA-Seq)



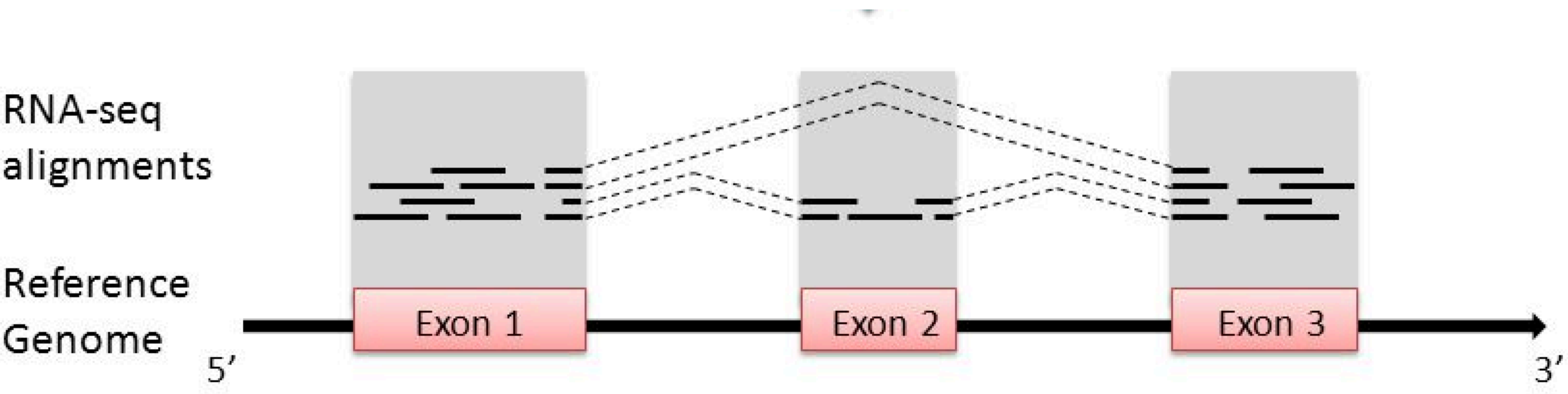
- reads are split within intros when mapped to genome
- aligners use transcript information or try to find splice events (STAR & TOPHAT)

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

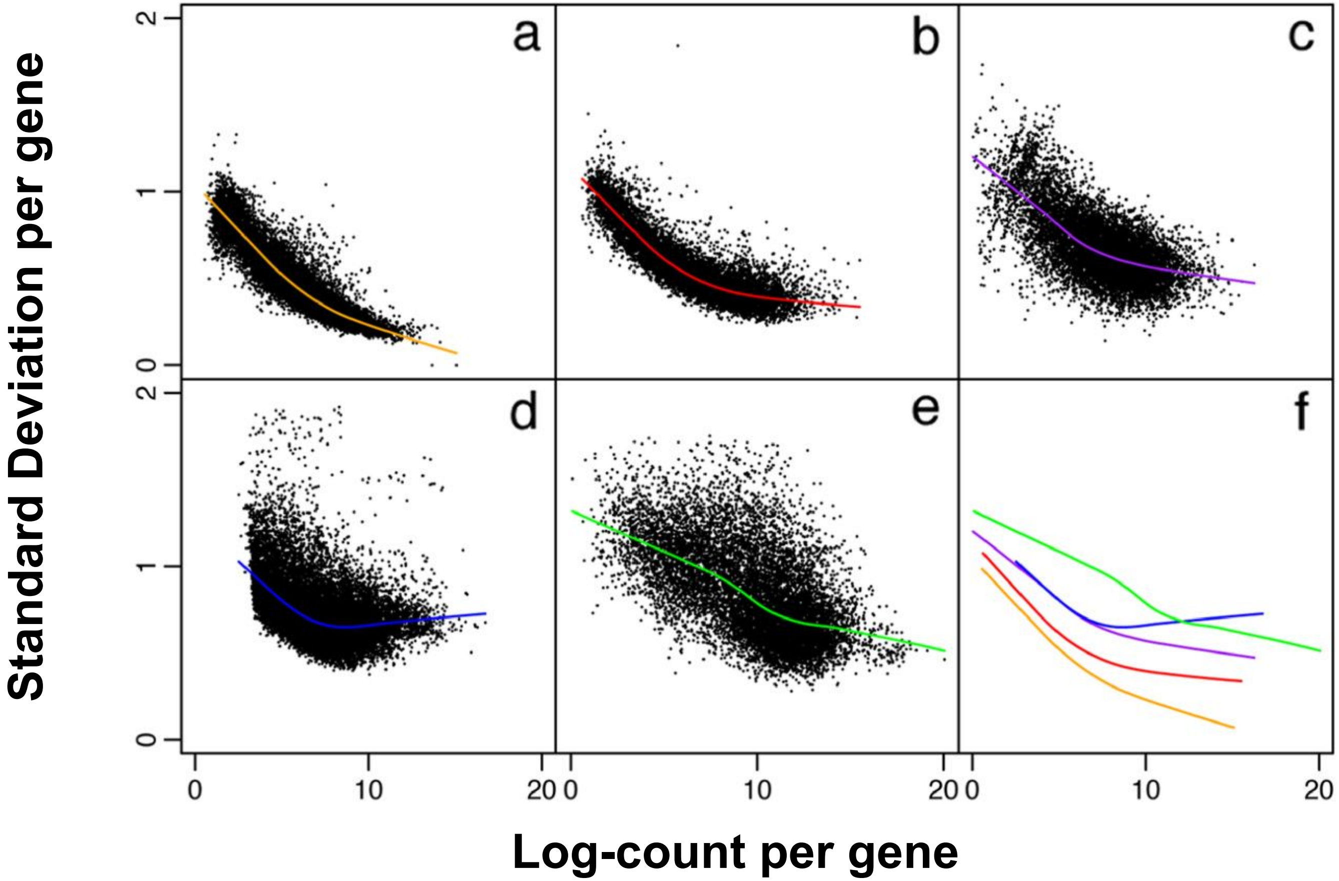
Gene Level - 17 reads
Exon level - exon 1 (8 reads), exon 2 (3 reads), exon 3 (6 reads)
Transcript Level - Exons 1,2 & 3 (10 reads) and exon 1 & 3 (7 reads) *
* complex computational methods required (RSe, or TopHAT needed for this)

Fragments per Kilobase (FPKM)

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

RNA-seq and Differential Analysis

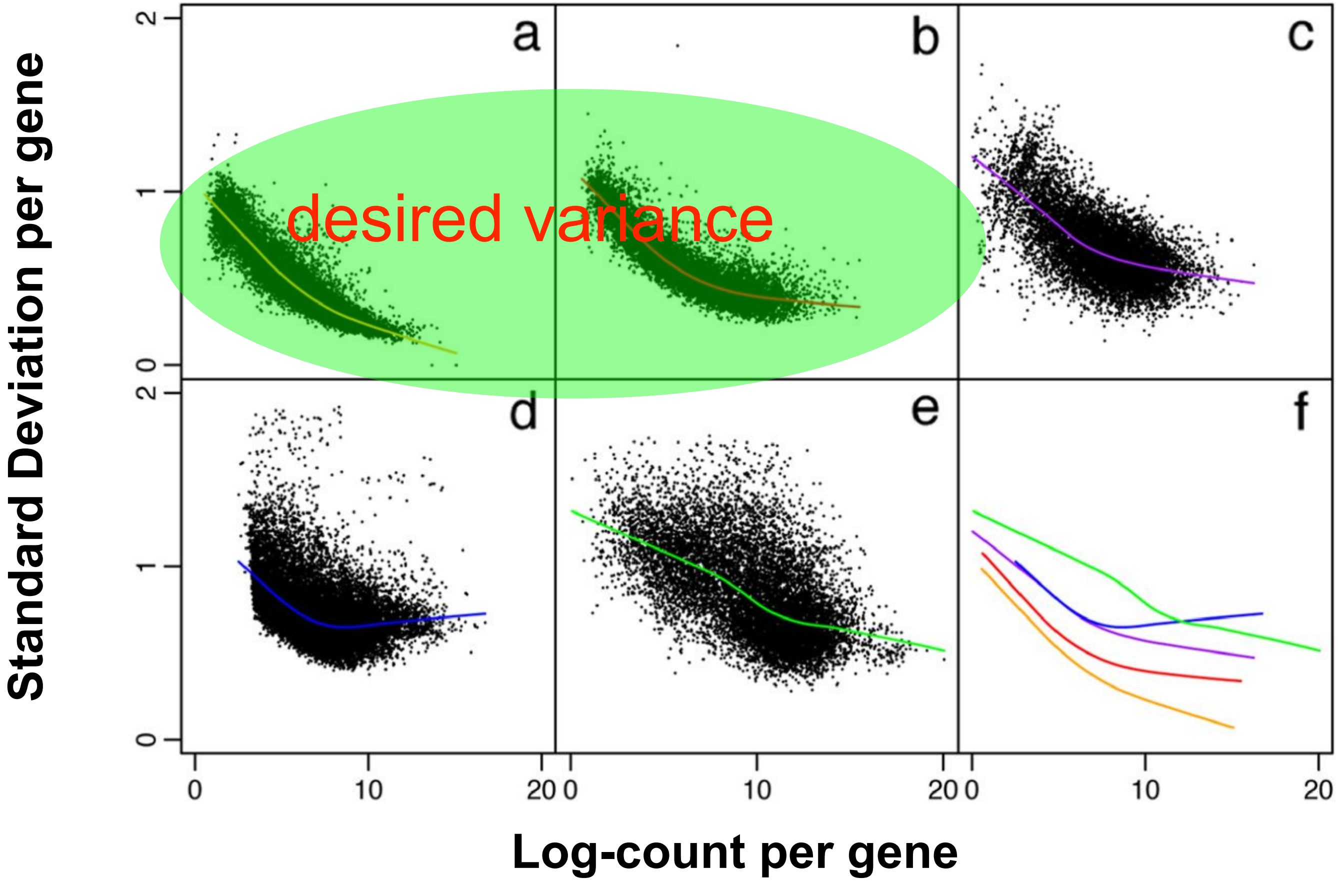
Arrays and RNA-seq have distinct distributions



VOOM analysis is necessary to make variance similar to arrays.

