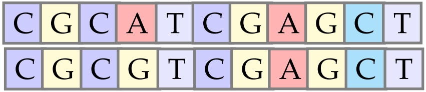
Pattern Recognition In Clinical Data

Saket Choudhary
Dual Degree Project

Guide: Prof. Santosh Noronha



June 30, 2014

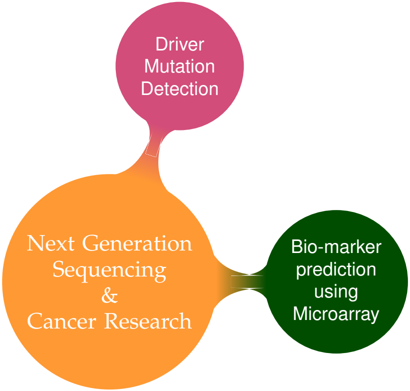
OBJECTIVE

Next Generation
Sequencing
&
Cancer Research

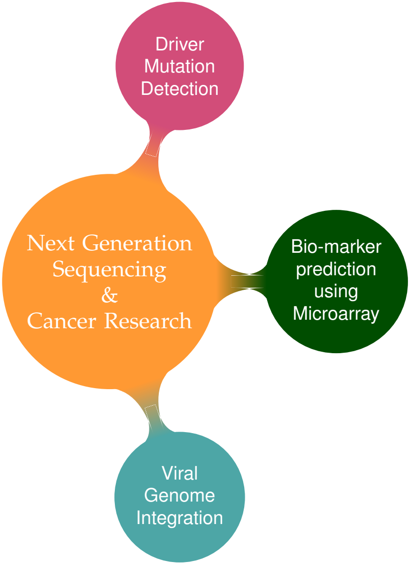
OBJECTIVE


Driver Mutation Detection
Next Generation Sequencing & Cancer Research

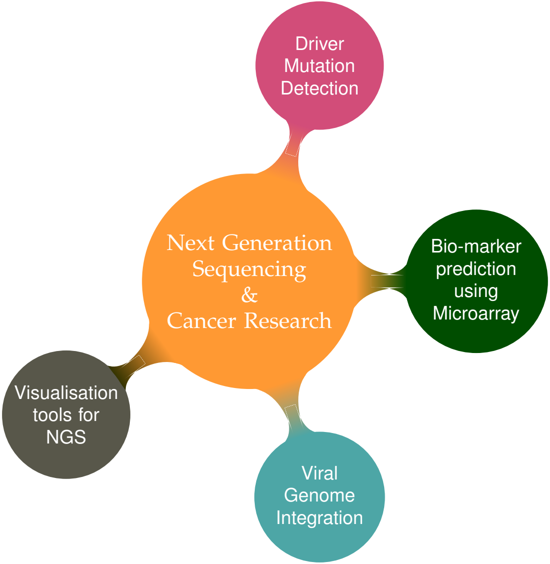
Objective


Driver Mutation Detection
Next Generation Sequencing & Cancer Research
Bio-marker prediction using Microarray

OBJECTIVE

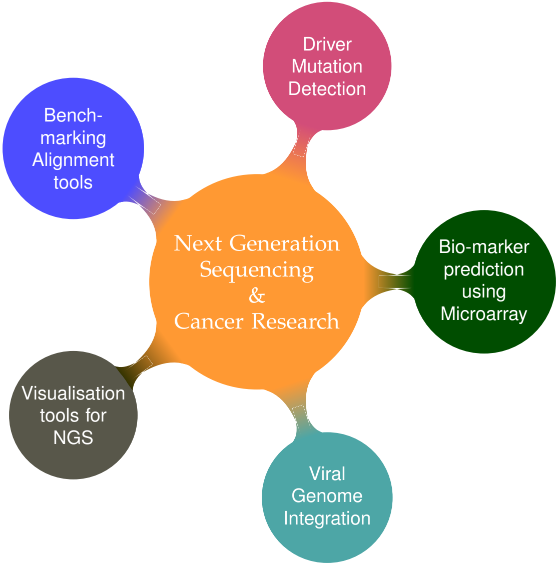

Driver Mutation Detection
Next Generation Sequencing & Cancer Research
Bio-marker prediction using Microarray
Viral Genome Integration

OBJECTIVE

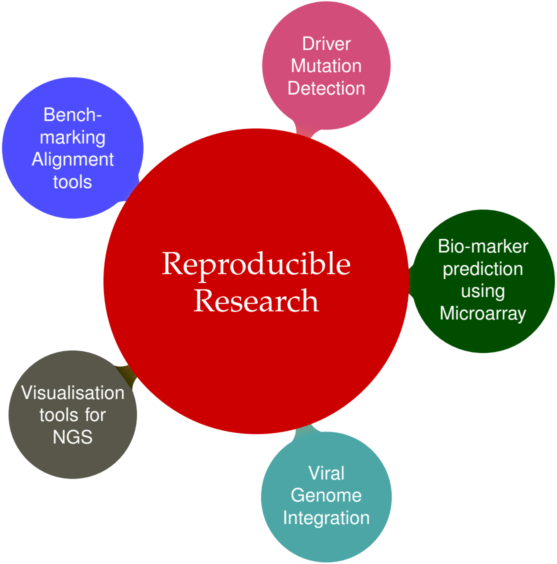


Driver Mutation Detection
Next Generation Sequencing & Cancer Research
Bio-marker prediction using Microarray
Visualisation tools for NGS
Viral Genome Integration

OBJECTIVE


Next Generation Sequencing & Cancer Research
Driver Mutation Detection
Bio-marker prediction using Microarray
Viral Genome Integration
Visualisation tools for NGS
Bench-marking Alignment tools

OBJECTIVE


Reproducible Research
Driver Mutation Detection
Bio-marker prediction using Microarray
Viral Genome Integration
Visualisation tools for NGS
Bench-marking Alignment tools

Workflows for Driver Mutation Detection

- Cancer = Lots of Mutations!
- Driver mutations confer selective advantage to the cell, being selected positively.
- Sites of driver mutation are targeted therapeutic sites, prognosis markers

Problem

- Multiple prediction tools
- Different score-range for prediction
- Non overlapping results, non-overlapping formats

Aim

Unify the various predictions, to help nail down the consensus

Workflows for Driver Mutation Detection

- Cancer = Lots of Mutations!
- Driver mutations confer selective advantage to the cell, being selected positively.
- Sites of driver mutation are targeted therapeutic sites, prognosis markers

Problem

- Multiple prediction tools
- Different score-range for prediction
- Non overlapping results, non-overlapping formats

Aim

Unify the various predictions, to help nail down the consensus

Workflows for Driver Mutation Detection

- Cancer = Lots of Mutations!
- Driver mutations confer selective advantage to the cell, being selected positively.
- Sites of driver mutation are targeted therapeutic sites, prognosis markers

Problem

- Multiple prediction tools
- Different score-range for prediction
- Non overlapping results, non-overlapping formats

Aim

Unify the various predictions, to help nail down the consensus

Workflows for Driver Mutation Detection

- Cancer = Lots of Mutations!
- Driver mutations confer selective advantage to the cell, being selected positively.
- Sites of driver mutation are targeted therapeutic sites, prognosis markers

Problem

- Multiple prediction tools
- Different score-range for prediction
- Non overlapping results, non-overlapping formats

Aim

Unify the various predictions, to help nail down the consensus

WORKFLOWS FOR DRIVER MUTATION DETECTION

- Cancer = Lots of Mutations!
- Driver mutations confer selective advantage to the cell, being selected positively.
- Sites of driver mutation are targeted therapeutic sites, prognosis markers

