Genomic data visualisation

Dr Jason Wong

UNSW AUSTRALIA

Never Stand Still | Medicine | Prince of Wales Clinical School

Introductory bioinformatics for human genomics workshop, UNSW

Day 1 – Thursday 29th January 2016

UNSW AUSTRALIA
ADULT CANCER PROGRAM

LOWY CANCER RESEARCH CENTRE

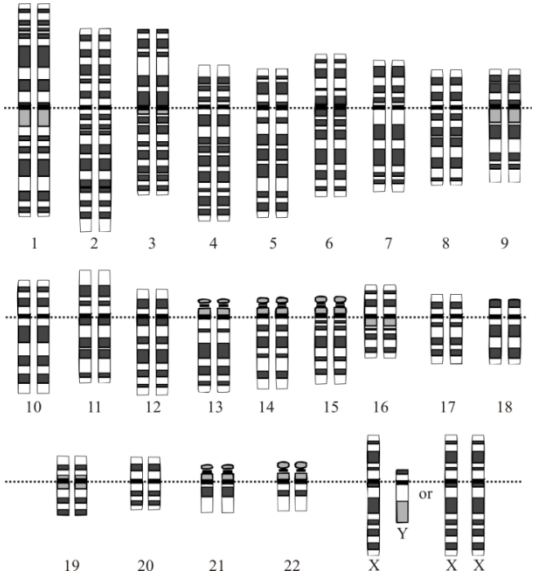
tcrn bridging the gap translational cancer research network

NSW GOVERNMENT

cancer institute NSW

Structure of human genome

- Consist of 23 pairs of chromosomes.
- Each chromosome is paired meaning that it is diploid.
- Each individual chromosome made up of double stranded DNA.
- Approximately ~3 billion bases in total.



5′TGACATGGGTACACATGACGGG 3′
3′ACTGTACCCATGTGTACTGCCC 5′

The size makes the genome difficult to visualise

Why visualise?

- Quality control (QC)
- To help interpret the data.
- Communicate results with others.

What to visualise?

- Two main types:

(1) Visualisation of individual genomic elements.

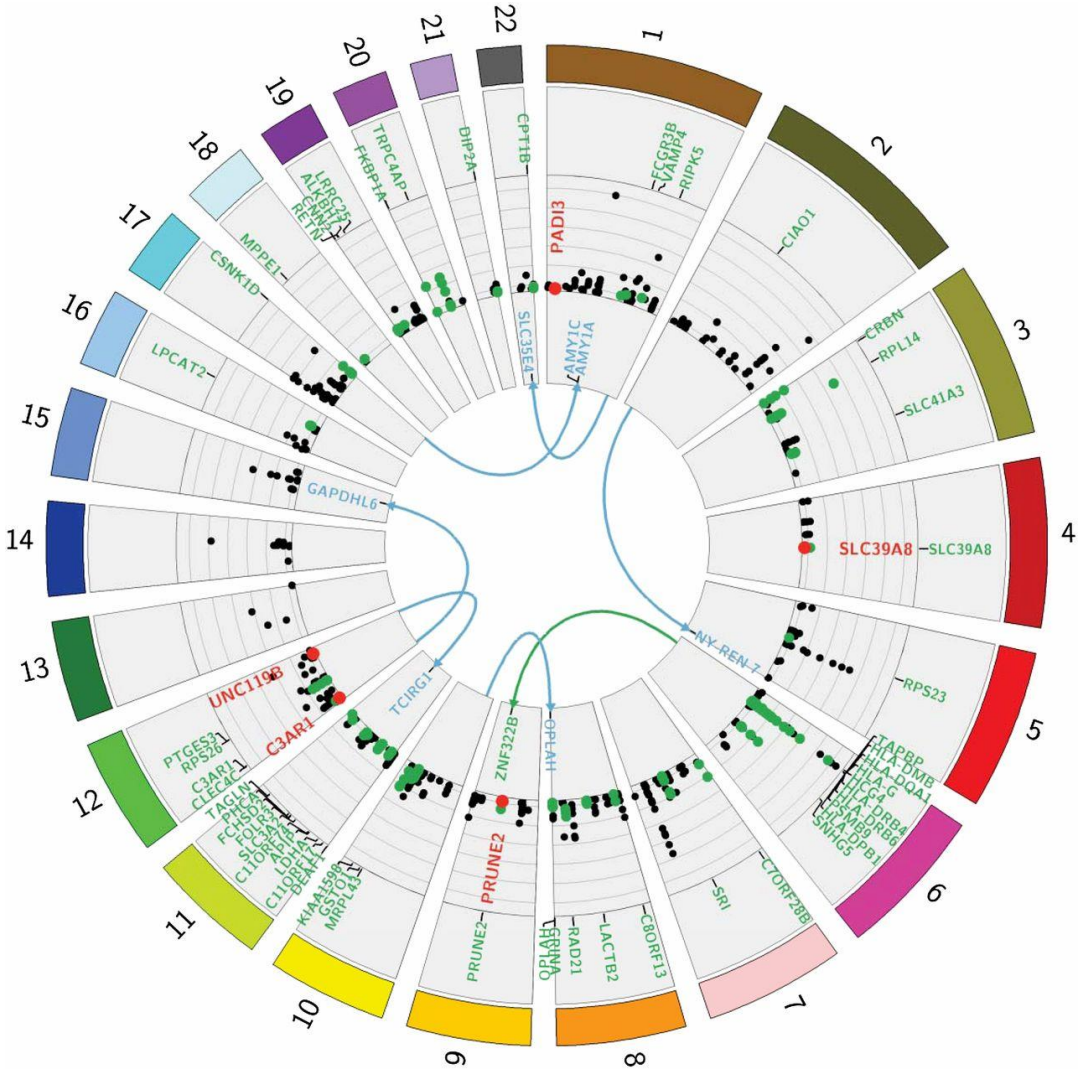
(For example using UCSC)

(2) Visualisation of summary of genomic elements.

What we will cover

- Visualisation of elements genome-wide
  - Circos plots
- Visualisation of summary of NGS data signals
  - Profile plots and heatmaps
- Gene expression data
  - Hierarchical clustering
  - Principal component analysis (PCA)
  - Differential expression (volcano plot)
  - Gene set enrichment analysis (GSEA)

Circos plot



- Provides a visual snapshot of the whole genome.
- Allow visualisation of relationships between different chromosomes.
- Uses tracks just like UCSC genome browser

http://circos.ca/

What making circos plots involve

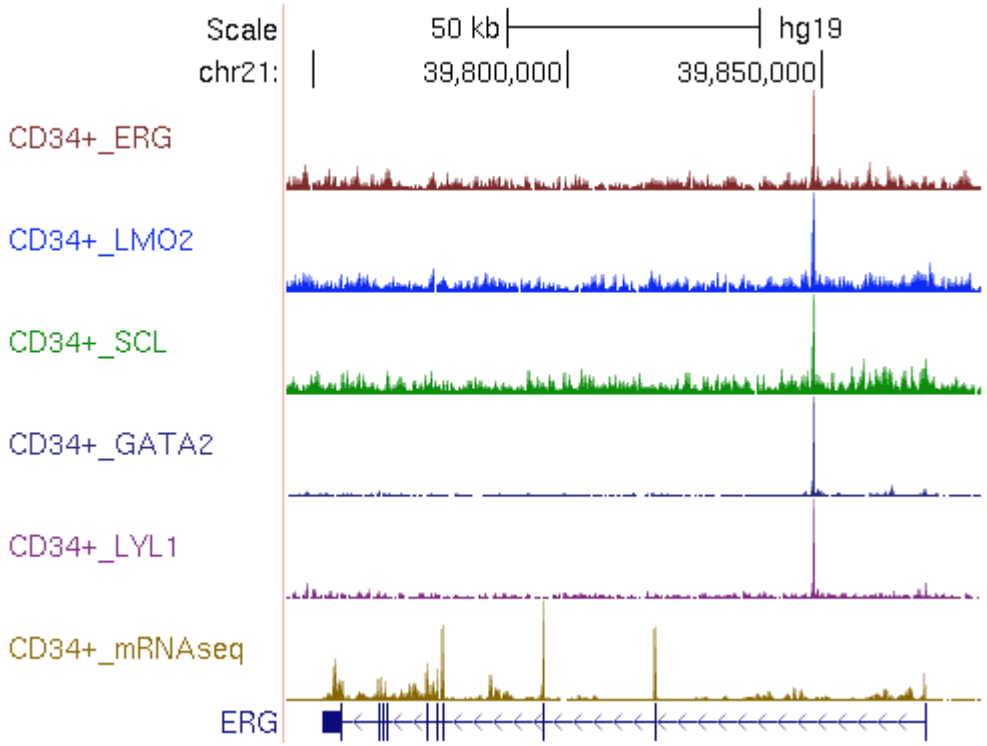
- Command line tool.
- Instructions are written in a text-based configuration file.