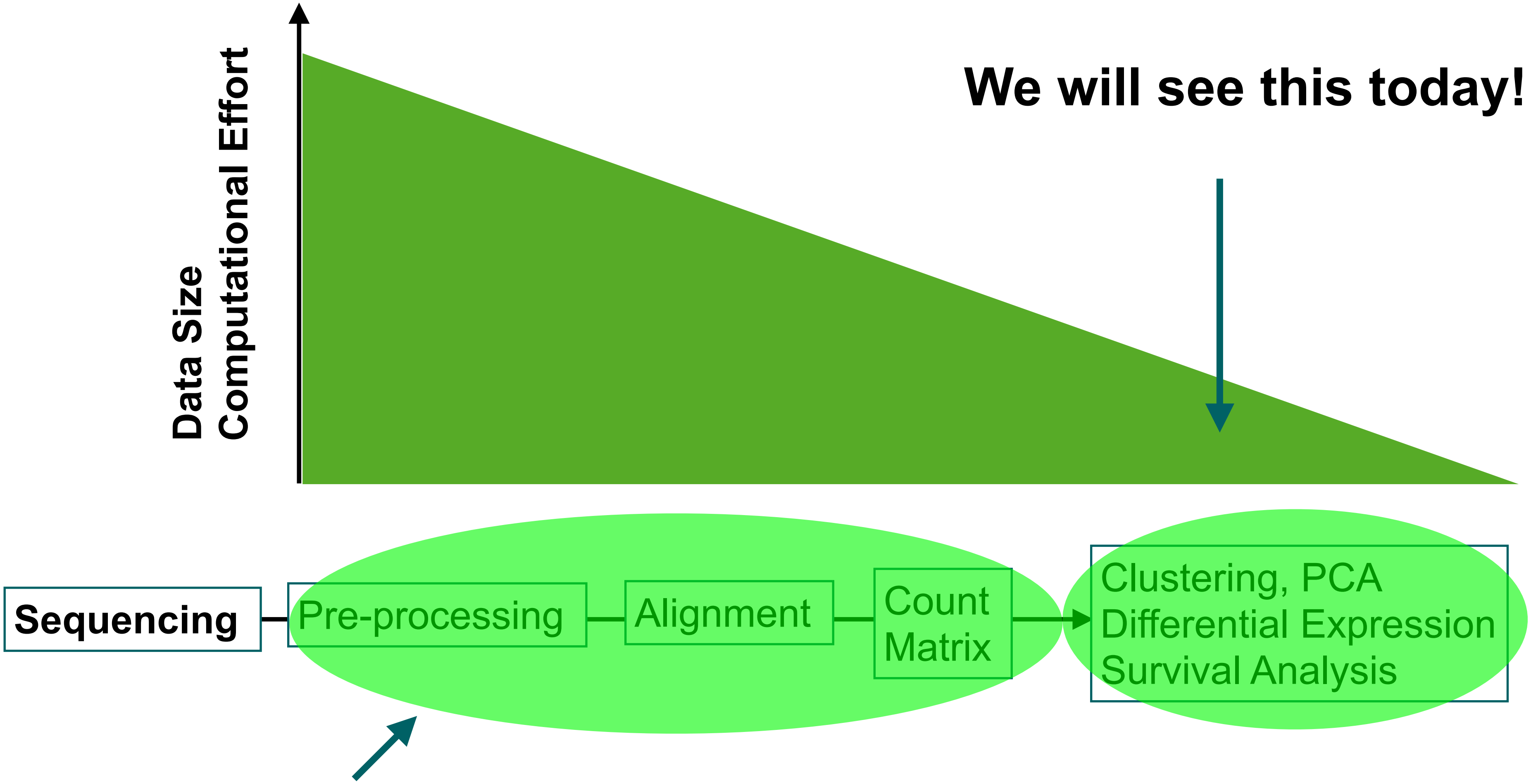
RNA-seq and Differential Analysis

Arrays and RNA-seq have distinct distributions



VOOM analysis is necessary to make variance similar to arrays.

Bioinformatics Pipeline / RNA-seq



Provided by TGCA or your Core Facility!

Hands on!

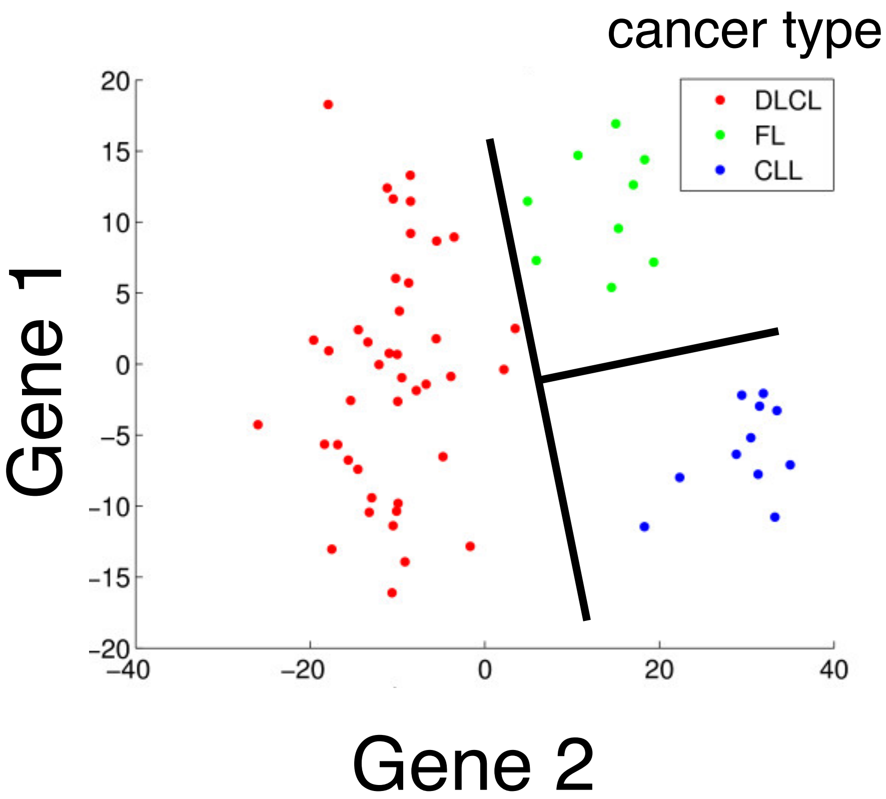
Personalized Medicine

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

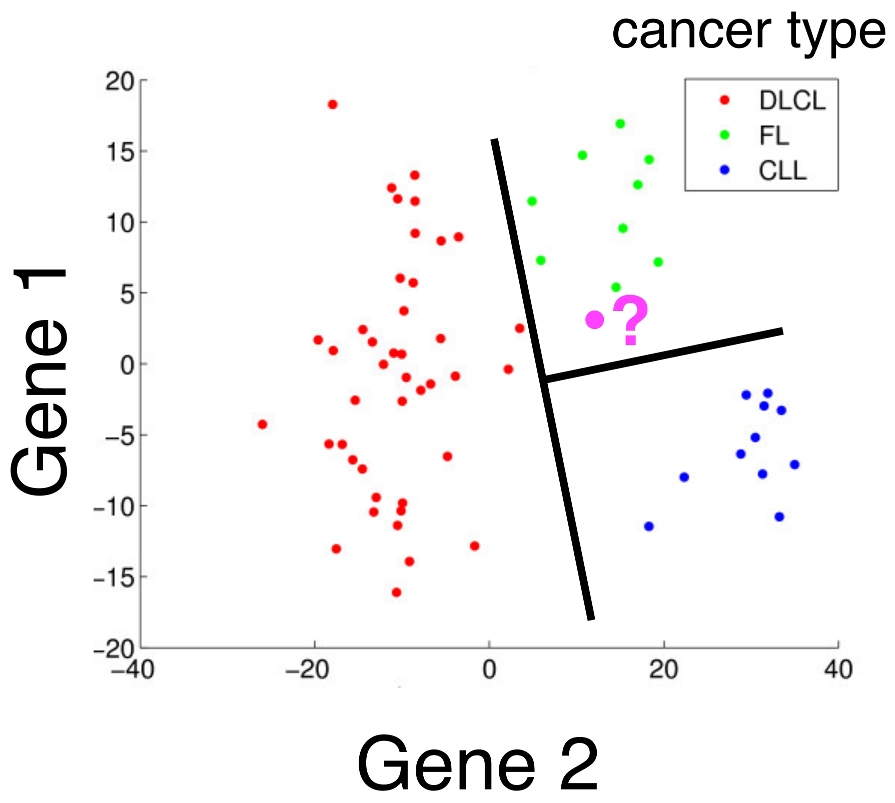
Expression matrix X (genes vs samples)

classification vector *Y* (diagnosis)

Find a function:

f(*x*) → *y*

Machine Learning - Classifier



Data

Expression matrix X (genes vs samples)

classification vector *Y* (diagnosis)