Problem

- Multiple prediction tools
- Different score-range for prediction
- Non overlapping results, non-overlapping formats

Aim

Unify the various predictions, to help nail down the consensus

WORKFLOWS FOR DRIVER MUTATION DETECTION

- Cancer = Lots of Mutations!
- Driver mutations confer selective advantage to the cell, being selected positively.
- Sites of driver mutation are targeted therapeutic sites, prognosis markers

Problem

- Multiple prediction tools
- Different score-range for prediction
- Non overlapping results, non-overlapping formats

Aim

Unify the various predictions, to help nail down the consensus

Workflows for Driver Mutation Detection

- Cancer = Lots of Mutations!
- Driver mutations confer selective advantage to the cell, being selected positively.
- Sites of driver mutation are targeted therapeutic sites, prognosis markers

Problem

- Multiple prediction tools
- Different score-range for prediction
- Non overlapping results, non-overlapping formats

Aim

Unify the various predictions, to help nail down the consensus

WORKFLOWS FOR DRIVER MUTATION DETECTION

Approach

- Wrap the tools in a toolbox using Galaxy
- Galaxy is a web based framework for running bioinformatic workflows, with focus on reproducibility of the analyses
- Combine all scores and render it as a heatmap. Easy way to pick up few target mutations

WORKFLOWS FOR DRIVER MUTATION DETECTION

Approach

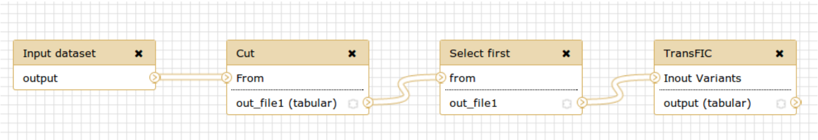
- Wrap the tools in a toolbox using Galaxy
- Galaxy is a web based framework for running bioinformatic workflows, with focus on reproducibility of the analyses
- Combine all scores and render it as a heatmap. Easy way to pick up few target mutations

WORKFLOWS FOR DRIVER MUTATION DETECTION

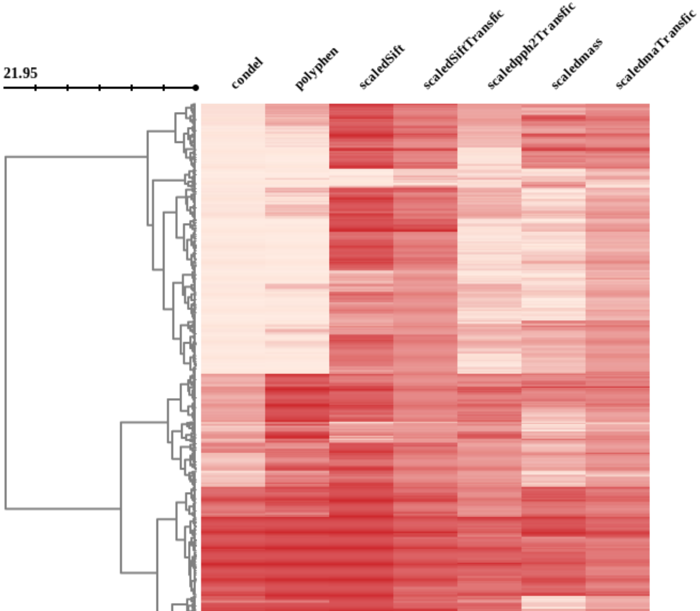
Approach

- Wrap the tools in a toolbox using Galaxy
- Galaxy is a web based framework for running bioinformatic workflows, with focus on reproducibility of the analyses
- Combine all scores and render it as a heatmap. Easy way to pick up few target mutations

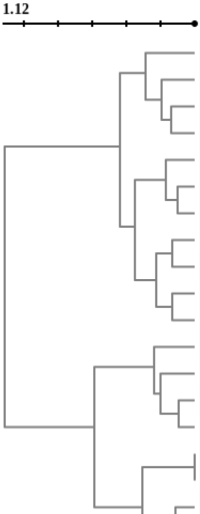
A sample workflow :



Here is how you **visualise**:



Here is how you **focus**:



condel	polyphen	scaledSift	scaledSiftTransfic	scaledpph2Transfic	scaledmass	scaledmaTransfic
0.051	0.416	1	0.882	0.466	0.419	0.548
0.054	0.35	0.96	0.692	0.433	0.508	0.565
0.069	0.438	0.96	0.694	0.46	0.648	0.641
0.067	0.483	0.99	0.774	0.471	0.565	0.597
0.075	0.477	0.86	0.624	0.47	0.327	0.517
0.077	0.485	0.96	0.694	0.472	0.317	0.513
0.061	0.431	0.98	0.731	0.439	0.365	0.478
0.072	0.446	0.89	0.624	0.44	0.489	0.709
0.084	0.496	0.9	0.633	0.468	0.521	0.6
0.068	0.411	0.95	0.663	0.432	0.378	0.646
0.075	0.477	0.86	0.634	0.459	0.327	0.665
0.046	0.266	0.95	0.685	0.415	0.705	0.617
0.066	0.418	0.96	0.689	0.436	0.892	0.657
0.044	0.299	0.97	0.702	0.381	0.87	0.693
0.033	0.217	0.97	0.707	0.381	0.889	0.655
0.035	0.269	0.73	0.585	0.415	0.74	0.667
0.035	0.269	0.73	0.585	0.415	0.74	0.667
0.047	0.322	0.82	0.607	0.439	0.663	0.621

Problems:

- No way to run multiple tools on a dataset without data-fiddling
- Lack of a way to combine these predictions
- Irreproducibility => What cut-offs used to filter drivers?(much more than this)

Solutions :

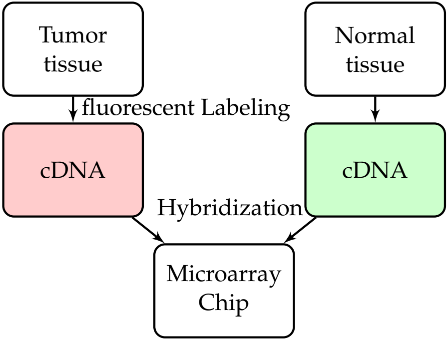
- Run multiple tools(in parallel) on the same dataset
- Combine predictions, visualise, focus
- Perfectly reproducible analyses

BIO-MARKER PREDICTION USING MICROARRAY DATA

Problem Definition

Given a set of gene expression values of two sets of patients: *normal* and cancer, *predict* a small subset of genes that could be used to differentiate these.

Microarrays



MICROARRAY: QUESTIONS WE ARE TRYING TO ANSWER

Questions

- Given expression data of 17000 genes, which of these genes are *differentially expressed*
- Among the differentially expressed set of genes, which genes show maximum association (+/-) with the *cohort*
- Is there a very small subset(5/10/20...) that can help differentiate the unknown samples