- Some web-based versions such as J-Circos

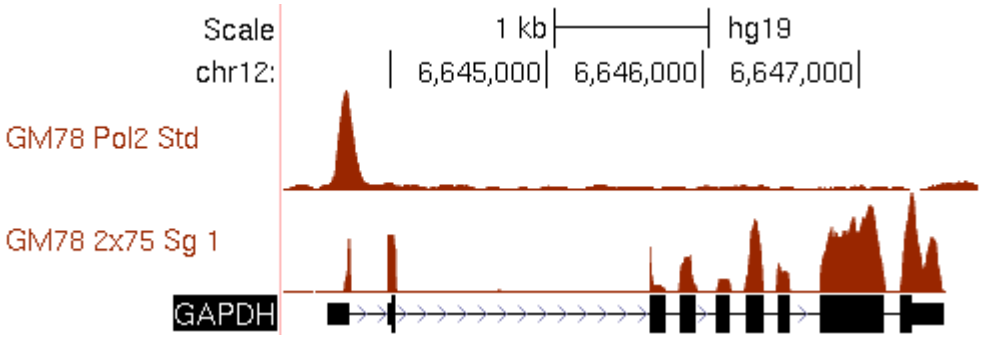
(http://www.australianprostatecentre.org/research/software/jcircos)

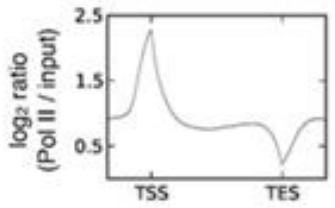
NGS profile data

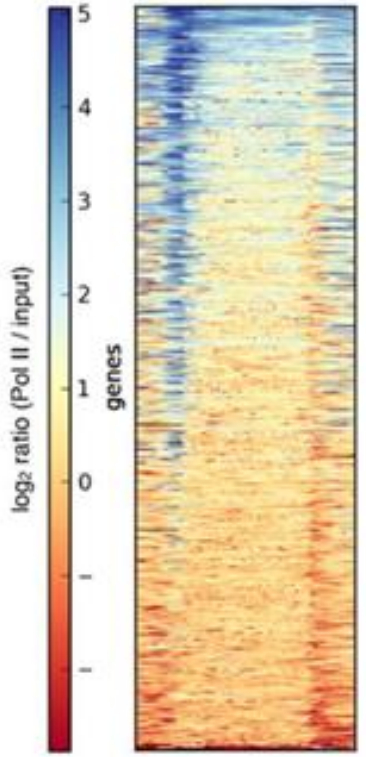
- Many types of NGS data such as RNA-seq and ChIP-seq provide quantitative information.
- How to look at signal across multiple loci?

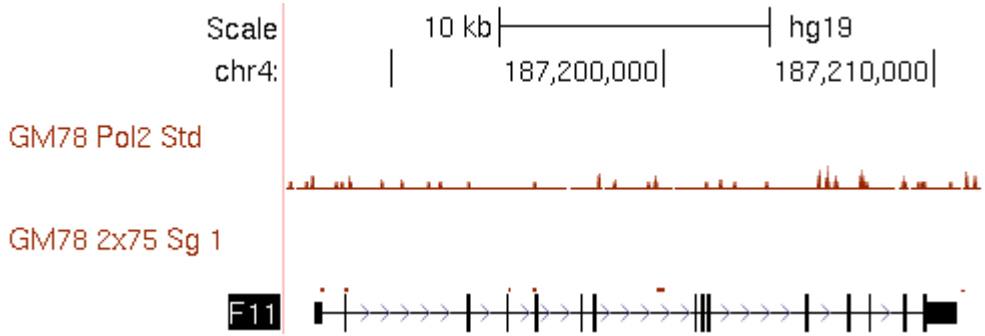


Example – PolII occupancy









Tools for making profiles

- deepTools (http://deeptools.ie-freiburg.mpg.de/)
  - Galaxy based tool
  - Functions to draw profiles as well as tools for QC.

- seqMINER (http://bips.u-strasbg.fr/)
  - Standalone Java-based tool
  - Limited functions but easy to use.
  - But sometimes a bit buggy…

deepTools

Table 1. Overview of currently available deepTools

Tool name	Type	Input files	Main output	Application
bamCorrelate	QC	2 or more BAM	Clustered heatmap of similarity measures	Determine Pearson or Spearman correlations between read distributions
bamFingerprint	QC	2 BAM	Diagnostic plot	Assess enrichment strength of a ChIP-seq sample versus a control
computeGCBias	QC	1 BAM	Diagnostic plots	Compare expected and observed GC distribution of reads
correctGCBias	Normalization	1 BAM	BAM or bigWig	Obtain GC-corrected read (coverage) file
bamCoverage	Normalization	1 BAM	bedGraph or bigWig	Obtain normalized read coverage of a single BAM
bamCompare	Normalization	2 BAM	bedGraph or bigWig	Normalize 2 BAM files to each other with a mathematical operation of Choice (fold change, log2 (ratio), sum, difference)
computeMatrix	Visualization	1 bigWig, min. 1 BED	gzipped table	Calculate the values for heatmaps and summary plots
profiler	Visualization	gzipped table from computeMatrix	xy-plot (summary plot)	Average profiles of read coverage for (groups of) genome regions
heatmapper	Visualization	gzipped table from computeMatrix	(Un)clustered heatmap or read coverages	Identify patterns of read coverages for genome regions

Here, we only indicate the main output files, but every data table underlying any image produced by deepTools can be downloaded and used in subsequent analyses. For a comparison of functionalities with previously published web servers, see Supplementary Table S1.

Ramirez et al. Nuc Acids Res. 42:W187-W191

Galaxy / deepTools | Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 0%

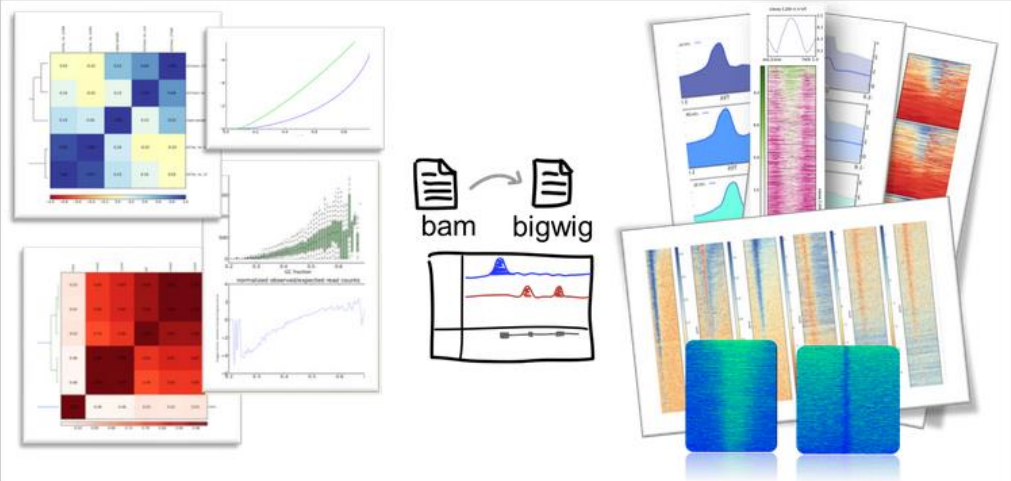
Tools

search tools

Get Data
Text Manipulation
Filter and Sort
Join, Subtract and Group
Statistics
deepTools

- heatmapper creates a heatmap for a score associated to genomic regions
- bigwigCompare normalizes and compares two bigWig files to obtain the ratio, log2ratio or difference
- bamCoverage generates a coverage bigWig file from a given BAM file. Multiple options are available to count reads and normalize coverage. (bam2bigwig)
- correctGCBias uses the output from computeGCBias to generate corrected BAM files
- bamCompare normalizes and compares two BAM files to obtain the ratio, log2ratio or difference. (bam2bigwig)
- bamFingerprint plots profiles of BAM files; useful for assesing ChIP signal strength
- profiler creates a profile plot for a score associated to genomic regions
- computeGCBias to see whether your samples should be normalized for GC bias

Welcome to the MPI-IE's Galaxy instance dedicated to the analysis of high-throughput sequencing data!