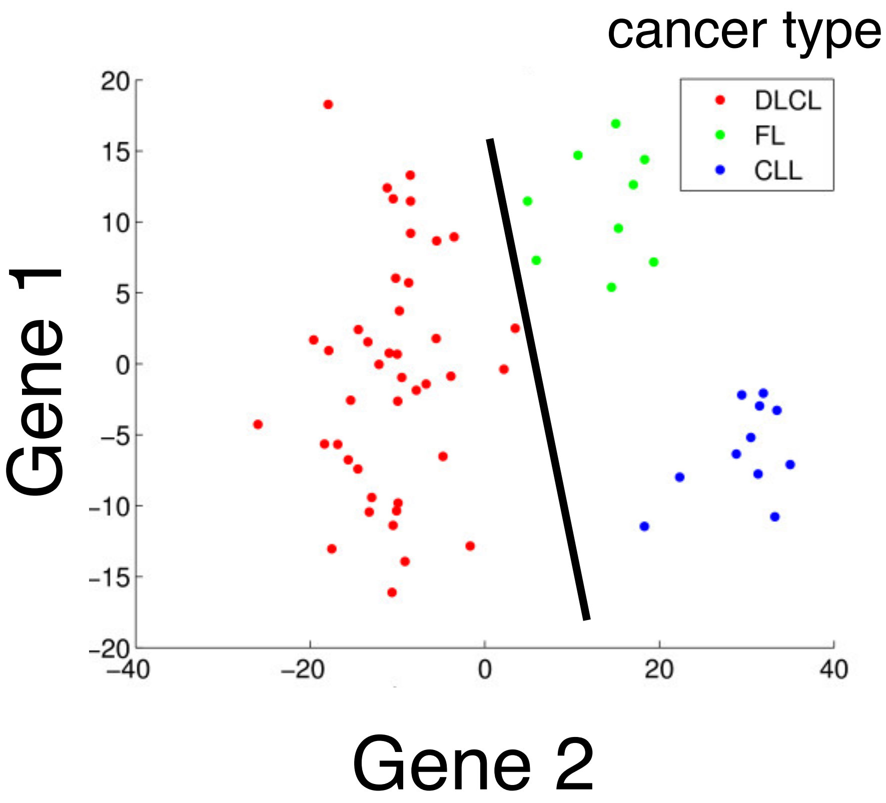
Find a function:

f(*x*) → *y*

For new patients X':

f(*x*') → *y*'

Linear Classifier



Linear Function:

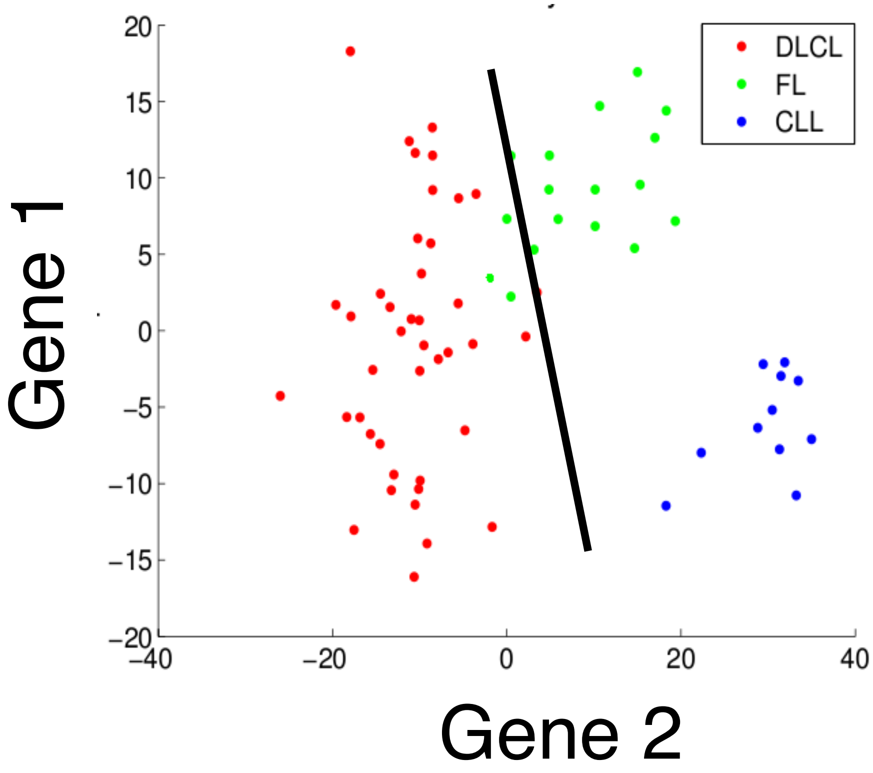
$$f\,(x\,,A) = a_0 + a_1x_1 + \ldots + a_Lx_L$$

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

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

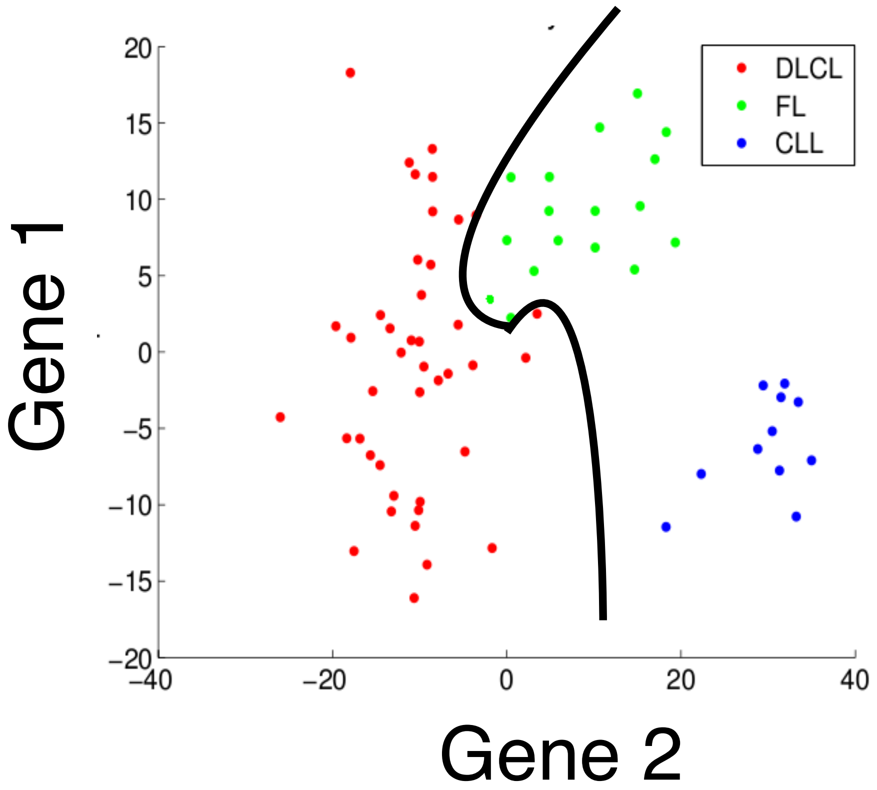
- Works for 2 classes only
  - Train a function for each cancer type
- Find coefficients *A*
  - estimated with neural networks or support vector machines

Linear Classifier - Problems



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

Nonlinear Classifier - Problems

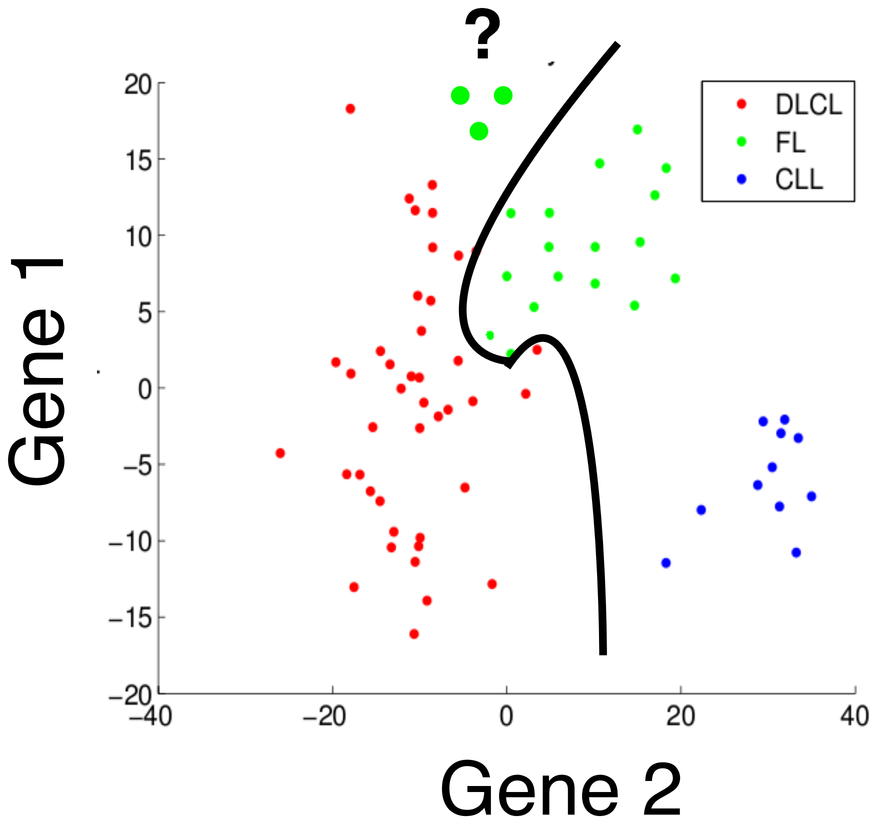


- Polinomial Function
- $f\ (x\ ,\ A) = a_0 + a_{11}x^3{}_1 + \ldots + a_{L1}x^3{}_L$