Microarray: Questions we are trying to answer

Questions

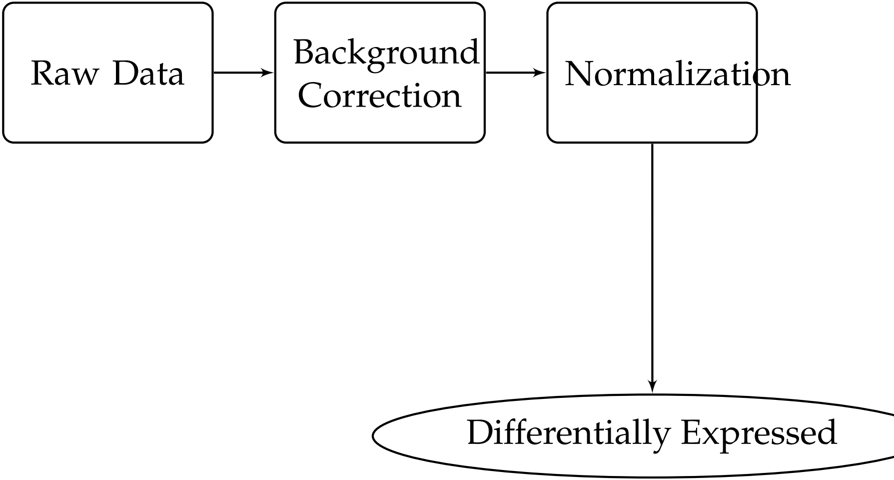
- Given expression data of 17000 genes, which of these genes are *differentially expressed*
- Among the differentially expressed set of genes, which genes show maximum association (+/-) with the *cohort*
- Is there a very small subset(5/10/20...) that can help differentiate the unknown samples

MICROARRAY: QUESTIONS WE ARE TRYING TO ANSWER

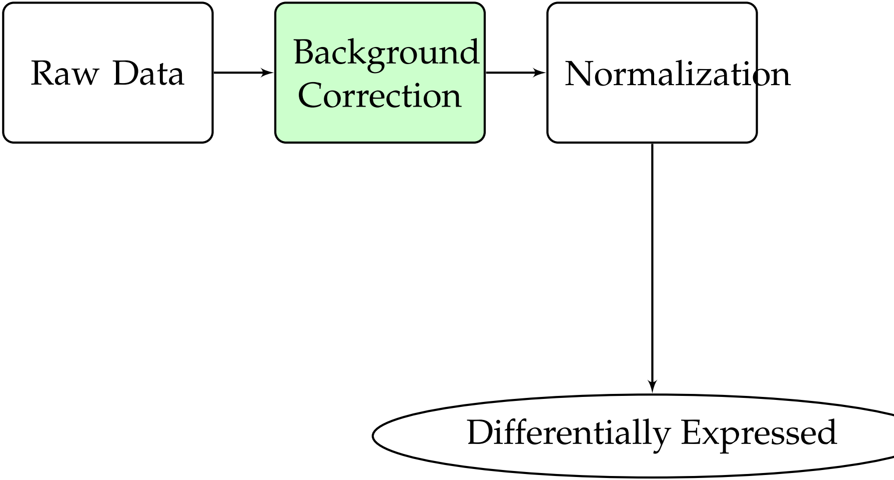
Questions

- Given expression data of 17000 genes, which of these genes are *differentially expressed*
- Among the differentially expressed set of genes, which genes show maximum association (+/-) with the *cohort*
- Is there a very small subset(5/10/20...) that can help differentiate the unknown samples

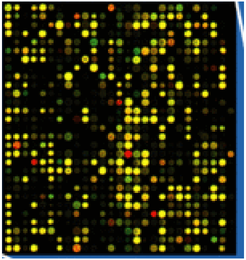
PRE-PROCESSING[STANDARD WORKFLOW]



BACKGROUND CORRECTION



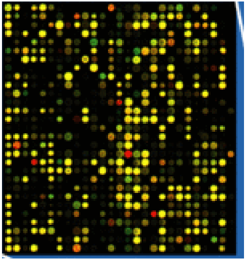
BACKGROUND CORRECTION



The Need

- Microarray spot intensities have two components: foreground + background
- Background may arise due to non-specific binding
- Important step to correct for ambient intensity around a spot

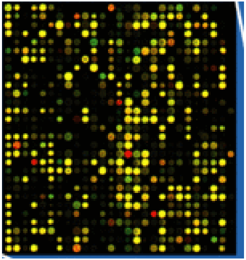
BACKGROUND CORRECTION



The Need

- Microarray spot intensities have two components: foreground + background
- Background may arise due to non-specific binding
- Important step to correct for ambient intensity around a spot

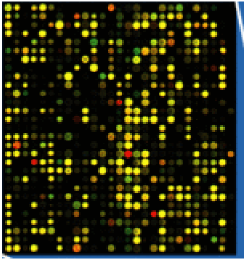
BACKGROUND CORRECTION



The Need

- Microarray spot intensities have two components: foreground + background
- Background may arise due to non-specific binding
- Important step to correct for ambient intensity around a spot

BACKGROUND CORRECTION



The Need

- Microarray spot intensities have two components: foreground + background
- Background may arise due to non-specific binding
- Important step to correct for ambient intensity around a spot

Näive approach: Subtract background intensities from the foreground
What's not right?: How does one interpret negative intensities?(Loss of information + bias)[Remember, background is itself measured from the nearby spots and not that one spot directly]
Alternate:

- Model observed [foreground-background] as sum of exponential (true) and normal (random noise)

$$S = B + T + S_b \tag{1}$$

S = foreground,
S_b = background
T = True signal

B = Random noise We model $S - S_b$ [observed intensity]

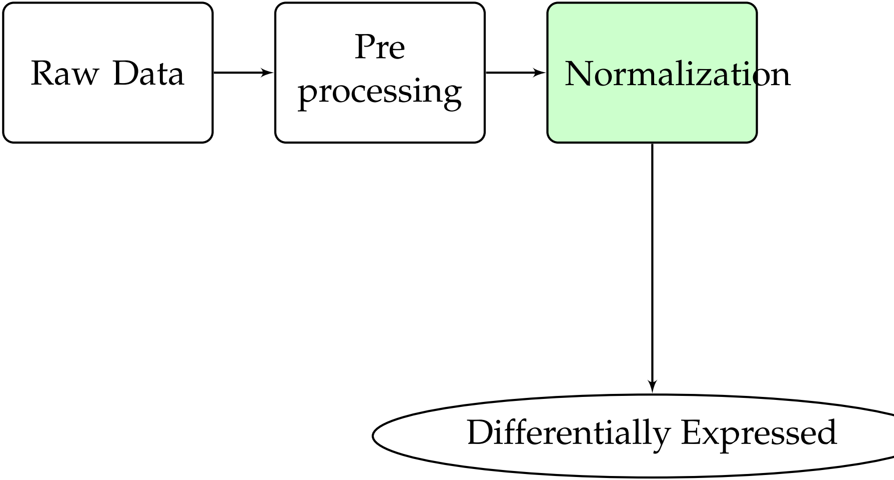
$$T \sim \frac{1}{\alpha} exp \frac{-t}{\alpha} \tag{2}$$

$t > 0$,

$$B \sim \mathcal{N}(\mu, \sigma^2) \tag{3}$$

μ, σ, α are unknowns
[Details later]

NORMALIZATION



NORMALISATION

The Need

- The expression levels of majority genes should be the same across arrays. This should be reflected in the overall intensity
- Adjust for effects arising due to array-to-array manufacture differences, different amounts of dye, different amount of hybridising sample etc

Objective

- Overall distribution of expression levels across arrays should be similar

Normalisation

The Need

- The expression levels of majority genes should be the same across arrays. This should be reflected in the overall intensity
- Adjust for effects arising due to array-to-array manufacture differences, different amounts of dye, different amount of hybridising sample etc

Objective

- Overall distribution of expression levels across arrays should be similar

NORMALISATION

The Need

- The expression levels of majority genes should be the same across arrays. This should be reflected in the overall intensity
- Adjust for effects arising due to array-to-array manufacture differences, different amounts of dye, different amount of hybridising sample etc

Objective

- Overall distribution of expression levels across arrays should be similar

Quantile Normalization

- Associate the highest value of dataset X to highest value of dataset Y, and so on...
- A Q-Q plot, thereafter would be a perfect diagonal