How to get started

Please make sure that you have read our Terms of Use. We also encourage you to subscribe to our deepTools mailing list (deeptools@googlegroups.com) where we will announce server maintenance times and program updates.

Every analysis will consist of the following 3 steps:

1. Data upload/import

- for *sample data*: go to "Shared Data" (**top menu**) --> "Data libraries" --> "Sample Data" --> Folder of choice --> "import to current history"

History

search datasets

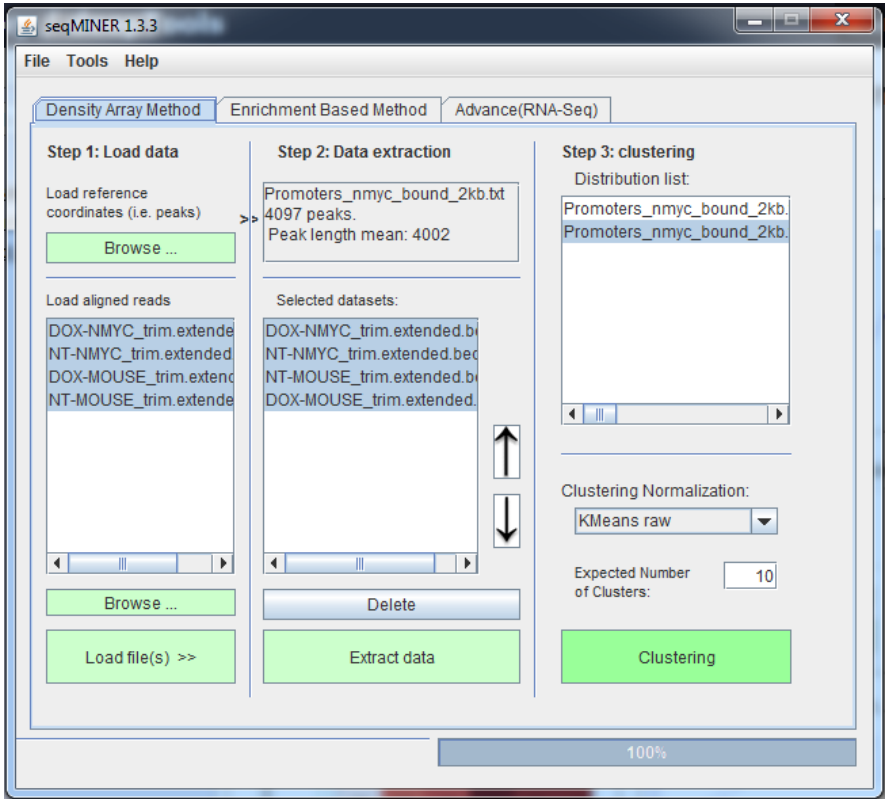
Unnamed history

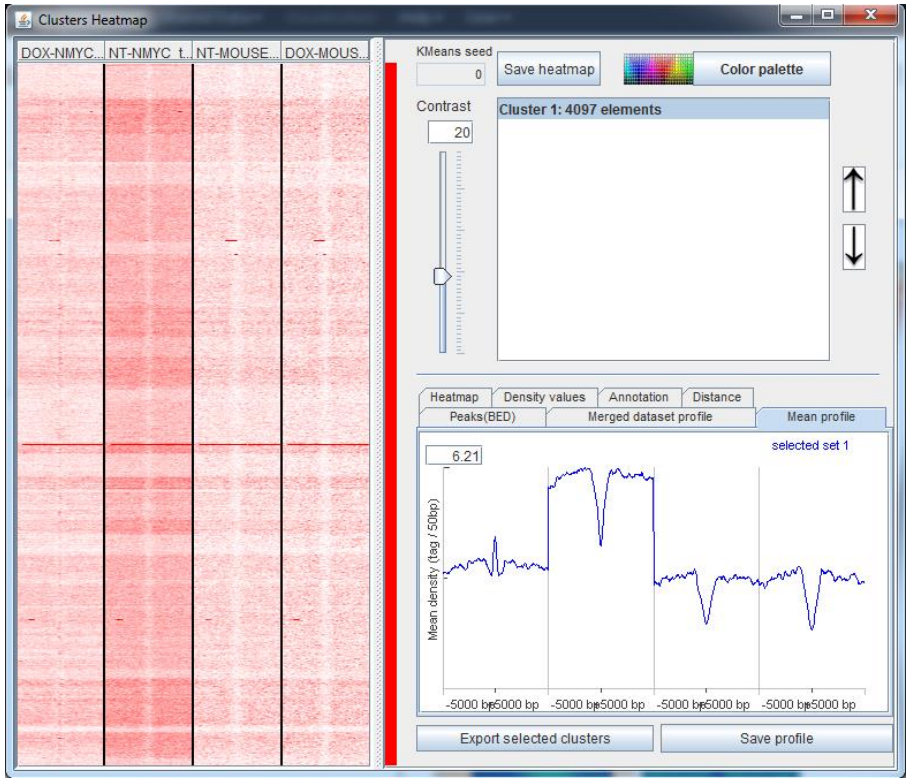
0 bytes

This history is empty. You can load your own data or get data from an external source

