



Politecnico
di Bari

Politecnico di Bari

Dipartimento di Ingegneria Elettrica e dell'Informazione
Corso di Laurea Magistrale in Ingegneria dei Sistemi Medicali



DIPARTIMENTO DI
INGEGNERIA ELETTRICA
E DELL'INFORMAZIONE

Bioinformatica Avanzata

RNA-Seq

Dr. Nicola Altini, Ph.D. Student

Dr. Simona De Summa, Bioinformatician, IRCCS "Giovanni Paolo II"

Prof. Eng. Vitoantonio Bevilacqua, Ph.D.



Anno Accademico 2019/2020

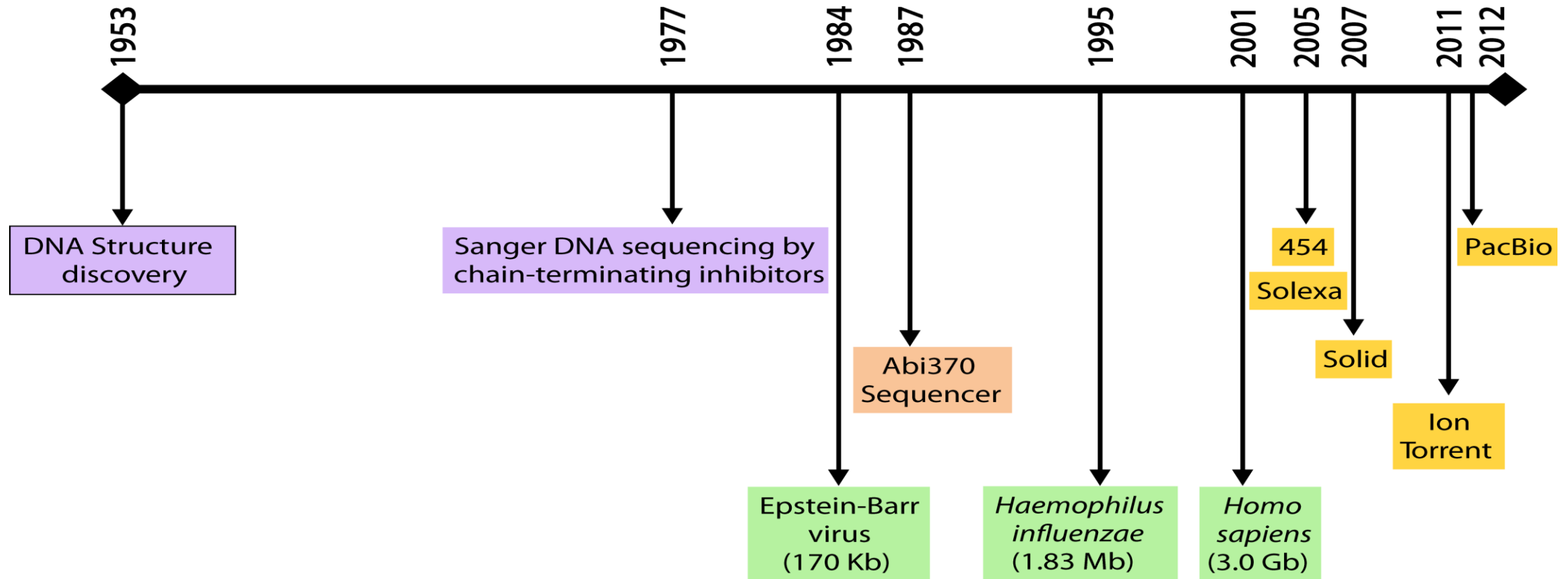


apulian
bioengineering
company

Outline

- Next-generation sequencing
- Sequencing strategies
- TCGA repository
- Differential expression analysis
- Negative binomial model
- Example: lncRNAs in prostate cancer in relation to Gleason score

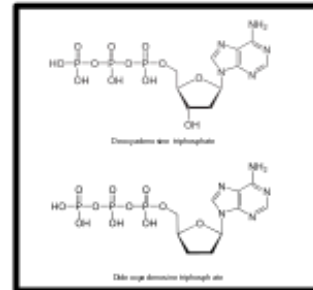
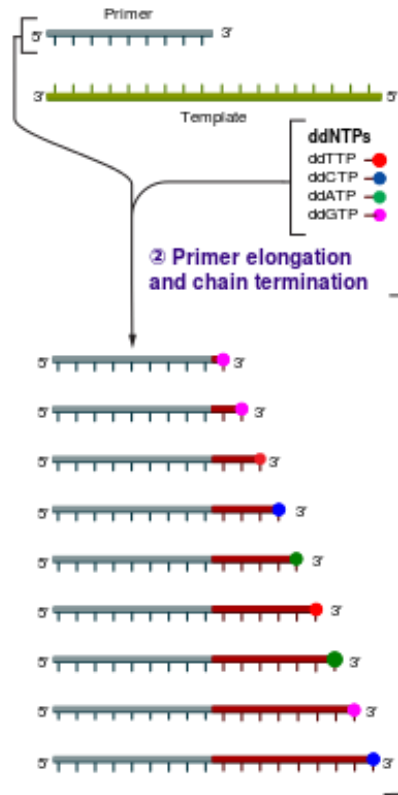
DNA sequencing timeline



Sanger method

① Reaction mixture

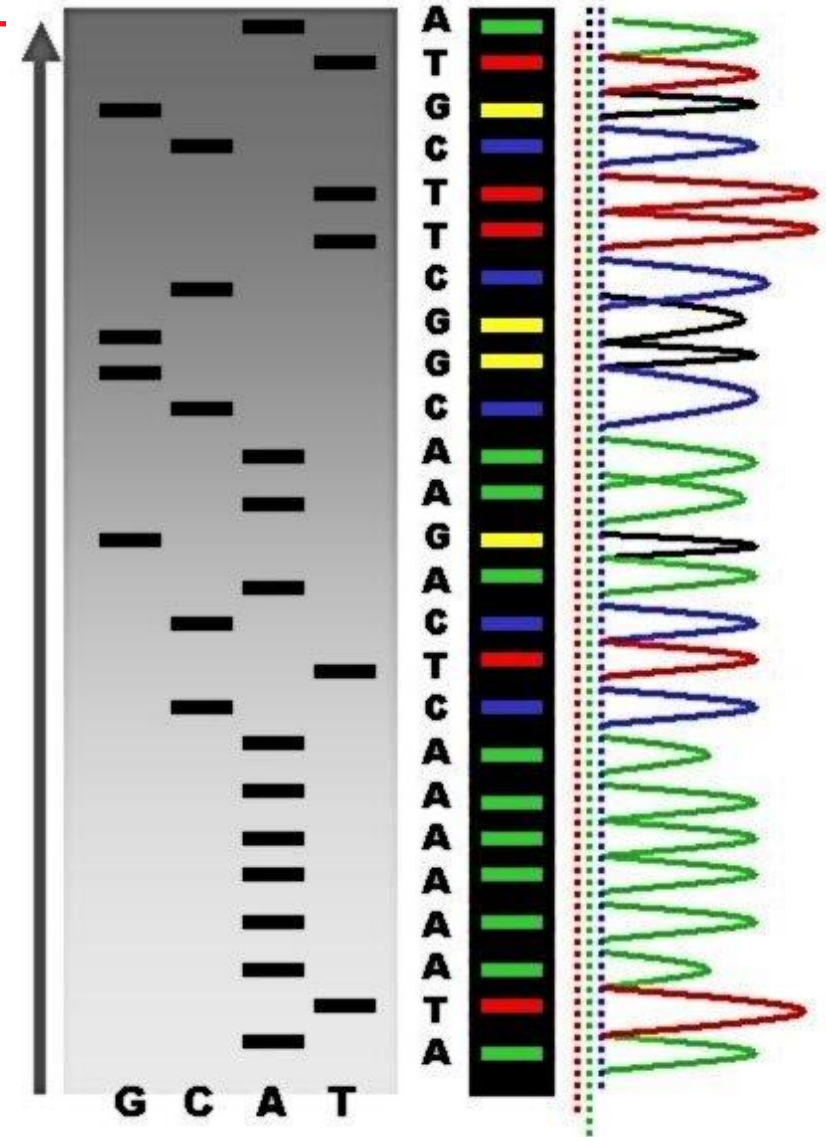
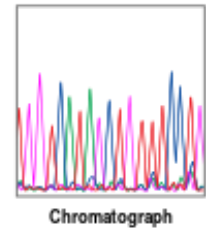
- ▶ Primer and DNA template
- ▶ DNA polymerase
- ▶ ddNTPs with flouochromes
- ▶ dNTPs (dATP, dCTP, dGTP, and dTTP)



③ Capillary gel electrophoresis separation of DNA fragments



④ Laser detection of flouochromes and computational sequence analysis

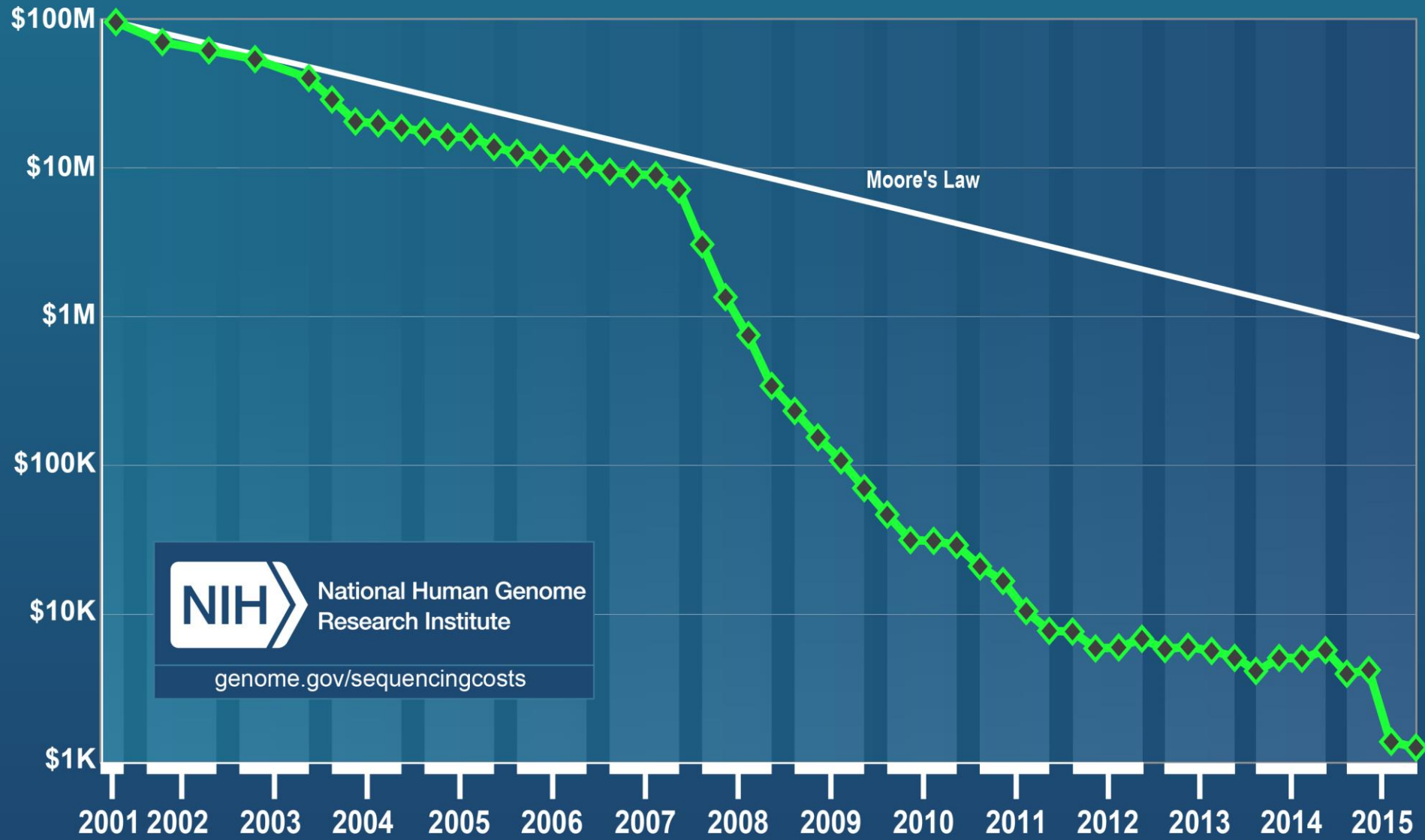


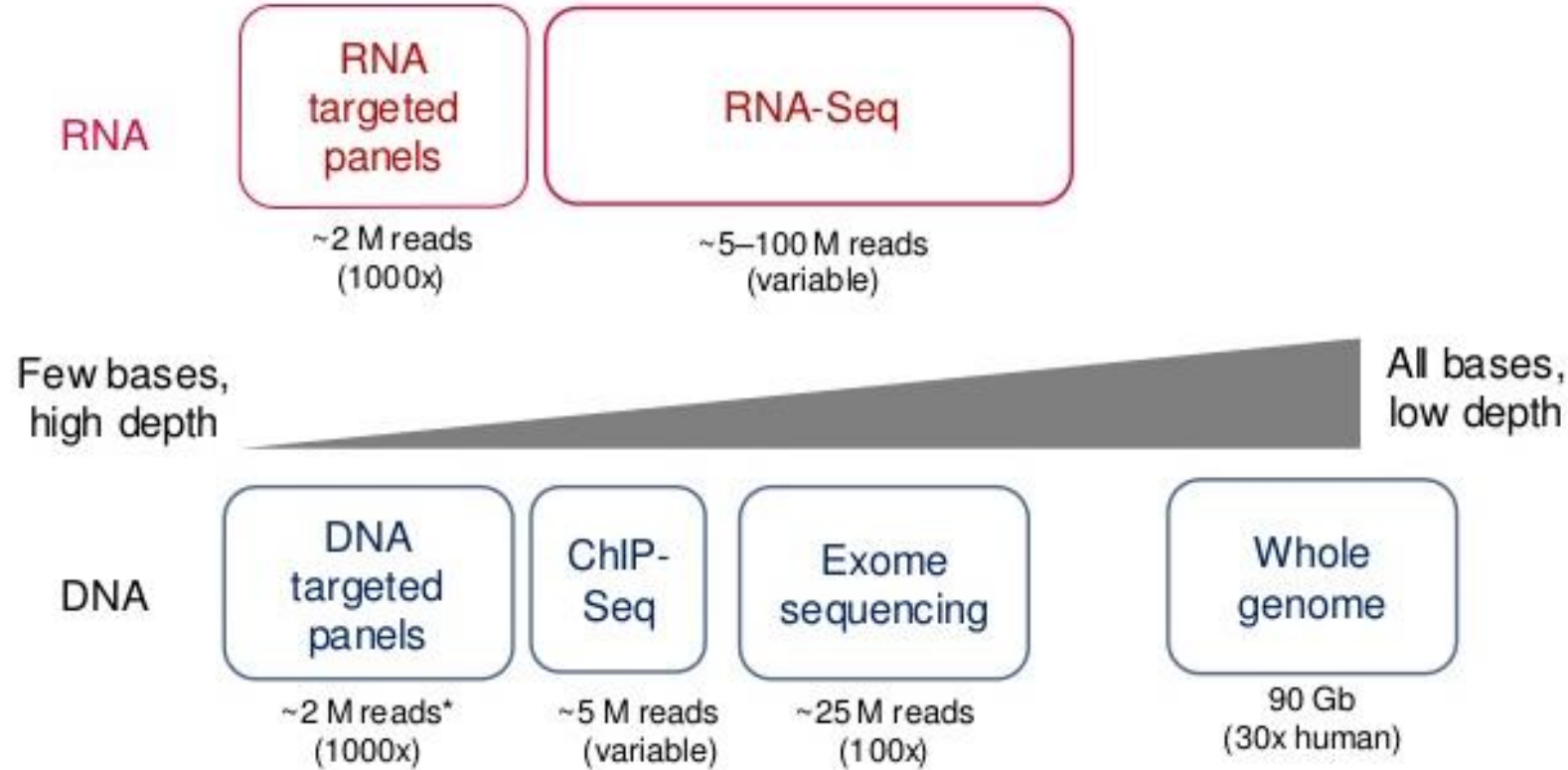
Sanger sequencing reached its technical limits

- Only modestly parallel (394 lanes/machine)
- Long read lengths (500-900 bp) & >99.9% correct
- Need to clone the DNA to obtain enough for sequencing reaction

- cost for typical Sanger sequencing is \$5-6/sample with reliable 500 bp of sequence