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

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

Nonlinear Classifier - Problems

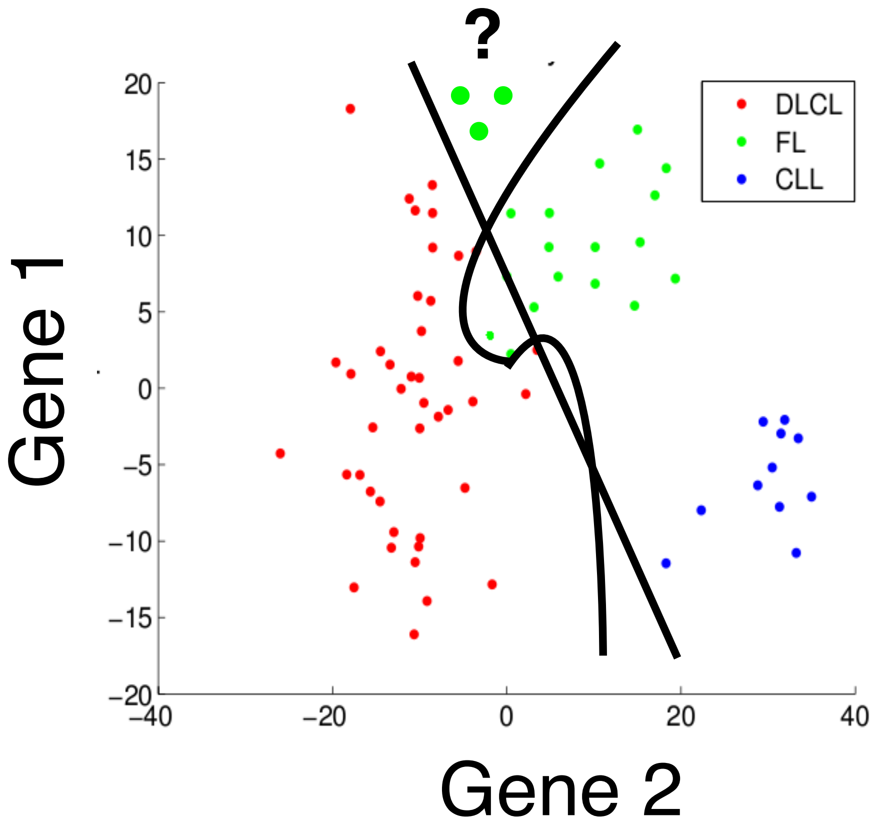


- Polinomial Function
- $f\,(x\,,\,A) = a_0 + a_{11}x^3{}_1 + \ldots + a_{L1}x^3{}_L$

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

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

Nonlinear Classifier - Problems



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

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

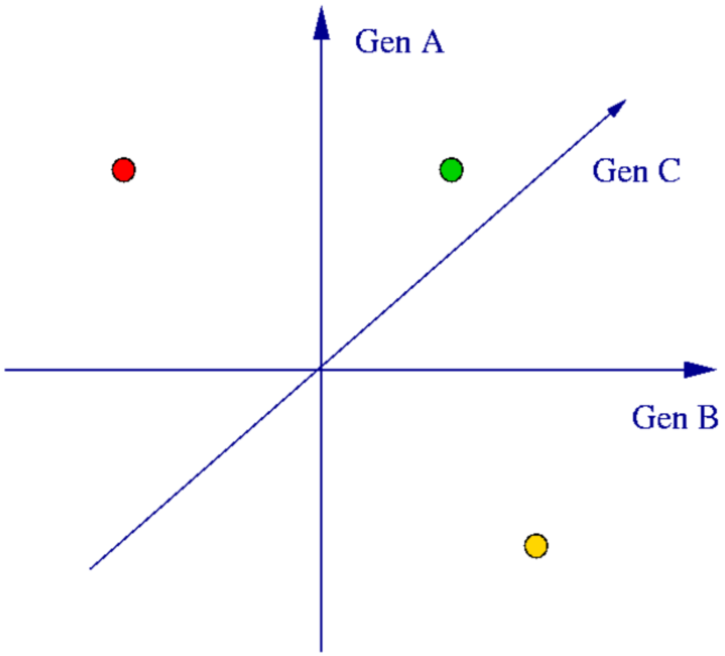
  $a_{12}x_1 + \ldots + a_{L2}x_L$
- Third order polynomial
- Problem: overfitting

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



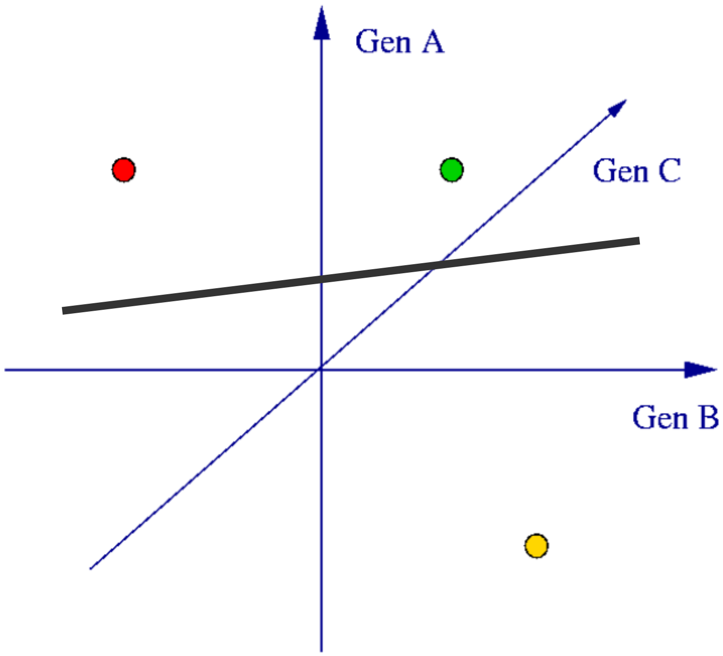
Sparse data

- three genes

- 2 patients with known cancer (red vs yellow)

- 1 unknown (green)

Curse of Dimensionality : Example



- Sparse data

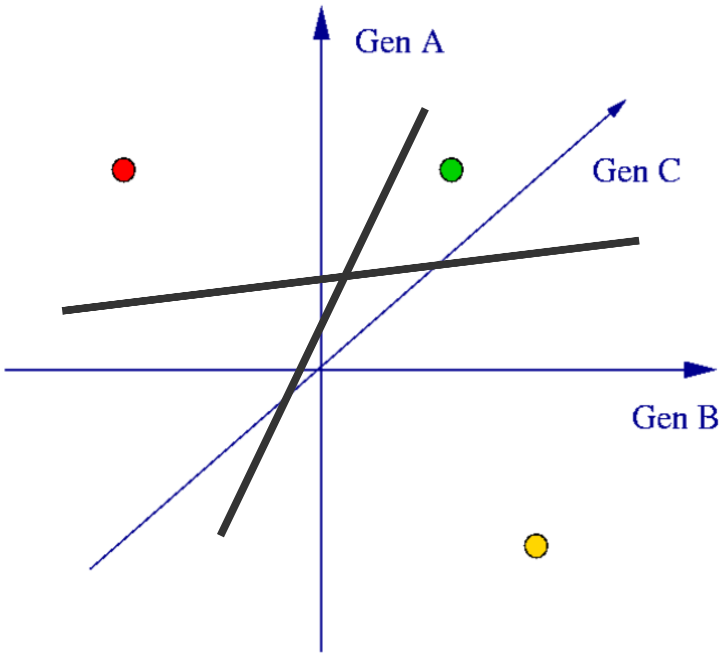
  - three genes

  - 2 patients with known cancer (red vs yellow)

  - 1 unknown (green)

Perfect classifier (on training)

Curse of Dimensionality : Example



- Sparse data

  - three genes

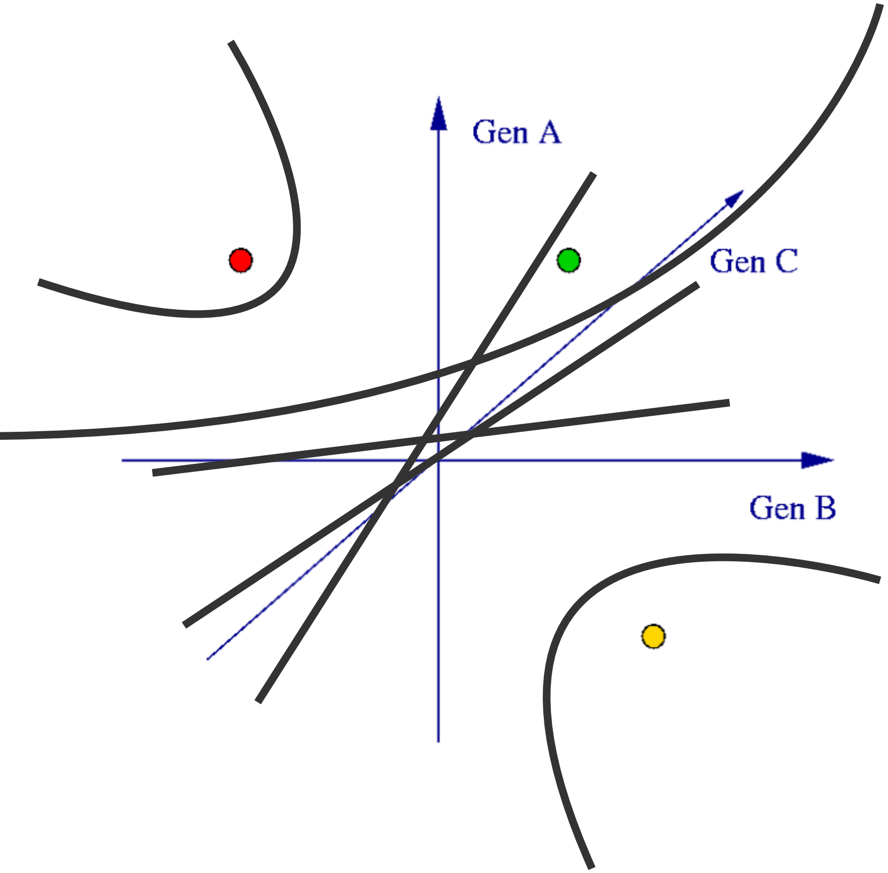
  - 2 patients with known cancer (red vs yellow)

  - 1 unknown (green)

Both are perfect classifiers (on training)

Hard to generalise!