UNSW THE UNIVERSITY OF NEW SOUTH WALES

seqMINER

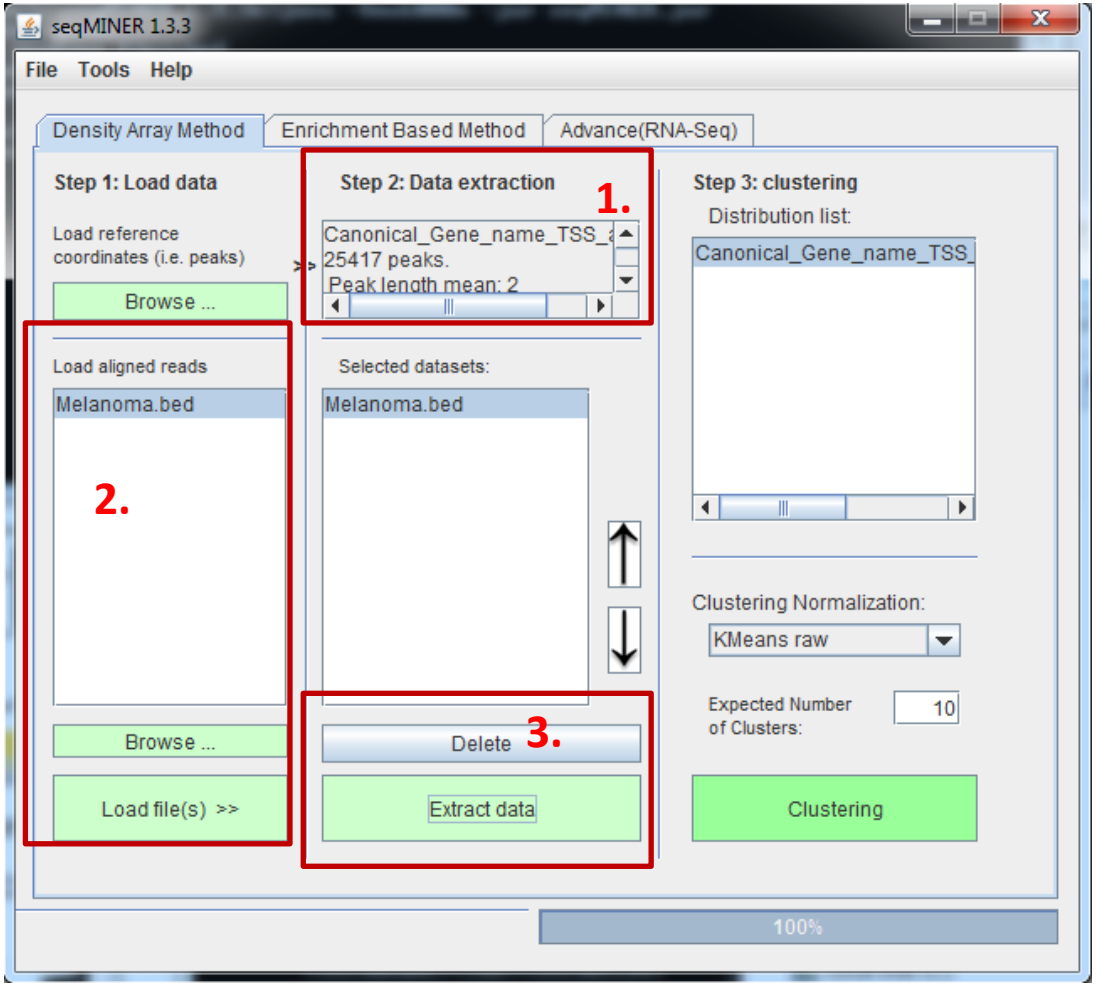


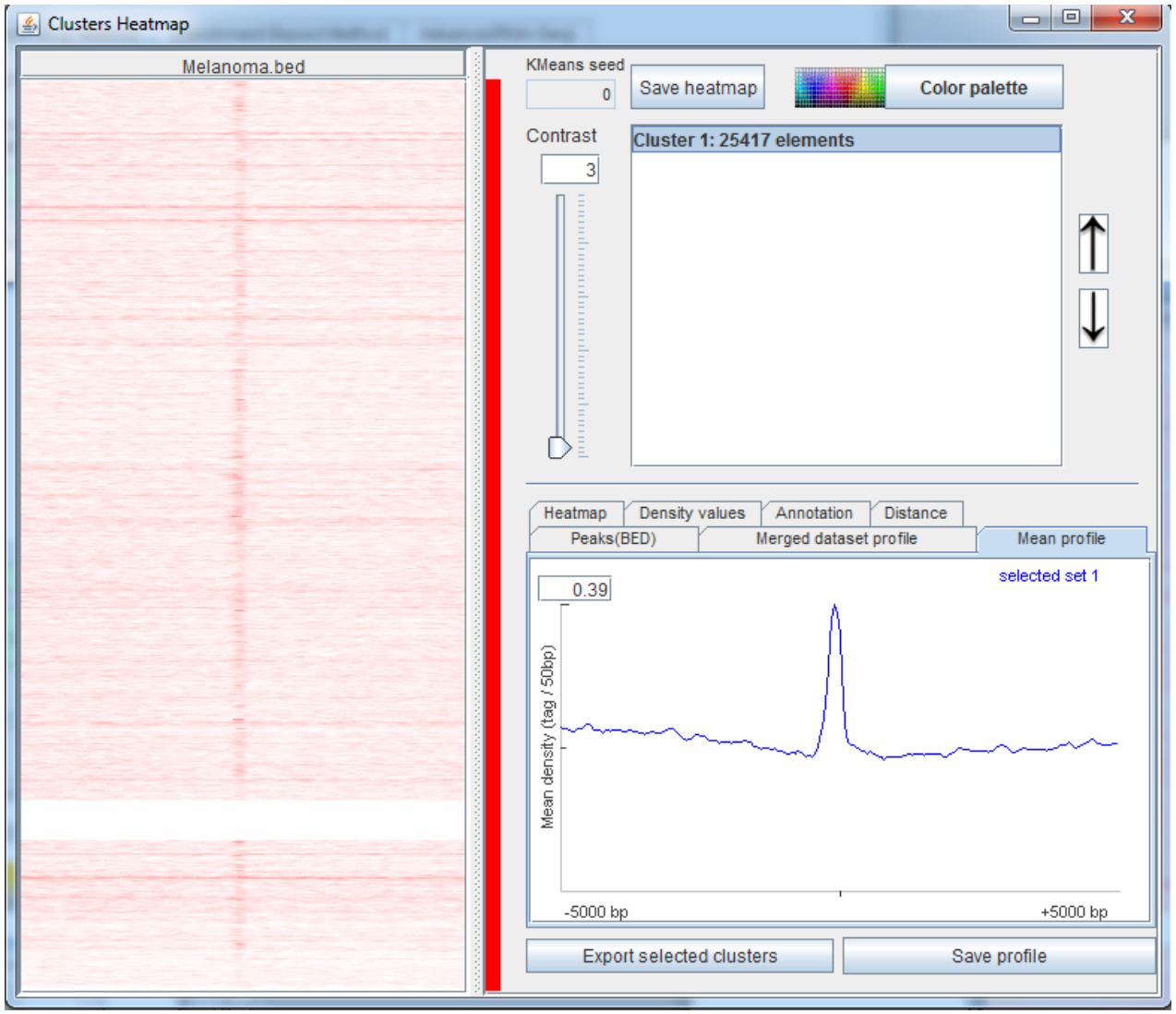


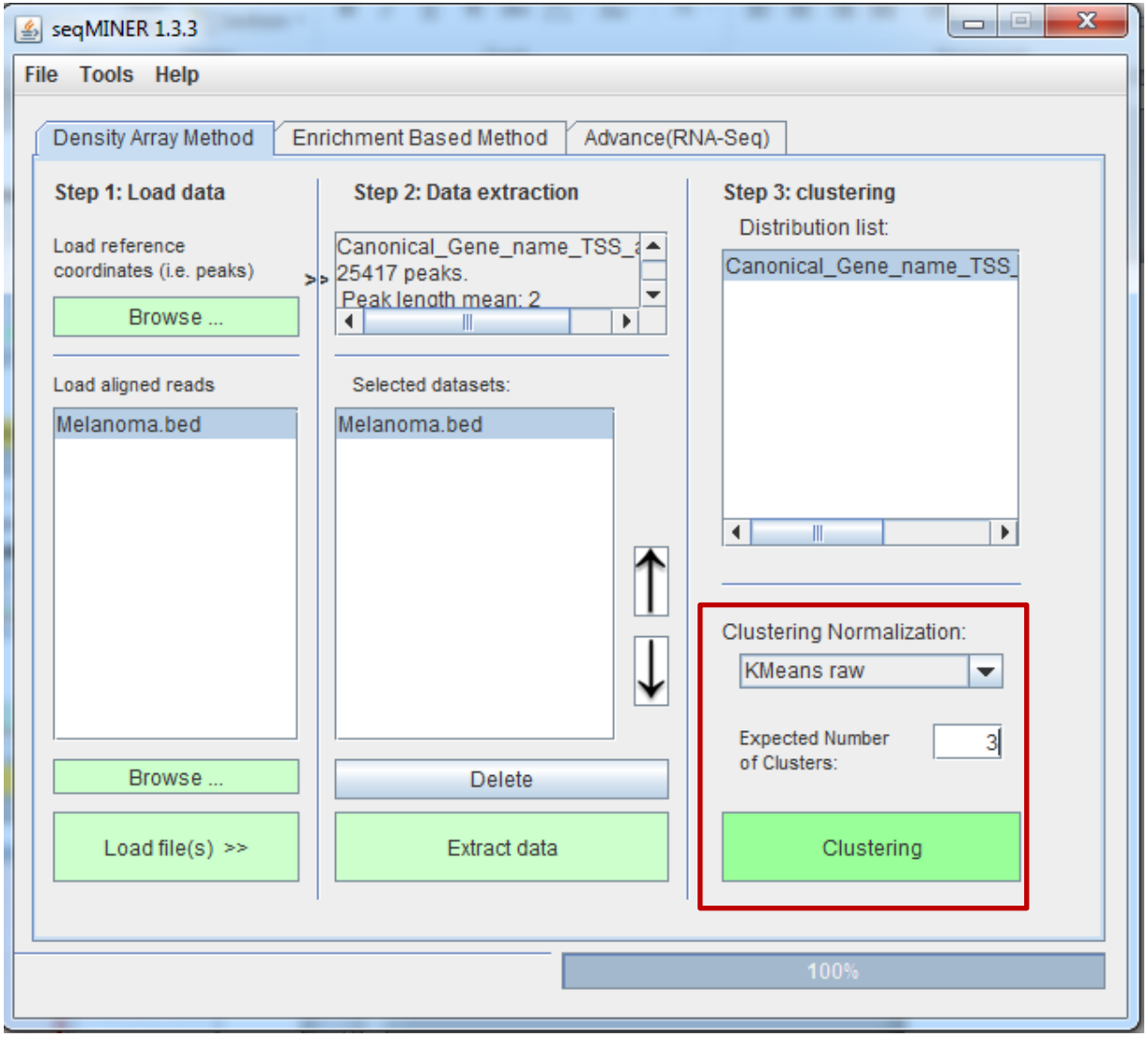
Ye et al. Nuc Acids Res. 39:e35

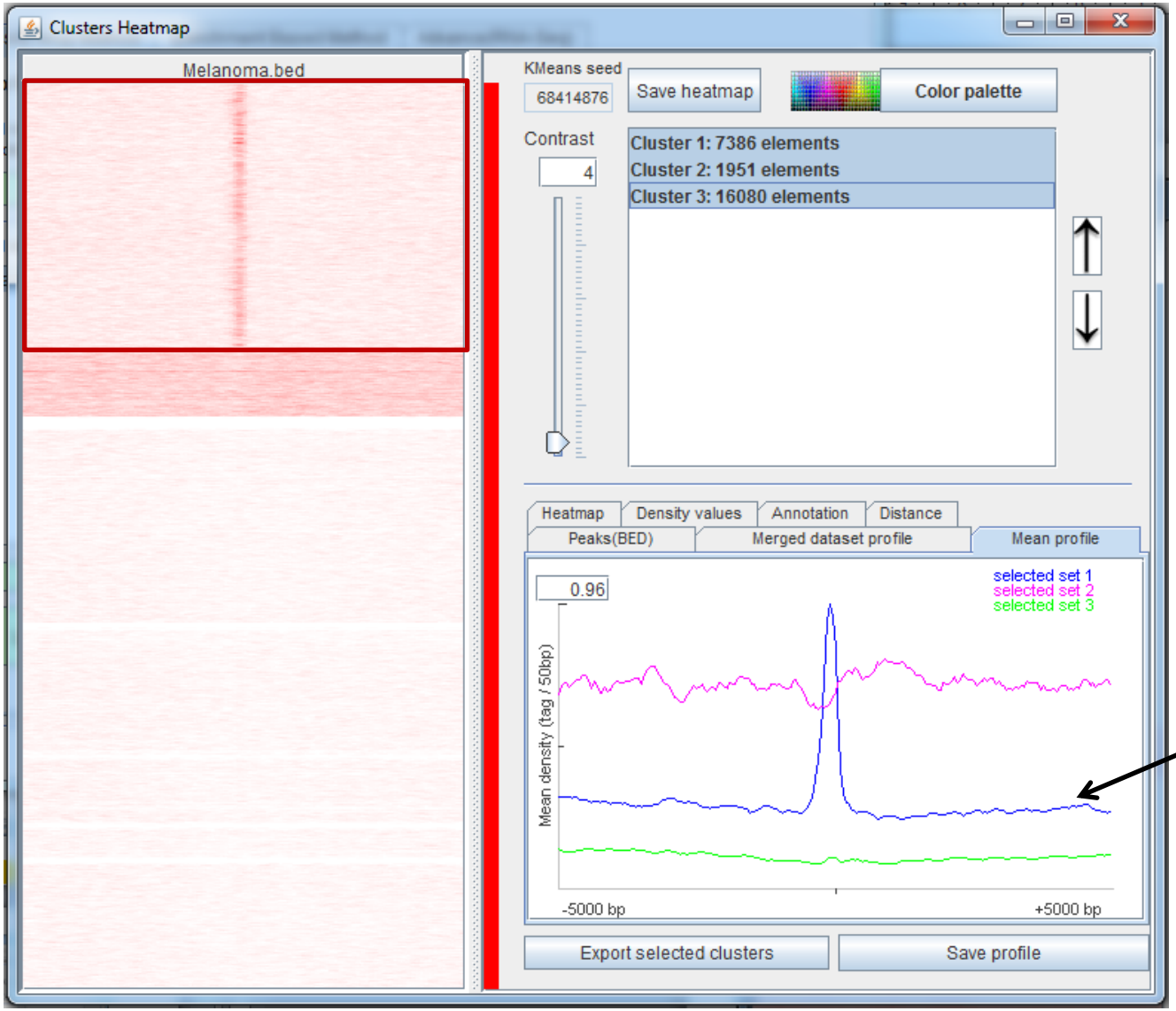
UNSW THE UNIVERSITY OF NEW SOUTH WALES

Example – distribution of melanoma mutation across gene promoters






seqMINER 1.3.3
File
Tools
Help
Density Array Method
Enrichment Based Method
Advance(RNA-Seq)
Step 1: Load data
Load reference coordinates (i.e. peaks)
Browse ...
Load aligned reads
Melanoma.bed
Browse ...
Load file(s) >>
Step 2: Data extraction