Quantile Normalization

- Associate the highest value of dataset X to highest value of dataset Y, and so on...
- A Q-Q plot, thereafter would be a perfect diagonal

NORMALIZATION

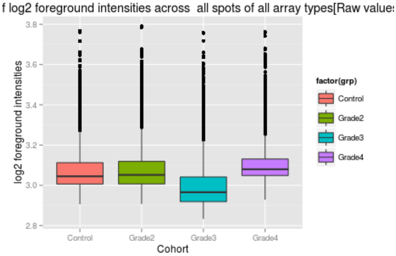


Figure: Raw intensities

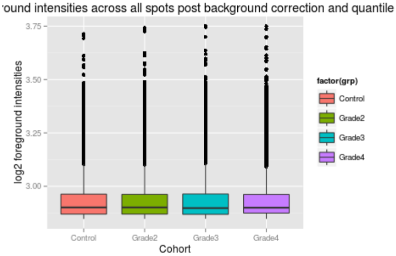
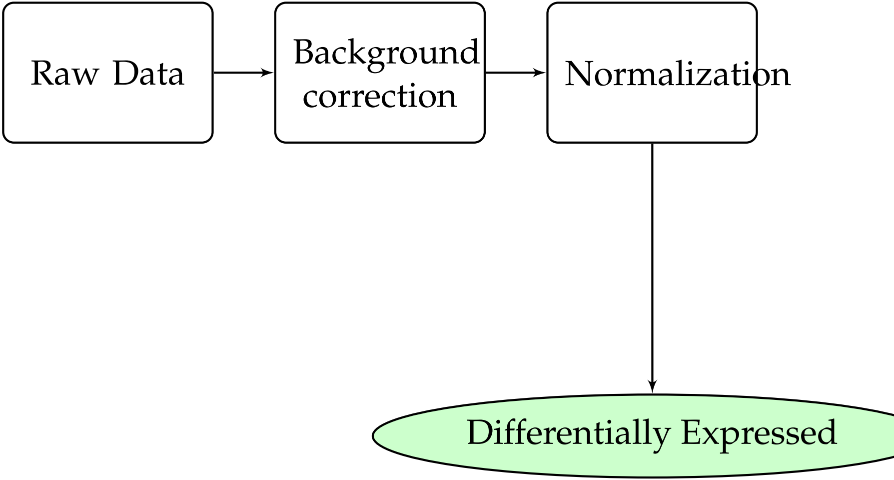


Figure: Normalized intensities

DIFFERENTIAL EXPRESSION



DIFFERENTIAL EXPRESSION I

Hypothesis

H_0: Gene X is not differentially expressed[Expression levels in the two cohorts are same]
H_1: Gene X is differentially expressed[up/down regulated]

- This is tested for **multiple** genes.[17000 of them].
- Any test statistic employed should be able to control for multiple testing. [Details later]

DIFFERENTIAL EXPRESSION II

We use a modified version of t-test. [Details later]
t-test :

$$z_i = \frac{\bar{x}_i^C - \bar{x}_i^D}{s_i} \tag{4}$$

$$s_i = \sqrt{\frac{sc_i^2}{N_C} + \frac{sd_i^2}{N_D}} \tag{5}$$

where sc_i and sd_i are the standard deviations with sample sizes N_C and N_D for the control and disease respectively.
This z_i statistic follows a t-distribution:

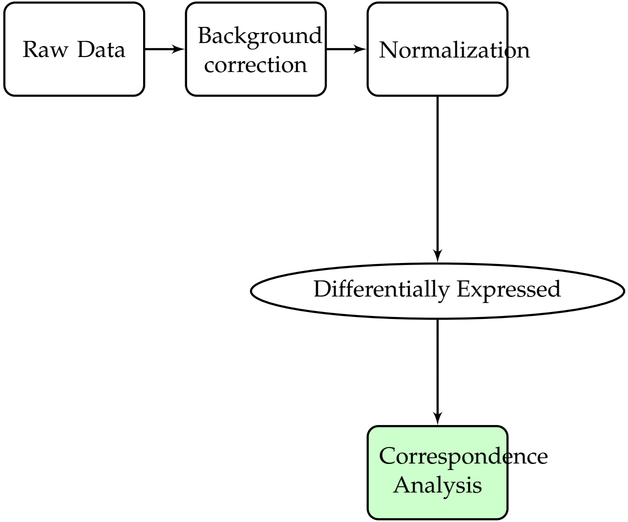
$$z_i \sim t_i \tag{6}$$

The associated p-value is given by:

DIFFERENTIAL EXPRESSION III

$$p - value = 2 * P(t_i \geq |z_i|) \quad (7)$$

SO FAR..


Raw Data
Background correction
Normalization
Differentially Expressed
Correspondence Analysis

DIMENSIONALITY REDUCTION

The Need

- The list of differentially expressed genes is too long, interpretation still not trivial
- How does one infer associations between the gene expressions and the cohorts?
  - p-values are not indicative of associations
  - log fold changes are (Ratio of average expression over cohorts) biologically important, they are already part of this long sublist, hence uninformative post the filtering step.

DIMENSIONALITY REDUCTION

The Need

- The list of differentially expressed genes is too long, interpretation still not trivial
- How does one infer associations between the gene expressions and the cohorts?
  - p-values are not indicative of associations
  - log fold changes are (Ratio of average expression over cohorts) biologically important, they are already part of this long sublist, hence uninformative post the filtering step.

DIMENSIONALITY REDUCTION

The Need

- The list of differentially expressed genes is too long, interpretation still not trivial
- How does one infer associations between the gene expressions and the cohorts?
  - p-values are not indicative of associations
  - log fold changes are (Ratio of average expression over cohorts) biologically important, they are already part of this long sublist, hence uninformative post the filtering step.

DIMENSIONALITY REDUCTION

The Need

- The list of differentially expressed genes is too long, interpretation still not trivial
- How does one infer associations between the gene expressions and the cohorts?
  - p-values are not indicative of associations
  - log fold changes are (Ratio of average expression over cohorts) biologically important, they are already part of this long sublist, hence uninformative post the filtering step.

Approach

- Project data in higher dimension(2000+ at times) to a lower dimension
- The data in lower-dimension should be a reflective of the higher-dimension data
- Try to determine that subset of genes that reveal information between the expression levels and associated cohort
- Try to avoid any kind of model assumptions

Approach

- Project data in higher dimension(2000+ at times) to a lower dimension
- The data in lower-dimension should be a reflective of the higher-dimension data
- Try to determine that subset of genes that reveal information between the expression levels and associated cohort
- Try to avoid any kind of model assumptions

Approach

- Project data in higher dimension(2000+ at times) to a lower dimension
- The data in lower-dimension should be a reflective of the higher-dimension data
- Try to determine that subset of genes that reveal information between the expression levels and associated cohort
- Try to avoid any kind of model assumptions

Approach

- Project data in higher dimension(2000+ at times) to a lower dimension
- The data in lower-dimension should be a reflective of the higher-dimension data
- Try to determine that subset of genes that reveal information between the expression levels and associated cohort
- Try to avoid any kind of model assumptions

Approach

- Project data in higher dimension(2000+ at times) to a lower dimension
- The data in lower-dimension should be a reflective of the higher-dimension data
- Try to determine that subset of genes that reveal information between the expression levels and associated cohort
- Try to avoid any kind of model assumptions

CORRESPONDENCE ANALYSIS

Underlying hypothesis

There is no association between the rows[genes] and columns[samples]

- Project data to first 2 or 3 **informative** coordinates
- Treats rows(genes) and columns(samples) equivalently
- Attempts to separate dissimilar objects from each other(both genes and samples simultaneously)
- Unlike the more *famous* PCA, reveals the association between genes and samples(biplots)