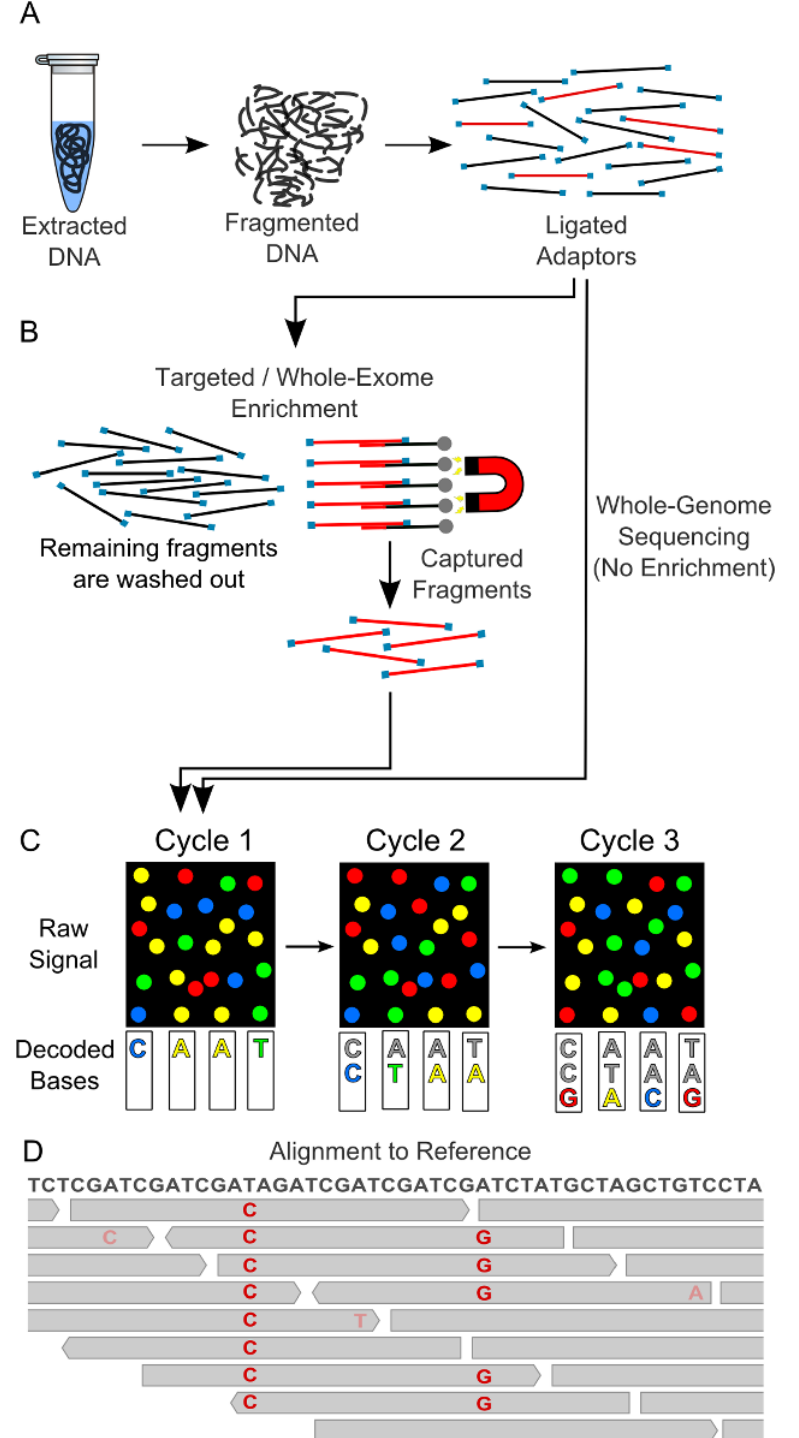
Cost per Genome



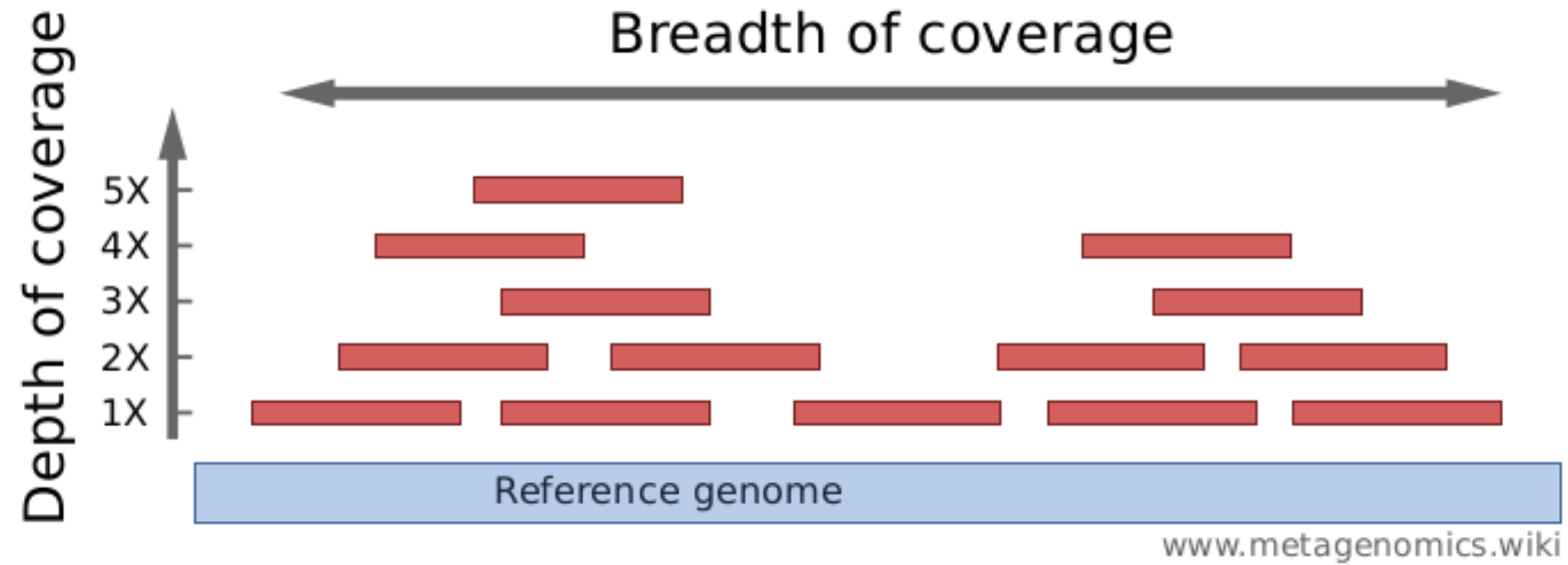


* Depends on gene number and sequencing depth

Gb = # reads x read length

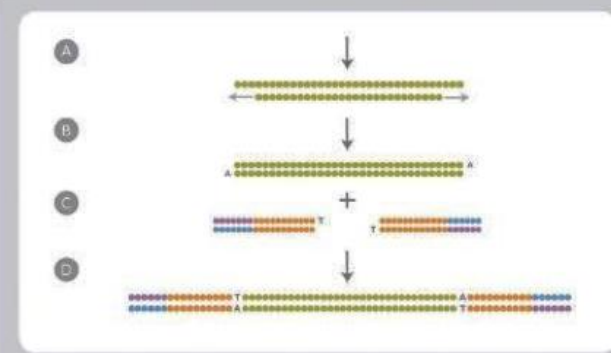


Important sequencing parameters



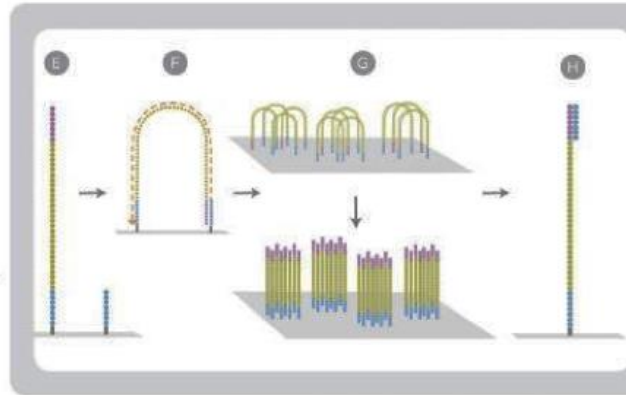
Illumina NGS

Library Prep:
~ 6 hours



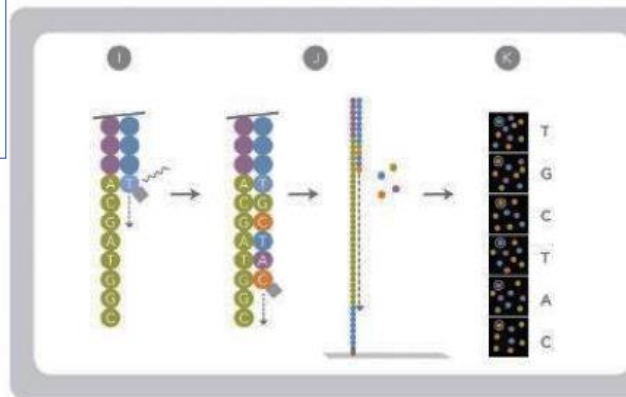
- A) Fragment DNA
- B) Repair ends/Add A overhang DNA
- C) Ligate adapters
- D) Select ligated DNA

Cluster generation
~ 6 hours



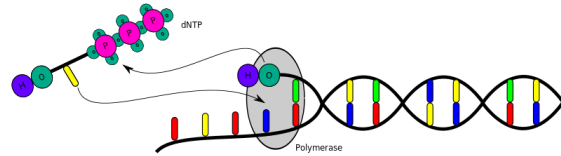
- E) Attach DNA to flow cell
- F) Bridge amplification
- G) Generate clusters
- H) Anneal sequencing primer

Sequencing
2-6 days

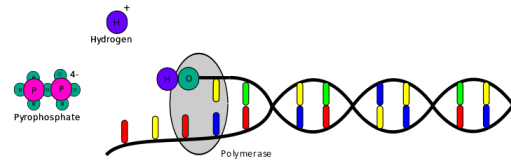


- I) Extend 1st base, read & deblock
- J) Repeat to extend strand
- K) Generate base calls

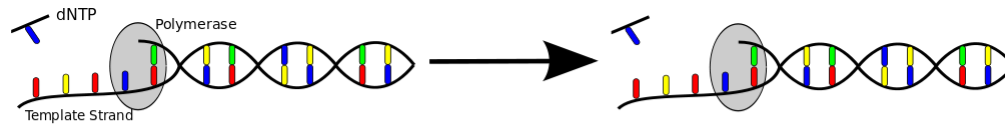
Ion Torrent – measures pH changes



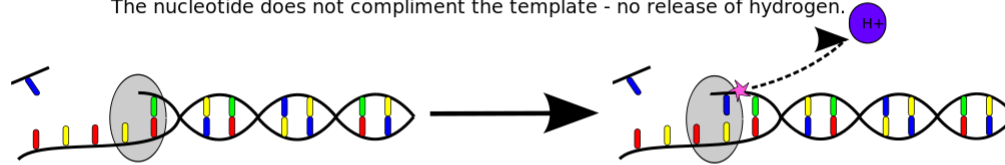
Polymerase integrates a nucleotide.



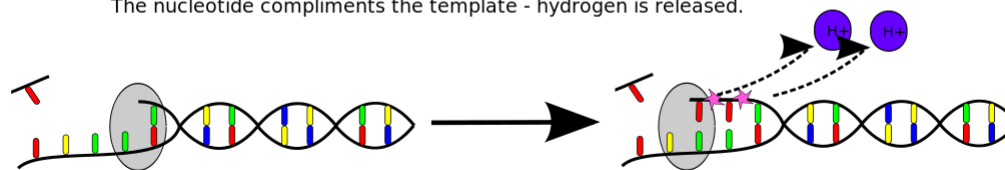
Hydrogen and pyrophosphate are released.



The nucleotide does not compliment the template - no release of hydrogen.



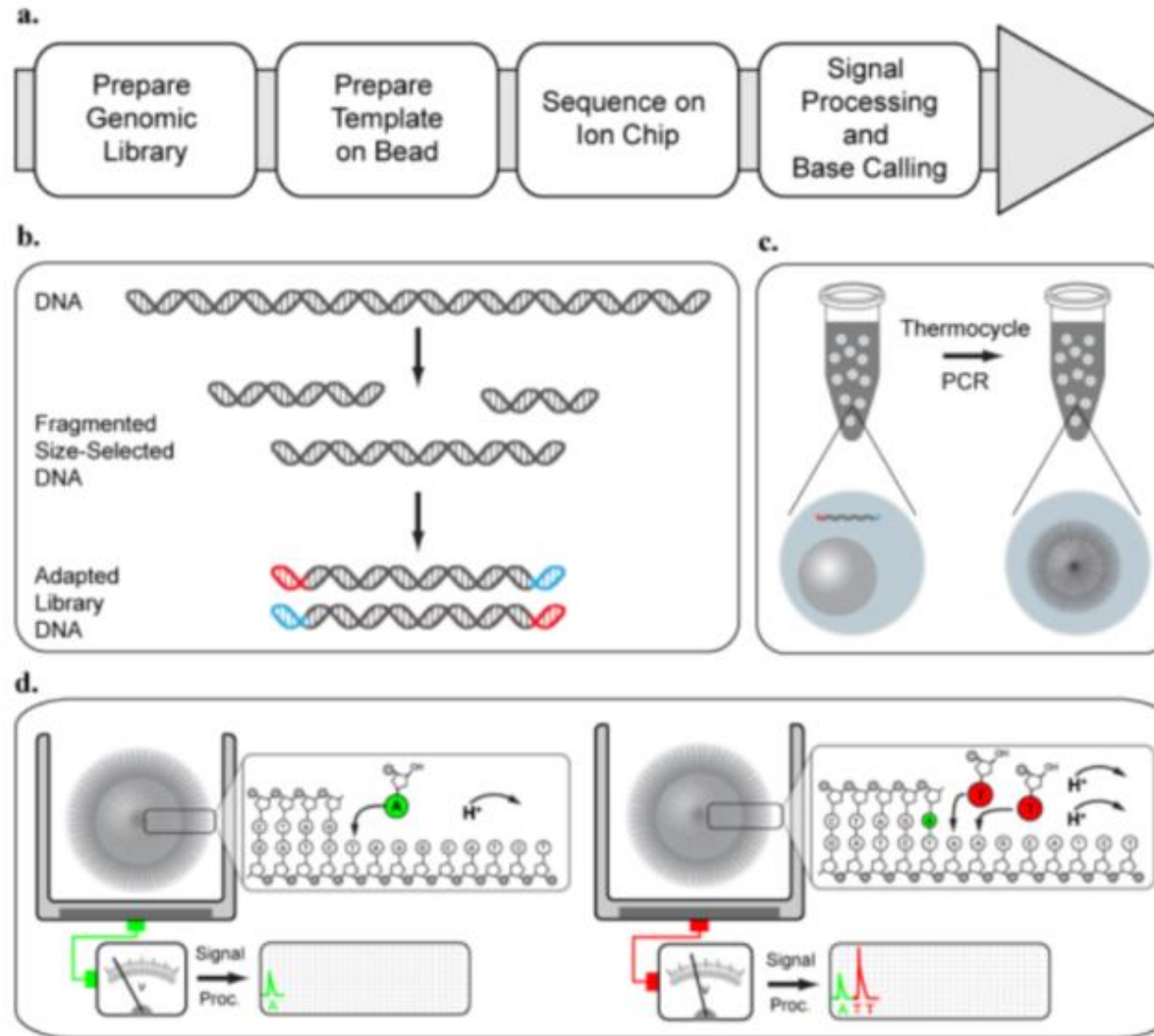
The nucleotide compliments the template - hydrogen is released.



The nucleotide compliments several bases in a row - multiple hydrogen ions are released.

Done on a semi-conductor chip

Ion Torrent workflow



RNA-Seq: a powerful approach

Microarray

Hybridization.
Scanning images.
Quantification.

Raw intensities

Preprocessing:
Background correction,
Normalization,
Summarization.

Expression levels of
Transcripts (continuous)

Statistical analysis

Cellular
functional/pathway
analysis

GO

KEGG

RNA-Seq

Sequencing.
Base call.

Short reads

Aligned to
reference genome,
known isoform & exon-
junction sequences.

Expression levels of
Transcripts (counts)

Statistical analysis

Differentially
expressed
transcripts

Novel
transcripts

Input: Genome, gene annotations, and sequencing reads

```
>AGTCGTAGCTAGTCGA
>GGATAGCGTGATCTAC
>TTTGTAGTCTATGCA
```

```
>TATGCTTATACGTT
>CGCACGGTAGCATG
>GTAGCGCTAGTACA
```

1. Align reads to genome

2. Normalize experiments

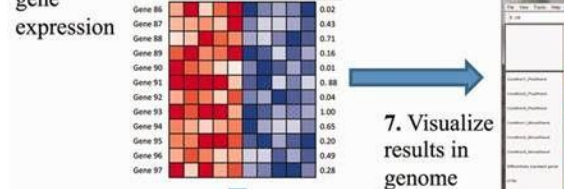
```
GACCAT
CCATTG
TAAGACCATTGC
```

```
CCAT
AGACC
TAAGACCATTGC
```

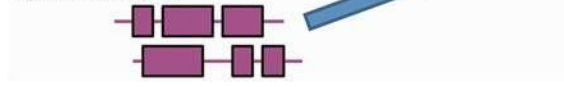
3. Assemble transcripts and identify transcript boundaries

4. Quantify transcript abundance

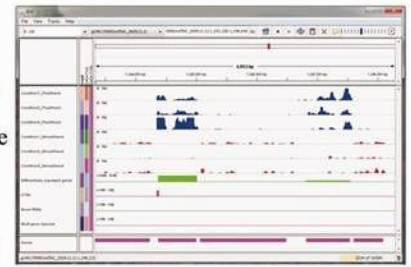
5. Test for differential gene expression



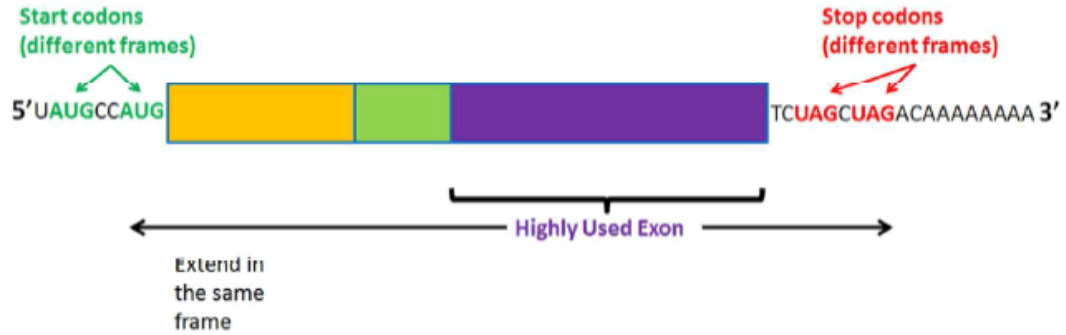
6. Characterize operon structures



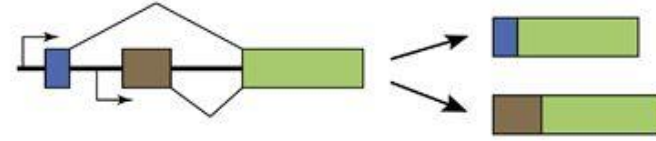
7. Visualize results in genome browser



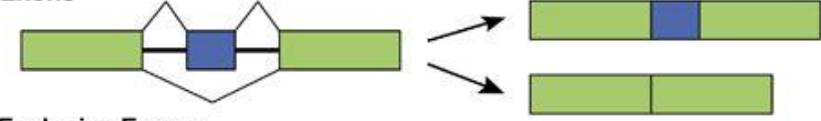
Transcript isoforms



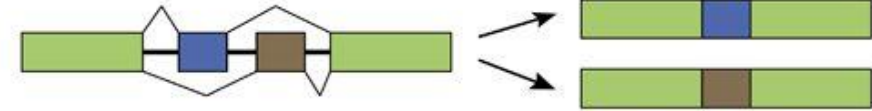
Alternative Promoters



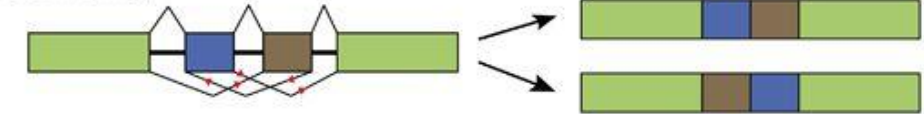
Cassette Exons



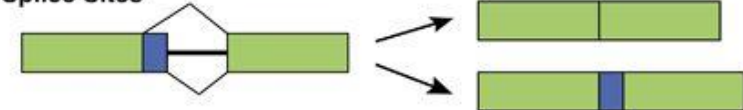
Mutually Exclusive Exons



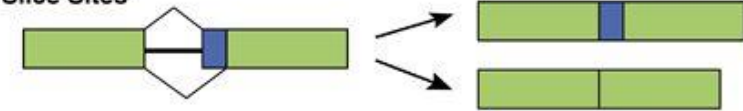
Exon Scrambling



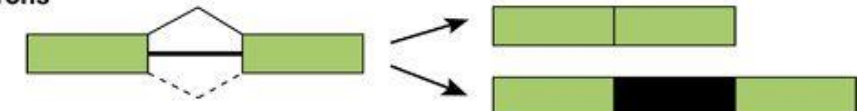
Alternative 5' Splice Sites



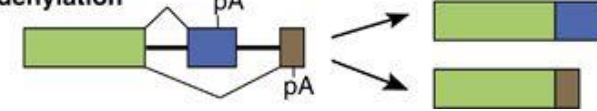
Alternative 3' Splice Sites

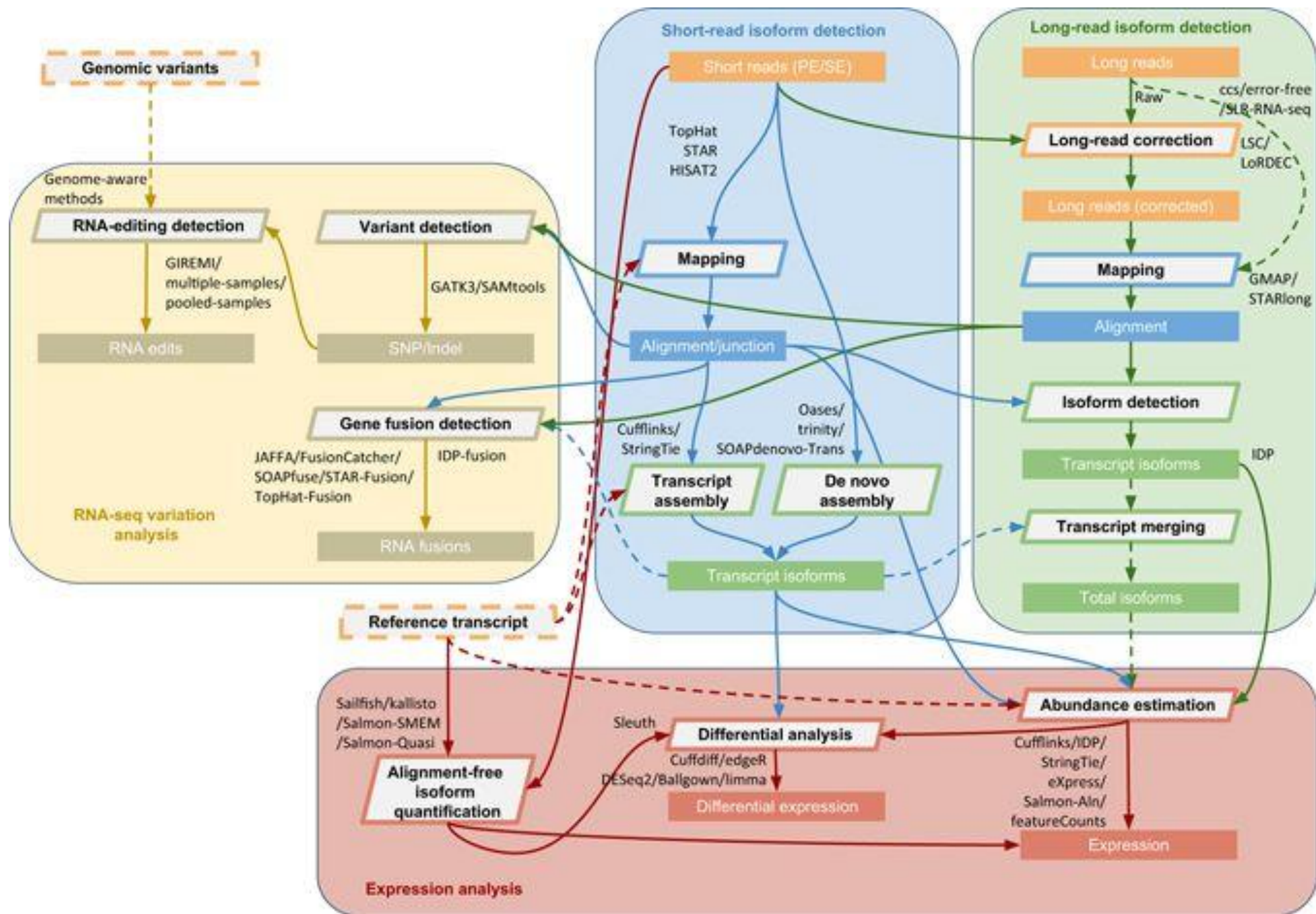


Retained Introns



Alternative Polyadenylation






Repositories

NCBI Resources How To Sign in to NCBI

SRA SRA Search Advanced Help



SG AACGCC TTGCATTAG TAA CGCC
AAAG AATAC CGTA
ATATTT GCAC

SRA

EMBL-EBI Services Research Training About us EMBL-EBI Hinxton

European Genome-phenome Archive

All

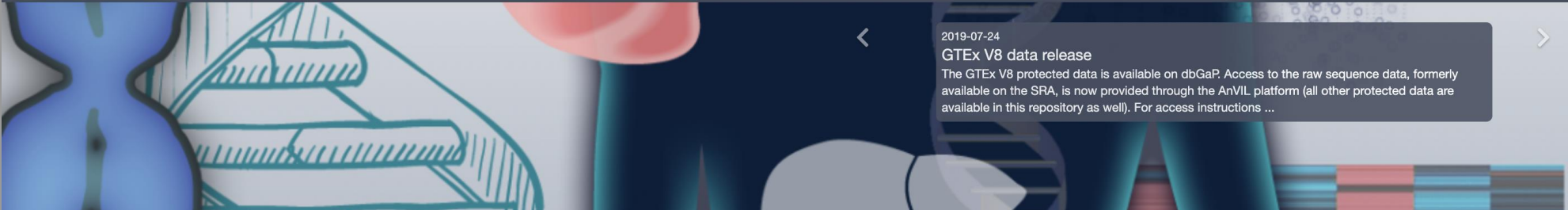
Examples: EGAS00000000001, Cancer

About GTEx Publications Access Biospecimens FAQs Contact

Home Datasets Expression QTLs & Browsers Sample Data Documentation Search Gene or SNP ID... Sign In

GTEx Portal

2019-07-24
GTEx V8 data release
The GTEx V8 protected data is available on dbGaP. Access to the raw sequence data, formerly available on the SRA, is now provided through the AnVIL platform (all other protected data are available in this repository as well). For access instructions ...



Soluzioni terapeutiche all'avanguardia per una migliore qualità di vita. 

**Publicato Atlante globale del Cancro
Svela radici comuni di tumori diversi**

- 
- 
- 

FRUTTO DI UN PROGETTO LANCIATO NEL 2006, HA ANALIZZATO LA «CARTA D'IDENTITÀ» DELLE NEOPLASIE DI 10MILA PAZIENTI E LE HA CATALOGATE IN 28 SOTTOGRUPPI BASANDOSI SULLE CARATTERISTICHE GENETICHE. È IL PUNTO DI PARTENZA PER CURE PIÙ PRECISE PER OGNI MALATO

di Vera Martinella



R.it

Medicina e Ricerca

- Home
- Alimentazione e Fitness
- Medicina e Ricerca
- Salute Seno
- Oncologia

Scopri di più su **Osservaltalia.it**



TECNOLOGIA

Pronto l'Atlante genetico dei tumori: un identikit per terapie...



Dalle docce per clochard al contatore riattivato, chi è don Corrado, l'elemosiniere di Papa Francesco



Le 10 tendenze n aperitivi più gettonate per l'estate

sanità informazione

- HOME
- LAVORO
- SALUTE
- FORMAZIONE
- METEO
- APPUNTAMENTI
- APICALI
- SPECIALI
- SERVE UN DOTTORE

Ricerca, creata mappa con identikit di 33 tipi di cancro. Rivoluzione per l'approccio terapeutico

Nasce l'Atlante genetico dei tumori, per individuare terapie più precise



 **Alimentazione e stile di vita**
DUE METÀ DELLA

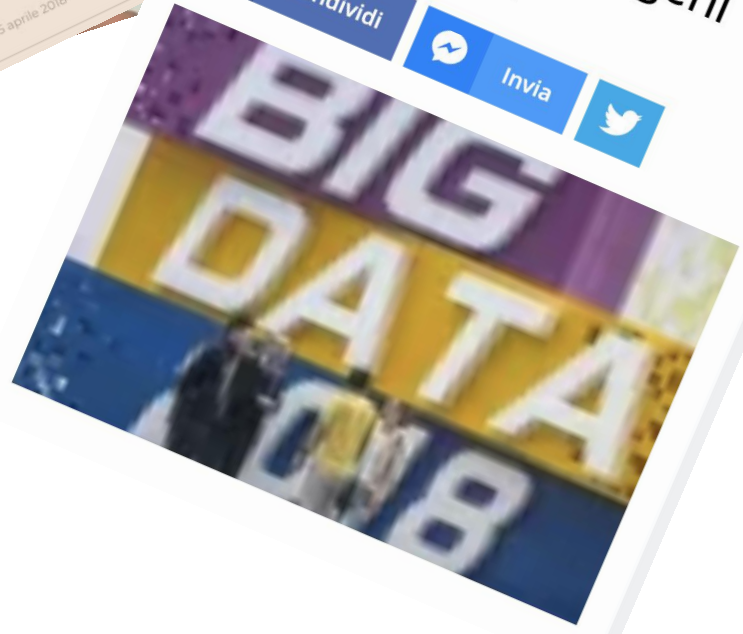
MAJORE
Elezioni, le parole chiave della Ue in crisi

Pronto l'Atlante genetico dei tumori: un identikit per terapie più precise

—di Francesca Cerati | 5 aprile 2018

Da big data alcuni geni tumori, studio

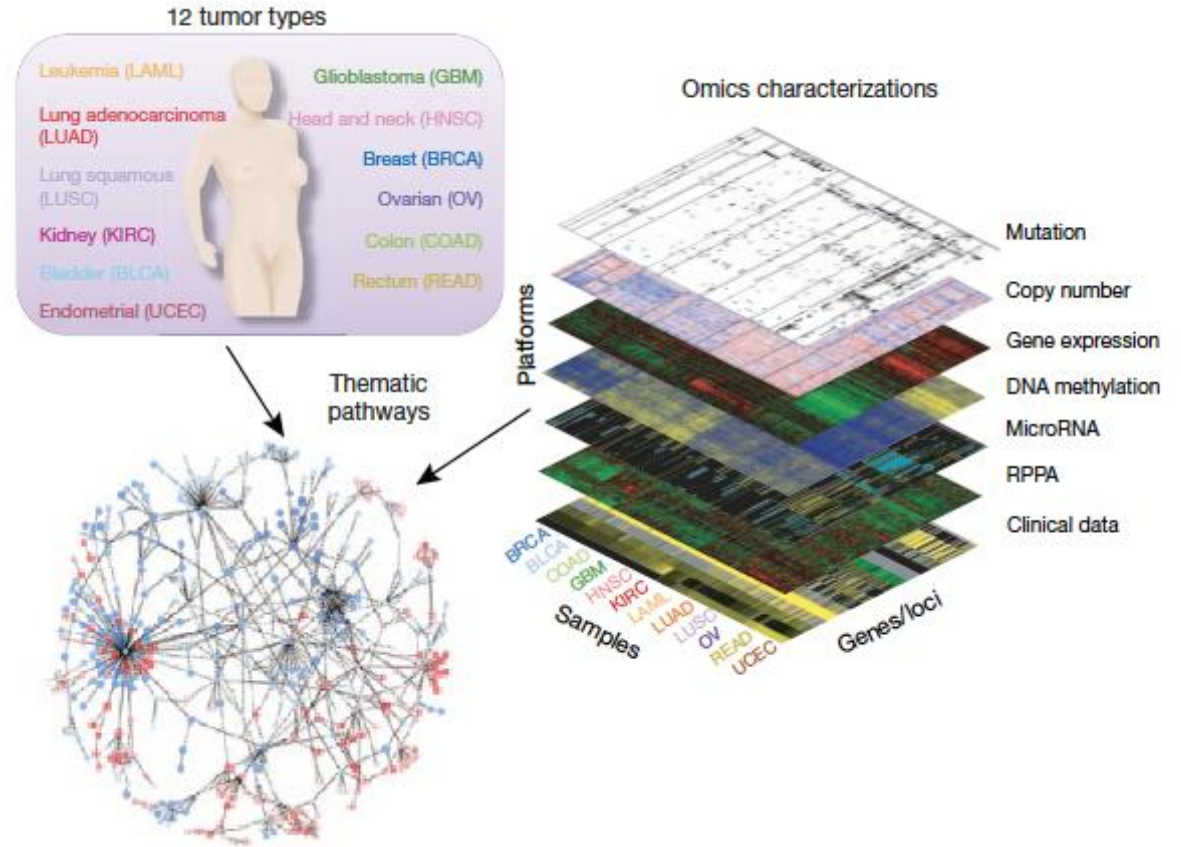
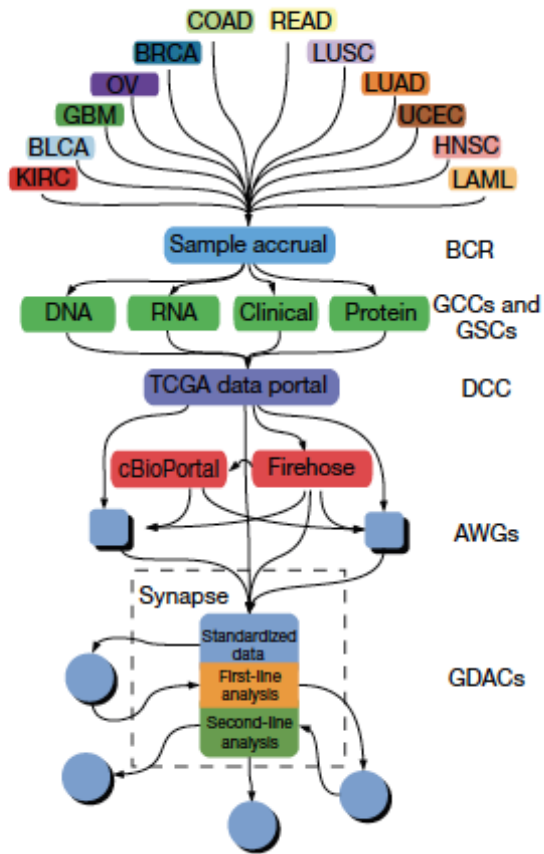
-  Condividi
-  Invia
- 



The Cancer Genome Atlas Pan-Cancer analysis project

The Cancer Genome Atlas Research Network¹, John N Weinstein^{2,3}, Eric A Collisson⁴, Gordon B Mills³, Kenna R Mills Shaw^{5,6}, Brad A Ozenberger⁷, Kyle Ellrott^{8,9}, Ilya Shmulevich¹⁰, Chris Sander¹¹ & Joshua M Stuart^{8,9}

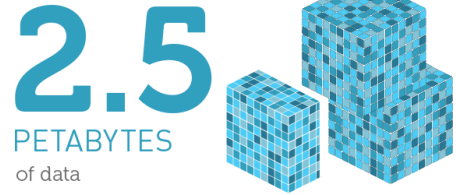
NATURE GENETICS | VOLUME 45 | NUMBER 10 | OCTOBER 2013



NATIONAL CANCER INSTITUTE THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over



To put this into perspective, **1 petabyte** of data is equal to



TCGA data describes



...including

...based on paired tumor and normal tissue sets collected from



...using



TCGA RESULTS & FINDINGS



**MOLECULAR
BASIS OF
CANCER**

Improved our understanding of the genomic underpinnings of cancer

For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.



**TUMOR
SUBTYPES**

Revolutionized how cancer is classified

TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.*



**THERAPEUTIC
TARGETS**

Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development

TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

THE TEAM



20
COLLABORATING
INSTITUTIONS
across the United States
and Canada

WHAT'S NEXT?

The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.



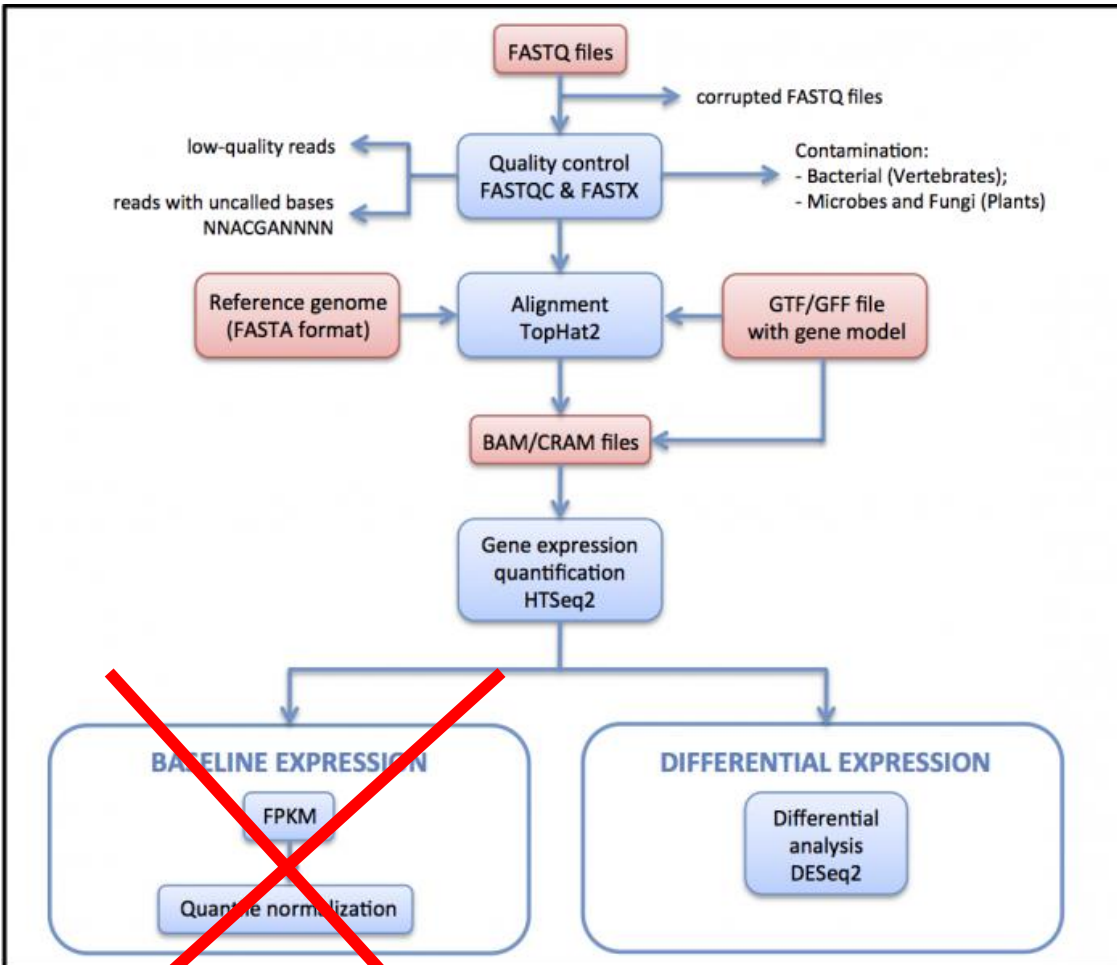
*TCGA's analysis of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, including a new subtype characterized by infection with Epstein-Barr virus.

What we have....

The Cancer Genome Atlas (TCGA)

<https://portal.gdc.cancer.gov/>





An update (12th August 2018):
You should abandon RPKM / FPKM normalisation. They are not ideal where cross-sample differential expression

Examples

- $$RPKM = \#MappedReads * \frac{1000bases * 10^6}{length\ of\ transcript * Total\ number\ of\ mapped\ reads}$$

- Example 1:

- 2500kb transcript with 900 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped):

- $$RPKM = 900 * \frac{1000 * 10^6}{2500 * 8.10^6} = 45$$

BAM headers: an essential part of a BAM file

```
@HD VN:1.0 GO:none SO:coordinate
@SQ SN:chrM LN:16571
@SQ SN:chr1 LN:247249719
@SQ SN:chr2 LN:242951149
[cut for clarity]
@SQ SN:chr9 LN:140273252
@SQ SN:chr10 LN:135374737
@SQ SN:chr11 LN:134452384
[cut for clarity]
@SQ SN:chr22 LN:49691432
@SQ SN:chrX LN:154913754
@SQ SN:chrY LN:57772954
@RG ID:20FUK.1 PL:illumina PU:20FUKAAXX100202.1 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.2 PL:illumina PU:20FUKAAXX100202.2 LB:Solexa-18484 SM:NA12878 CN:BI
@RG ID:20FUK.3 PL:illumina PU:20FUKAAXX100202.3 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.4 PL:illumina PU:20FUKAAXX100202.4 LB:Solexa-18484 SM:NA12878 CN:BI
@RG ID:20FUK.5 PL:illumina PU:20FUKAAXX100202.5 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.6 PL:illumina PU:20FUKAAXX100202.6 LB:Solexa-18484 SM:NA12878 CN:BI
@RG ID:20FUK.7 PL:illumina PU:20FUKAAXX100202.7 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.8 PL:illumina PU:20FUKAAXX100202.8 LB:Solexa-18484 SM:NA12878 CN:BI
@PG ID:BWA VN:0.5.7 CL:tk
@PG ID:GATK TableRecalibration VN:1.0.2864
20FUKAAXX100202:1:1:12730:189900 163 chrM 1 60 101M = 282 381
GATCACAGGTCTATCACCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTA...[more bases]
?BA@A>BBBBACBBAC@BBCBBCBC@BC@CAC@:BBCBBCACAACBABCBCBAC...[more quals]
RG:Z:20FUK.1 NM:i:1 SM:i:37 AM:i:37 MD:Z:72G28 MQ:i:60 PG:Z:BWA UQ:i:33
```

Required: Standard header

Essential: contigs of aligned reference sequence. Should be in karyotypic order.

Essential: read groups. Carries platform (PL), library (LB), and sample (SM) information. Each read is associated with a read group

Useful: Data processing tools applied to the reads



	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

	Control1	Control2	KnockDown
TSPAN6	11	16	
TNMD	1	0	
DPM1	435	743	
SCYL3	203	218	
C1orf112	216	643	
FGR	2365	5011	
CFH	6	1	
FUCA2	380	865	
...	
NFYA	888	827	

- Reads (BAM file) are counted for each gene model (gtf file) using HTSeq-count:

HTSeq

www.huber.embl.de/users/anders/HTSeq

For each position i in the read, a set $S(i)$ is defined as the set of all features overlapping position i .

Then, consider the set S , which is (with i running through all position within the read)

- the union of all the sets $S(i)$ for mode union.
- the intersection of all the sets $S(i)$ for mode intersection-strict.
- the intersection of all non-empty sets $S(i)$ for mode intersection-nonempty.

Sequencing count data

	control-1	control-2	control-3	treated-1
FBgn0000008	78	46	43	47
FBgn0000014	2	0	0	0
FBgn0000015	1	0	1	0
FBgn0000017	3187	1672	1859	2445
FBgn0000018	369	150	176	288
[...]				

Sequencing counting rules

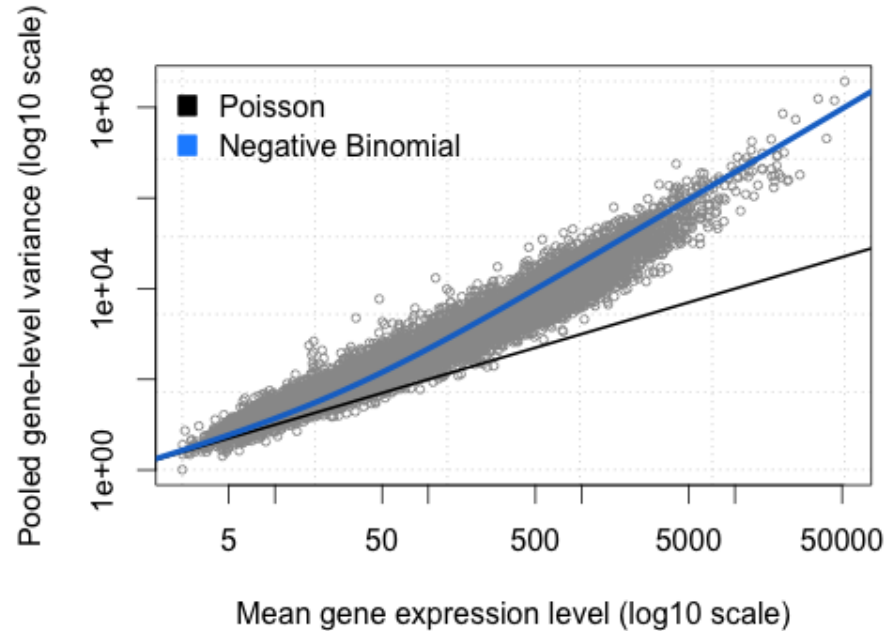
- Count reads, not base-pairs
- Count each read at most once.
- Discard a read if
 - it cannot be uniquely mapped
 - its alignment overlaps with several genes
 - the alignment quality score is bad
 - (for paired-end reads) the mates do not map to the same gene

In RNA-Seq, noise (and hence power) depends on count level.

Why?

A FIRST INTUITION

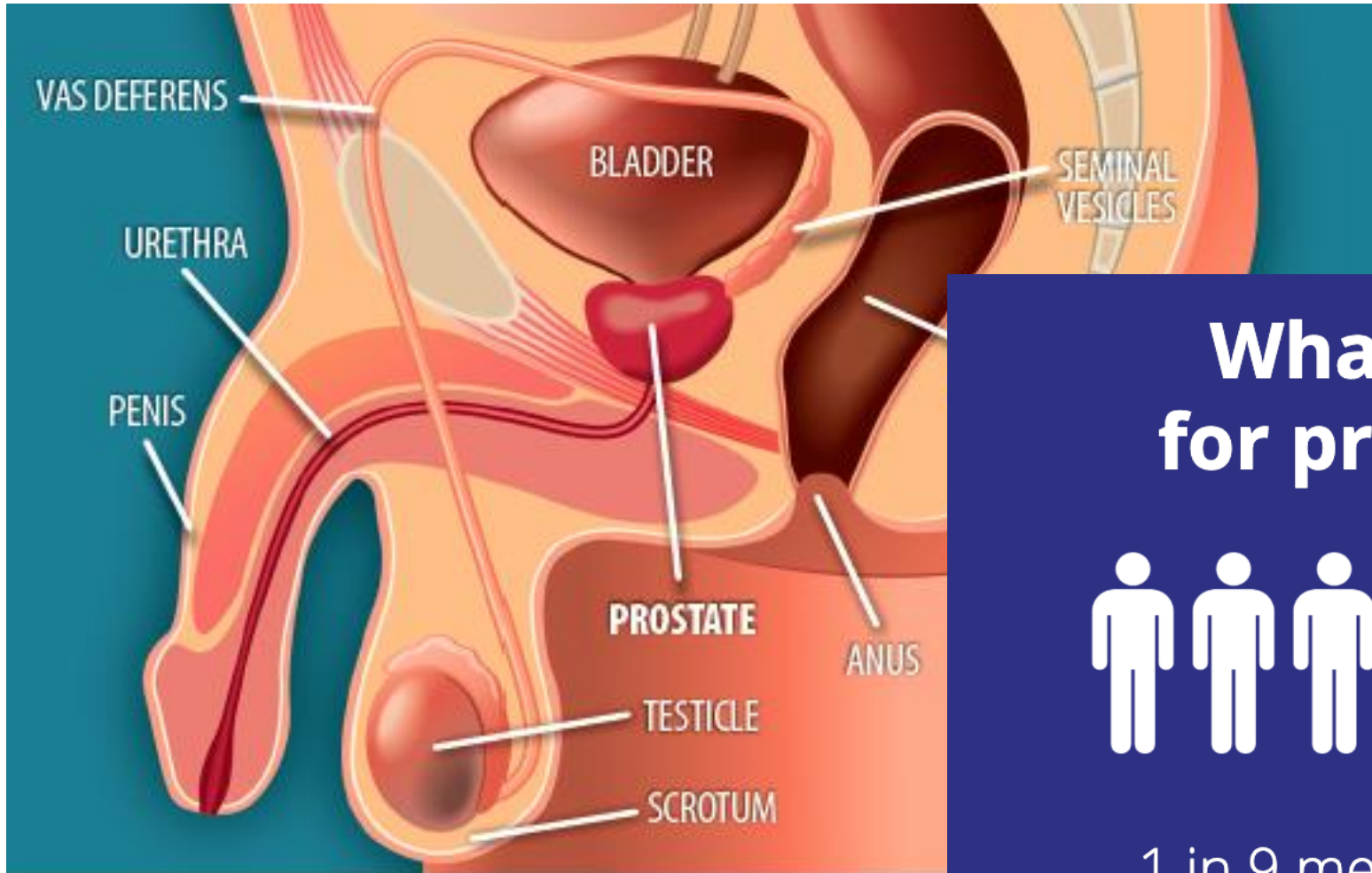
1. In a standard sequencing experiment (RNA-Seq), we map the sequencing reads to the reference genome and count how many reads fall within a given gene (or exon).
2. This means that the input for the statistical analysis are discrete non-negative integers (“counts”) for each gene in each sample.
3. The total number of reads for each sample tends to be in the millions, while the counts per gene vary considerably but tend to be in the tens, hundreds or thousands. Therefore, the chance of a given read to be mapped to any specific gene is rather small.
4. Discrete events that are sampled out of a large pool with low probability sounds very much like a **Poisson process**. And indeed it is. In fact, earlier iterations of RNA-Seq analysis modeled sequencing data as a Poisson distribution. There is one problem, however. The **variability of read counts** in sequencing experiments tends to be **larger than the Poisson distribution allows**.



If we assume that our samples are biological replicates, it is not surprising that the **same transcript is present at slightly different levels in each sample**, even under the same conditions. It is obvious that the variance of counts is generally greater than their mean, especially for genes expressed at a higher level. This phenomenon is called “**overdispersion**”. The NB distribution is similar to a Poisson distribution but has an extra parameter called the “clumping” or “dispersion” parameter. It is like a Poisson distribution with more variance (overdispersion) we observe in sequencing data. In the framework of the NB distribution, it is accounted for by allowing Gamma-distributed uncertainty about the expected counts (the Poisson rate) for each gene.

Example: lncRNAs in prostate cancer in relation to Gleason score

Prostate cancer

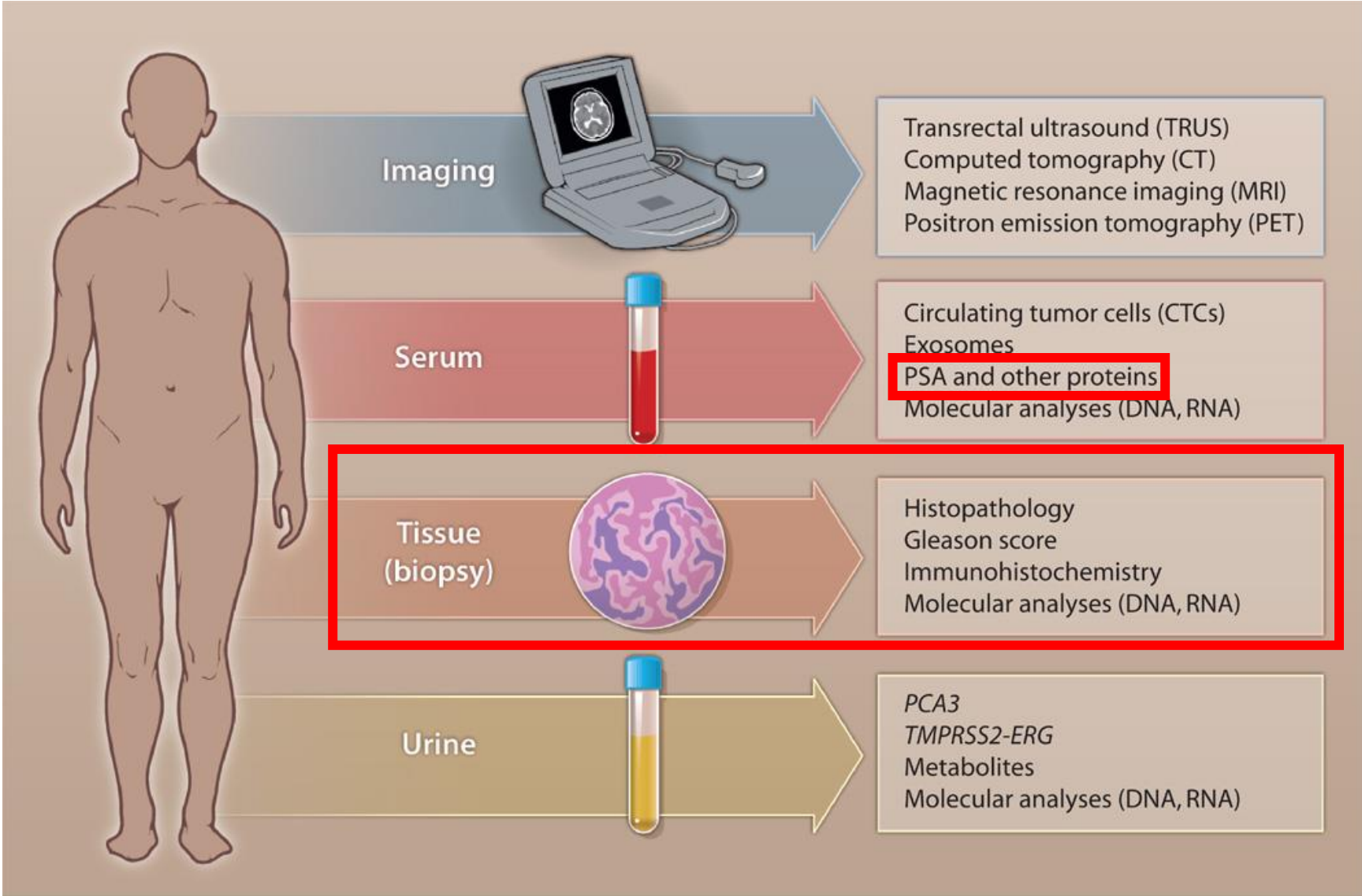


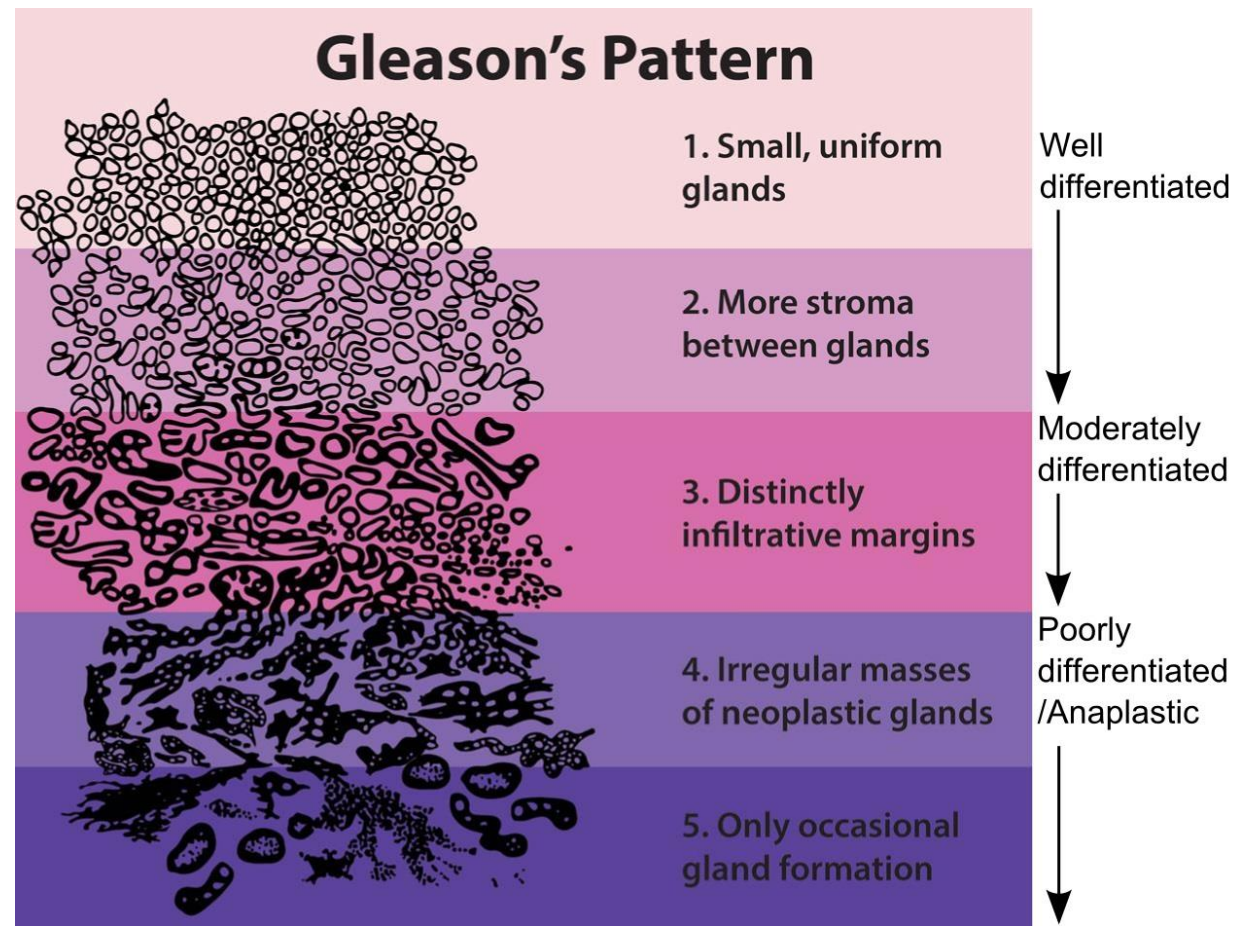
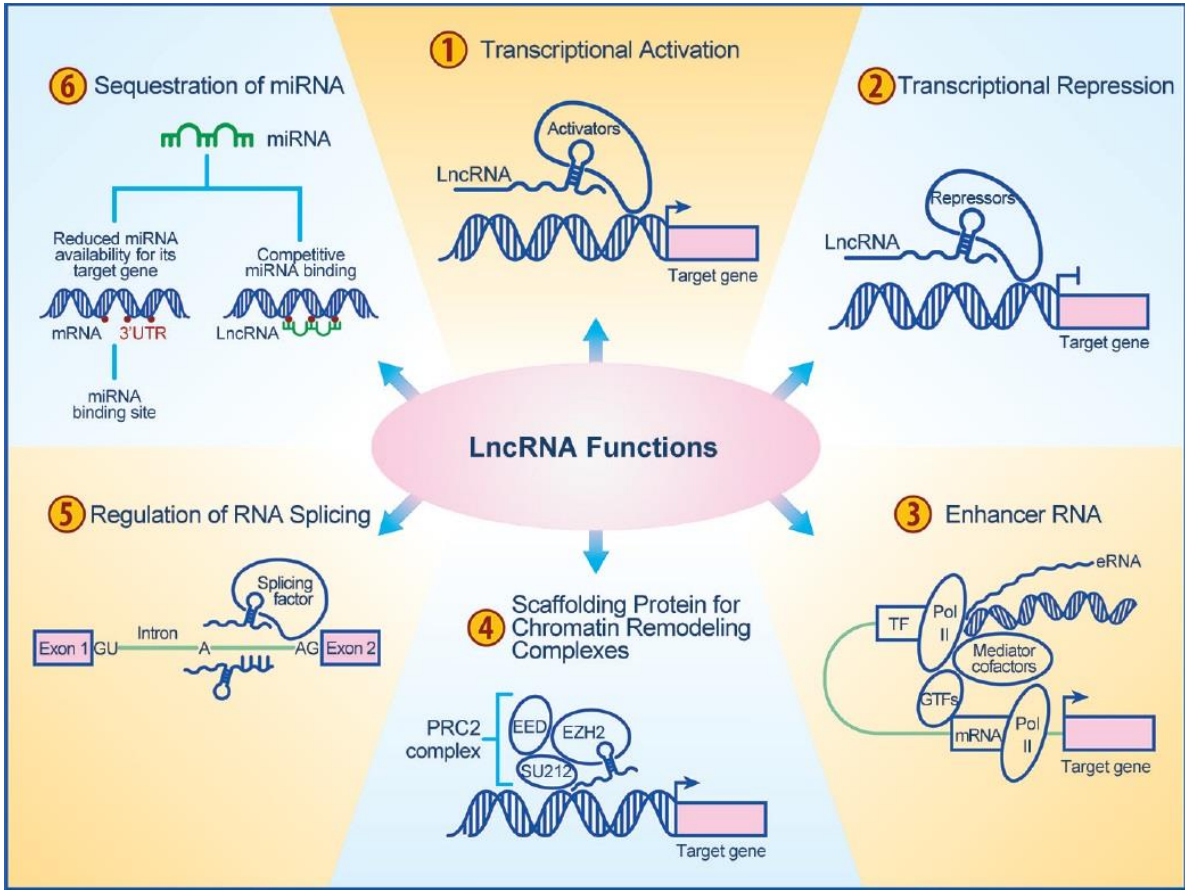
**What is your risk
for prostate cancer?**



1 in 9 men will develop prostate
cancer within their lifetime²

Prostate cancer diagnostics: present and future





Gleason Score	Rate of Cancer Growth
6 or Less	Average
3+4=7	Moderately Faster
4+3=7	Very Fast
8-10	Extremely Fast

Outline

- **Example: IncRNAs in prostate cancer in relation to Gleason score**
 - Input Data preparation
 - MA Plots
 - Dimensionality Reduction
 - PCA
 - t-SNE
 - Statistical distributions
 - Poisson
 - Negative binomial
 - Statistical significance
 - p -values
 - Significant genes

Prostate cancer – read counts

499 patients (samples)

500 genes (features)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	TCG	TCG	TCG	TCG	TCG	TCG	TCG	TCG	TCG	TCG	TCG	TCG	TCG	TCG	TCG	TCG	TCG	TCG	TCG	TCG
1 ENSG00000250529	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2 ENSG00000223985	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
3 ENSG00000257845	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
4 ENSG00000216560	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5 ENSG00000257378	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	1	0
6 ENSG00000226733	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
7 ENSG00000215231	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8 ENSG00000231210	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	2	0	0	0
9 ENSG00000281769	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
10 ENSG00000229005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11 ENSG00000258548	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
12 ENSG00000249396	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
13 ENSG00000276707	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14 ENSG00000278060	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0
15 ENSG00000228262	0	0	0	0	0	0	8	0	0	1	0	6	0	1	3	0	2	0	1	2
16 ENSG00000233639	0	0	0	0	0	0	3	0	0	0	0	2	0	0	2	0	0	0	0	0

Prostate cancer – clinical data

	1 sample_ID	2 sample_file	3 gleason
1	'TCGA-2A-A8VL-01A-21R-A37L-07'	'TCGA-2A-A8VL-01A-21R-A37L-07'	'low'
2	'TCGA-2A-A8VO-01A-11R-A37L-07'	'TCGA-2A-A8VO-01A-11R-A37L-07'	'low'
3	'TCGA-2A-A8VT-01A-11R-A37L-07'	'TCGA-2A-A8VT-01A-11R-A37L-07'	'high'
4	'TCGA-2A-A8VV-01A-11R-A37L-07'	'TCGA-2A-A8VV-01A-11R-A37L-07'	'low'
5	'TCGA-2A-A8VX-01A-11R-A37L-07'	'TCGA-2A-A8VX-01A-11R-A37L-07'	'high'
6	'TCGA-2A-A8W1-01A-11R-A37L-07'	'TCGA-2A-A8W1-01A-11R-A37L-07'	'high'
7	'TCGA-2A-A8W3-01A-11R-A37L-07'	'TCGA-2A-A8W3-01A-11R-A37L-07'	'high'
8	'TCGA-2A-AAYF-01A-11R-A41O-07'	'TCGA-2A-AAYF-01A-11R-A41O-07'	'low'
9	'TCGA-2A-AAYO-01A-11R-A41O-07'	'TCGA-2A-AAYO-01A-11R-A41O-07'	'low'
10	'TCGA-2A-AAYU-01A-11R-A41O-07'	'TCGA-2A-AAYU-01A-11R-A41O-07'	'low'
11	'TCGA-4L-AA1F-01A-11R-A41O-07'	'TCGA-4L-AA1F-01A-11R-A41O-07'	'high'
12	'TCGA-CH-5737-01A-11R-1580-07'	'TCGA-CH-5737-01A-11R-1580-07'	'high'
13	'TCGA-CH-5738-01A-11R-1580-07'	'TCGA-CH-5738-01A-11R-1580-07'	'low'
14	'TCGA-CH-5739-01A-11R-1580-07'	'TCGA-CH-5739-01A-11R-1580-07'	'low'
15	'TCGA-CH-5740-01A-11R-1580-07'	'TCGA-CH-5740-01A-11R-1580-07'	'low'
16	'TCGA-CH-5741-01A-11R-1580-07'	'TCGA-CH-5741-01A-11R-1580-07'	'high'
17	'TCGA-CH-5743-01A-21R-1580-07'	'TCGA-CH-5743-01A-21R-1580-07'	'low'
18	'TCGA-CH-5744-01A-11R-1580-07'	'TCGA-CH-5744-01A-11R-1580-07'	'high'
19	'TCGA-CH-5745-01A-11R-1580-07'	'TCGA-CH-5745-01A-11R-1580-07'	'low'
20	'TCGA-CH-5746-01A-11R-1580-07'	'TCGA-CH-5746-01A-11R-1580-07'	'low'

Dimensionality Reduction

- Prostate cancer count table
 - 500 features
 - 499 samples
- Lot of features to handle in transcriptomic data
- Example: Pasilla dataset
 - 14,599 features
 - 7 samples
- For visualization and analysis purposes it is good to reduce the dimensionality. Possible techniques:
 - PCA
 - t-SNE

PCA

- **Principal component analysis (PCA)** is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.
- This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.
- The resulting vectors (each being a linear combination of the variables and containing n observations) are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

PCA – MATLAB

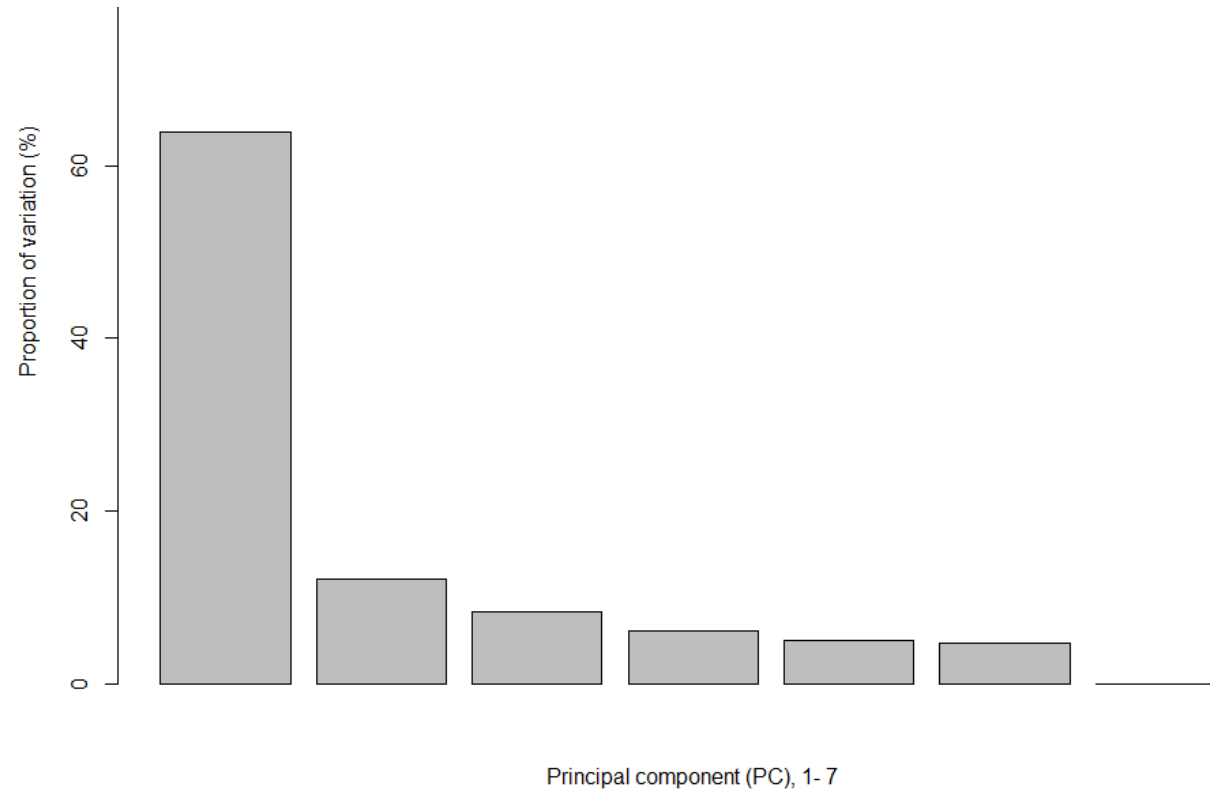
```
coeff = pca(X)
```

- returns the principal component coefficients, also known as loadings, for the n -by- p data matrix X .
- Rows of X correspond to observations and columns correspond to variables. There are n observations and p variables.
- The coefficient matrix is p -by- p . Each column of `coeff` contains coefficients for one principal component, and the columns are in descending order of component variance. By default, `pca` centers the data and uses the singular value decomposition (SVD) algorithm.

PCA – Scree plot

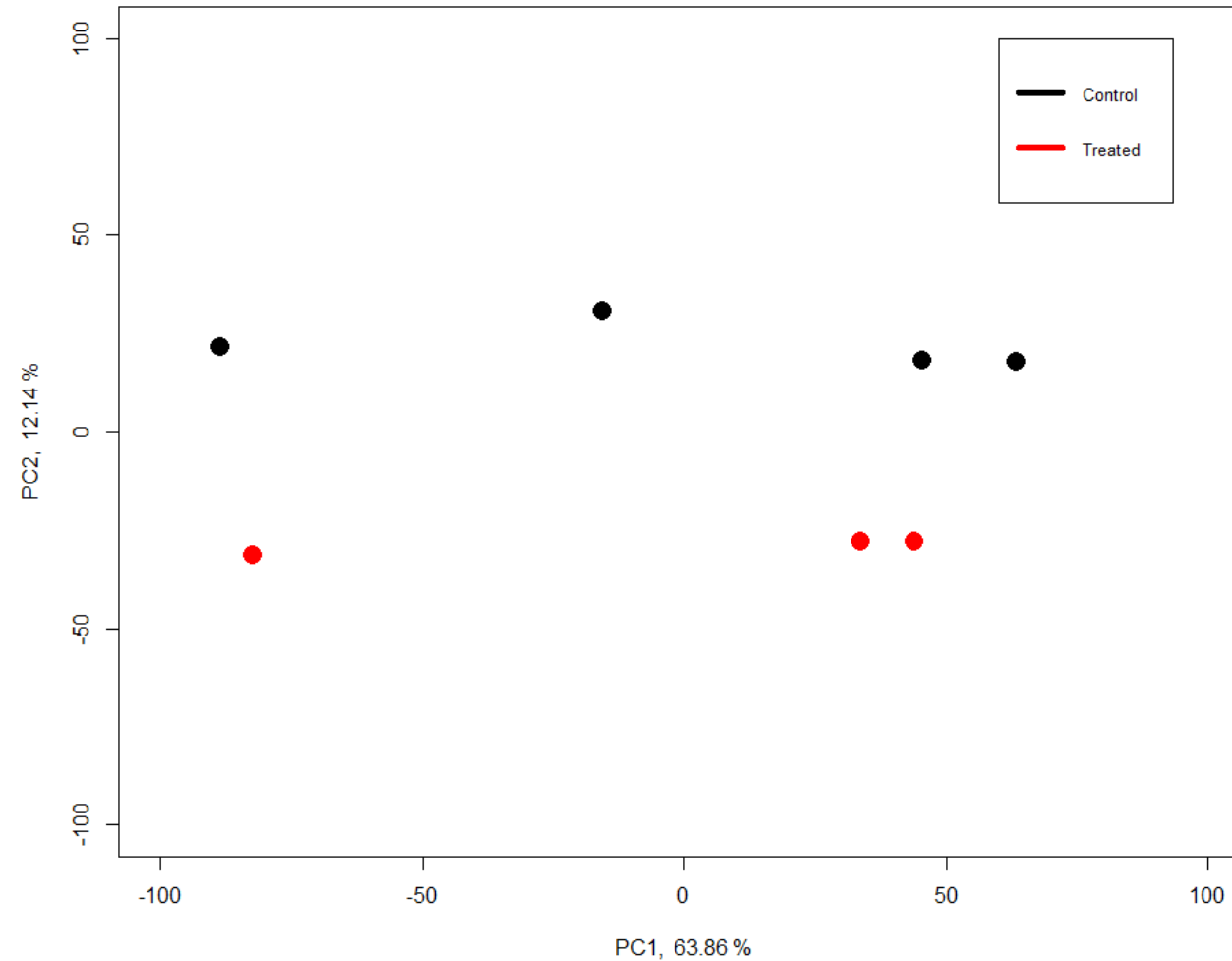
Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	63.3517	27.6242	22.87721	19.44818	17.72155	17.10358	2.966e-13
Proportion of Variance	0.6386	0.1214	0.08328	0.06018	0.04997	0.04655	0.000e+00
Cumulative Proportion	0.6386	0.7600	0.84330	0.90348	0.95345	1.00000	1.000e+00



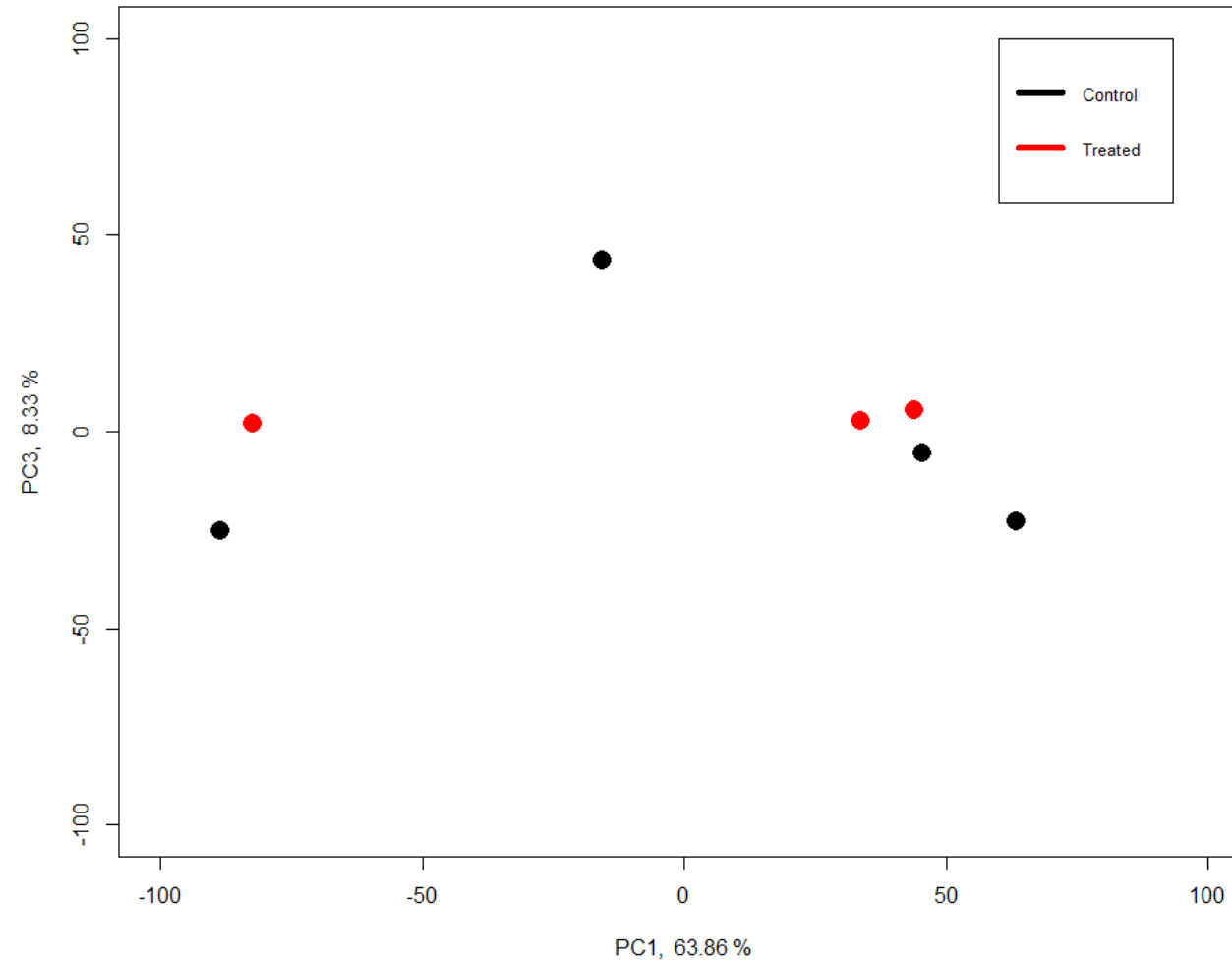
PCA – bi-plot PC1 and PC2

Principal components analysis bi-plot



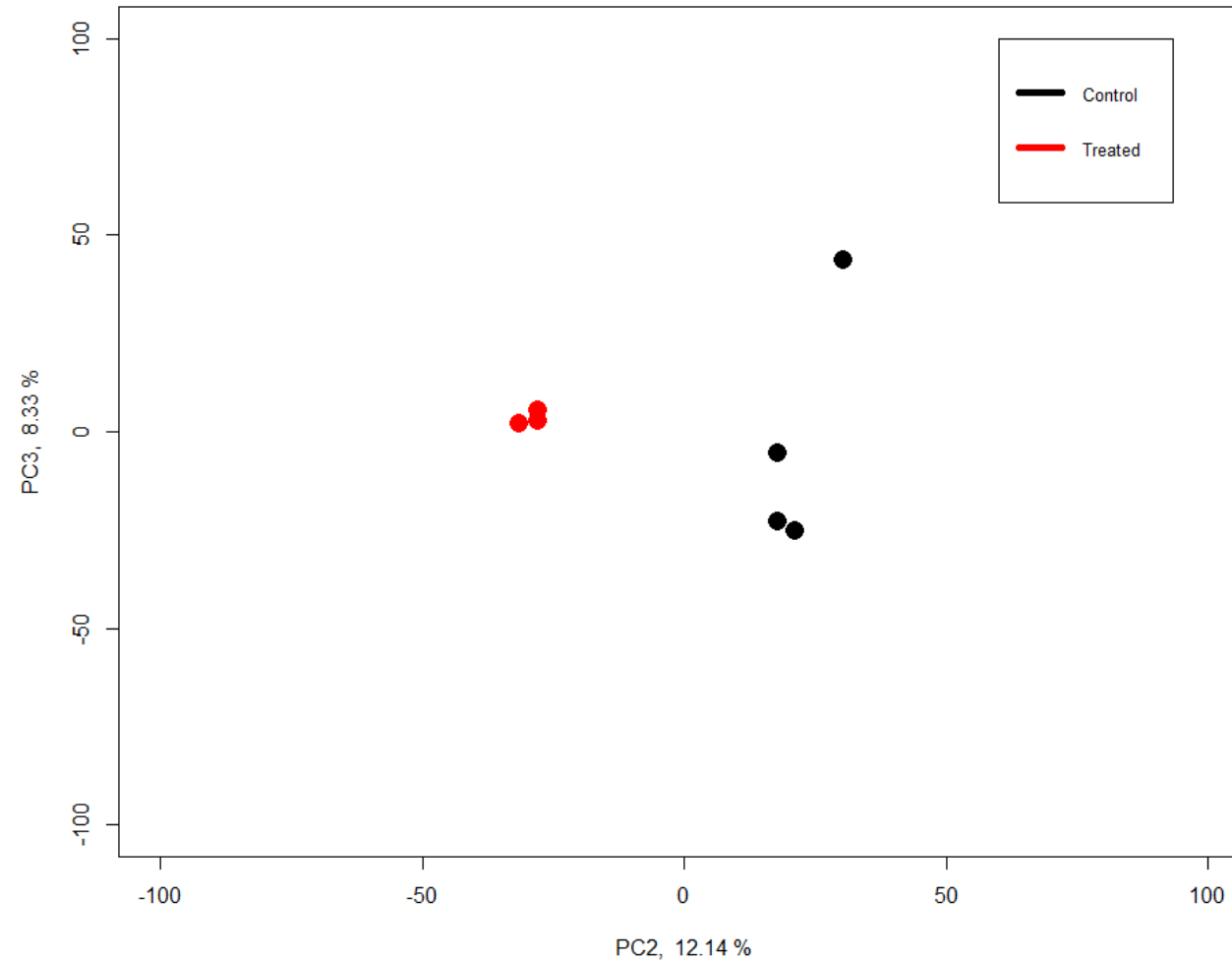
PCA – bi-plot PC1 and PC3

Principal components analysis bi-plot

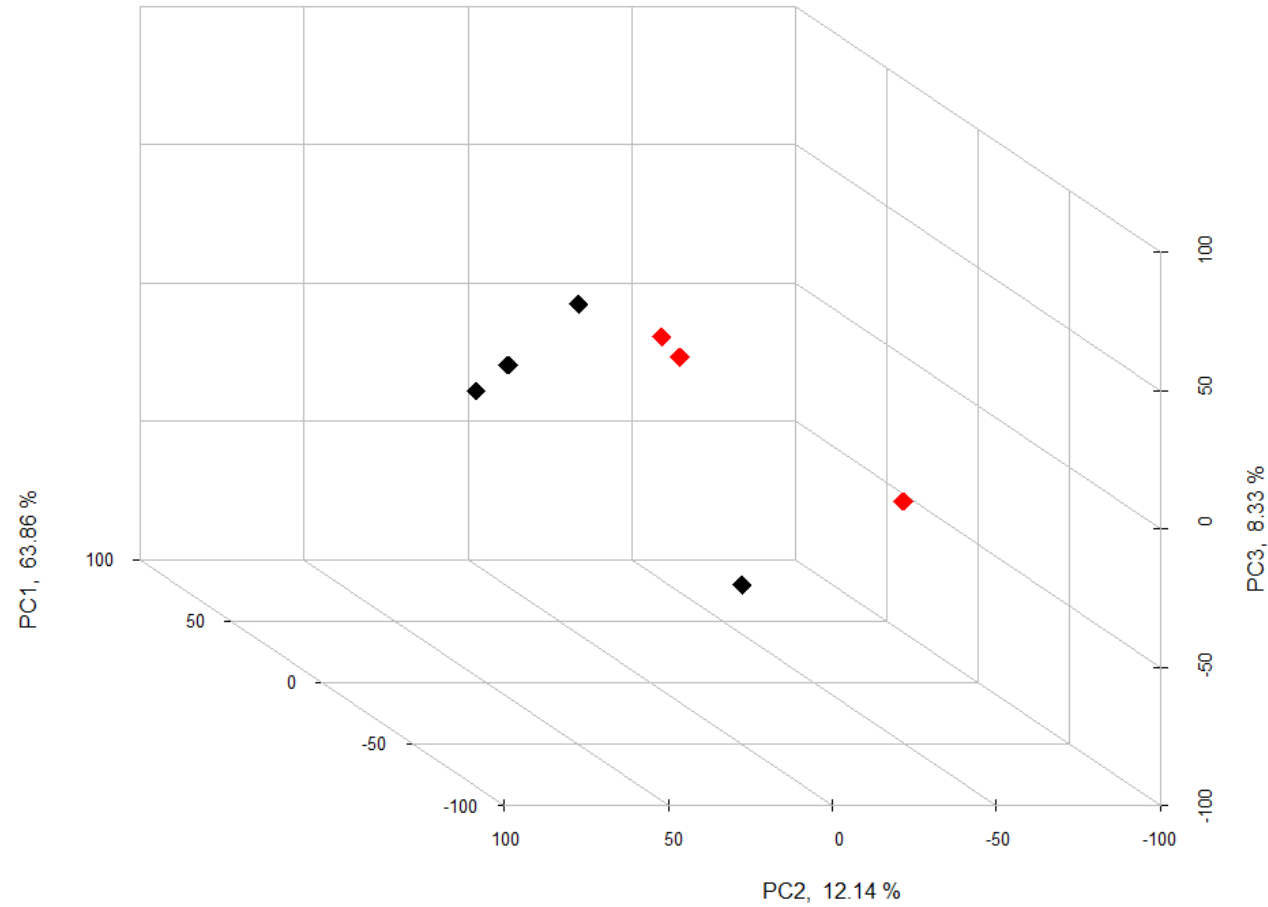


PCA – bi-plot PC2 and PC3

Principal components analysis bi-plot



PCA – tri-plot PC1, PC2 and PC3



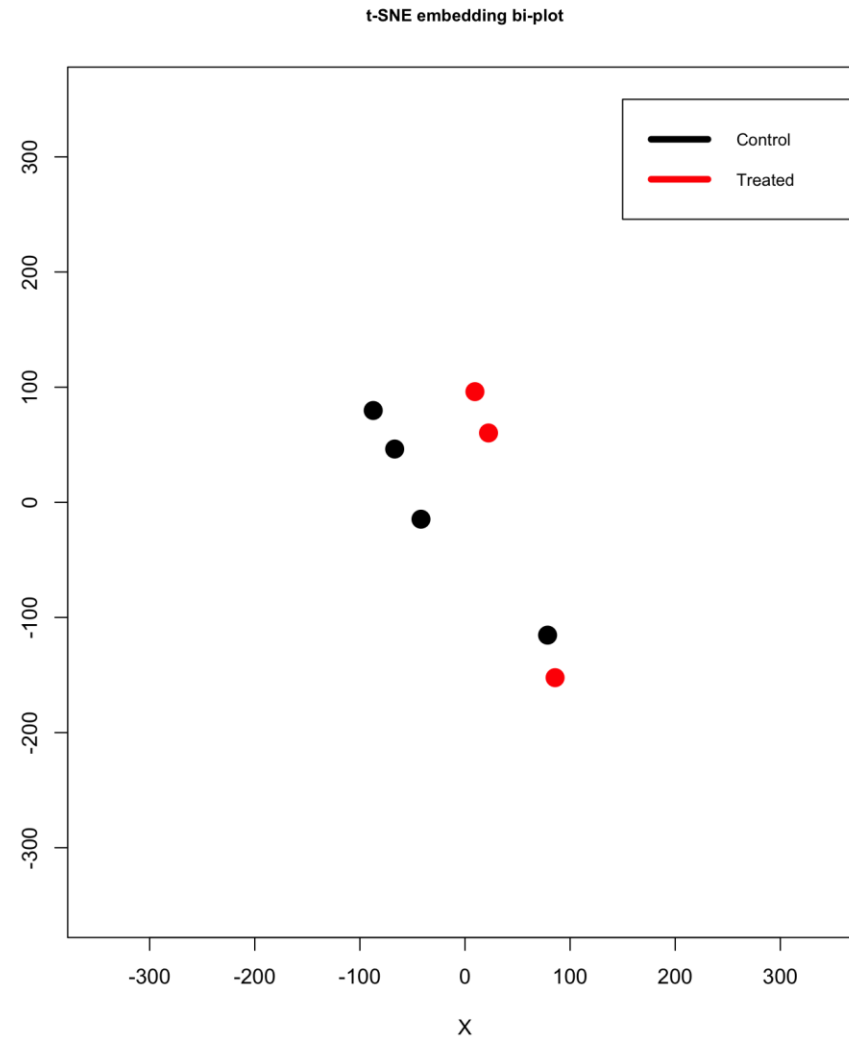
t-SNE

- It is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of 2 or 3 dimensions.
- Specifically, it models each high-dimensional object by a 2D or 3D point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.

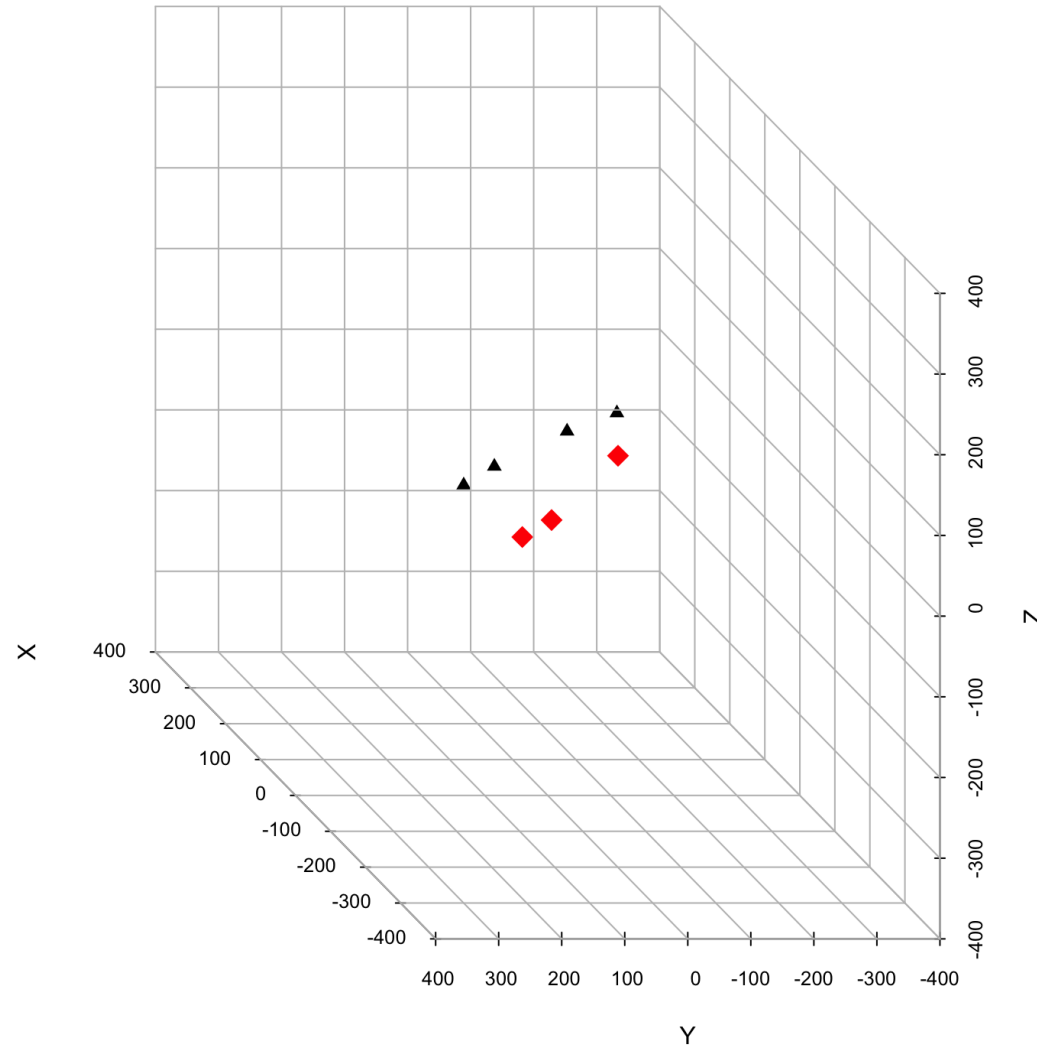
t-SNE

- The t-SNE algorithm comprises two main stages:
 1. t-SNE constructs a probability distribution over pairs of high-dimensional objects in such a way that similar objects have a high probability of being picked while dissimilar points have an extremely small probability of being picked.
 2. t-SNE defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the Kullback–Leibler divergence between the two distributions with respect to the locations of the points in the map.
- Hyper-parameter: **perplexity**
 - It is basically the effective number of neighbors for any point, and t-SNE works relatively well for any value between 5 and 50. Larger perplexities will take more global structure into account, whereas smaller perplexities will make the embeddings more locally focused.

t-SNE bi-plot



t-SNE tri-plot



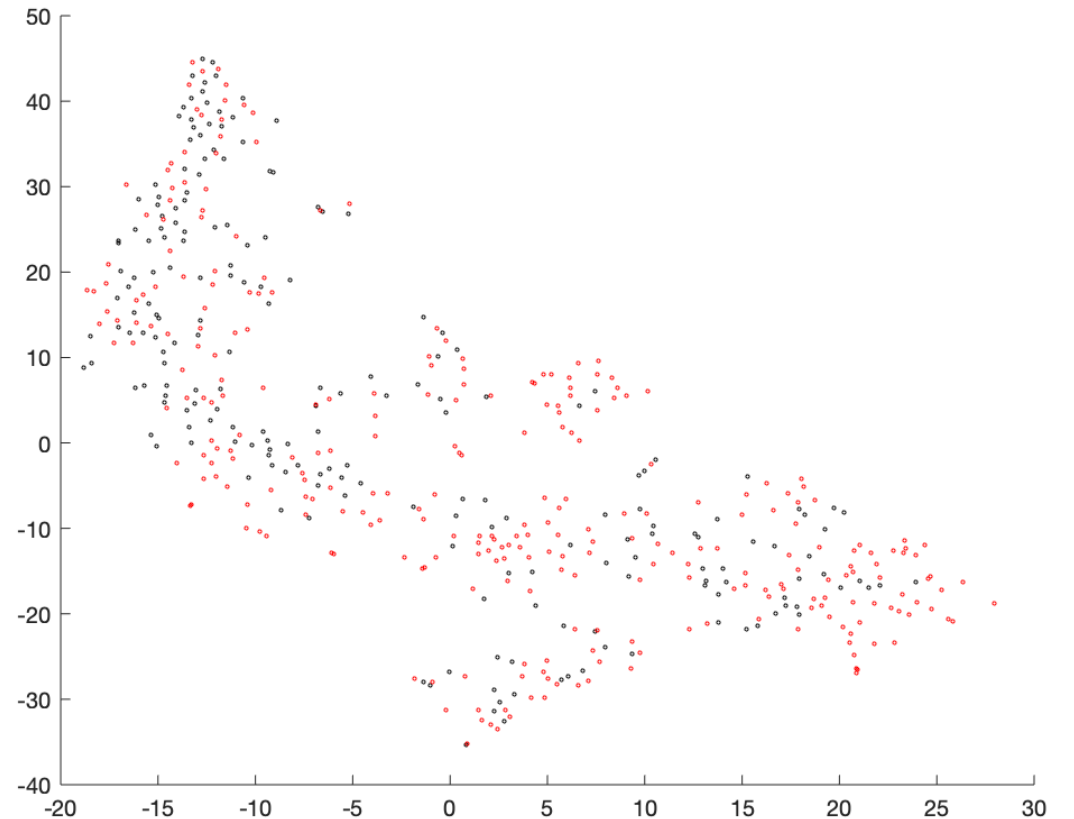
t-SNE – MATLAB

$Y = \text{tsne}(X)$

returns a matrix of two-dimensional embeddings of the high-dimensional rows of X .

X – Data points specified as an n -by- m matrix, where each row is one m -dimensional point.

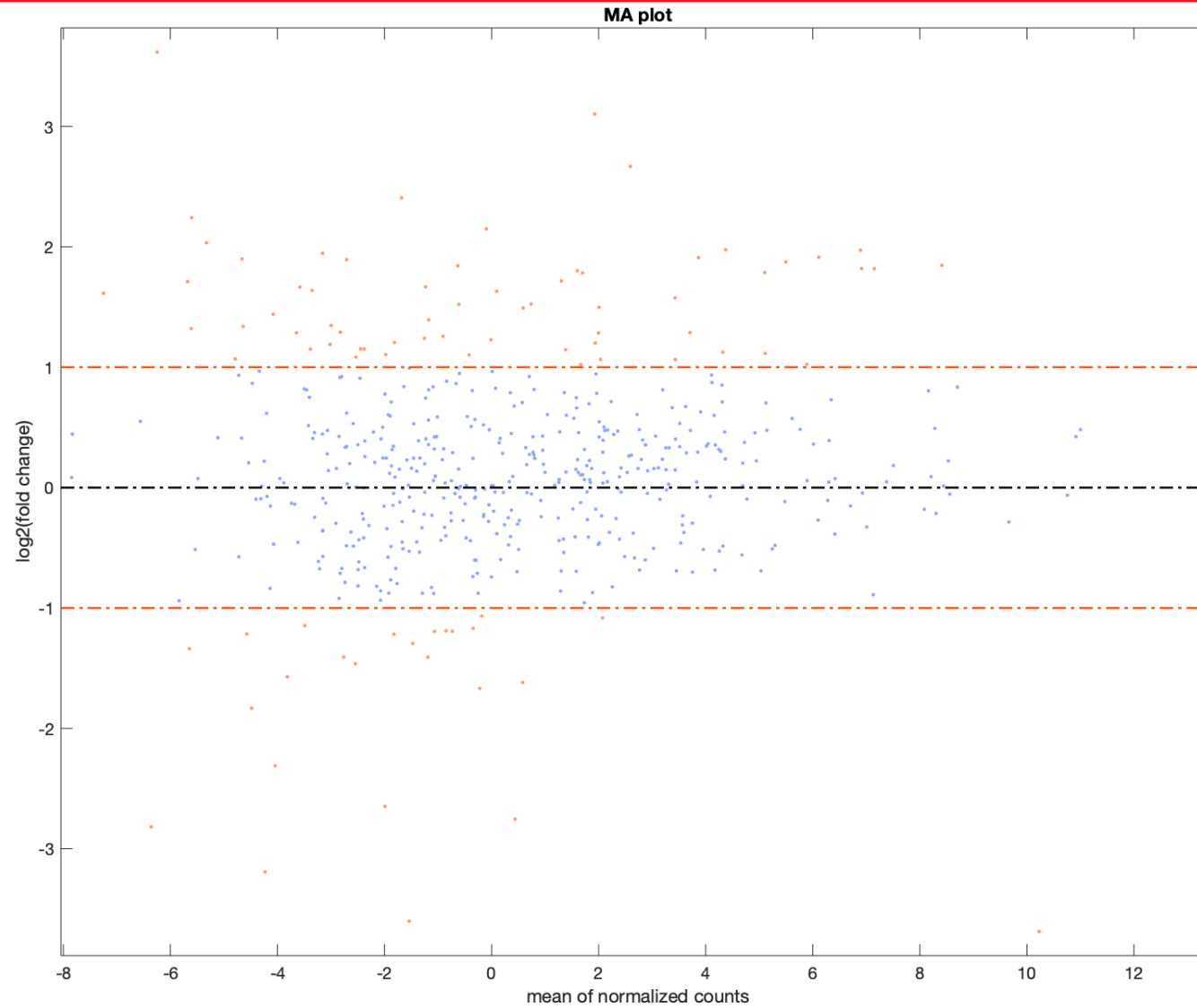
Y – Embedded points returned as an n -by- NumDimensions matrix. Each row represents one embedded point.



Fold Change

- Fold change is a measure describing how much a quantity changes between an original and a subsequent measurement.
- It is defined as the ratio between the two quantities; for quantities A and B, then the fold change of B with respect to A is B/A .
- In our case, we define Fold Change as:
$$\text{foldChange} = \text{meanHigh} ./ \text{meanLow};$$
- You can look at the difference of the gene expression among two conditions, by calculating the fold change (FC) for each gene, i.e. the ratio between the counts in the **high gleason** group over the counts in the **low gleason** group.
- Generally these ratios are considered in the \log_2 scale, so that any change is symmetric with respect to zero.
- A ratio of $1/2$ or $2/1$ corresponds to -1 or $+1$ in the log scale.

MA plot



Hypothesis Testing

- Based on a count table, we want to detect differentially expressed genes between different conditions.
 - How can we detect genes for which the counts of reads change between conditions more systematically than as expected by chance?
- We would like to use statistical testing to decide whether, for a given gene, an observed difference in read counts is significant, that is, whether it is greater than what would be expected just due to natural random variation.
- Null hypothesis H_0 :
 - the gene g is not differentially expressed between the conditions
- Alternative hypothesis H_1 :
 - the gene g is differentially expressed between the conditions

Hypothesis Testing

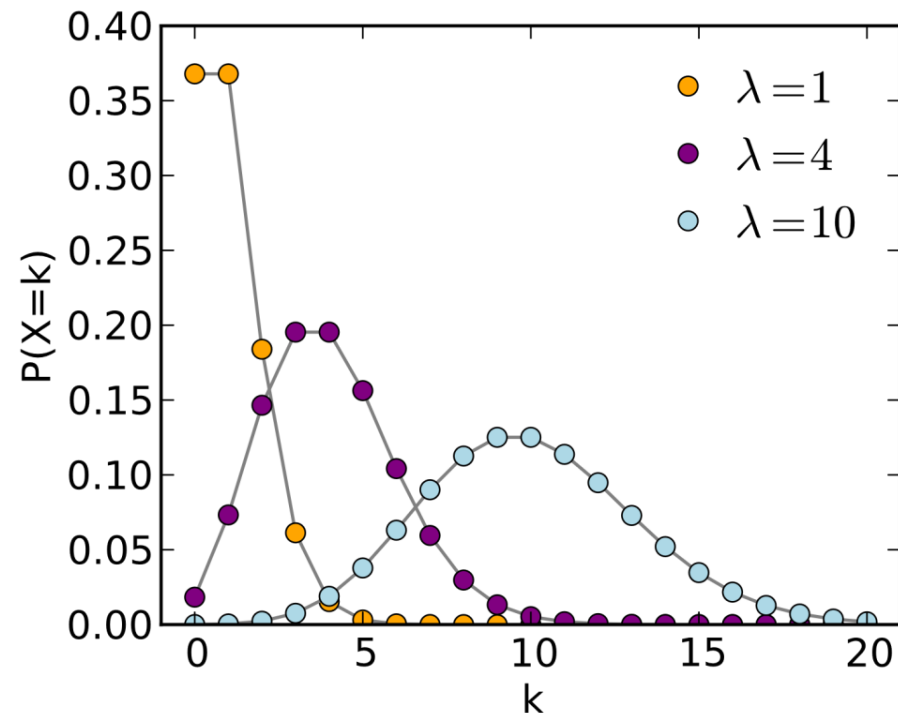
- How to quantify the difference?
- The statistical tests do not give a simple answer of whether the hypothesis is true or not. What a statistical test determines is how likely that null hypothesis is to be true.
- After a test statistic is computed, it is often converted to a p -value. Then the difference is quantified in terms of the p -value.
- If the p -value is small then the null hypothesis is deemed to be untrue and it is rejected in favour of the alternative.
- The p -value is the probability of seeing a result as extreme or more extreme than the observed data, when the null hypothesis is true.
- It is a usual convention in biology to use a critical p -value of 0.05.

Type of errors in tests

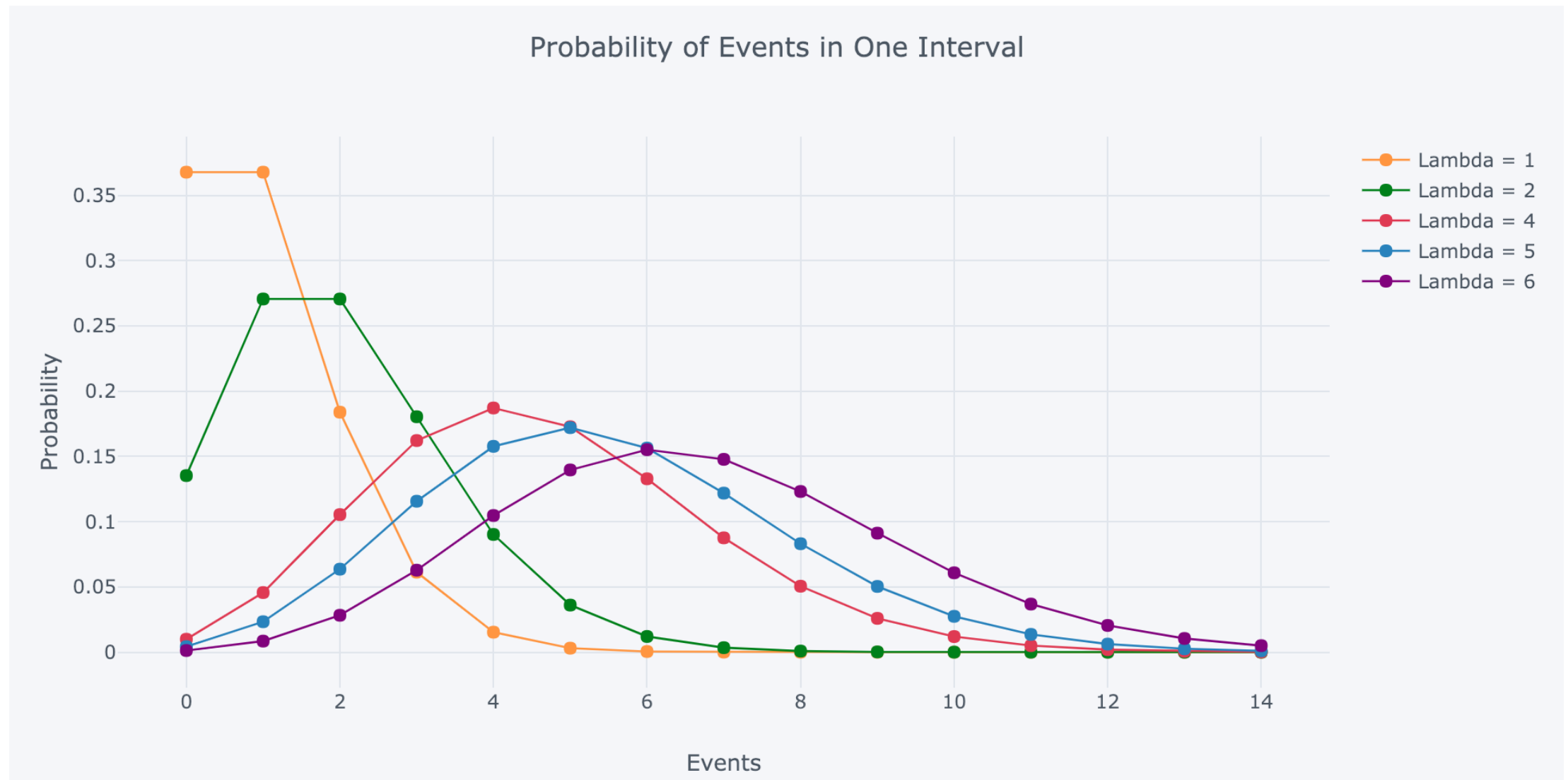
Table of error types		Null hypothesis (H_0) is	
		True	False
Decision about null hypothesis (H_0)	Don't reject	Correct inference (true negative) (probability = $1 - \alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1 - \beta$)

Poisson distribution

- The **Poisson distribution** is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event.



Poisson distribution



Poisson distribution

- We denote Poisson distribution with

$$Pois(\lambda)$$

- $\lambda \in \mathbb{R}^+$ is the *rate*

- **Probability mass function (pmf)**

$$f(k; \lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- **Mean**

$$\mu = \lambda$$

- **Variance**

$$\sigma^2 = \lambda$$

Negative binomial distribution

- The **negative binomial distribution** is a discrete probability distribution of the number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of failures (denoted r) occurs.
- For example, if we define a 1 as failure, all non-1s as successes, and we throw a dice repeatedly until 1 appears the 3rd time ($r = 3$ failures), then the probability distribution of the number of non-1s that appeared will be a negative binomial distribution.

Negative binomial distribution

- We denote negative binomial distribution with:

$$NB(r, p)$$

- $r > 0$ is the number of failures until the experiment is stopped
- $p \in (0,1)$ is the probability of success for each experiment
- k is the number of successes

- **Probability mass function (pmf)**

$$f(k; r, p) = \Pr(X = k) = \binom{k + r - 1}{k} (1 - p)^r p^k$$

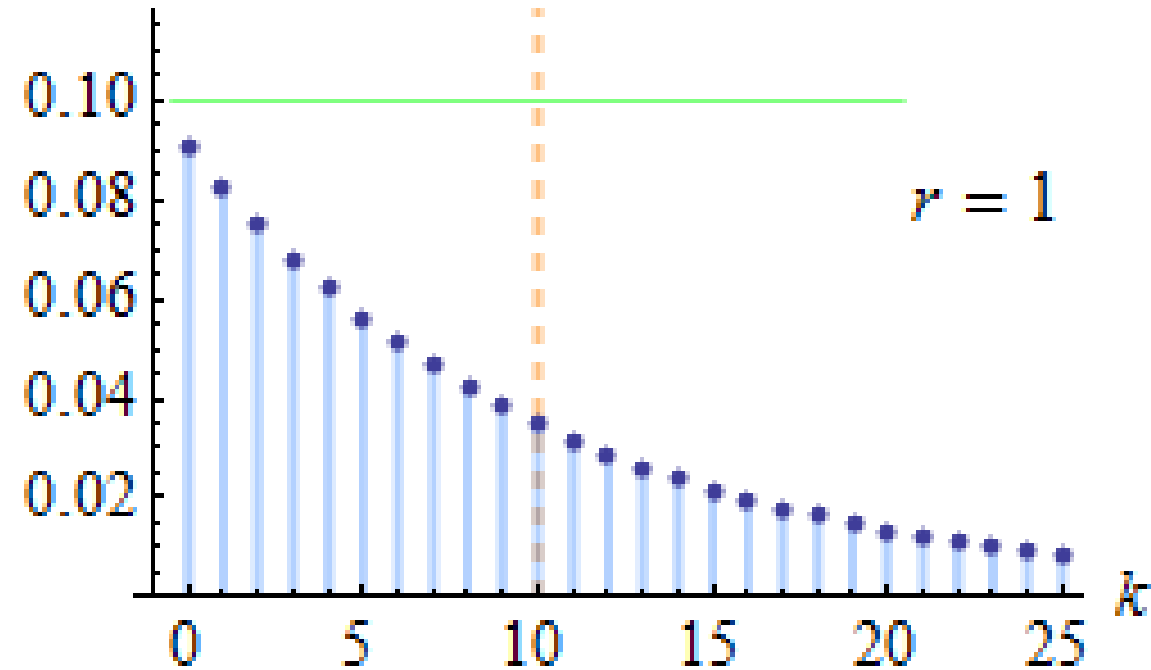
- **Mean**

$$\mu = \frac{pr}{1 - p}$$

- **Variance**

$$\sigma^2 = \frac{pr}{(1 - p)^2} = \mu + \frac{\mu^2}{r}$$

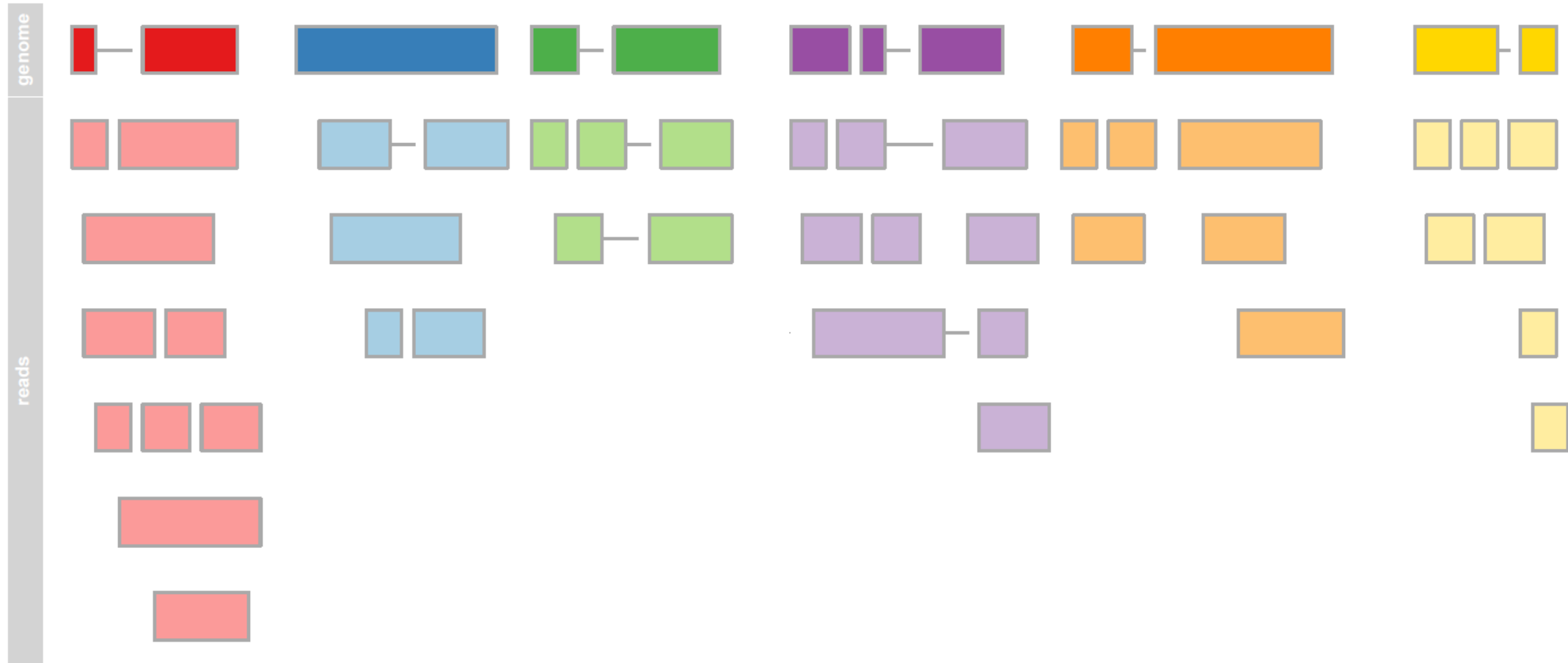
Negative binomial distribution



Modeling count data



Modeling count data



Modeling count data

K_g : sequence reads assigned to a particular region g

p_g : proportion of DNA fragments arising from the region g

$K_g = 0, 1, 2, \dots, n$ ($n + 1$ possible discrete values)

n : total number of sequenced reads

We want to estimate $\Pr(K_g = k)$

If there are k aligned reads to region g , then must be $n - k$ not aligned to region g .

The probability of this is given by: $p^k(1 - p)^{n-k}$

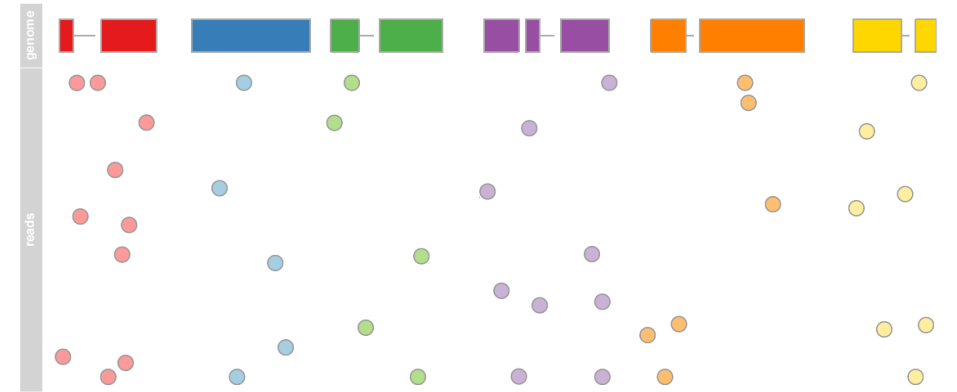
The number of ways of arranging k successes in n trials is:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Then, the probability of k successes in n independent trials is:

$$P(K_g = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

This is the Binomial distribution, which we denote $K_g \sim \text{Bin}(n, p)$

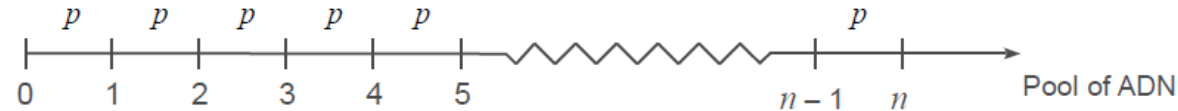


Modeling count data

Poisson approximation

As the number of sequenced reads becomes large and the probability p shrinks, the Binomial distribution can be approximated by the Poisson distribution:

- split the pool of DNA fragments into a large number of n intervals, each with a small probability p of success and such that the total number of successes $K_g \sim \text{Bin}(n, p)$



- the number of successes, occurring in the interval $[0, n]$, occur independently and at a constant average rate $\lambda = np$. Then

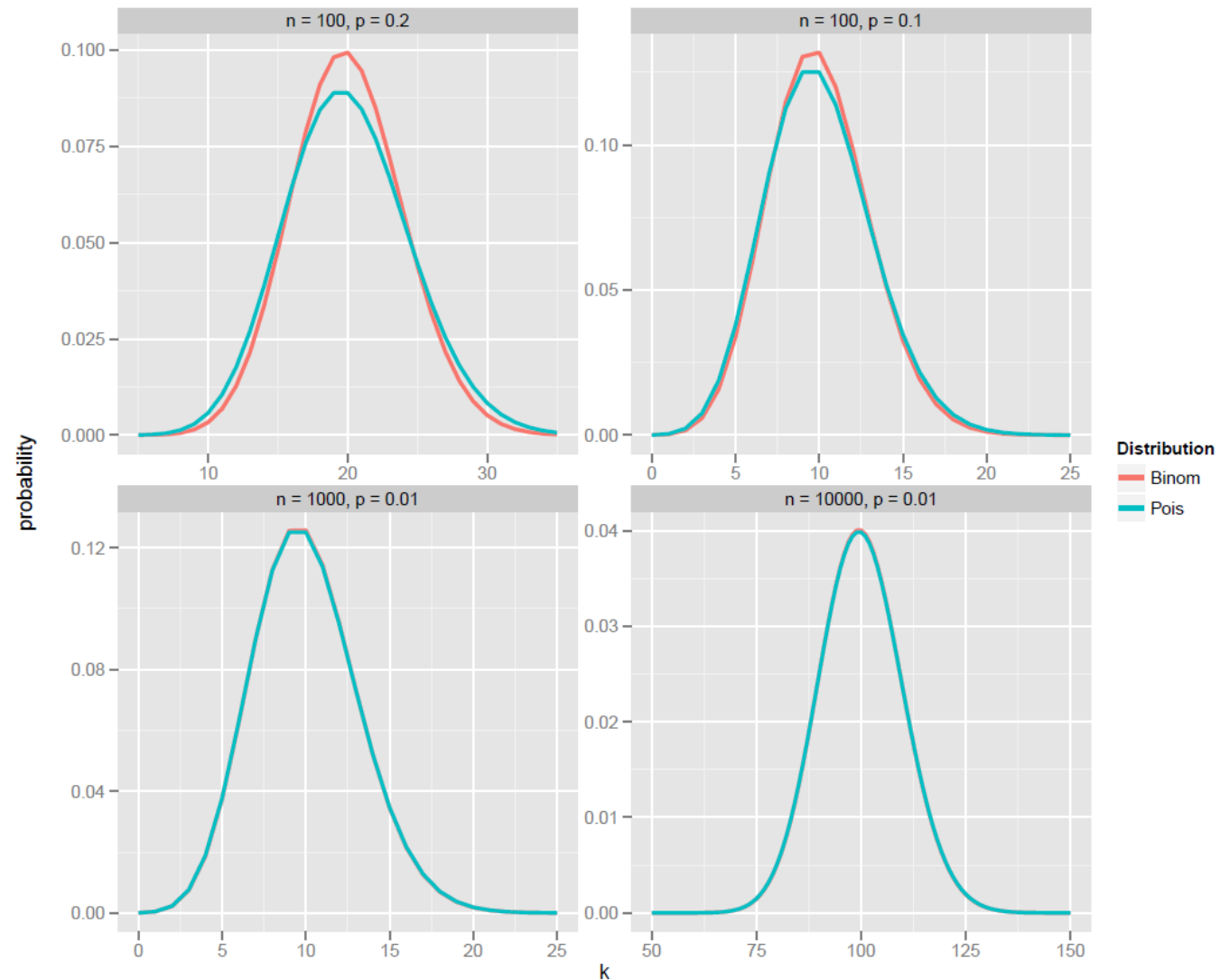
$$\begin{aligned}\lambda &\rightarrow np \\ n &\rightarrow \infty \\ p &\rightarrow 0\end{aligned}$$

$$\text{Bin}(n, p) \rightarrow \text{Pois}(\lambda)$$

$$\begin{aligned}P(K_g = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &\approx \frac{\lambda^k}{k!} e^{-\lambda}\end{aligned}$$

This is the **Poisson distribution** with parameters λ . We write $K_g \sim \text{Pois}(\lambda)$.

Modeling count data



Modeling count data

When the variance in a Poisson model is greater than the mean, the counts are said to be “overdispersed” with respect to a Poisson distribution.

To model overdispersed counts, the Poisson distribution model can be modified as

$$\mathcal{L}(K_g|\epsilon) \sim \text{Pois}(\theta), \quad \text{with } \theta = \lambda\epsilon$$

where ϵ is a nonnegative multiplicative random-effect term to model individual heterogeneity (Winkelmann 2008).

By placing a gamma distribution (see Box 2) prior on ϵ with $\alpha = \beta = r$, $\epsilon \sim \text{Gamma}(r, r)$, we have a negative binomial distribution for K_g , $K_g \sim \text{NB}(\alpha, p)$, parameterized by probability parameter $p = \lambda/(\lambda + r)$ and dispersion parameter $\alpha = r$ (see Box 3). The mean and variance is given by:

$$E(K_g) = \lambda \quad \text{and} \quad \text{Var}(K_g) = \lambda + \phi\lambda^2$$

where $\phi = 1/r$ is the inverse dispersion parameter. Thus $\text{Var}(K_g) \geq E(K_g)$ and we obtain a model for overdispersed counts. Notice that, as ϕ decreases to 0, the variance of K_g approaches the usual Poisson variance λ (i.e. np_g).

Biological noise: overdispersion

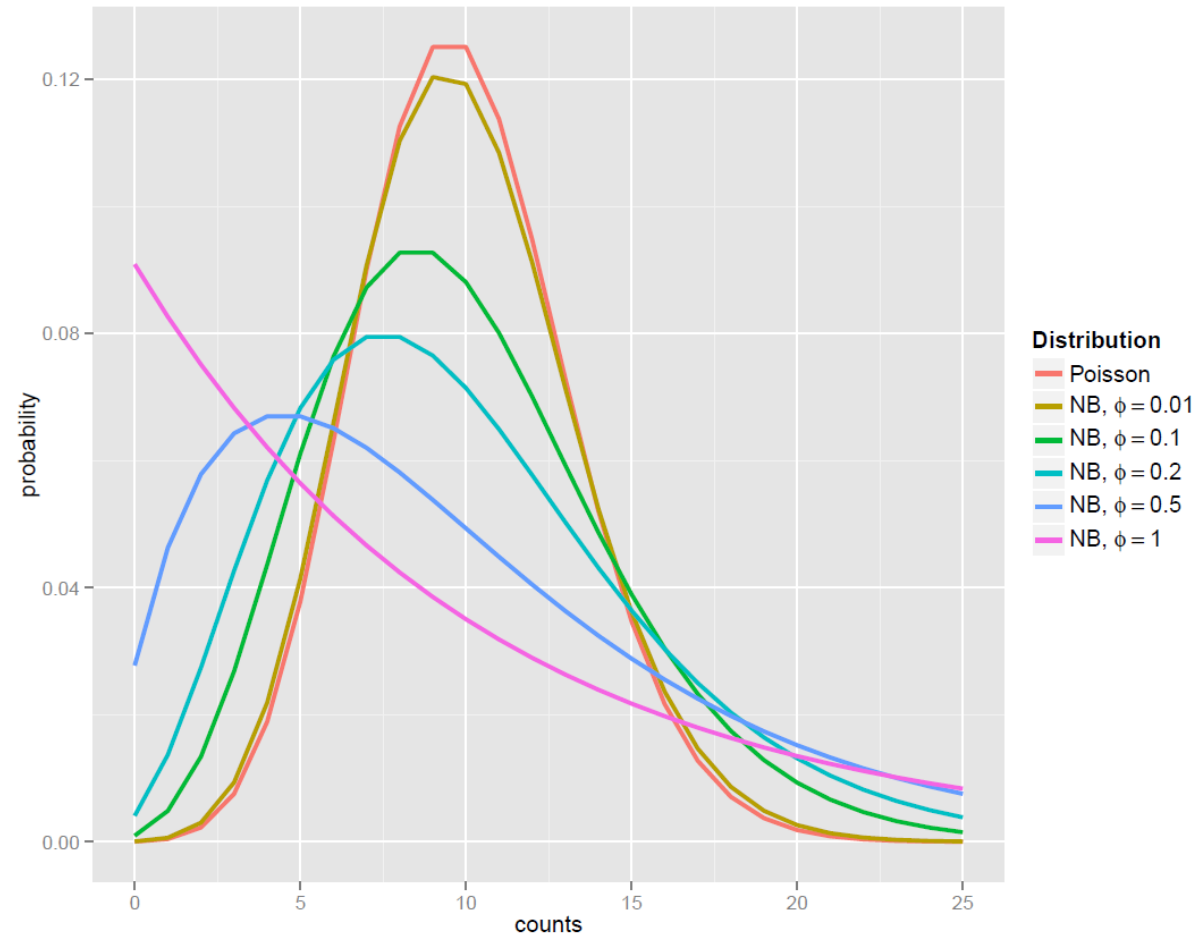


Figure 24: Poisson and negative binomial distributions, all with a mean value equal to 10. As the dispersion parameter, ϕ goes to zero, the negative binomial distribution converges to a Poisson. The discrete probabilities are joined by lines for easing the visualization.

Bioinformatics Toolbox – `nbintest()`

- `test = nbintest(X, Y)` performs a hypothesis test that two independent samples of short-read count data, in each row of `X` and `Y`, come from distributions with equal means under the assumptions that:
 - Short-read counts are modeled using the negative binomial distribution.
 - Variance and mean of data in each row are linked through a regression function along all the rows.
- `test` is a `NegativeBinomialTest` object with two-sided p -values stored in the `pValue` property.
- Use this function when you want to perform an unpaired hypothesis test for short-read count data (from high-throughput assays such as RNA-Seq or ChIP-Seq) with small sample sizes (in the order of tens at most). For instance, use this function to decide if observed differences in read counts between two conditions are significant for given genes.

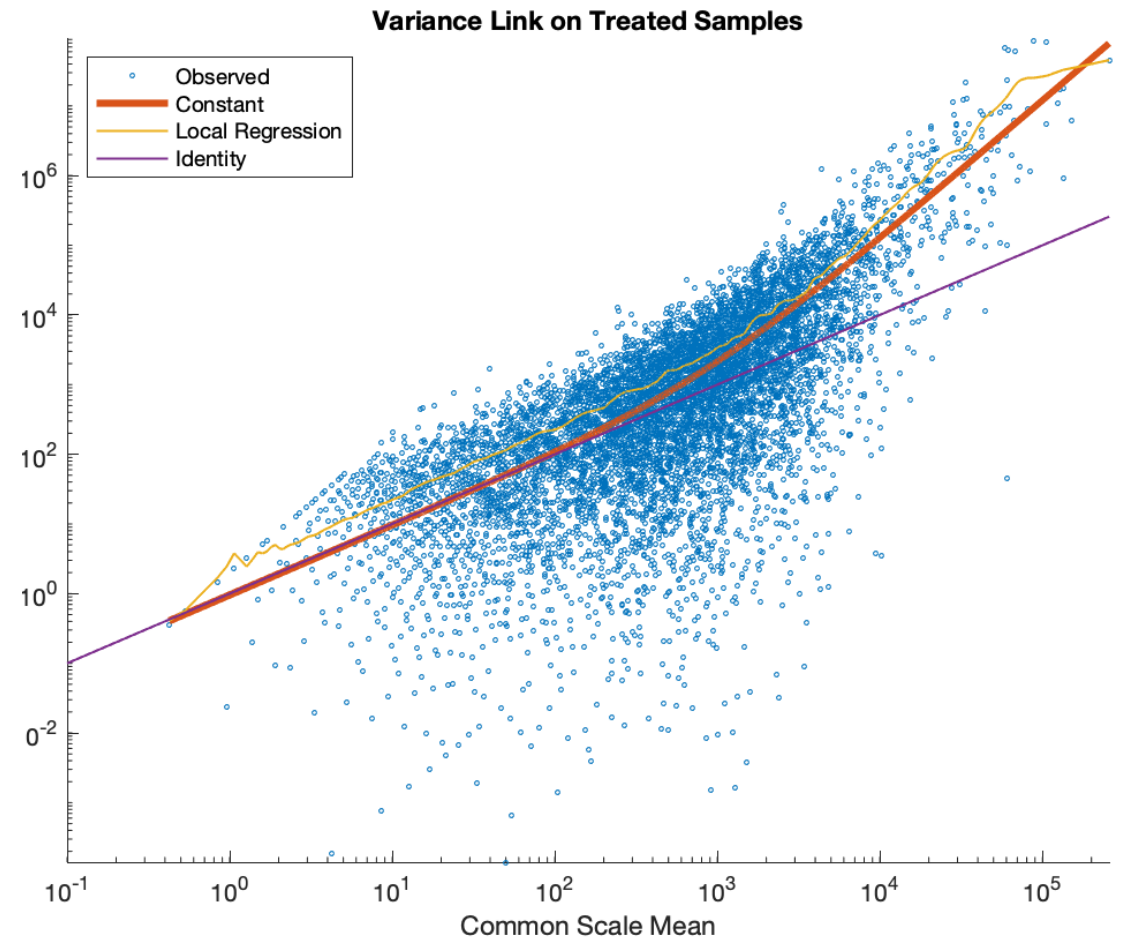
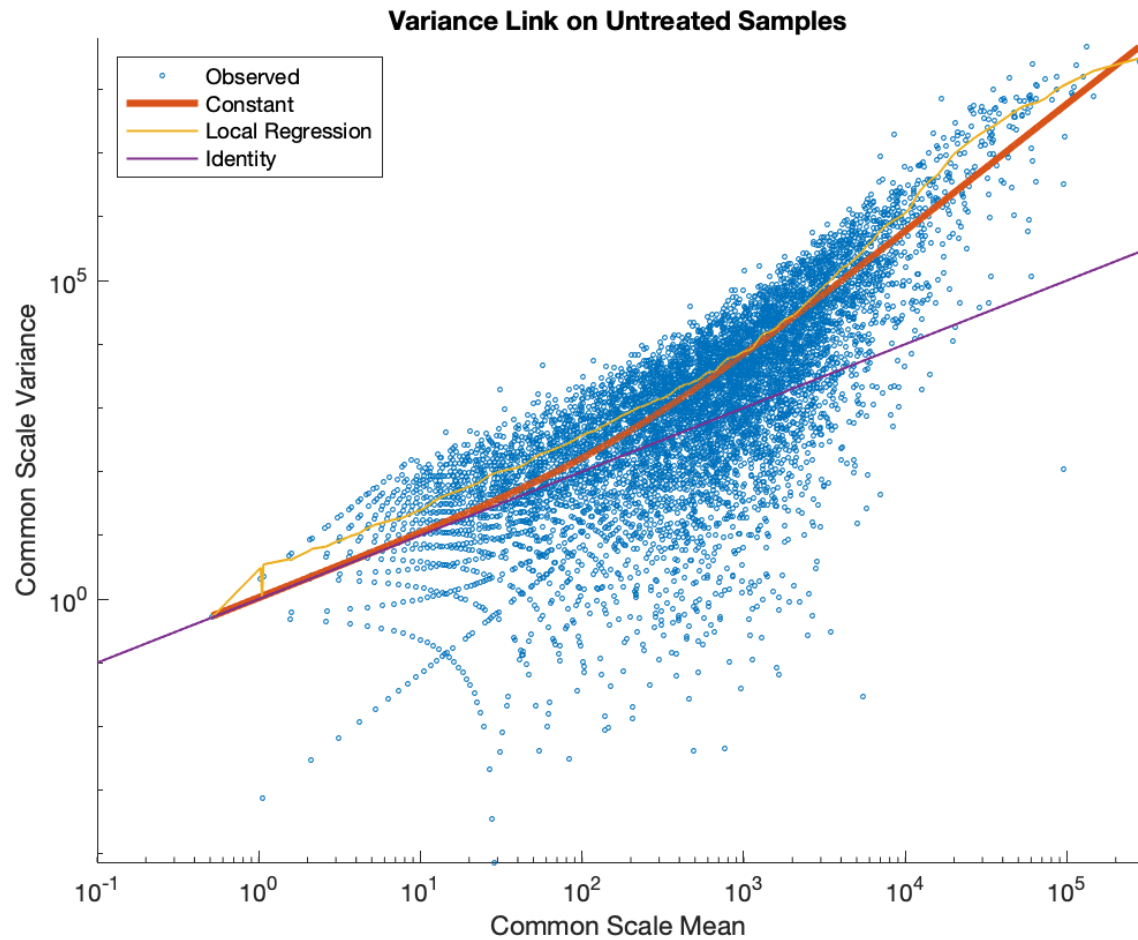
Bioinformatics Toolbox – `nbintest()`

'VarianceLink'

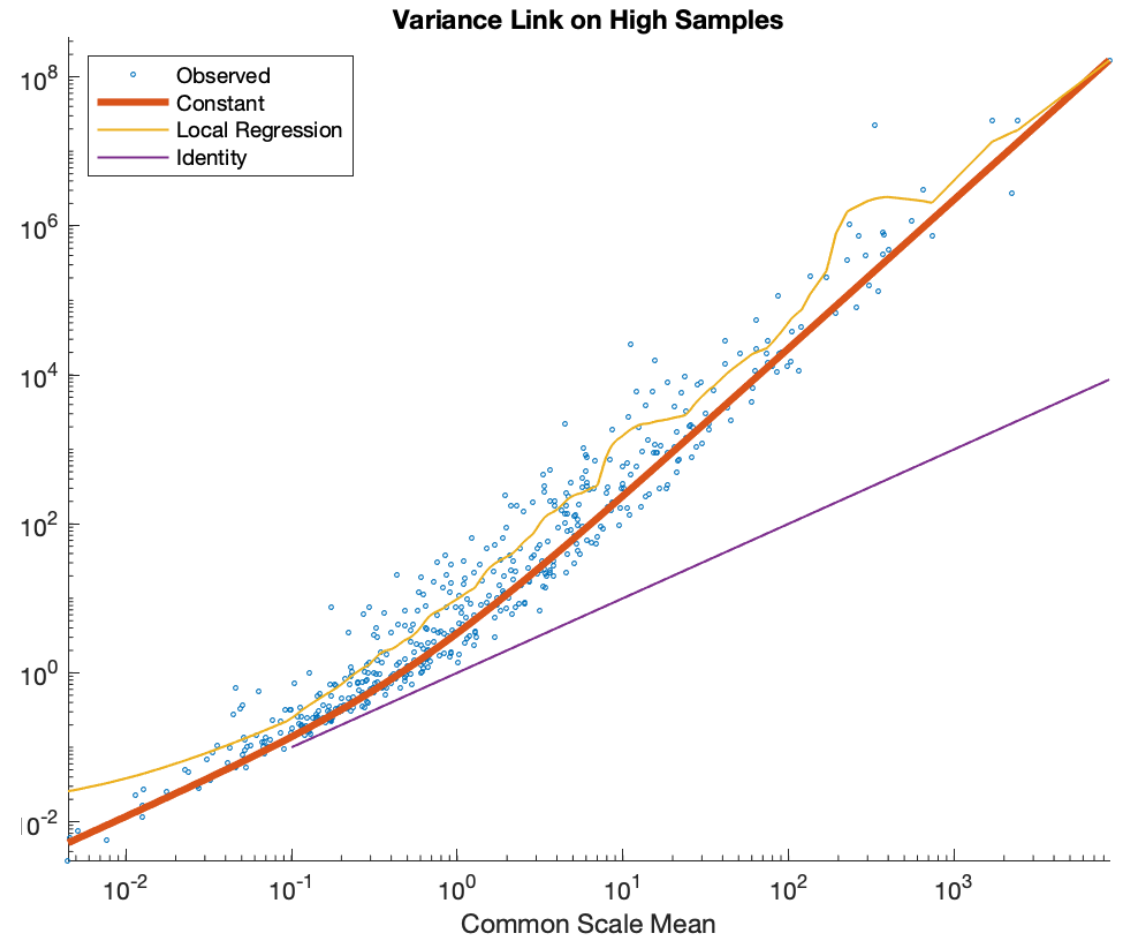
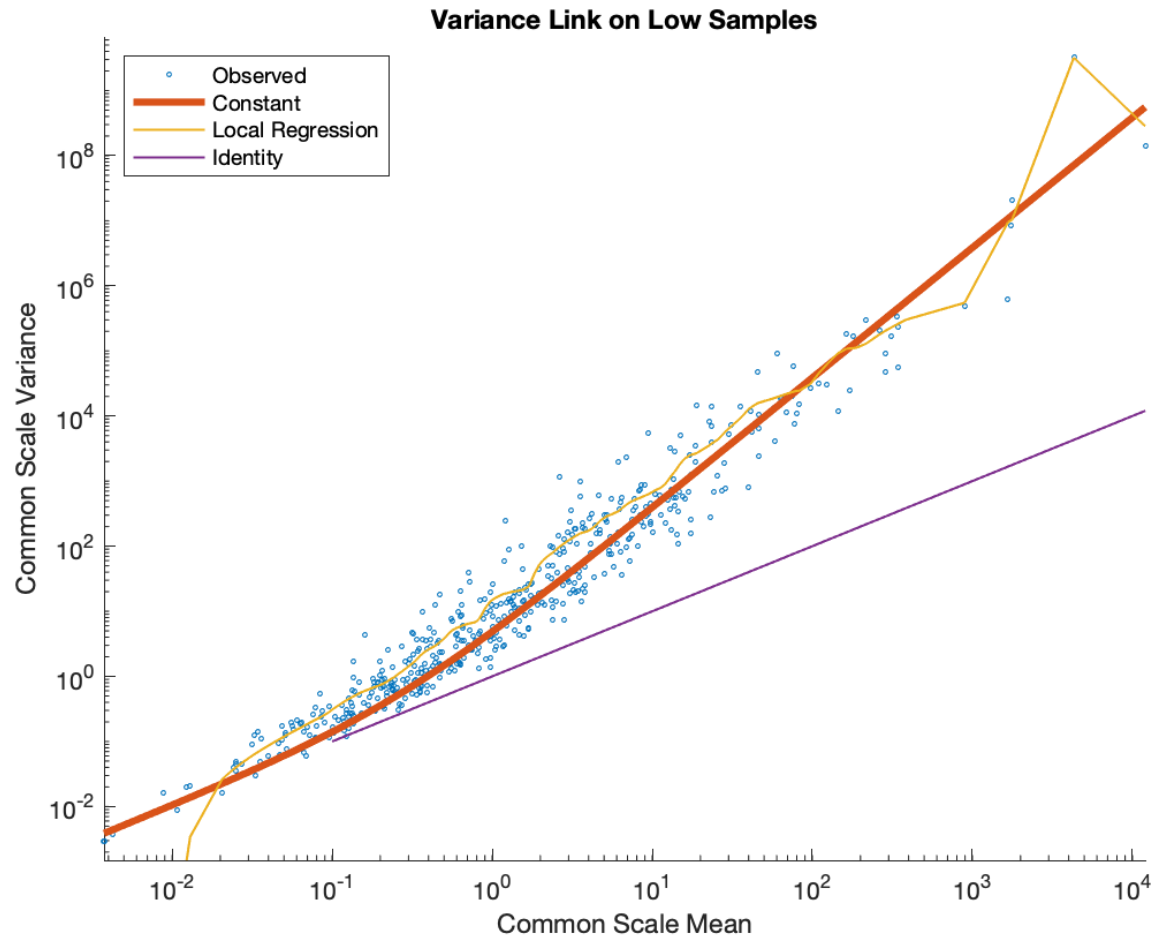
Linkage type between the variance and mean

- 'LocalRegression' The variance is the sum of the shot noise term (mean) and a locally regressed nonparametric smooth function of the mean as described in. This option is the default. Use this option if your data is overdispersed and has more than 1000 rows (genes).
- 'Constant' The variance is the sum of the shot noise term (mean) and a constant multiplied by the squared mean. This method uses all the rows in the data to estimate the constant. Use this option if your data is overdispersed and has less than 1000 rows.
- 'Identity' The variance is equal to the mean as described in. Counts are therefore modeled by the **Poisson distribution** individually for each row of X and Y. Use this option if your data has few genes and the regression between the variance and mean is not possible because of very small number of samples or replicates. This option is not recommended for overdispersed data.

Variance Link – pasilla



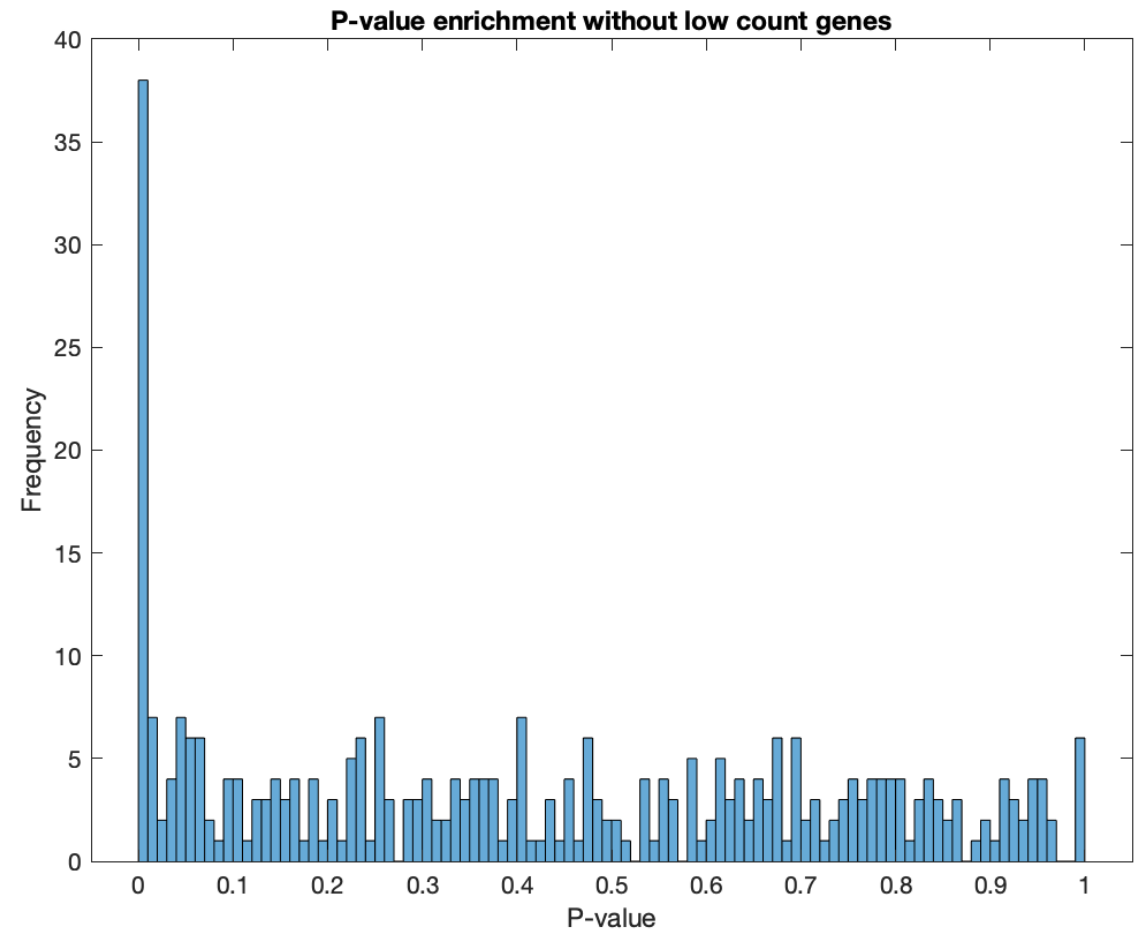
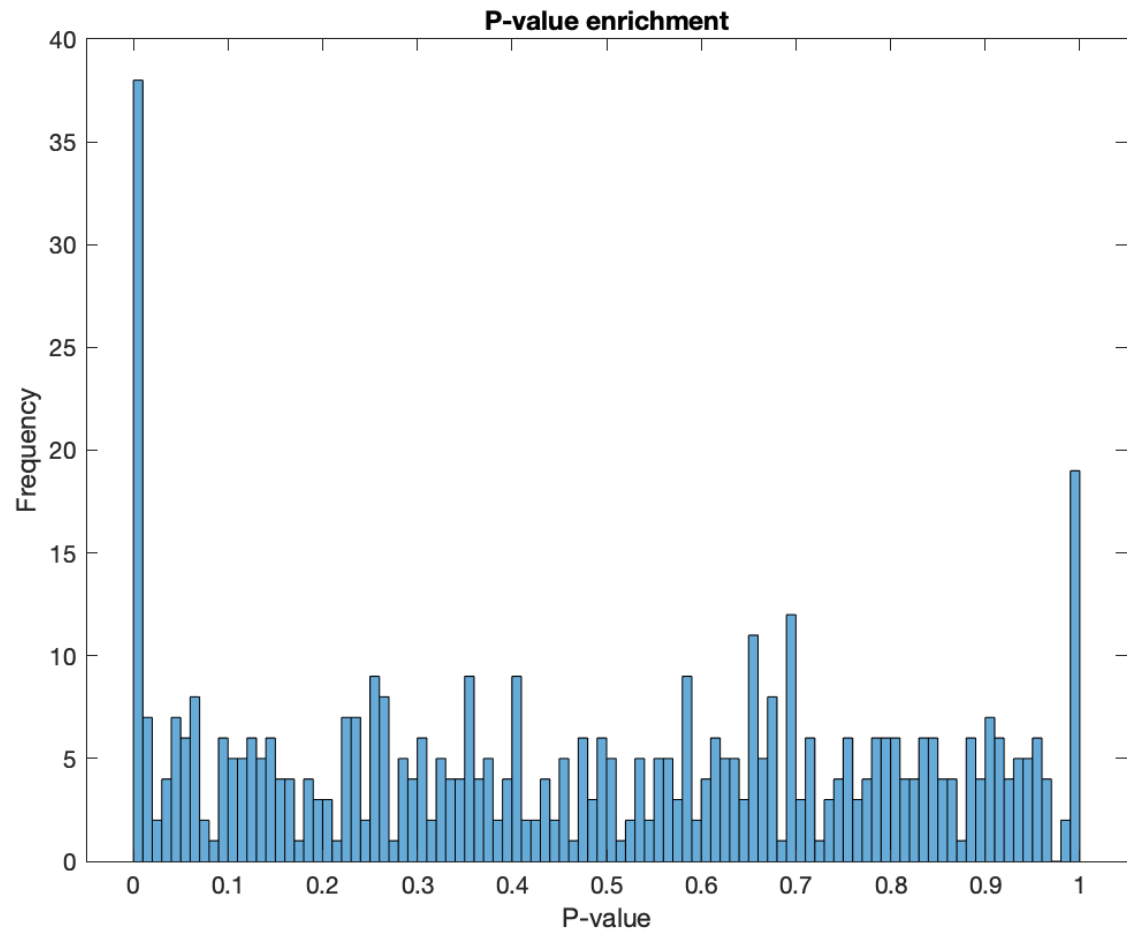
Variance Link – prostate cancer



p -values

- The output of `nbintest` includes a vector of p -values. A p -value indicates the probability that a change in expression as strong as the one observed (or even stronger) would occur under the null hypothesis, i.e. the conditions have no effect on gene expression.
- In the histogram of the p -values we observe an enrichment of low values (due to differentially expressed genes), whereas other values are uniformly spread (due to non-differentially expressed genes).
- The enrichment of values equal to 1 are due to genes with very low counts.

p -values histogram



Adjusted p -values

- Use `mafdr` function for obtaining adjusted p -values

```
FDR = mafdr(PValues)
```

returns `FDR` that contains a positive false discovery rate (p_{FDR}) for each entry in `PValues` using the procedure introduced by Storey (2002).

`PValues` contains one p -value for each feature (for example, a gene) in a data set.

- Optional parameter '`BHFDR`' — Flag to use linear step-up procedure

Flag to use the linear step-up procedure introduced by Benjamini and Hochberg (1995), specified as the comma-separated pair consisting of '`BHFDR`' and `true` or `false`.

The default value is `false`, that is, the function uses the procedure introduced by Storey (2002).

Benjamini-Hochberg adjustment

- The **Benjamini-Hochberg** (BH) adjustment is a statistical method that provides an adjusted p -value answering the following question: what would be the fraction of false positives if all the genes with adjusted p -values below a given threshold were considered significant?
- Set a threshold of 0.1 for the adjusted p -values, equivalent to consider a 10% false positives as acceptable, and identify the genes that are significantly expressed by considering all the genes with adjusted p -values below this threshold.
- We define a significant gene if adjusted p -value is less than 0.1

```
% compute the adjusted P-values (BH correction)
padj = mafdr(tLocal.pValue, 'BHFDR', true);
% add to the existing table
geneTable.pvalue = tLocal.pValue;
geneTable.padj = padj;
% create a table with significant genes
sig = geneTable.padj < 0.1;
```

Significant Genes

	1	2	3	4	5	6	7
	meanBase	meanHigh	meanLow	foldChange	log2FC	pvalue	padj
1 ENSG00000269900	2.3177e+03	333.8243	4.3016e+03	0.0776	-3.6877	0	0
2 ENSG00000281131	412.3767	645.3852	179.3682	3.5981	1.8472	1.2308e-20	3.0771e-...
3 ENSG00000235123	171.2898	266.9165	75.6632	3.5277	1.8187	1.1820e-13	1.9700e-...
4 ENSG00000258162	147.3062	234.7539	59.8586	3.9218	1.9715	4.9304e-12	6.1630e-...
5 ENSG00000275173	0.6489	0.0988	1.1990	0.0824	-3.6015	4.4600e-11	4.4600e-...
6 ENSG00000226779	8.7947	15.1989	2.3904	6.3584	2.6687	1.5877e-09	1.3231e-...
7 ENSG00000235984	85.0298	134.3791	35.6805	3.7662	1.9131	1.0660e-08	7.6144e-...
8 ENSG00000225258	145.1493	226.2200	64.0785	3.5304	1.8198	1.2720e-07	7.9499e-...
9 ENSG00000253369	54.9525	86.3528	23.5521	3.6665	1.8744	2.7079e-07	1.5044e-...
10 ENSG00000183535	2.0214	0.5220	3.5209	0.1483	-2.7539	5.3606e-06	2.6803e-...
11 ENSG00000254695	1.2058	1.9680	0.4436	4.4364	2.1494	5.9909e-06	2.7232e-...
12 ENSG00000259457	434.1445	556.3867	311.9022	1.7838	0.8350	8.5056e-06	3.5440e-...
13 ENSG00000261538	25.7713	41.0953	10.4473	3.9336	1.9758	1.4281e-05	5.4927e-...
14 ENSG00000243081	6.2161	11.1360	1.2962	8.5914	3.1029	1.8614e-05	6.6478e-...
15 ENSG00000278626	297.2310	377.9431	216.5189	1.7455	0.8037	3.9420e-05	0.0013
16 ENSG00000254080	41.1710	63.8399	18.5020	3.4504	1.7868	8.6679e-05	0.0027

Significant Genes

	1	2	3	4	5	6	7
	meanBase	meanHigh	meanLow	foldChange	log2FC	pvalue	padj
10 ENSG00000183535	2.0214	0.5220	3.5209	0.1483	-2.7539	5.3606e-06	2.6803e-...
11 ENSG00000254695	1.2058	1.9680	0.4436	4.4364	2.1494	5.9909e-06	2.7232e-...
12 ENSG00000259457	434.1445	556.3867	311.9022	1.7838	0.8350	8.5056e-06	3.5440e-...
13 ENSG00000261538	25.7713	41.0953	10.4473	3.9336	1.9758	1.4281e-05	5.4927e-...
14 ENSG00000243081	6.2161	11.1360	1.2962	8.5914	3.1029	1.8614e-05	6.6478e-...
15 ENSG00000278626	297.2310	377.9431	216.5189	1.7455	0.8037	3.9420e-05	0.0013
16 ENSG00000254080	41.1710	63.8399	18.5020	3.4504	1.7868	8.6679e-05	0.0027
17 ENSG00000258077	0.4264	0.7175	0.1352	5.3060	2.4076	1.6657e-04	0.0049
18 ENSG00000233508	17.9213	28.3133	7.5293	3.7604	1.9109	2.4633e-04	0.0068
19 ENSG00000251152	2.9183	4.4752	1.3613	3.2875	1.7170	7.0543e-04	0.0186
20 ENSG00000280703	3.8910	6.0291	1.7529	3.4394	1.7822	9.5195e-04	0.0238
21 ENSG00000259443	3.6446	5.6643	1.6249	3.4860	1.8016	0.0012	0.0285
22 ENSG00000243479	63.0556	84.5346	41.5766	2.0332	1.0238	0.0020	0.0452
23 ENSG00000229807	11.5330	15.6013	7.4646	2.0900	1.0635	0.0023	0.0489
24 ENSG00000235736	0.1386	0.2201	0.0571	3.8537	1.9463	0.0023	0.0489
25 ENSG00000266120	0.1871	0.2949	0.0794	3.7155	1.8935	0.0033	0.0658
26 ENSG00000228055	0.0812	0.0272	0.1352	0.2015	-2.3114	0.0039	0.0714
27 ENSG00000251003	1.2448	1.8820	0.6076	3.0973	1.6310	0.0039	0.0714
28 ENSG00000260760	0.3651	0.1005	0.6297	0.1596	-2.6474	0.0045	0.0768
29 ENSG00000253643	12.4192	18.6023	6.2361	2.9830	1.5768	0.0043	0.0768
30 ENSG00000280623	1.0315e+04	8.6166e+03	1.2012e+04	0.7173	-0.4793	0.0054	0.0897

References

- MATLAB Documentation
 - RNA-Seq Example. URL: <https://it.mathworks.com/help/bioinfo/ug/identifying-differentially-expressed-genes-from-rna-seq-data.html>
- Wikipedia
 - Negative binomial distribution. URL: https://en.wikipedia.org/wiki/Negative_binomial_distribution
 - Poisson distribution. URL: https://en.wikipedia.org/wiki/Poisson_distribution
- StatQuest with Josh Starmer
 - URL: <https://www.youtube.com/channel/UCtYLUtgS3k1Fg4y5tAhLbw>
- Ignacio Gonzalez (2014), Tutorial Statistical analysis of RNA-Seq data. URL: <http://www.nathalievialaneix.eu/doc/pdf/tutorial-rnaseq.pdf>
- Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10), R106.
- Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., ... & Brown, J. B. (2011). The developmental transcriptome of *Drosophila melanogaster*. *Nature*, 471(7339), 473.
- Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3), 479-498.
- Benjamini et al. Controlling the false discovery rate: a practical and powerful approach to multiple testing. 1995. *Journal of the Royal Statistical Society, Series B57* (1):289-300.