Curse of Dimensionality : Example



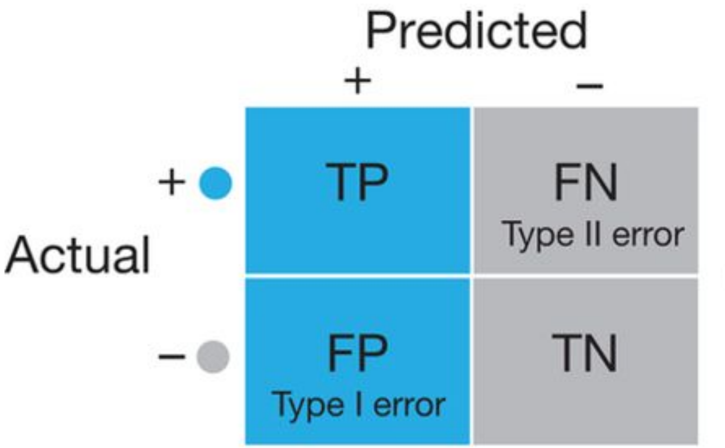
- There are millions of perfect linear classifiers
- And even more non-linear classifiers!

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 -)

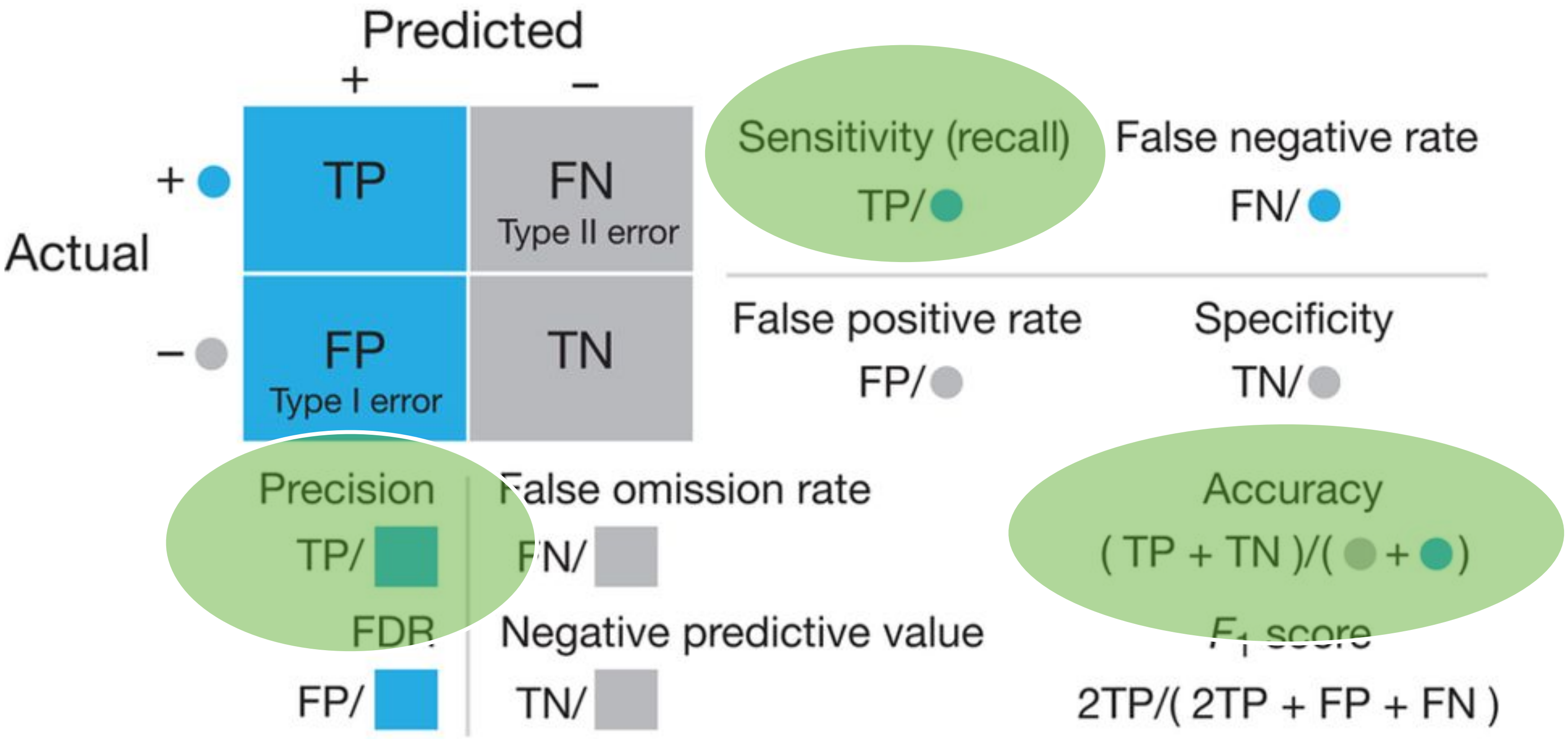


		Predicted	
		+	–
Actual	+ ●	TP	FN Type II error
	– ●	FP Type I error	TN

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

Classifier Evaluation

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



		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/⬜

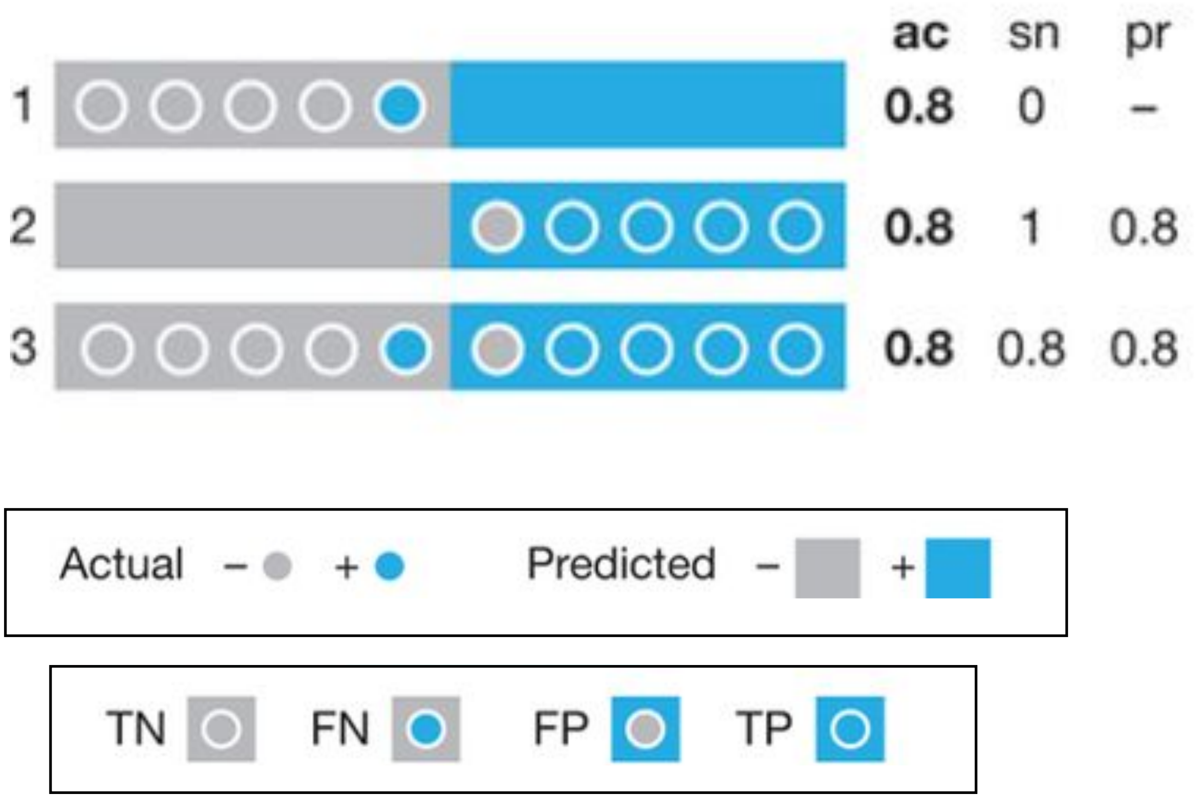
Accuracy
(TP + TN)/(⚪ + 🔵)