Canonical_Gene_name_TSS_
25417 peaks.
Peak length mean: 2
Selected datasets:
Melanoma.bed
Delete
Extract data
Step 3: clustering
Distribution list:
Canonical_Gene_name_TSS_
Clustering Normalization:
KMeans raw
Expected Number of Clusters:
3
Clustering
100%



Cluster 1 are mainly highly expressed genes.

Visualisation of gene expression data

- Gene expression data is typically generated by microarray or RNA-seq.

- Both used to generate expression level of mRNA in a sample

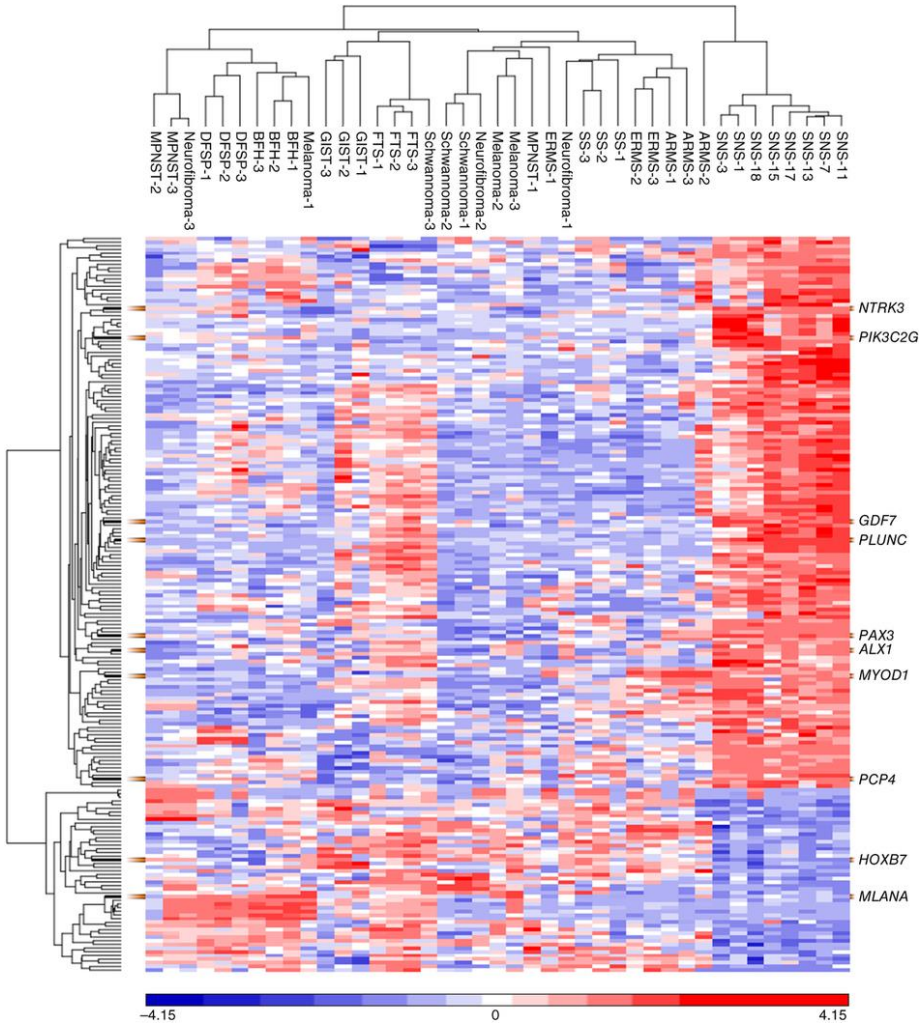
Things to consider for gene expression analysis

- Gene expression analysis is complex.