CORRESPONDENCE ANALYSIS

Underlying hypothesis

There is no association between the rows[genes] and columns[samples]

- Project data to first 2 or 3 **informative** coordinates
- Treats rows(genes) and columns(samples) equivalently
- Attempts to separate dissimilar objects from each other(both genes and samples simultaneously)
- Unlike the more *famous* PCA, reveals the association between genes and samples(biplots)

CORRESPONDENCE ANALYSIS

Underlying hypothesis

There is no association between the rows[genes] and columns[samples]

- Project data to first 2 or 3 **informative** coordinates
- Treats rows(genes) and columns(samples) equivalently
- Attempts to separate dissimilar objects from each other(both genes and samples simultaneously)
- Unlike the more *famous* PCA, reveals the association between genes and samples(biplots)

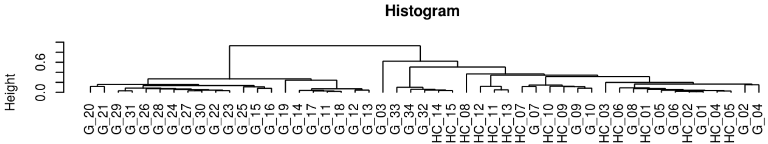
CORRESPONDENCE ANALYSIS

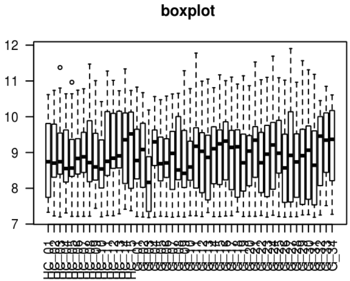
Underlying hypothesis

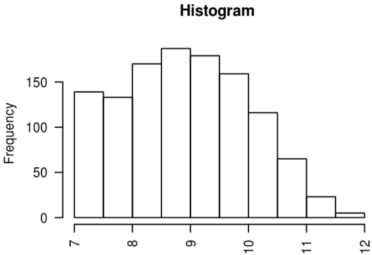
There is no association between the rows[genes] and columns[samples]

- Project data to first 2 or 3 **informative** coordinates
- Treats rows(genes) and columns(samples) equivalently
- Attempts to separate dissimilar objects from each other(both genes and samples simultaneously)
- Unlike the more *famous* PCA, reveals the association between genes and samples(biplots)

CLUSTERING

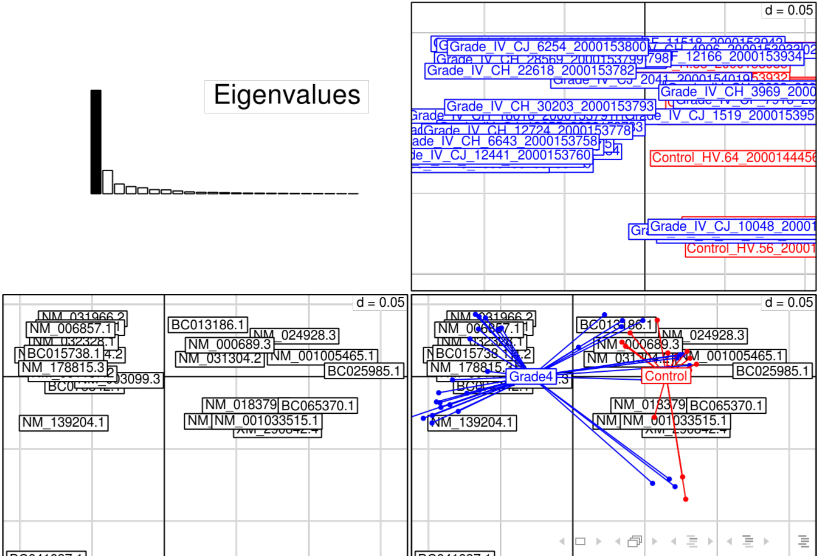

Histogram
Height
0.0
0.6
G_20 G_21 G_29 G_31 G_26 G_28 G_24 G_27 G_30 G_22 G_23 G_25 G_15 G_16 G_19 G_14 G_17 G_11 G_18 G_12 G_13 G_03 G_33 G_34 G_32 HC_14 HC_15 HC_08 HC_12 HC_11 HC_13 HC_07 G_07 HC_10 HC_09 G_09 G_10 HC_03 HC_06 G_08 HC_01 G_05 G_06 HC_02 G_01 HC_04 HC_05 G_02 G_04


boxplot
12
11
10
9
8
7


Histogram
Frequency
150
100
50
0
7
8
9
10
11
12

INTERPRETING BIPLOTS

The output of a CA is a biplot:



INTERPRETING BIPLOTS

- The distance on biplot are proportional to χ^2 distances in the original higher dimension
- The farther away a point is from the centroid, the higher is that row's contribution to the value of statistic
- Associations between the rows and columns is given by the angle made by lines joining the centroid to the points(acute=positive, right=no association)
- Thus we focus on points along the end of the axes. Positive regulation is indicated by genes appearing in the upper half.

INTERPRETING BIPLOTS

- The distance on biplot are proportional to χ^2 distances in the original higher dimension
- The farther away a point is from the centroid, the higher is that row's contribution to the value of statistic
- Associations between the rows and columns is given by the angle made by lines joining the centroid to the points(acute=positive, right=no association)
- Thus we focus on points along the end of the axes. Positive regulation is indicated by genes appearing in the upper half.

INTERPRETING BIPLOTS

- The distance on biplot are proportional to χ^2 distances in the original higher dimension
- The farther away a point is from the centroid, the higher is that row's contribution to the value of statistic
- Associations between the rows and columns is given by the angle made by lines joining the centroid to the points(acute=positive, right=no association)
- Thus we focus on points along the end of the axes. Positive regulation is indicated by genes appearing in the upper half.

INTERPRETING BIPLOTS

- The distance on biplot are proportional to χ^2 distances in the original higher dimension
- The farther away a point is from the centroid, the higher is that row's contribution to the value of statistic
- Associations between the rows and columns is given by the angle made by lines joining the centroid to the points(acute=positive, right=no association)
- Thus we focus on points along the end of the axes. Positive regulation is indicated by genes appearing in the upper half.

INTERPRETING BIPLOTS

- The distance on biplot are proportional to χ^2 distances in the original higher dimension
- The farther away a point is from the centroid, the higher is that row's contribution to the value of statistic
- Associations between the rows and columns is given by the angle made by lines joining the centroid to the points(acute=positive, right=no association)
- Thus we focus on points along the end of the axes. Positive regulation is indicated by genes appearing in the upper half.

INTERPRETING BIPLOTS

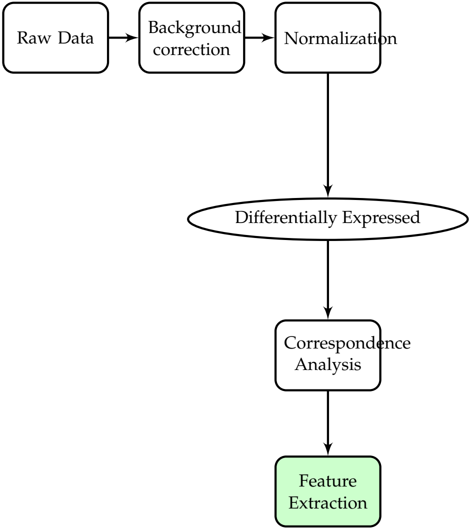
In PCA the distance between the projected points are euclidean, whereas CA takes into account the chi-squared distances. This is relevant here, since we are dealing with expression values and we are concerned with the **levels** and not the absolute values. for example consider :

CA vs PCA

$A = 1, 2, 3$
$B = 10, 25, 34$

Are A,B related/same?