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

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

Classifier Evaluation

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

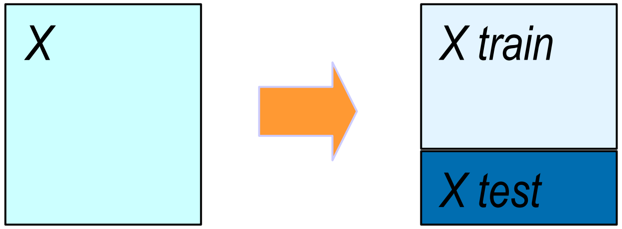


	ac	sn	pr
1	0.8	0	–
2	0.8	1	0.8
3	0.8	0.8	0.8

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

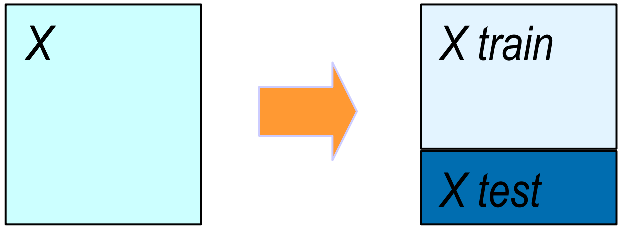
Classifier Evaluation

- 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



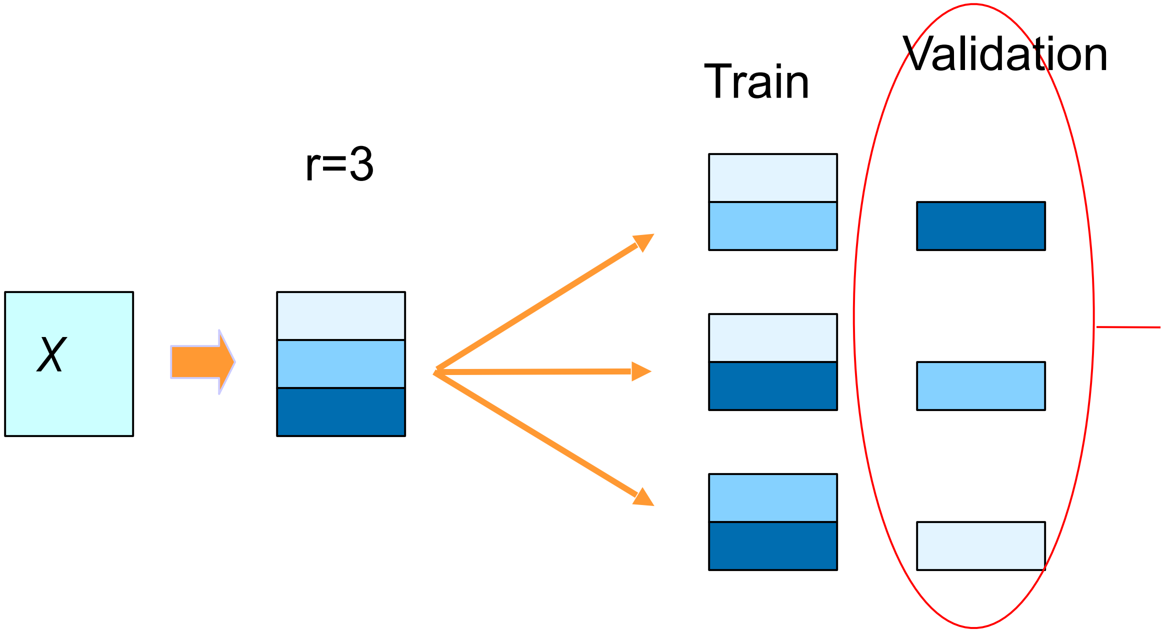
Classifier Evaluation

- 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



mean and std of accuracy, precision and sensitivity

Elastic Net

Is based on a linear function:

$f\,(x\,,A) = a_0{+}a_1x_1{+}...{+}a_Lx_L$

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

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

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.

3. Check if group of patients are associated to survival, tumour grade or any other clinical variable? You can use the **table** function for some of these analysis. (next week!)

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

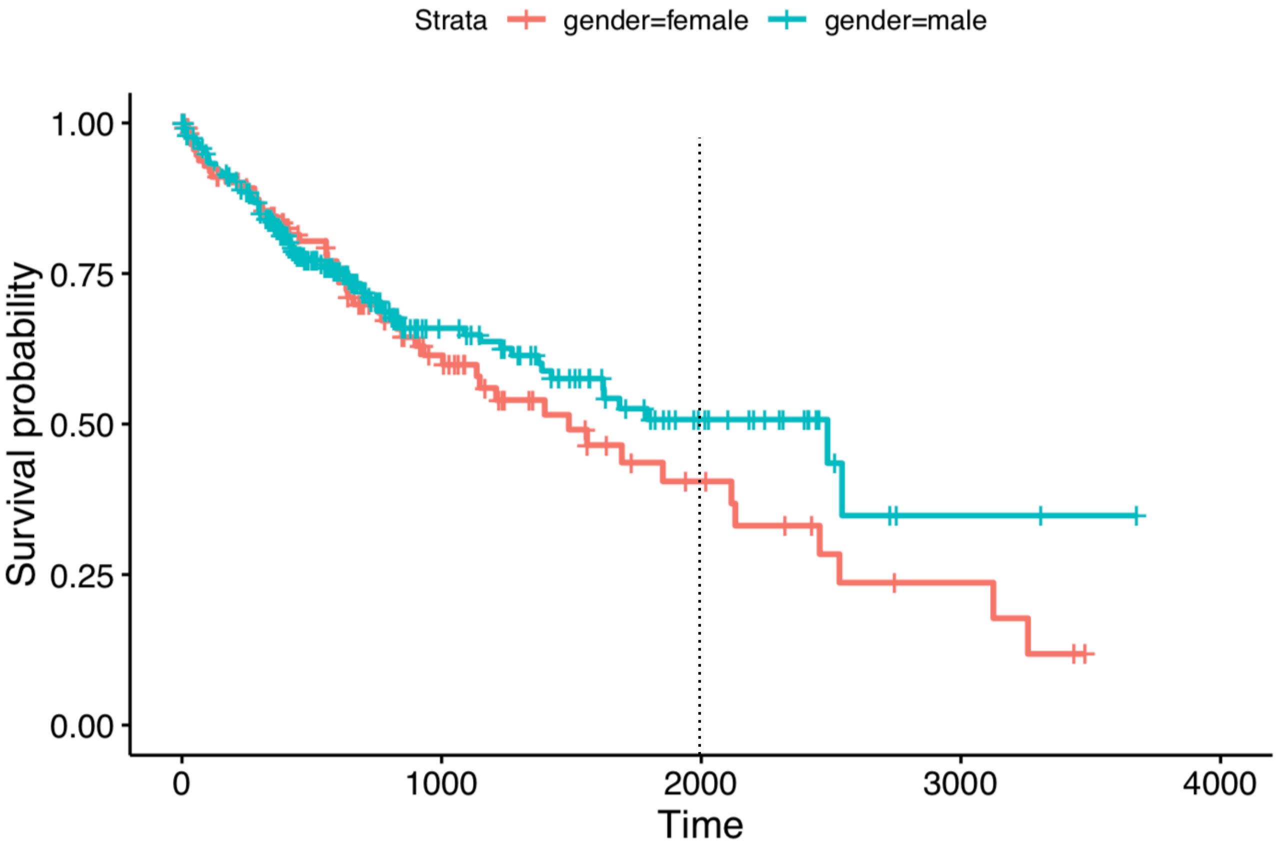
Data:

- **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*

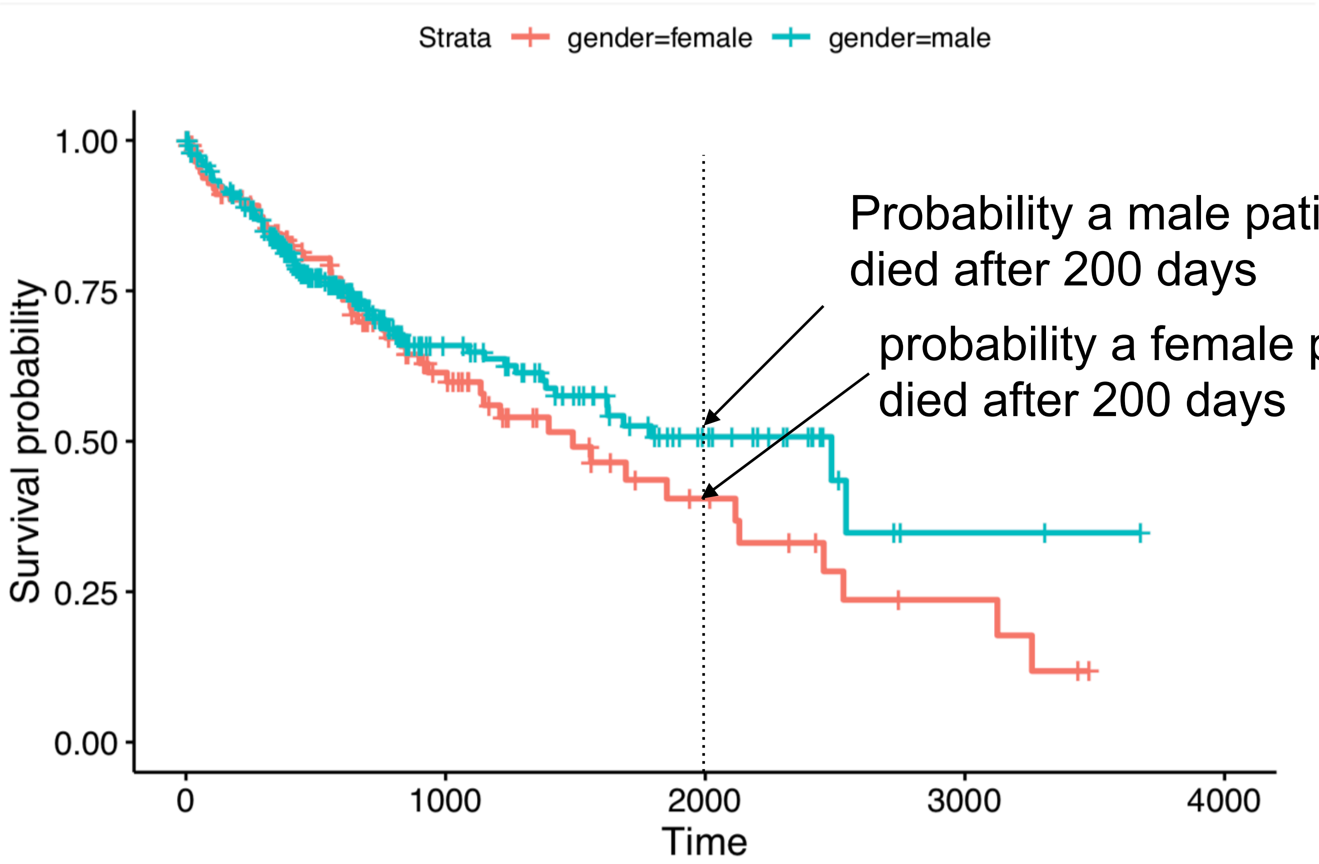
Kaplan-Meier plot

Survival of LIHC patients - male vs. Female



Kaplan-Meier plot

Survival of LIHC patients - male vs. Female



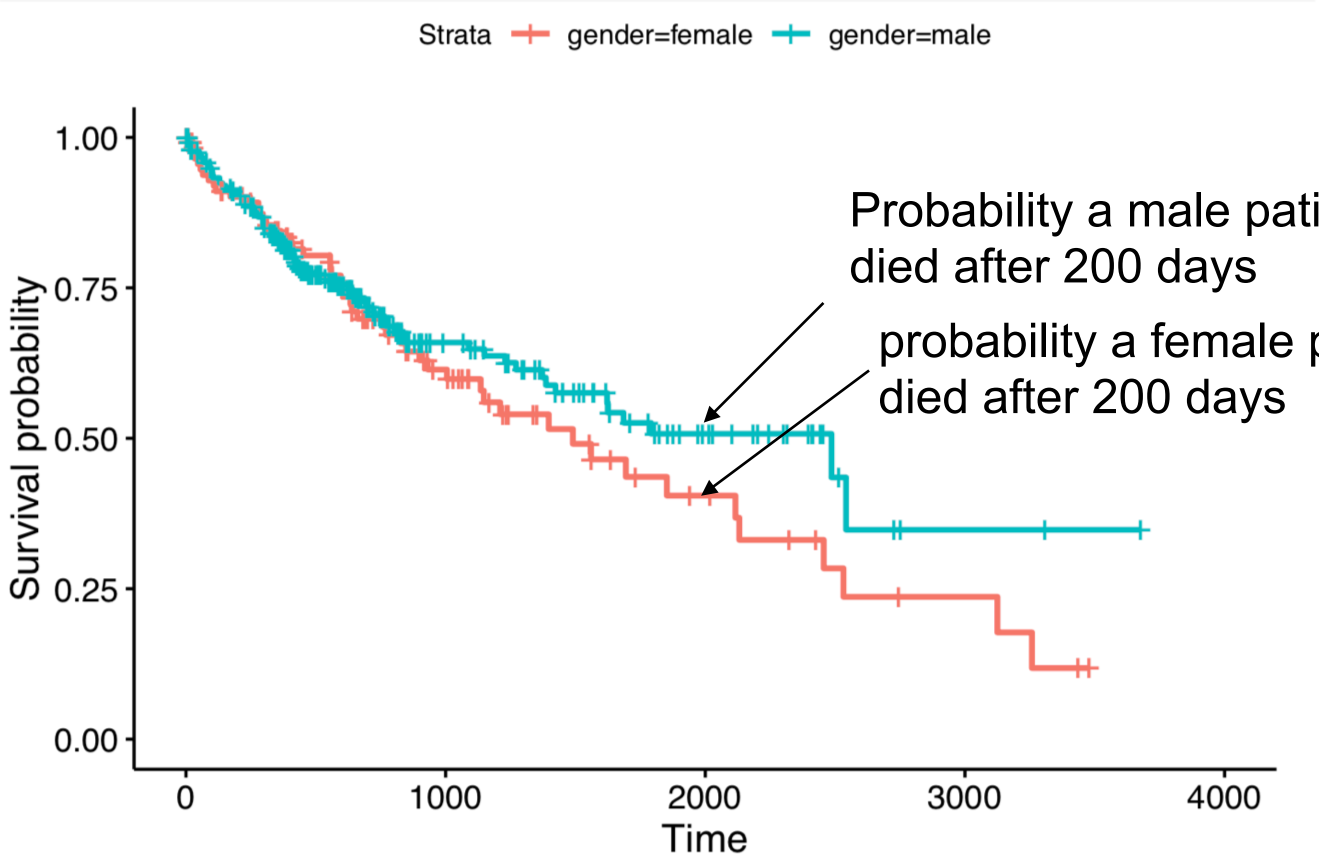
Probability a male patient died after 200 days

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?

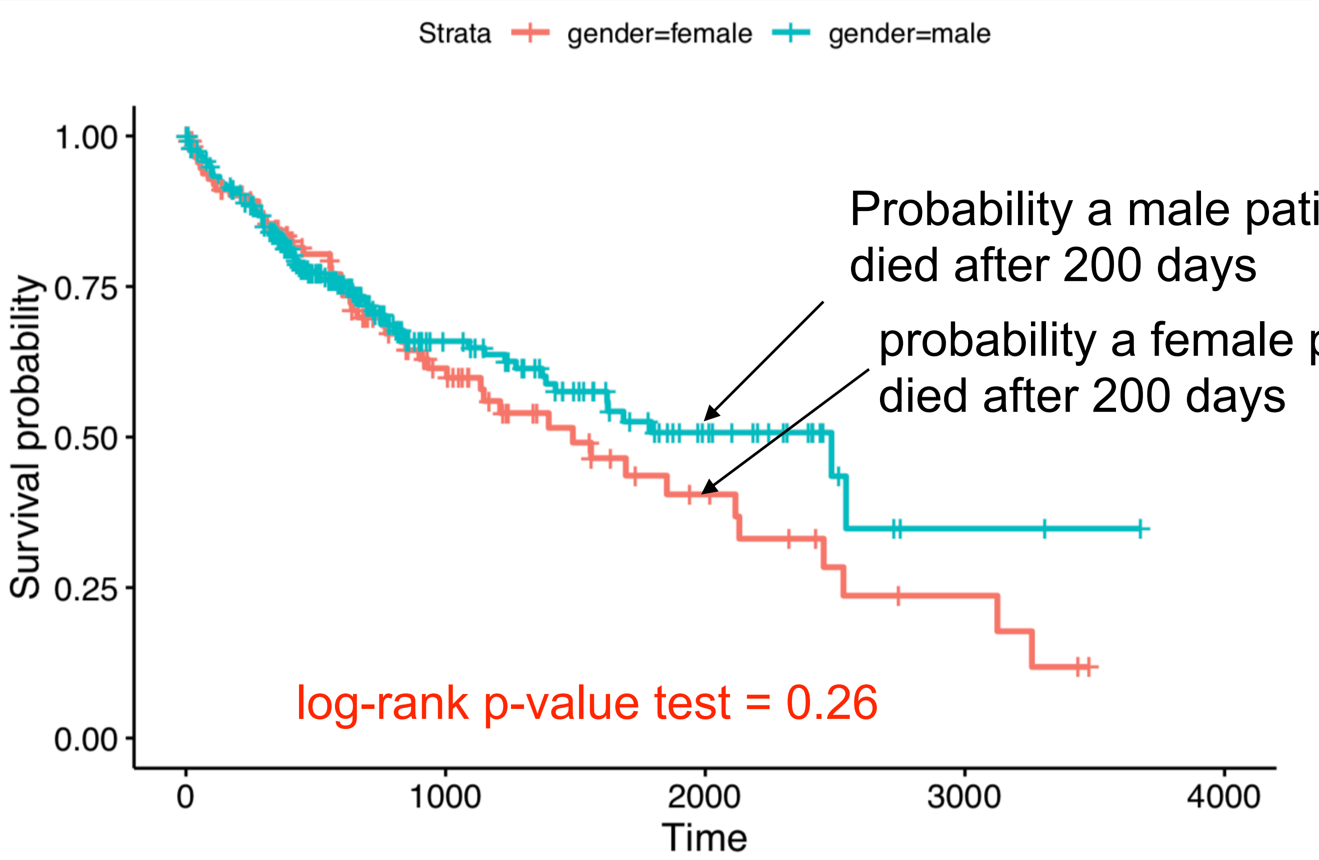


Probability a male patient died after 200 days

probability a female patient died after 200 days

Log-rank test

Is the survival difference significant?