- The following usually needs to be considered in addition to visualisation:
  - Data normalisation
  - Batch effect removal
  - Appropriate statistical model for differential gene expression analysis

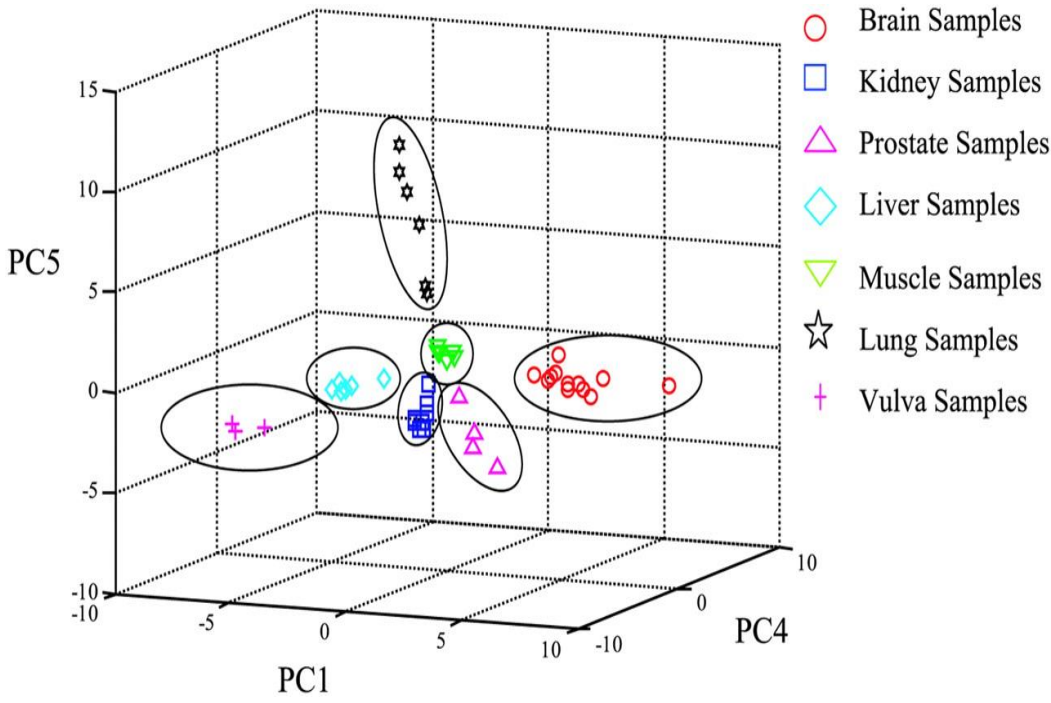
Hierarchical clustering

- Grouping of samples and/or genes based on similarity.
- Only major parameter is how to measure similarity.
- Effective for seeing how samples are different and whether clusters genes have similar expression profiles.

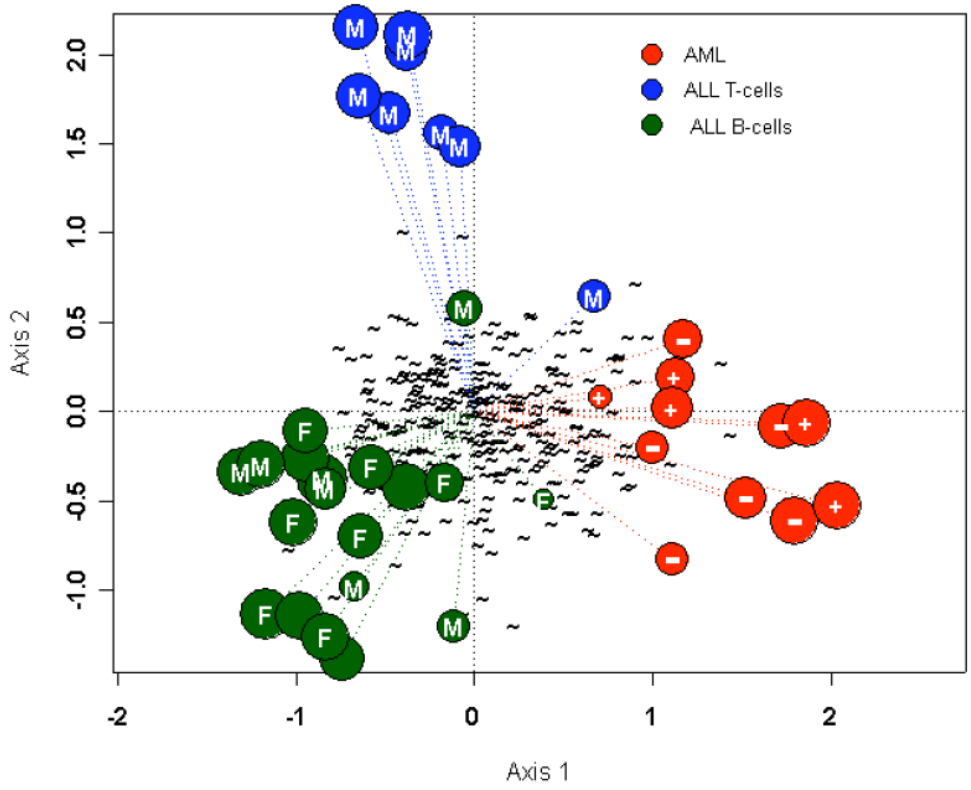


Principal component analysis (PCA)

- Reduces dimension of the data. (i.e. 20,000 genes into 3D).
- These new dimensions are represented as principal components (PC).
- Each PC captures a certain % of variation between samples such that PC1 captures the most.



Variations of PCA for gene expression analysis exists, such as the GE-biplot.

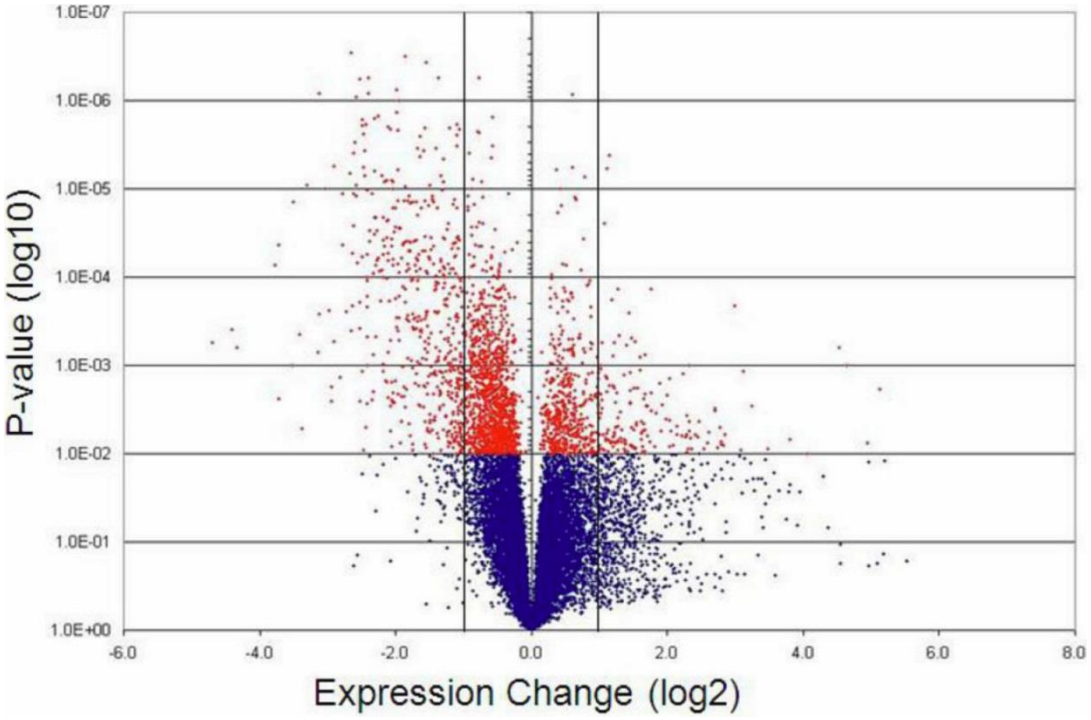


Data is transformed such that sample and genes are plot on the same axes and their directionality can be directly interpreted.

Pittelkow and Wilson (2003) Stat App Genetics Mol Biol 2:6

Visualising differential gene expression

- Each gene is a fold change and a p-value.
- Plotting these gives a volcano plot.

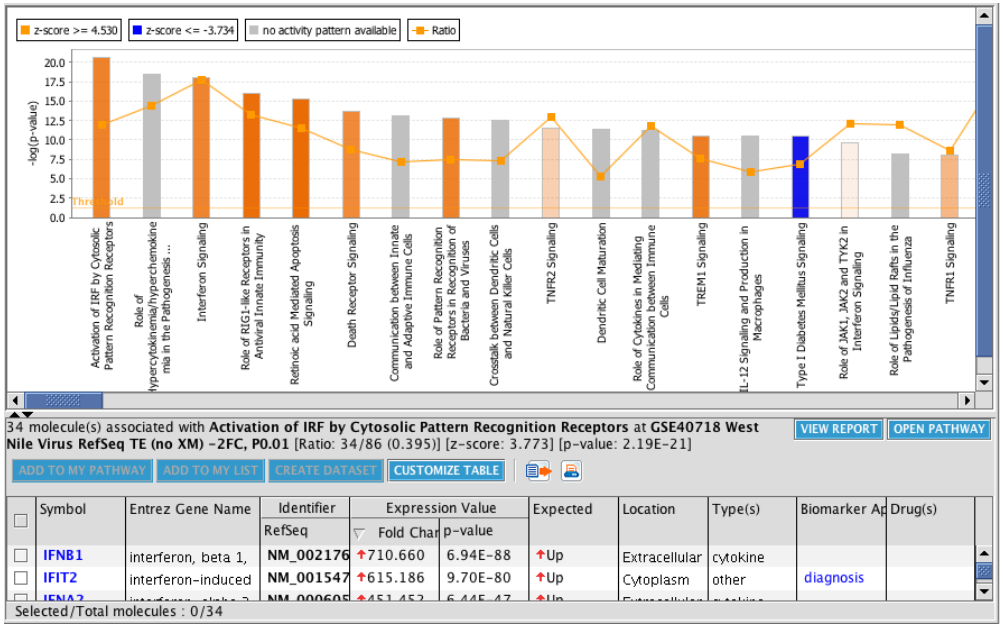


How to make these plots

- *R* is possibly most commonly used among bioinformaticians.
- Commercial packages such as Partek Genome Studio can also be handy for gene expression data visualisation.
- BUT always a good idea to consult a bioinformatician/biostatisticians who is familiar with gene expression analysis.

Pathway analysis

- Often the end product of gene expression analysis is a list of significant genes.
- Its difficult to look at each gene individually.
- It is usually more meaningful to see if they below to particular biological pathways.



Ingenuity IPA

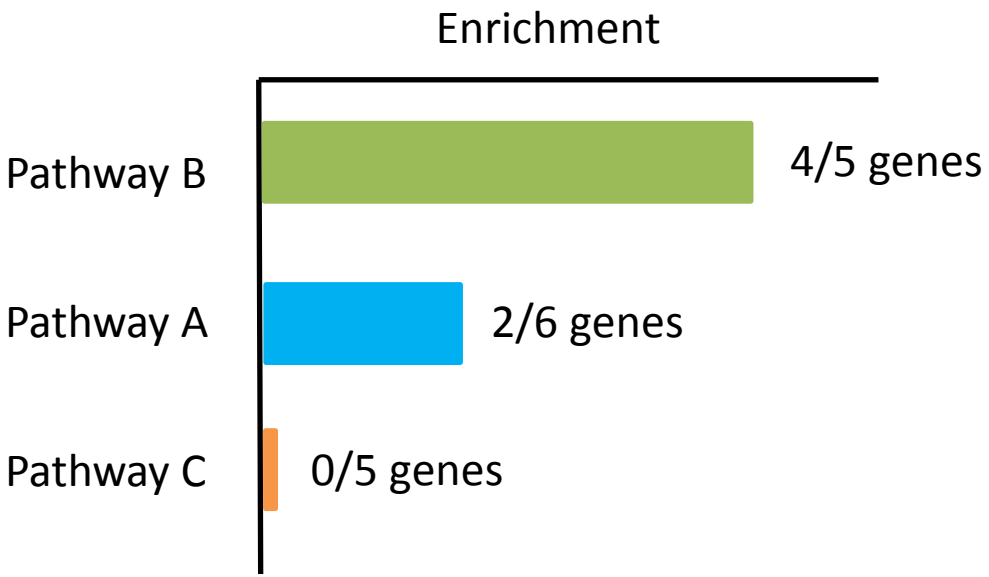
Pathway analysis example

My gene list
- Gene A
- Gene C
- Gene F
- Gene G
- Gene K

Pathway A
- **Gene A**
- Gene B
- Gene H
- **Gene K**
- Gene P
- Gene X

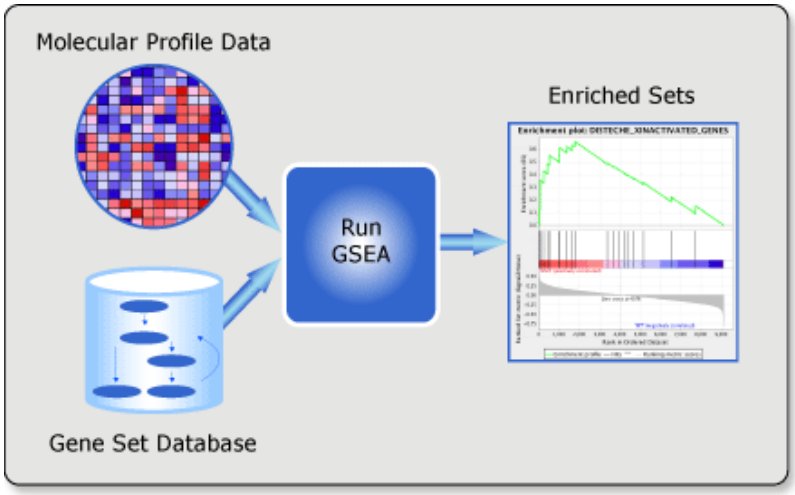
Pathway B
- **Gene C**
- **Gene F**
- **Gene G**
- **Gene K**
- Gene Z

Pathway C
- Gene B
- Gene R
- Gene S
- Gene T
- Gene U



Gene set enrichment analysis (GSEA)

- However, sometimes it is difficult to define a list of significant genes.
  - Cutoff might be arbitrary
  - Small sample size can make p-values difficult to interpret



- GSEA is an alterative to standard pathway analysis.

Interpreting GSEA plots

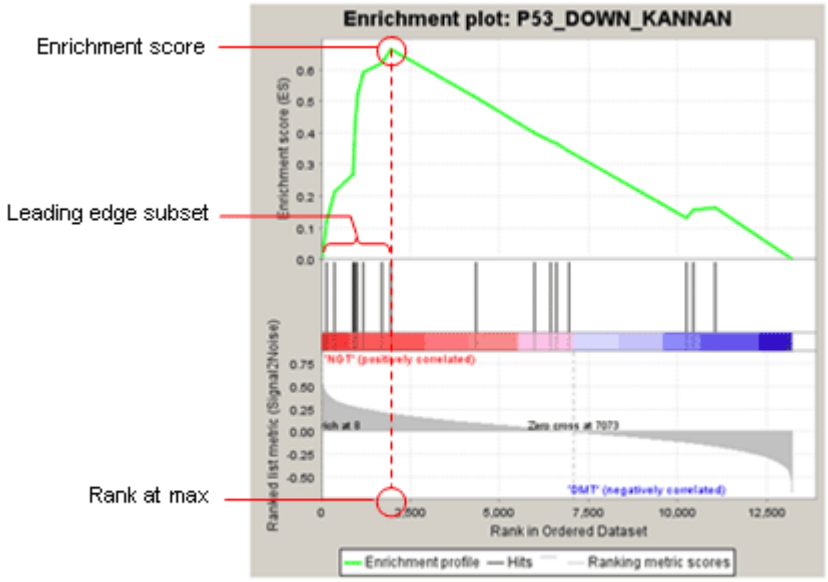
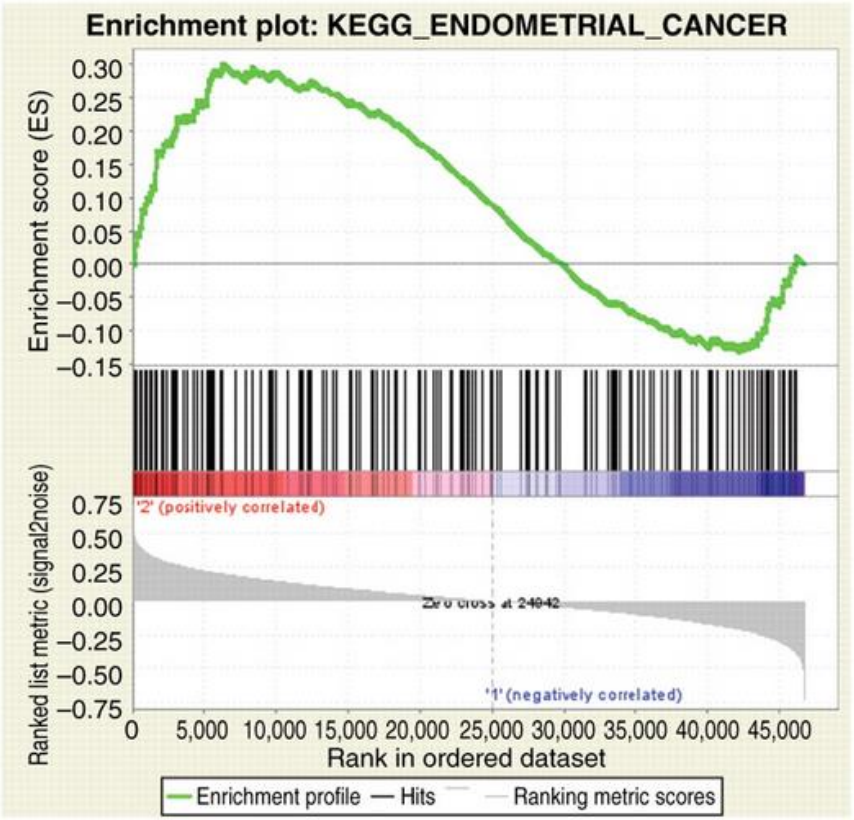
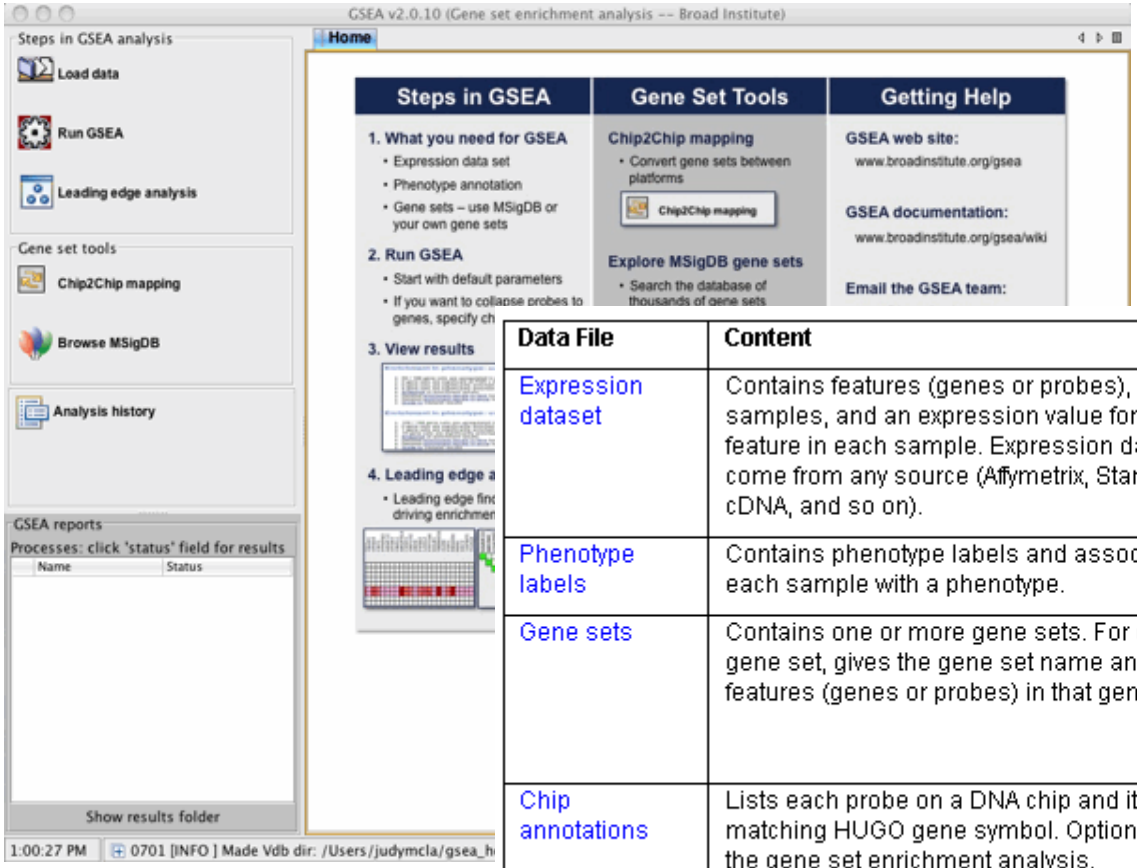


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List



Running GSEA



Data File	Content	Format	Source
Expression dataset	Contains features (genes or probes), samples, and an expression value for each feature in each sample. Expression data can come from any source (Affymetrix, Stanford cDNA, and so on).	res, gct, pcl, or txt	You create the file.
Phenotype labels	Contains phenotype labels and associates each sample with a phenotype.	cls	You create the file or have GSEA create it for you.
Gene sets	Contains one or more gene sets. For each gene set, gives the gene set name and list of features (genes or probes) in that gene set.	gmx or gmt	You use the files on the Broad ftp site, export gene sets from the Molecular Signature Database (MSigDb) or create your own gene sets file.
Chip annotations	Lists each probe on a DNA chip and its matching HUGO gene symbol. Optional for the gene set enrichment analysis.	Chip	You use the files on the Broad ftp site, download the files from the GSEA web site, or create your own chip file.

http://software.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html

- Do you have any other data types that need to be visualised?

Further reading

- Circos
  - www.ncbi.nlm.nih.gov/pubmed/19541911
  - https://www.adelaide.edu.au/bioinformatics-hub/seminars-workshops/notes-handouts/circos.pdf
- deepTools
  - www.ncbi.nlm.nih.gov/pubmed/24799436
- seqMiner
  - www.ncbi.nlm.nih.gov/pubmed/21177645
- Clustering gene expression data
  - www.ncbi.nlm.nih.gov/pubmed/16333293
- GSEA
  - www.ncbi.nlm.nih.gov/pubmed/16199517