SO FAR..


Raw Data
Background correction
Normalization
Differentially Expressed
Correspondence Analysis
Feature Extraction

FEATURE EXTRACTION & CLASSIFICATION

The Need

- Given the shortlist of genes showing association with the cohorts, we need to identify the subset of most informative genes
- CA does not answer this question. A panel of genes all exhibiting positive/negative association with the cohorts might not be too informative collectively
- Genes whose expression levels are themselves correlated, being in the same panel are less informative

Feature Extraction & Classification

The Need

- Given the shortlist of genes showing association with the cohorts, we need to identify the subset of most informative genes
- CA does not answer this question. A panel of genes all exhibiting positive/negative association with the cohorts might not be too informative collectively
- Genes whose expression levels are themselves correlated, being in the same panel are less informative

Feature Extraction & Classification

The Need

- Given the shortlist of genes showing association with the cohorts, we need to identify the subset of most informative genes
- CA does not answer this question. A panel of genes all exhibiting positive/negative association with the cohorts might not be too informative collectively
- Genes whose expression levels are themselves correlated, being in the same panel are less informative

FEATURE EXTRACTION & CLASSIFICATION

The Need

- Given the shortlist of genes showing association with the cohorts, we need to identify the subset of most informative genes
- CA does not answer this question. A panel of genes all exhibiting positive/negative association with the cohorts might not be too informative collectively
- Genes whose expression levels are themselves correlated, being in the same panel are less informative

Approach

- Choose a classification algorithm
- Start with all features, determine the coefficients for the model
- Eliminate the least informative feature
- Re-train the model, cross validate
- Repeat till you end up with required set of features

Approach

- Choose a classification algorithm
- Start with all features, determine the coefficients for the model
- Eliminate the least informative feature
- Re-train the model, cross validate
- Repeat till you end up with required set of features

Approach

- Choose a classification algorithm
- Start with all features, determine the coefficients for the model
- Eliminate the least informative feature
- Re-train the model, cross validate
- Repeat till you end up with required set of features

Approach

- Choose a classification algorithm
- Start with all features, determine the coefficients for the model
- Eliminate the least informative feature
- Re-train the model, cross validate
- Repeat till you end up with required set of features

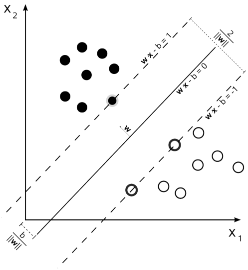
Approach

- Choose a classification algorithm
- Start with all features, determine the coefficients for the model
- Eliminate the least informative feature
- Re-train the model, cross validate
- Repeat till you end up with required set of features

Approach

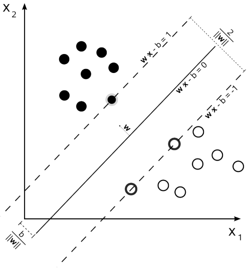
- Choose a classification algorithm
- Start with all features, determine the coefficients for the model
- Eliminate the least informative feature
- Re-train the model, cross validate
- Repeat till you end up with required set of features

SVM



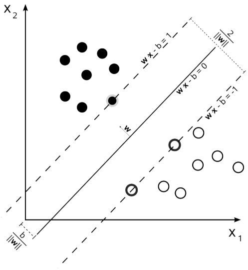
- Search for a hyperplane that best separates the data, maximising the margin of separation
- Data is assumed to be linearly separable (can be made to work irrespective of that)
- Given the high dimension of input, it is safe to assume that at that number of dimensions our data is linearly separable

SVM



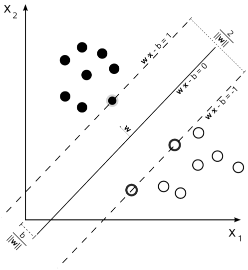
- Search for a hyperplane that best separates the data, maximising the margin of separation
- Data is assumed to be linearly separable (can be made to work irrespective of that)
- Given the high dimension of input, it is safe to assume that at that number of dimensions our data is linearly separable

SVM



- Search for a hyperplane that best separates the data, maximising the margin of separation
- Data is assumed to be linearly separable (can be made to work irrespective of that)
- Given the high dimension of input, it is safe to assume that at that number of dimensions our data is linearly separable

SVM



- Search for a hyperplane that best separates the data, maximising the margin of separation
- Data is assumed to be linearly separable (can be made to work irrespective of that)
- Given the high dimension of input, it is safe to assume that at that number of dimensions our data is linearly separable

SVM

Recursive feature elimination with k-fold cross validation

- Determine the rankings of each feature by training a SVM on given data
- Randomly partition data in k equally sized subsets
- The data with n feature is trained on $k-1$ subsets and validated using the remaining 1 set.
- this training process is repeated k times, such that each of the k subsamples are used exactly once as validation dataset
- These k results are then averaged for determining the specificity
- Eliminate the feature with least weight and repeat

SVM

Recursive feature elimination with k-fold cross validation

- Determine the rankings of each feature by training a SVM on given data
- Randomly partition data in *k* equally sized subsets
- The data with *n* feature is trained on $k-1$ subsets and validated using the remaining 1 set.
- this training process is repeated *k* times, such that each of the *k* subsamples are used exactly once as validation dataset
- These *k* results are then averaged for determining the specificity
- Eliminate the feature with least weight and repeat

SVM

Recursive feature elimination with k-fold cross validation

- Determine the rankings of each feature by training a SVM on given data
- Randomly partition data in k equally sized subsets
- The data with n feature is trained on $k - 1$ subsets and validated using the remaining 1 set.
- this training process is repeated k times, such that each of the k subsamples are used exactly once as validation dataset
- These k results are then averaged for determining the specificity
- Eliminate the feature with least weight and repeat

SVM

Recursive feature elimination with k-fold cross validation

- Determine the rankings of each feature by training a SVM on given data
- Randomly partition data in k equally sized subsets
- The data with n feature is trained on $k-1$ subsets and validated using the remaining 1 set.
- this training process is repeated k times, such that each of the k subsamples are used exactly once as validation dataset
- These k results are then averaged for determining the specificity
- Eliminate the feature with least weight and repeat

SVM

Recursive feature elimination with k-fold cross validation

- Determine the rankings of each feature by training a SVM on given data
- Randomly partition data in k equally sized subsets
- The data with n feature is trained on $k-1$ subsets and validated using the remaining 1 set.
- this training process is repeated k times, such that each of the k subsamples are used exactly once as validation dataset
- These k results are then averaged for determining the specificity
- Eliminate the feature with least weight and repeat

SVM

Recursive feature elimination with k-fold cross validation

- Determine the rankings of each feature by training a SVM on given data
- Randomly partition data in k equally sized subsets
- The data with n feature is trained on $k - 1$ subsets and validated using the remaining 1 set.
- this training process is repeated k times, such that each of the k subsamples are used exactly once as validation dataset
- These k results are then averaged for determining the specificity
- Eliminate the feature with least weight and repeat

SVM

Recursive feature elimination with k-fold cross validation

- Determine the rankings of each feature by training a SVM on given data
- Randomly partition data in k equally sized subsets
- The data with n feature is trained on $k-1$ subsets and validated using the remaining 1 set.
- this training process is repeated k times, such that each of the k subsamples are used exactly once as validation dataset
- These k results are then averaged for determining the specificity
- Eliminate the feature with least weight and repeat

CONCLUSIONS

- Developed a whole workflow to arrive at the final list of bio-markers
- Need to be tested for biological significance, previous literature reports
- Results generated dynamically, perfectly reproducible

CONCLUSIONS

- Developed a whole workflow to arrive at the final list of bio-markers
- Need to be tested for biological significance, previous literature reports
- Results generated dynamically, perfectly reproducible

CONCLUSIONS

- Developed a whole workflow to arrive at the final list of bio-markers
- Need to be tested for biological significance, previous literature reports
- Results generated dynamically, perfectly reproducible

CONCLUSIONS

- Developed a whole workflow to arrive at the final list of bio-markers
- Need to be tested for biological significance, previous literature reports
- Results generated dynamically, perfectly reproducible

VISUALISATION TOOLS

The power of the unaided mind is highly overrated. The real powers come from devising external aids that enhance cognitive abilities.

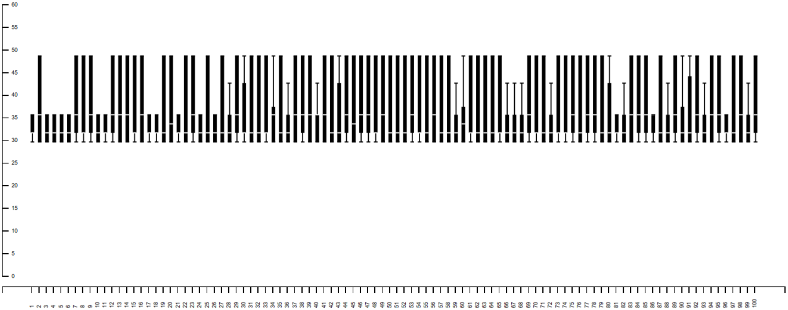
Donald Norman

PHRED SCORE VIEWER

fastq format

@SEQ_ID
GATTTGGGGTTCAAA
+
!"*((((***+))

PHRED SCORE VIEWER



Need/Motivation

- Cross-platform viewer for visualising the quality of fastq reads
- No commands required, user-friendly for biologists

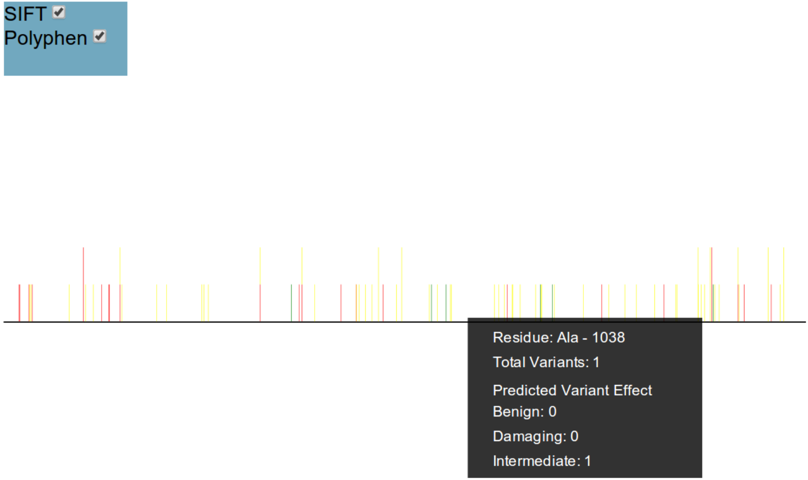
Need/Motivation

- Cross-platform viewer for visualising the quality of fastq reads
- No commands required, user-friendly for biologists

Need/Motivation

- Cross-platform viewer for visualising the quality of fastq reads
- No commands required, user-friendly for biologists

HUMAN GENETIC VARIATION VIEWER



HUMAN GENETIC VARIATION VIEWER

Need/Motivation

- Comprehensive visualisation of catalogue of protein variants
- Could be used to discover patterns with respect to mutation sites, frequency

HUMAN GENETIC VARIATION VIEWER

Need/Motivation

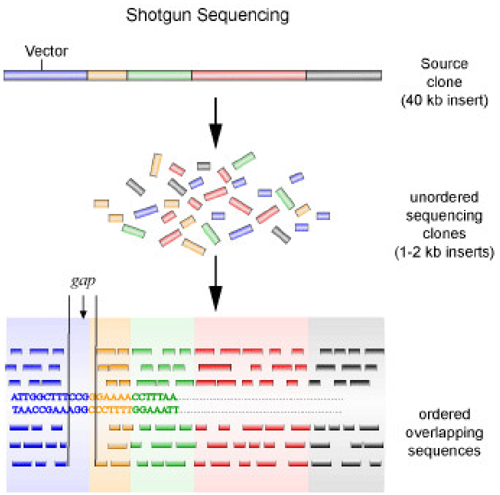
- Comprehensive visualisation of catalogue of protein variants
- Could be used to discover patterns with respect to mutation sites, frequency

HUMAN GENETIC VARIATION VIEWER

Need/Motivation

- Comprehensive visualisation of catalogue of protein variants
- Could be used to discover patterns with respect to mutation sites, frequency

NEXT GENERATION SEQUENCING



VIRAL GENOME DETECTION

Cervical cancers have been proven to be associated with Human Papillomavirus(HPV)

Cervical cancer datasets from Indian women was put through an analysis to detect :

1. Any possible HPV integration
2. Sites of HPV integration

Who Cares?

- Prognosis
- Replacing whole genome sequencing, by targeted sequencing at the sites where these virus have been detected in a cohort of samples, thus speeding up the whole process.

Viral Genome Detection

Cervical cancers have been proven to be associated with Human Papillomavirus(HPV)
Cervical cancer datasets from Indian women was put through an analysis to detect :

1. Any possible HPV integration
2. Sites of HPV integration

Who Cares?

- Prognosis
- Replacing whole genome sequencing, by targeted sequencing at the sites where these virus have been detected in a cohort of samples, thus speeding up the whole process.

Viral Genome Detection

Cervical cancers have been proven to be associated with Human Papillomavirus(HPV)

Cervical cancer datasets from Indian women was put through an analysis to detect :

1. Any possible HPV integration
2. Sites of HPV integration

Who Cares?

- Prognosis
- Replacing whole genome sequencing, by targeted sequencing at the sites where these virus have been detected in a cohort of samples, thus speeding up the whole process.

Figure: Detecting Virus Genomes

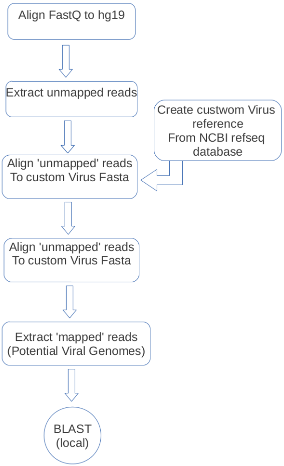
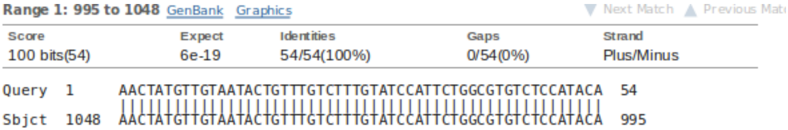


Figure: Aligned HPV genomes



Range 1: 995 to 1048 GenBank Graphics ▼ Next Match ▲ Previous Matc

Score	Expect	Identities	Gaps	Strand
100 bits(54)	6e-19	54/54(100%)	0/54(0%)	Plus/Minus

```
Query  1     AACTATGTTGTAATACTGTTTGTCTTTGTATCCATTCTGGCGTGTCTCCATACA  54
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1048  AACTATGTTGTAATACTGTTTGTCTTTGTATCCATTCTGGCGTGTCTCCATACA  995
```

BWA v/s BWA-PSSM I

BWA-PSSM is uses quality score matrices to *improve* the alignment.

