# STAT540 Project Work Flow

This document contains the workflow of the project. Note that the steps mentioned here can be altered or additional analysis can be carried out as we progress along the project. Please remember to document all your R scripts preferably with text-editor like notepad++. We also need to track the project progress, results and problems in some form of document accessible/shared to all the members.

# Phase 1:

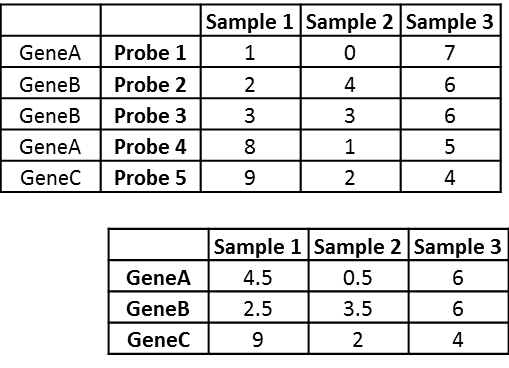
## Getting Ready

* The raw data for the microarray are available in the GEO repository (<http://www.ncbi.nlm.nih.gov/geo/>) and the dataset ID is **GSE35988.** Download the raw data files.
* Collect the necessary MetaData (in a spreadsheet) (Look into supplementary information and table of the paper and also in the GEO database entries). We need information like
  + Sample Name
  + Sample Type (Which type of Tissue is the sample collected from)
  + Tumor Type (Primary Tumor or Metastatic Tumor or Normal Prostate Tumor)
  + GSM ID (i.e. sample ID in the GEO database)
  + Platform ID (GPL ID)
  + Other clinical information associated with the sample

# Microarray Pre-Processing & Normalization

* Hunt for the R-package (in Bioconductor <http://www.bioconductor.org/> ) that is able to pre-process the data we have (Note: We have microarray data from Agilent platform).
  + We can start from a R-package LIMMA (<http://www.bioconductor.org/packages/2.12/bioc/html/limma.html>) as it is one of the most popular package from microarray pre-processing and can be used for multi-platform.
  + Read the user-manual and start pre-processing.
  + We should remember that the our dataset has been assayed using two versions of Agilent platform (they might have different set of probes)
  + **GPL6480**: Agilent-014850 Whole Human Genome Microarray 4x44K G4112F
  + **GPL6848**: Agilent-012391 Whole Human Genome Oligo Microarray G4112A
  + So the samples belonging to GPL6480 and GPL6848 should be pre-processed separately.
  + Get the probe-level annotation for both the platforms and include them in the probe level expression matrix.
  + After compiling the probe level expression into a matrix, we should check if there are any **Missing values**. (There are some methods and R-packages built to deal with Missing Values. If anyone has ideas which one would be the appropriate to solve it? If not we can completely discard the probes with missing values.)
  + Do some sanity checks to see if there are any defective probes. If some probes looks really bizarre than discard them.
  + We need to focus on the following issues at this stage
  + Background Correction
  + Normalization within an array (sample)
  + Normalization between arrays (across all the samples in the dataset)
  + Now merge the probe-level expression matrix of both the platforms GPL6480 and GPL6848.
  + Each probe is mapped to a particular gene or locus in the genome. Convert the probe-level annotation to gene-level annotation (i.e. get HUGO Gene Symbols but remember that all the genes may not be well annotated and have a HUGO Gene Symbol)
  + Multiple probes may be mapped to a single gene. So take the average expression values of all the probes mapped to a particular from respective samples.

For Example:



* Again generate various plots to visualize data and see how the data looks like
* Look if the data needs further normalization steps!
* Discard genes which does not have HUGO gene symbols as those are useless in our downstream analysis.

# Batch Correction

* We need to first look at how the samples in the dataset are clustered (using some PCA analysis). Ideally the Normal, Primary and Metastatic samples must be clustered separately according to their sample labels (or phenotype).
* Next, retrieve the sample processing (experimentation) dates which will be available in the raw data for each of the array.
* Again cluster the samples, and look for some batch effects. Ideally the samples must cluster according to their phenotype (Normal, Primary and Metastatic) and not according to the processing dates.
* If batch effect is well evident, we can use tools like COMBAT (<http://www.bu.edu/jlab/wp-assets/ComBat/Abstract.html>) to correct of the batch effect
* After processing the COMBAT, again cluster the samples and this time it should cluster according to the sample phenotype i.e. **COMBAT should not destroy the biological differences between different classes of tumors**.
* If not we may have to decide on what possible steps we can take or skip this step.
* Finally we will have our Gene Expression Matrix ready for further analysis.

# Phase 2:

## Meta-Analysis of the dataset

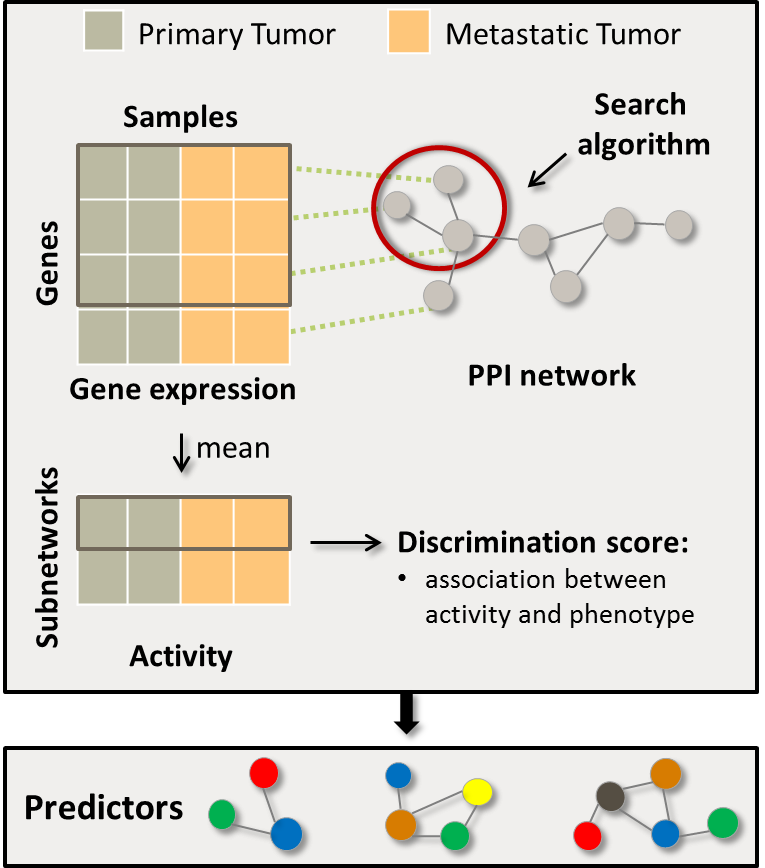
* How are the sample grouped according to the phenotype (clustering, PCA)
* Are the different phenotypes (samples) correlation? How does the correlation look?
* Find Differentially Expressed genes (using simple T-test or even Rank based Wilcoxon Test)
* Do the significantly differentially expressed genes have any significant correlation?
* We may brainstorm for additional questions?
* **Note:** We should bare in our mind that our dataset has
  + normal tissue and primary tumor are sampled from prostate
  + whereas metastatic tumor tissue samples are from different tissues/organs of body like Lung, Lymph Node, Kidney, Pancreases, and many more
* This might be an advantage as well as an disadvantage for us
* **Advantage:** Different biological process or pathway might be responsible for metastasis to different organ. But we assume that this tissue-level heterogeneity in the metastatic samples might over represent the common metastatic process and the tissue specific signals may remain diluted.
* **Disadvantage**: As we are comparing the primary vs metastasis which is like comparing gene-expression in prostate tissue with that other tissue tissues. So the differentially expressed gene might represent this tissue level difference rather than the difference brought about by difference in the phenotype.
* But to our rescue (Stuart et al. 2004) has published a list of genes that are tissue specific in prostate. So we can use this resource to filter out the stromal cells present in our gene-list. But we must be very careful as not to remove any important genes.

Stuart RO et al. 2004. In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. Proceedings of the National Academy of Sciences of the United States of America. 101:615–20. doi: 10.1073/pnas.2536479100.

# Phase 3:

## OptDis (Subnetwork Algorithm)

* OptDis integrates the Gene Expression data with Protein-Protein Interaction (PPI) Network to discover sets of genes (subnetwork) that discriminate between the tumor phenotypes.
* OptDis requires huge processing power so might not be feasible in a home computer. So we will be using computational resources at our lab (Collin’s Lab) to perform this task. But I can give a demo to everyone on how things are done when we reach at this point.
* Then we can compare the genes discovered from OptDis with the differentially expressed genes to test our hypothesis



# Phase 4:

## Function and Pathway Enrichment:

* The genes discovered from the OptDis will be analyzed for any functional and pathway enrichment
* For this we can use Ingenuity Pathway Analysis (IPA: <http://www.ingenuity.com/>). Collin’s Lab has the full licence for this tool. You may sign-up for a free-trial to get familiar to the tool. We can certain give some demo later on.
* DAVID (<http://david.abcc.ncifcrf.gov/>) also has similar functionality and is Free but is not very comprehensive as IPA.

## Network Visualization:

* We can hunt for some Cytoscape (http://www.cytoscape.org/) packages or R packages to visualize the some of the interesting sun-networks.

# Phase 5:

## Data Interpretation and Inferring the Biological Mechanism

* This would totally depend upon our biological knowledge on the processes involved in the metastasis in prostate cancer.
* We need to dig into some literatures to give our data some meaningful conclusion.
* We can also compare the list of genes discovered and pathways with some of the results that I already have that have been generated through similar analysis in other prostate cancer datasets.