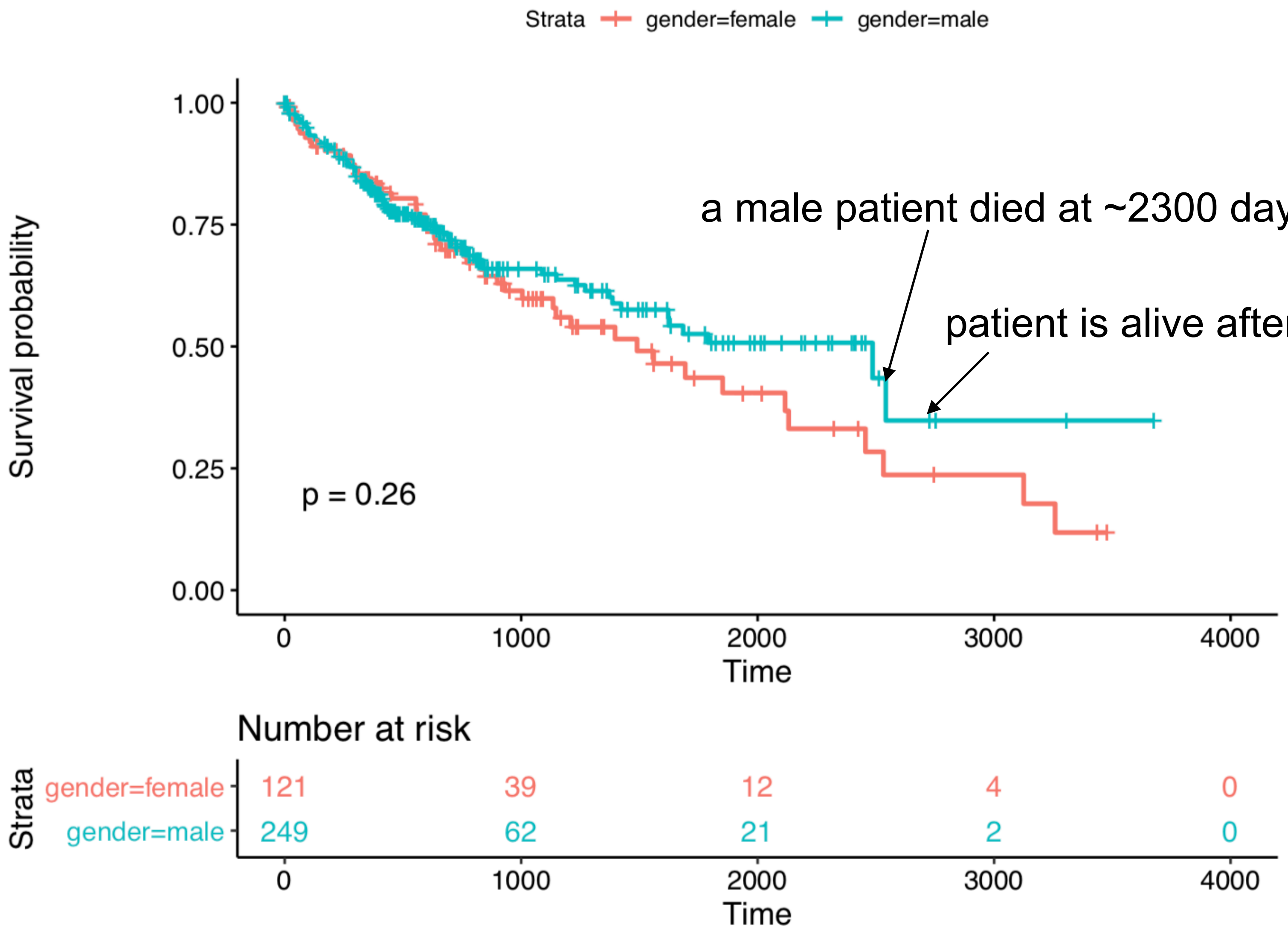


Probability a male patient died after 200 days

probability a female patient died after 200 days

log-rank p-value test = 0.26

Kaplan-Meier plot



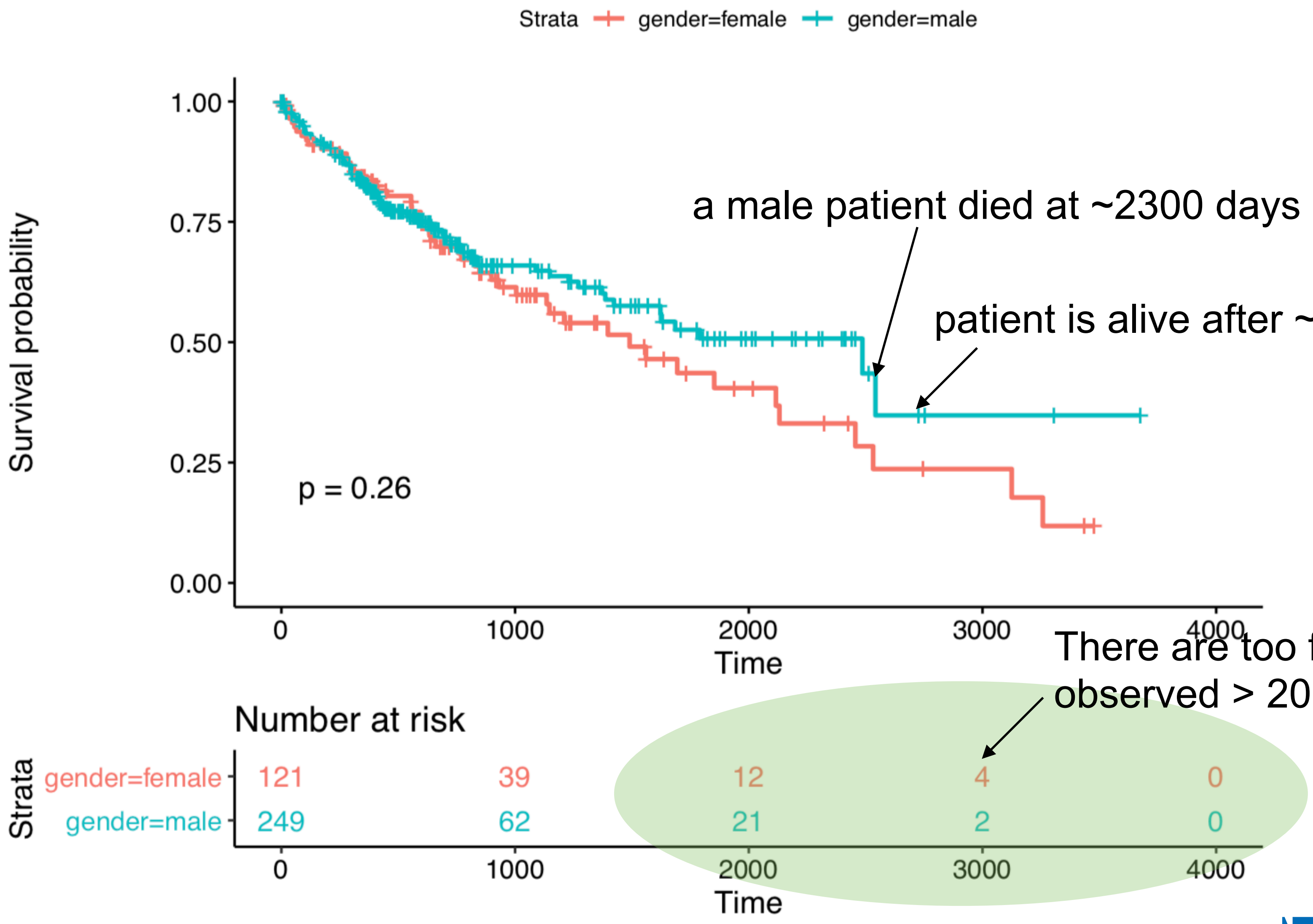
a male patient died at ~2300 days

patient is alive after ~2500 days

Number at risk

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

Kaplan-Meier plot



Number at risk

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

a male patient died at ~2300 days

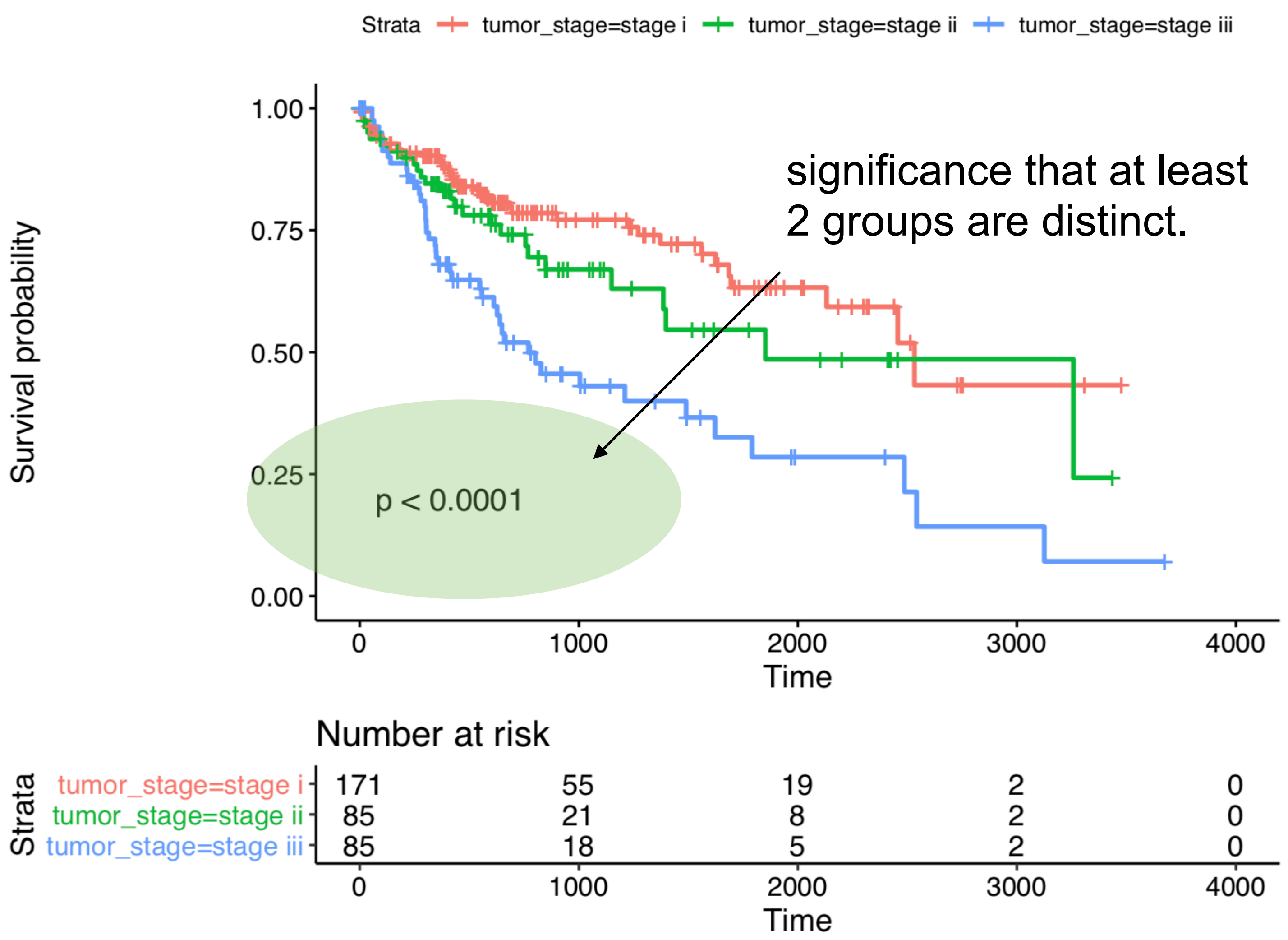
patient is alive after ~2500 days

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.

$p < 0.0001$

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

Time

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!!!

Hands on!

www.costalab.org

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY