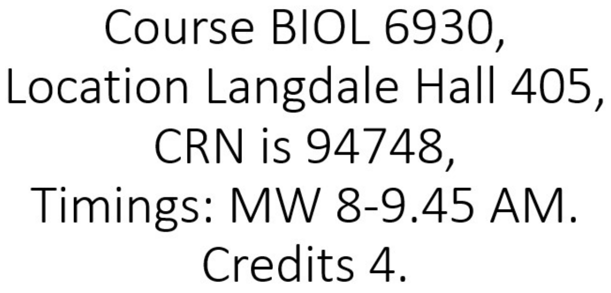
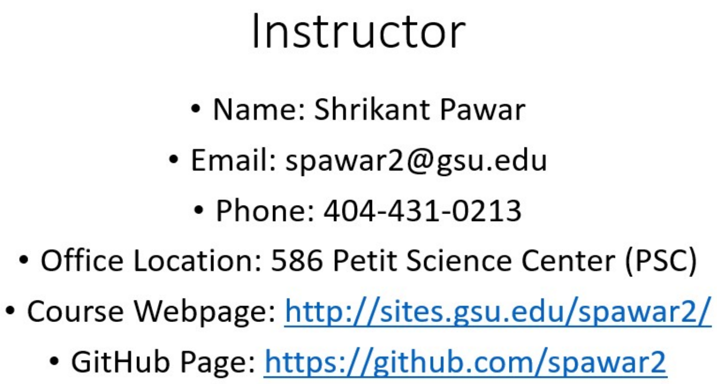
1 R programming (6900) Basics of NGS and Micro-array analysis

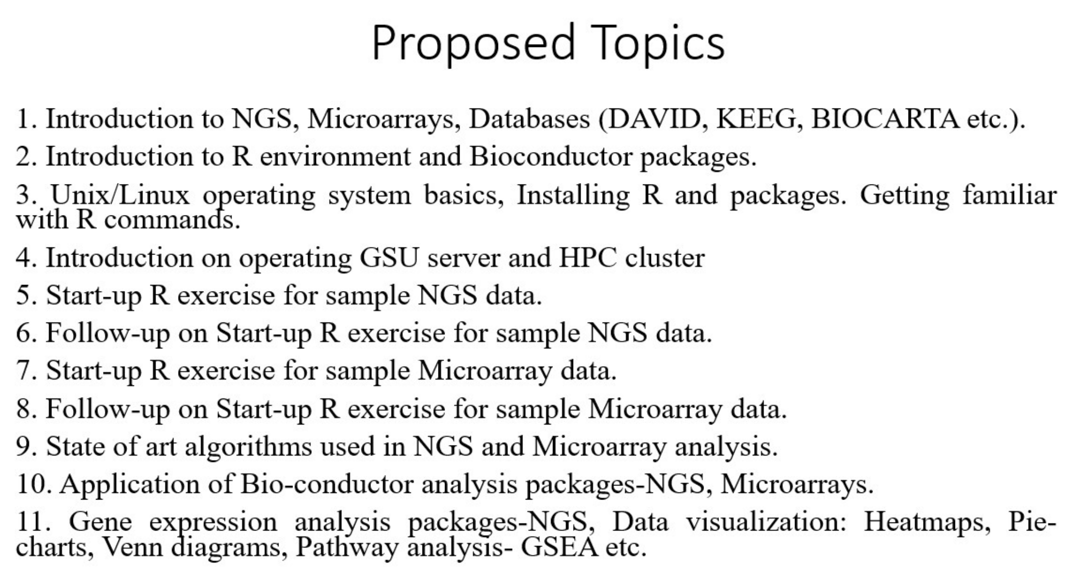


Course BIOL 6930,
Location Langdale Hall 405,
CRN is 94748,
Timings: MW 8-9.45 AM.
Credits 4.



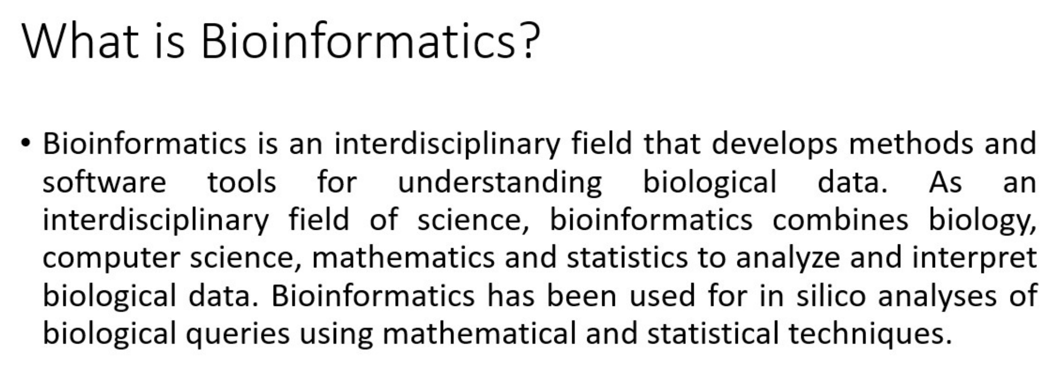
Instructor

- Name: Shrikant Pawar
- Email: spawar2@gsu.edu
- Phone: 404-431-0213
- Office Location: 586 Petit Science Center (PSC)
- Course Webpage: http://sites.gsu.edu/spawar2/
- GitHub Page: https://github.com/spawar2



Proposed Topics

1. Introduction to NGS, Microarrays, Databases (DAVID, KEEG, BIOCARTA etc.).
2. Introduction to R environment and Bioconductor packages.
3. Unix/Linux operating system basics, Installing R and packages. Getting familiar with R commands.
4. Introduction on operating GSU server and HPC cluster
5. Start-up R exercise for sample NGS data.
6. Follow-up on Start-up R exercise for sample NGS data.
7. Start-up R exercise for sample Microarray data.
8. Follow-up on Start-up R exercise for sample Microarray data.
9. State of art algorithms used in NGS and Microarray analysis.
10. Application of Bio-conductor analysis packages-NGS, Microarrays.
11. Gene expression analysis packages-NGS, Data visualization: Heatmaps, Pie-charts, Venn diagrams, Pathway analysis- GSEA etc.



What is Bioinformatics?

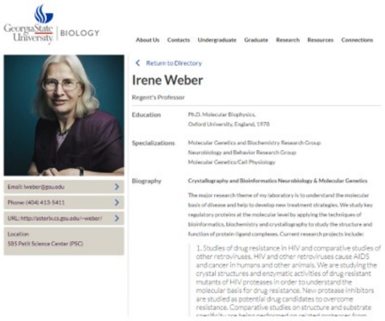
- Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines biology, computer science, mathematics and statistics to analyze and interpret biological data. Bioinformatics has been used for in silico analyses of biological queries using mathematical and statistical techniques.

What is Bioinformatics?

https://www.youtube.com/watch?v=v1cTNhiZ2_c

Structural Bioinformatics:

- Protein structure prediction is another important application of bioinformatics. The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it.
- http://asterix.cs.gsu.edu/~weber/



GeorgiaState University | BIOLOGY
About Us Contacts Undergraduate Graduate Research Resources Connections
Return to Directory

Irene Weber

Regent's Professor

Education: Ph.D. Molecular Biophysics, Oxford University, England, 1978

Specializations: Molecular Genetics and Biochemistry Research Group; Neurobiology and Behavior Research Group; Molecular Genetics/Cell Physiology

Biography: **Crystallography and Bioinformatics Neurobiology & Molecular Genetics**

The major research theme of my laboratory is to understand the molecular basis of disease and help to develop new treatment strategies. We study key regulatory proteins at the molecular level by applying the techniques of bioinformatics, biochemistry and crystallography to study the structure and function of protein-ligand complexes. Current research projects include:

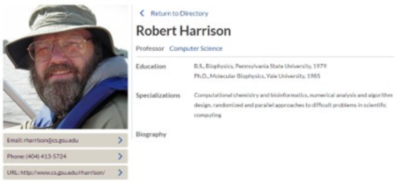
> 1. Studies of drug resistance in HIV and comparative studies of other retroviruses. HIV and other retroviruses cause AIDS and cancer in humans and other animals. We are studying the crystal structures and enzymatic activities of drug resistant mutants of HIV proteases in order to understand the molecular basis for drug resistance. New protease inhibitors are studied as potential drug candidates to overcome resistance. Comparative studies on structure and substrate specificity are being performed on related proteases from

Email: iweber@gsu.edu

Phone: (404) 413-5411

URL: http://asterix.cs.gsu.edu/~weber/

Location
585 Petit Science Center (PSC)



Robert Harrison

Professor Computer Science

Education: B.S., Biophysics, Pennsylvania State University, 1979; Ph.D., Molecular Biophysics, Yale University, 1985

Specializations: Computational chemistry and bioinformatics, numerical analysis and algorithm design, randomized and parallel approaches to difficult problems in scientific computing

Biography

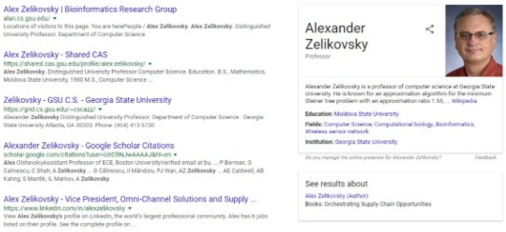
Email: rharrison@cs.gsu.edu

Phone: (404) 413-5724

URL: http://www.cs.gsu.edu/rharrison/

Network and Systems Biology

- Network analysis seeks to understand the relationships within biological networks such as metabolic or protein–protein interaction networks.
- Systems biology involves the use of computer simulations of cellular subsystems (such as the networks of metabolites and enzymes that comprise metabolism, signal transduction pathways and gene regulatory networks) to both analyze and visualize the complex connections of these cellular processes.
- http://alan.cs.gsu.edu/NGS/?q=cscazz



Alex Zelikovsky | Bioinformatics Research Group
alan.cs.gsu.edu/ ▾
Locations of visitors to this page. You are herePeople / **Alex Zelikovsky**. **Alex Zelikovsky**. Distinguished University Professor. Department of Computer Science

Alex Zelikovsky - Shared CAS
https://shared.cas.gsu.edu/profile/alex-zelikovsky/ ▾
Alex Zelikovsky. Distinguished University Professor Computer Science. Education. B.S., Mathematics, Moldova State University, 1980 M.S., Computer Science ...

Zelikovsky - GSU C.S. - Georgia State University
https://grid.cs.gsu.edu/~cscazz/ ▾
Alexander **Zelikovsky** Distinguished University Professor. Department of Computer Science · Georgia State University Atlanta, GA 30303. Phone: (404) 413 5730

Alexander Zelikovsky - Google Scholar Citations
scholar.google.com/citations?user=UzCRNJwAAAAJ&hl=en ▾
Alex OlshevskyAssistant Professor of ECE, Boston UniversityVerified email at bu. ... P Berman, G Calinescu, C Shah, A **Zelikovsky** ... G Călinescu, II Măndoiu, PJ Wan, AZ **Zelikovsky** ... AE Caldwell, AB Kahng, S Mantik, IL Markov, A **Zelikovsky**.

Alex Zelikovsky - Vice President, Omni-Channel Solutions and Supply ...
https://www.linkedin.com/in/alexzelikovsky ▾
View **Alex Zelikovsky's** profile on LinkedIn, the world's largest professional community. Alex has 6 jobs listed on their profile. See the complete profile on ...

Alexander Zelikovsky
Professor

Alexander Zelikovsky is a professor of computer science at Georgia State University. He is known for an approximation algorithm for the minimum Steiner tree problem with an approximation ratio 1.55, ... Wikipedia

Education: Moldova State University

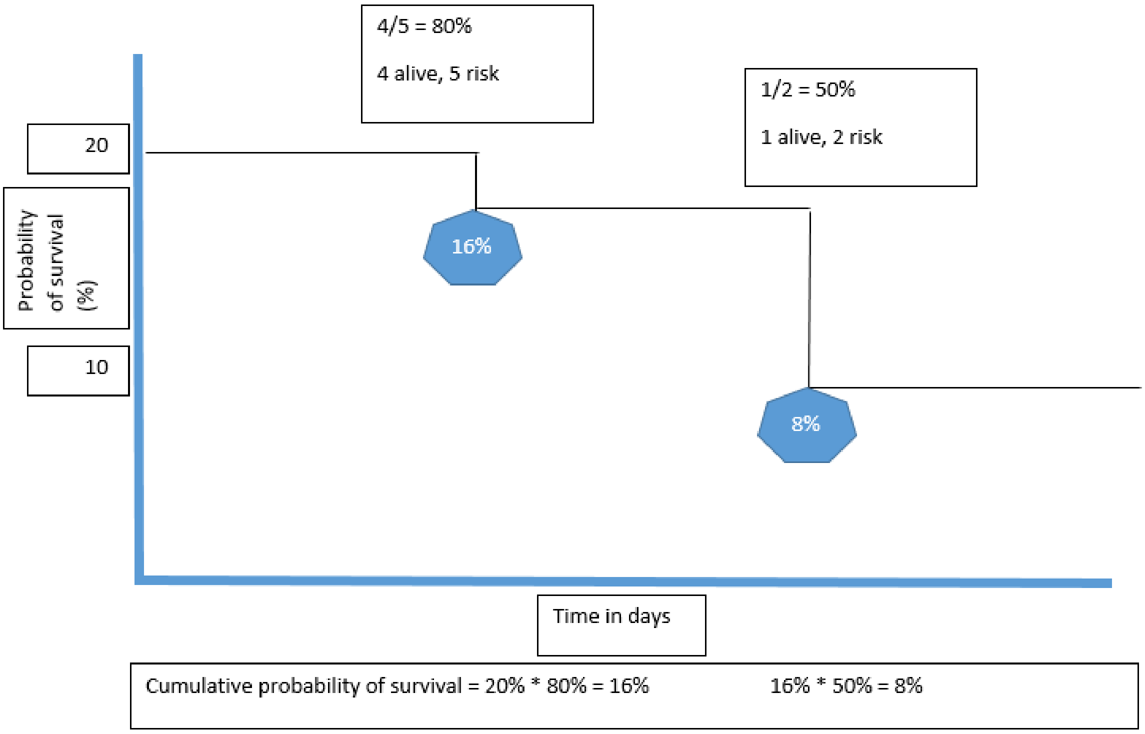
Fields: Computer Science, Computational biology, Bioinformatics, Wireless sensor network

Institution: Georgia State University

Do you manage the online presence for Alexander Zelikovsky? Feedback

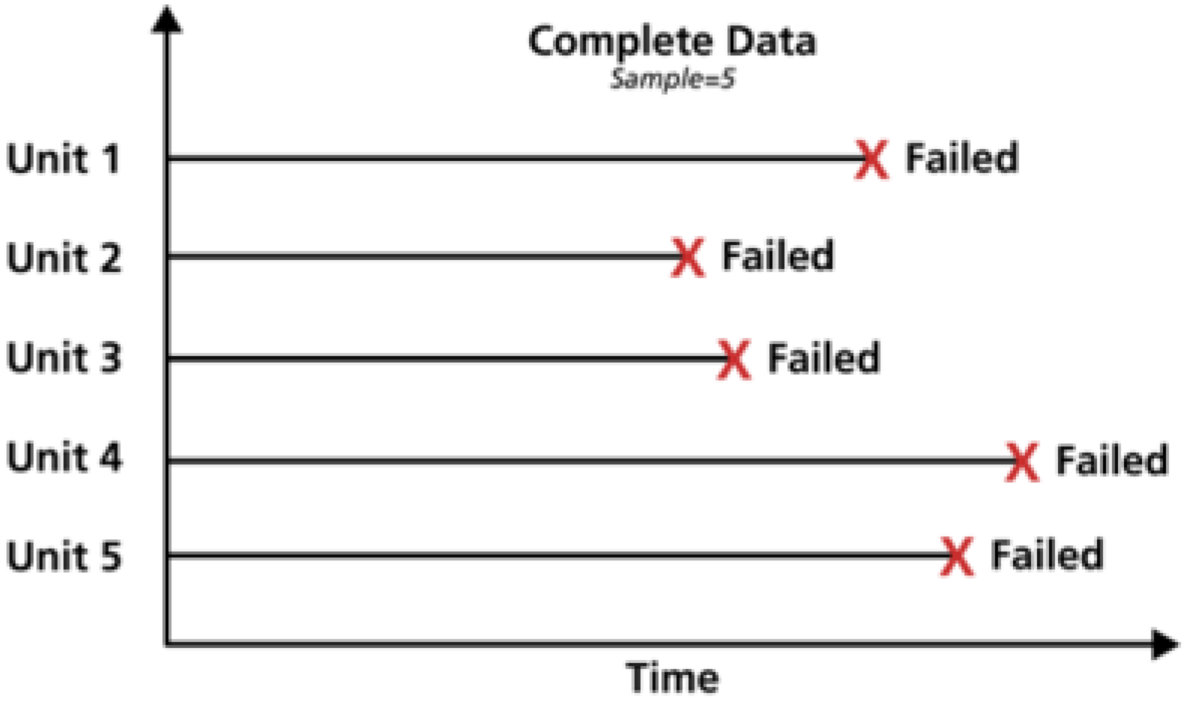
See results about
Alex Zelikovsky (Author)
Books: Orchestrating Supply Chain Opportunities

1.1 Survival Plots: Kaplan Meier Analysis

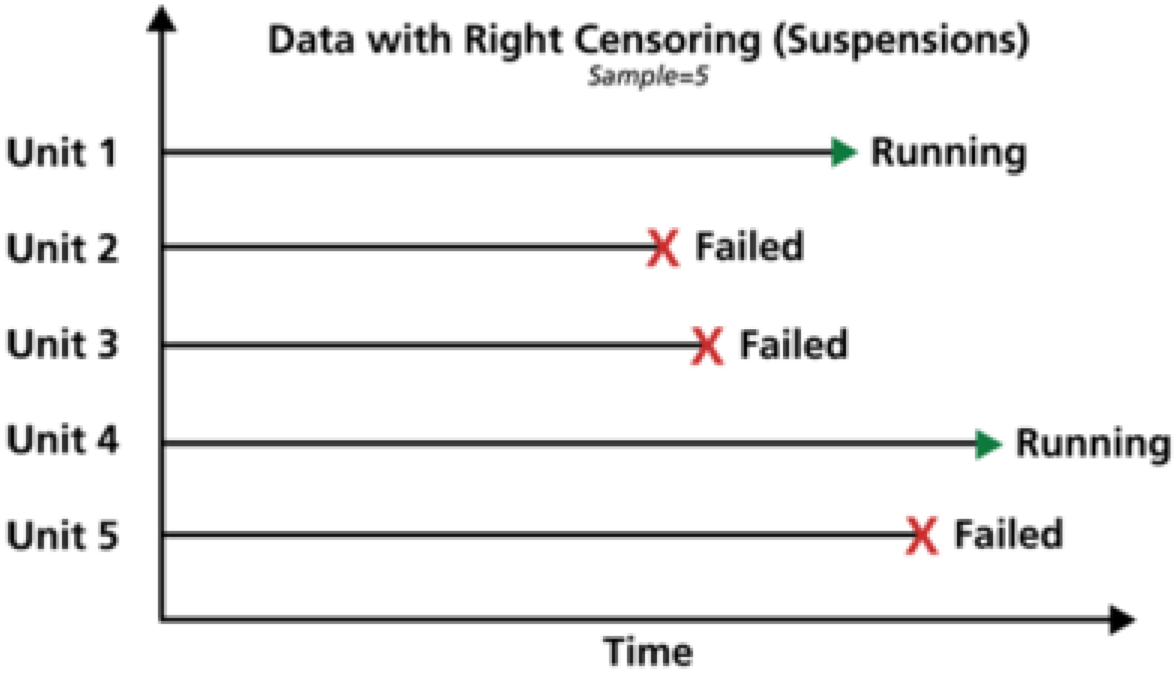


The first thing to do is to use function Surv() to build the standard survival object. The variable t1 records the time to death or the censored time. A plus sign after the time in the print out indicates censoring. The formula instructs the survfit() function to fit a model with intercept only.

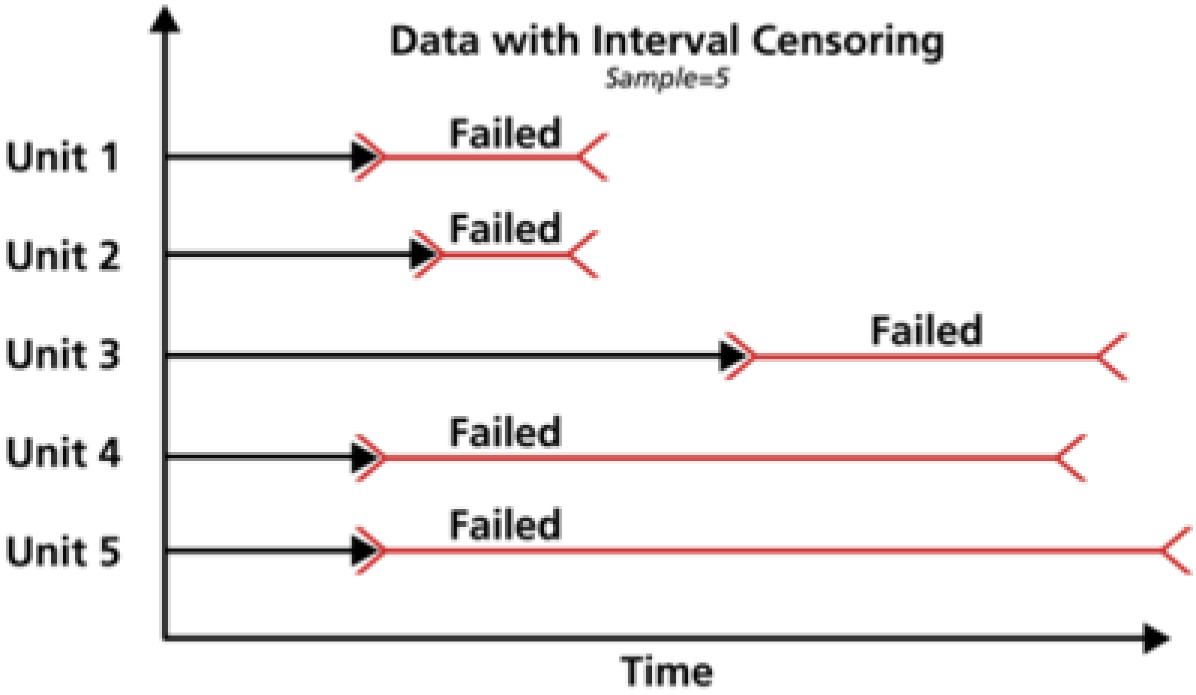
1.2 What is censoring? Understanding your data is extremely important



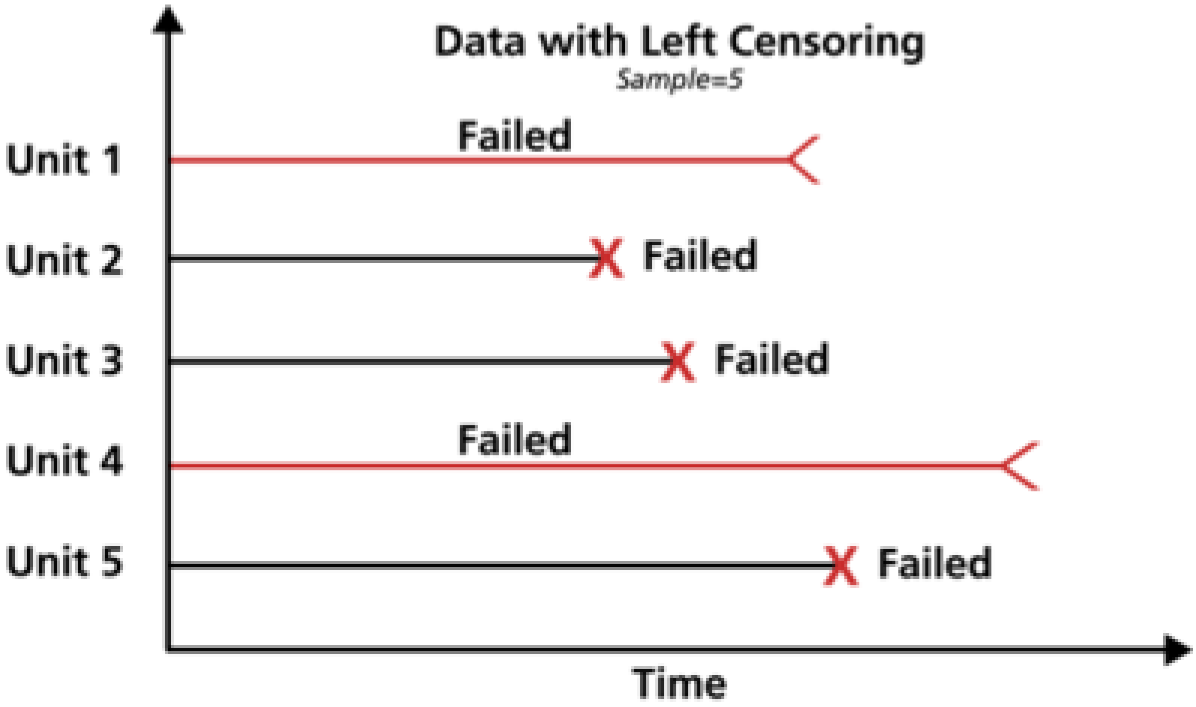
Complete data means that the value of each sample unit is observed or known. For example, if we had to compute the average test score for a sample of ten students, complete data would consist of the known score for each student. Likewise in the case of life data analysis, our data set (if complete) would be composed of the times-to-failure of all units in our sample. For example, if we tested five units and they all failed (and their times-to-failure were recorded), we would then have complete information as to the time of each failure in the sample.



The most common case of censoring is what is referred to as right censored data, or suspended data. In the case of life data, these data sets are composed of units that did not fail. For example, if we tested five units and only three had failed by the end of the test, we would have right censored data (or suspension data) for the two units that did not failed. The term right censored implies that the event of interest (i.e., the time-to-failure) is to the right of our data point. In other words, if the units were to keep on operating, the failure would occur at some time after our data point (or to the right on the time scale).

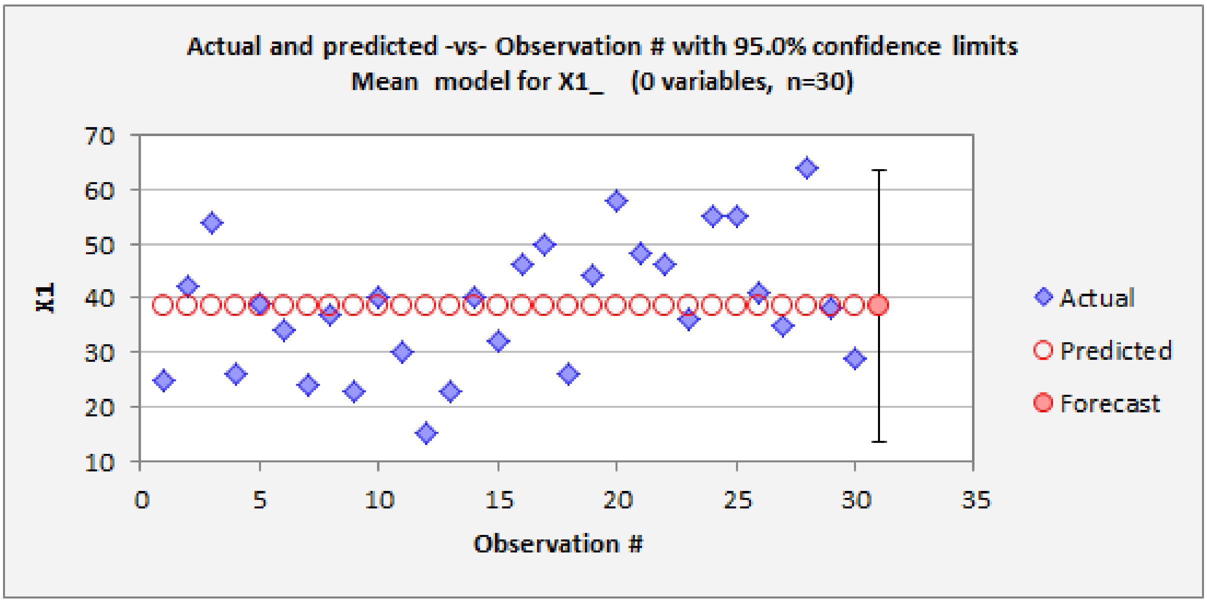


The second type of censoring is commonly called interval censored data. Interval censored data reflects uncertainty as to the exact times the units failed within an interval. This type of data frequently comes from tests or situations where the objects of interest are not constantly monitored. For example, if we are running a test on five units and inspecting them every 100 hours, we only know that a unit failed or did not fail between inspections. Specifically, if we inspect a certain unit at 100 hours and find it operating, and then perform another inspection at 200 hours to find that the unit is no longer operating, then the only information we have is that the unit failed at some point in the interval between 100 and 200 hours. This type of censored data is also called inspection data by some authors.



The third type of censoring is similar to the interval censoring and is called left censored data. In left censored data, a failure time is only known to be before a certain time. For instance, we may know that a certain unit failed sometime before 100 hours but not exactly when. In other words, it could have failed any time between 0 and 100 hours. This is identical to interval censored data in which the starting time for the interval is zero.

1.3 Using Intercepts?

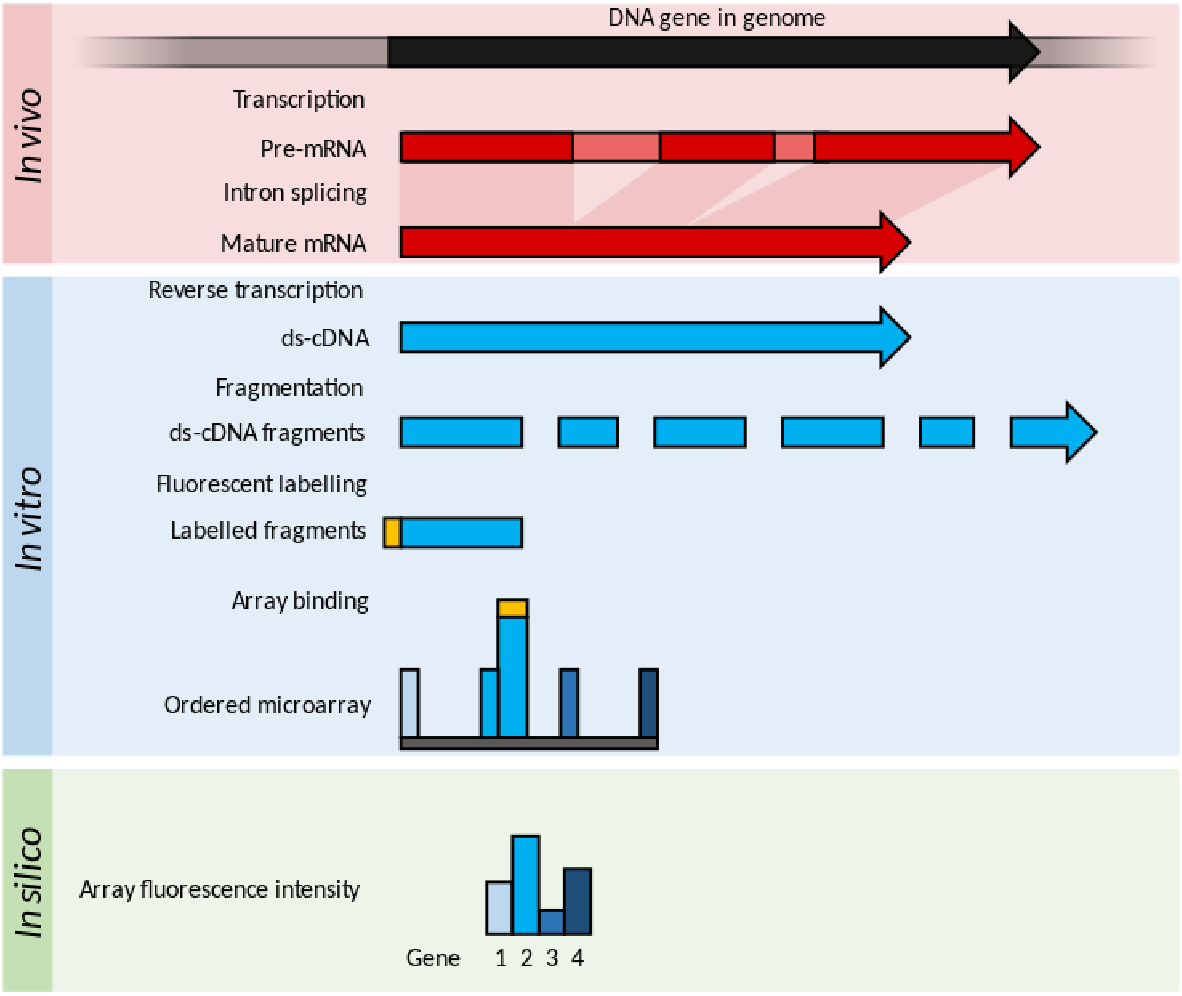


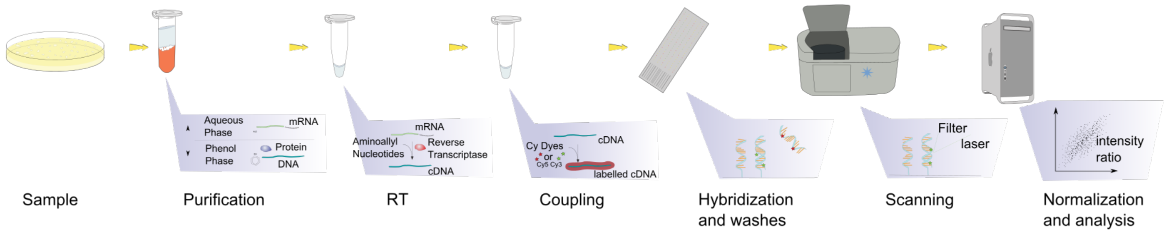
1.4 R exercise on KM plot

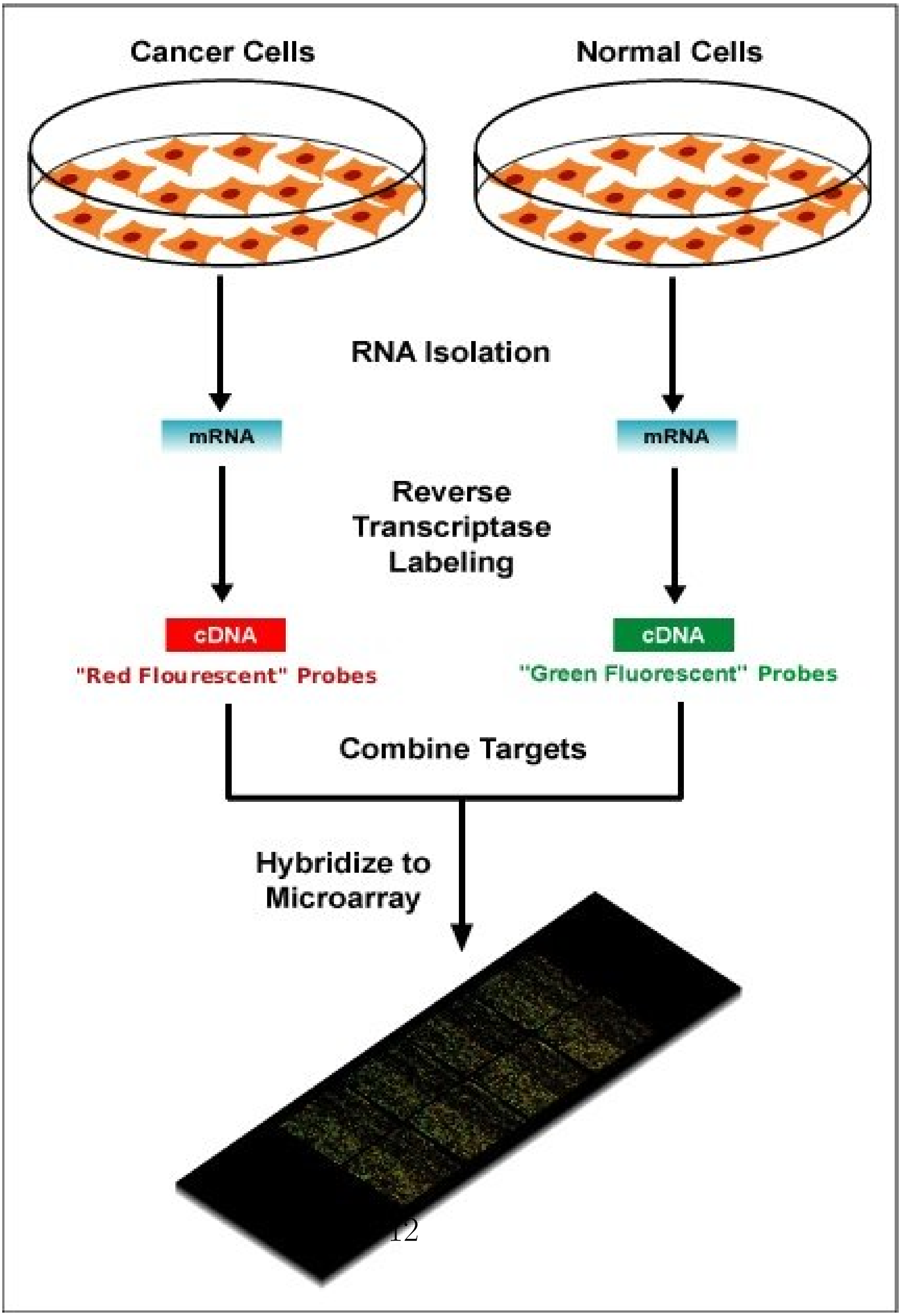
1.5 Microarray Technique:

A DNA microarray (biochip) is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome.

General technique of performing Microarray


In vivo
DNA gene in genome
Transcription
Pre-mRNA
Intron splicing
Mature mRNA
In vitro
Reverse transcription
ds-cDNA
Fragmentation
ds-cDNA fragments
Fluorescent labelling
Labelled fragments
Array binding
Ordered microarray
In silico
Array fluorescence intensity
Gene
1 2 3 4


Aqueous Phase
mRNA
Phenol Phase
Protein
DNA
mRNA
Aminoallyl Nucleotides
Reverse Transcriptase
cDNA
Cy Dyes
or
Cy5 Cy3
cDNA
labelled cDNA
Filter
laser
intensity ratio
Sample
Purification
RT
Coupling
Hybridization and washes
Scanning
Normalization and analysis


Cancer Cells
Normal Cells
RNA Isolation
mRNA
mRNA
Reverse
Transcriptase
Labeling
cDNA
"Red Flourescent" Probes
cDNA
"Green Fluorescent" Probes
Combine Targets
Hybridize to
Microarray

One-channel vs two-channel microarrays:

Two-color microarrays or two-channel microarrays are typically hybridized with cDNA prepared from two samples to be compared (e.g. diseased tissue versus healthy tissue) and that are labeled with two different fluorophores.

Fluorescent dyes commonly used for cDNA labeling include Cy3, which has a fluorescence emission wavelength of 570 nm (corresponding to the green part of the light spectrum), and Cy5 with a fluorescence emission wavelength of 670 nm (corresponding to the red part of the light spectrum). The two Cy-labeled cDNA samples are mixed and hybridized to a single microarray that is then scanned in a microarray scanner to visualize fluorescence of the two fluorophores after excitation with a laser beam of a defined wavelength

Data analysis: 1. Image analysis 2. Data processing: background subtraction (based on global or local background), determination of spot intensities and intensity ratios, visualisation of data (e.g. see MA plot), and log-transformation of ratios, global or local normalization of intensity ratios, and segmentation into different copy number regions using step detection algorithms. 3. Class discovery analysis: This analytic approach, sometimes called unsupervised classification or knowledge discovery, tries to identify whether microarrays (objects, patients, mice, etc.) or genes cluster together in groups. Identifying naturally existing groups of objects. 4. Class prediction analysis: This approach, called supervised classification, establishes the basis for developing a predictive model into which future unknown test objects can be input in order to predict the most likely class membership of the test objects. 5. Hypothesis-driven statistical analysis: Identification of statistically significant changes in gene expression are commonly identified using the t-test, ANOVA, Bayesian method[29] MannWhitney test methods tailored to microarray data sets, which take into account multiple comparisons. 6. Network-based methods: Statistical methods that take the underlying structure of gene networks into account, representing either associative or causative interactions or dependencies among gene products.

1.6 R exercise on Microarray analysis

Please refer to the GitHub account for all the in-class code repositories for this analysis. The link is as follows: https://github.com/spawar2/Cl
Exercise-1

1.7 RNA sequencing concepts:

Why Is RNA-Seq Better Than Microarrays? There are several advantages RNA-seq has over microarrays:

With RNA-seq you can interrogate more than just differential gene expression. Although there are microarrays available for exon-level and microRNA analysis, most users are still interested in basic, probably 3 biased, differential gene expression. With RNA-seq you can look at coding and non-coding RNA, at splicing and allele specific expression, and possibly soon at full-length cDNA sequences, eliminating the need to infer or assemble isoforms.

Microarrays are also biased, as we have to decide what content to place on the array. Since RNA-seq does not use probes or primers, the data suffer from much lower biases (although I do not mean to say RNA-seq has none).

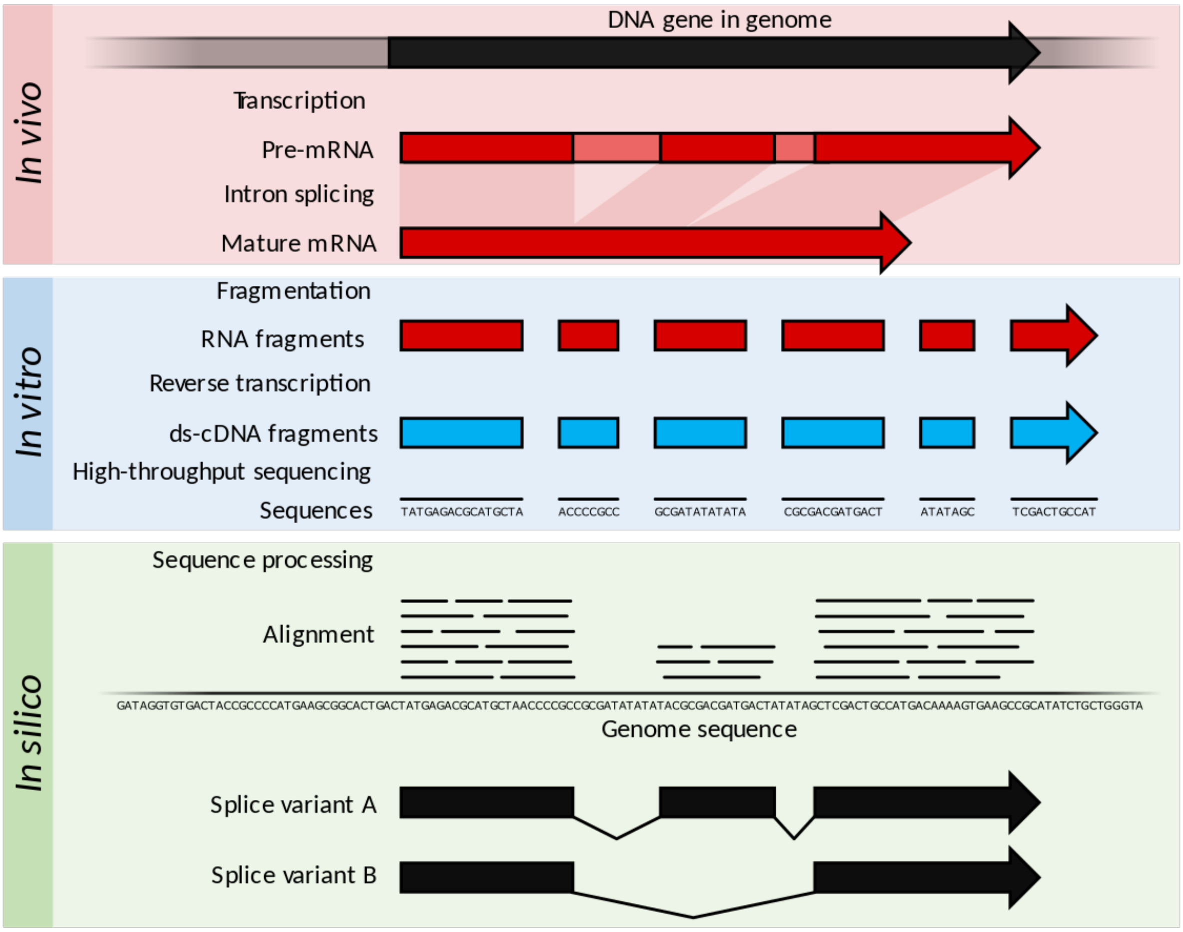
RNA-seq provides digital data in the form of aligned read-counts, resulting in a very wide dynamic range, improving the sensitivity of detection for rare transcripts.

It is also very cost-competitive to microarrays, as today, between 6-30 samples can be multiplexed in a single Illumina sequencing lane.

Lastly, you can reanalyze an RNA-seq dataset as more information about the transcriptome becomes available. If a paper is published showing an interesting splice-variant in a similar system to the one you work on, then you might want to go back and look at that splicing in your samples; and youd already have the data to do so.

How Does RNA-Seq Work? There are many methods for performing an RNA-seq experiment. In fact, the techniques are evolving so rapidly it can be difficult to decide which one to use. A basic choice is between 1) random-primed cDNA synthesis from double-stranded cDNA or 2) RNA-ligation methods. Most people use the first method and then need to make a further choice between a strand-specific protocol and one that is not. The method used most in my lab is Illuminas TruSeq RNA-seq, which is a random-primed cDNA synthesis non-strand-specific protocol.

Following is the process of generating RNA seq data:



Once you have a sequencing library, it is sequenced to a specified depth, which is dependent on what you want to do with the data. These reads are aligned to the genome or transcriptome and are counted to determine differential gene expression or further analyzed to determine splicing and isoform expression. Most people are sequencing RNA using paired-end 50-100bp methods. The exception is microRNA sequencing, as this only requires single-end 36bp sequencing in most cases.

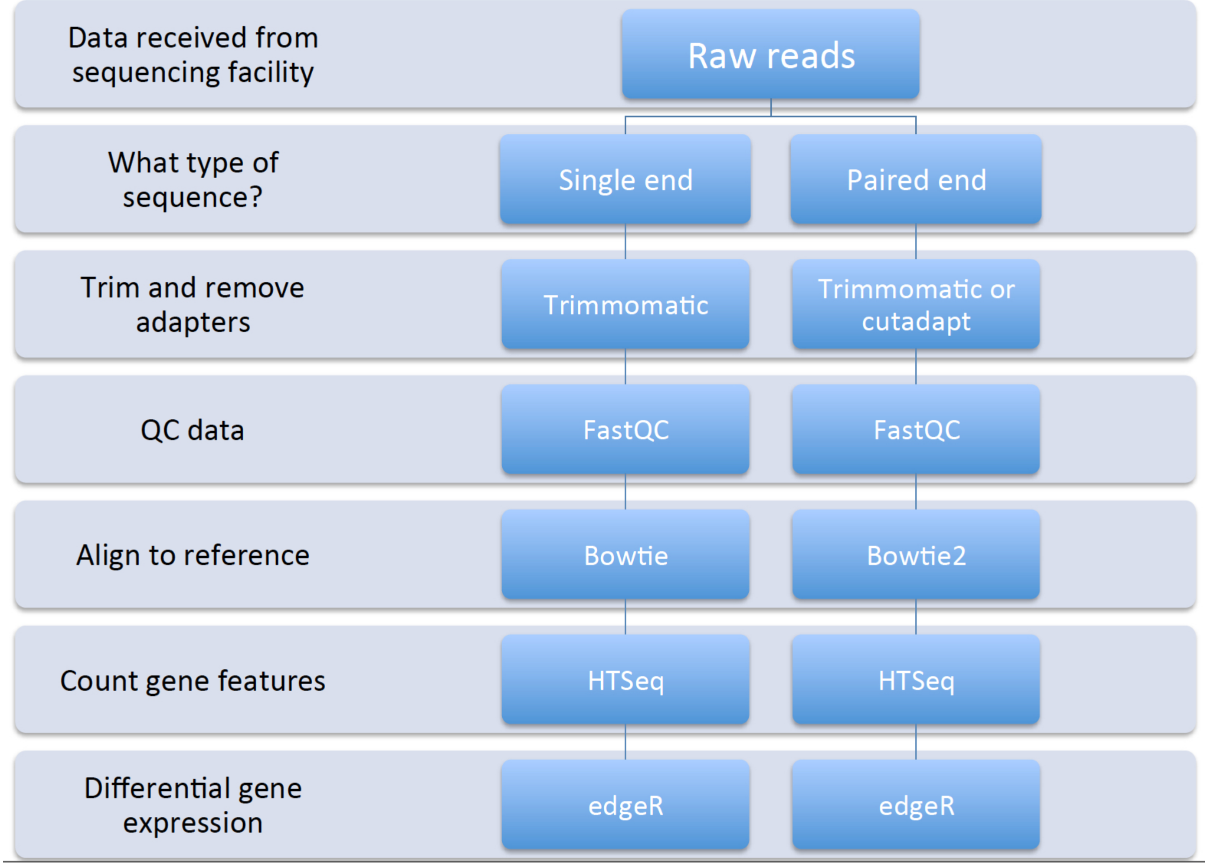
Library preparation:

RNA Isolation: RNA is isolated from tissue and mixed with deoxyribonuclease (DNase). DNase reduces the amount of genomic DNA. The amount of RNA degradation is checked with gel and capillary electrophoresis and is used to assign an RNA integrity number to the sample. This RNA quality and the total amount of starting RNA are taken into consideration during the subsequent library preparation, sequencing, and analysis steps.

RNA selection/depletion: To analyze signals of interest, the isolated RNA can either be kept as is, filtered for RNA with 3' polyadenylated (poly(A)) tails to include only mRNA, depleted of ribosomal RNA (rRNA), and/or filtered for RNA that binds specific sequences.

cDNA synthesis: DNA sequencing technology is more mature, so the RNA is reverse transcribed to cDNA. Reverse transcription results in loss of strandedness, which can be avoided with chemical labelling. Fragmentation and size selection are performed to purify sequences that are the appropriate length for the sequencing machine. The RNA, cDNA, or both are fragmented with enzymes, sonication, or nebulizers. Fragmentation of the RNA reduces 5' bias of randomly primed-reverse transcription and the influence of primer binding sites, with the downside that the 5' and 3' ends are converted to DNA less efficiently. Fragmentation is followed by size selection, where either small sequences are removed or a tight range of sequence lengths are selected. Because small RNAs like miRNAs are lost, these are analyzed independently. The cDNA for each experiment can be indexed with a hexamer or octamer barcode, so that these experiments can be pooled into a single lane for multiplexed sequencing.

Following is the workflow for analyzing RNA seq data from raw files:



Transcriptome assembly: De novo: This approach does not require a reference genome to reconstruct the transcriptome, and is typically used if the genome is unknown, incomplete, or substantially altered compared to the reference

Genome guided: This approach relies on the same methods used for DNA alignment, with the additional complexity of aligning reads that cover non-continuous portions of the reference genome.