@read

ACT

+

III

Assuming Sanger encoded quality scores, all the base positions have a phred score of (73-33=40) . Given an error model of the sequencing platform, it is possible to come up with a matrix like:

	A	T	G	C
A				
T				
G				
C				

BWA v/s BWA-PSSM II

for all possible phred scores, which assigns to each possible score and a given nuclotide a score given by (i,j), emphasizing the probability that an observed nucleotide by the sequencer is indeed the same nucleotide

- Simulate genomes with different error rates and insertion-deletion ratios
- Simulate reads from the genomes
- Align reads to reference

A ROC curve can be plotted since the number of reads that are expected to match is known apriori.

BWA v/s BWA-PSSM III

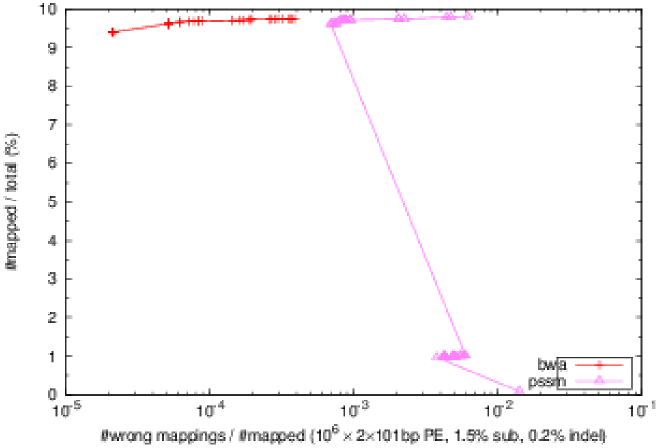


Figure: ROC curve for BWA v/s BWA-PSSM mappings

WRAP UP

- Developed a toolbox for driver mutation prediction.
  - Open Sourced
  - Deployed to be used by community
- Predicted a set of bio-markers for Glioma
  - Pending validation (literature, biological)
- Determined presence of HPV sequences in Cervical cancers
- Tools for Visualisation
  - Phred quality viewer
  - Human Genetic Variation Viewer

WRAP UP

- Developed a toolbox for driver mutation prediction.
  - Open Sourced
  - Deployed to be used by community
- Predicted a set of bio-markers for Glioma
  - Pending validation (literature, biological)
- Determined presence of HPV sequences in Cervical cancers
- Tools for Visualisation
  - Phred quality viewer
  - Human Genetic Variation Viewer

WRAP UP

- Developed a toolbox for driver mutation prediction.
  - Open Sourced
  - Deployed to be used by community
- Predicted a set of bio-markers for Glioma
  - Pending validation (literature, biological)
- Determined presence of HPV sequences in Cervical cancers
- Tools for Visualisation
  - Phred quality viewer
  - Human Genetic Variation Viewer

WRAP UP

- Developed a toolbox for driver mutation prediction.
  - Open Sourced
  - Deployed to be used by community
- Predicted a set of bio-markers for Glioma
  - Pending validation (literature, biological)
- Determined presence of HPV sequences in Cervical cancers
- Tools for Visualisation
  - Phred quality viewer
  - Human Genetic Variation Viewer

WRAP UP

- Developed a toolbox for driver mutation prediction.
  - Open Sourced
  - Deployed to be used by community
- Predicted a set of bio-markers for Glioma
  - Pending validation (literature, biological)
- Determined presence of HPV sequences in Cervical cancers
- Tools for Visualisation
  - Phred quality viewer
  - Human Genetic Variation Viewer

WRAP UP

- Developed a toolbox for driver mutation prediction.
  - Open Sourced
  - Deployed to be used by community
- Predicted a set of bio-markers for Glioma
  - Pending validation (literature, biological)
- Determined presence of HPV sequences in Cervical cancers
- Tools for Visualisation
  - Phred quality viewer
  - Human Genetic Variation Viewer

WRAP UP

- Developed a toolbox for driver mutation prediction.
  - Open Sourced
  - Deployed to be used by community
- Predicted a set of bio-markers for Glioma
  - Pending validation (literature, biological)
- Determined presence of HPV sequences in Cervical cancers
- Tools for Visualisation
  - Phred quality viewer
  - Human Genetic Variation Viewer

APPENDIX

Appendix

Differential Expression Statistics I

Smyth et al. suggested linear models for modelling microarray experiments. N set of samples, gene g with gene expression level y_g :

$$y_g^T = (y_{g1}, y_{g2}, ..., y_{gn}) \tag{8}$$

$$E(y_g) = X\alpha_g \tag{9}$$

Where X is the design matrix and α_g is an unknown coefficient vector.

$$var(y_g) = W_g\sigma_g^2 \tag{10}$$

where W_g is a weight matrix, and σ_g^2 represents unknown genewise variance. Consider β_g as the log-fold change for gene g.

DIFFERENTIAL EXPRESSION STATISTICS II

Assume the contrast to be tested is $\beta_g = c^T \alpha_g$ where c^T is a contrast matrix like X. Since α_g is unknown, given the response vectors and X it is possible to fit a linear model to obtain an estimate of coefficient vector as $\hat{\alpha_g}$ such that the covariance is given by:

$$var(\hat{\alpha_g}) = V_g \sigma_g^2 \tag{11}$$

where V_g is independent from σ_g^2 and is positive definite. Thus the estimate of β_g is given by $\hat{\beta}_g = c^T \alpha_g$ Assuming $\hat{\beta}_g$ to be normally distributed without forcing the normal distribution on y_g. $\hat{\beta}_g$ is assumed to be normally distributed with mean β_g and can be approximated as :

$$\hat{\beta}_g | \beta_g, \sigma_g^2 \sim \mathcal{N}(\beta_g, v_g \sigma^2) \tag{12}$$

DIFFERENTIAL EXPRESSION STATISTICS III

where

$$v_g = c^T V_g c \tag{13}$$

the variance s_g^2 is assumed to follow a scaled χ^2 distribution.

$$s_g^2 | \sigma_g^2 \sim \frac{\sigma_g^2}{d_g} \chi_{d_g}^2 \tag{14}$$

where d_g represents the residual degrees of freedom for gene g. Under the above assumptions, the statistic t_g follows a t-distribution with d_g degrees of freedom:

$$t_g = \frac{\hat{\beta}_g}{s_g \sqrt{v_g}}$$

Differential Expression Statistics IV

Information Pooling:
Given we are fitting linear models to thousands of genes, we could make use of this parallel structure fitting same model to the gene. We focus on β_{gj} and σ_g using a prior distribution model to focus how they change across genes :

$$\frac{1}{\sigma_g^2} = \frac{1}{d_0 s_0^2} \chi^2_{d0} \tag{15}$$

Let p_j = proportion of differentially expressed genes :

$$P(\beta_{gj} \neq 0) = p_j \tag{16}$$

Thus updating our prior information(prio obs. equals zero with variance v_0):

$$\beta_{gj} | \sigma_g^2, \beta_{gj} \neq 0 \sim N(0, v_0 \sigma_g^2) \tag{17}$$

DIFFERENTIAL EXPRESSION STATISTICS V

Posterior mean of $\frac{1}{\sigma_g^2}$ is given by $\frac{1}{\hat{s_g^2}}$:

$$\hat{s_g^2} = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g} \tag{18}$$

Thus the moderated t-statistic :

$$\hat{t_{gj}} = \frac{\hat{\beta_{gj}}}{s_g \sqrt{v_{gj}}} \tag{19}$$

has $d_0 + d_g$ degrees of freedom.

CORRESPONDENCE ANALYSIS I

Let $N = IxJ$ denote the data matrix. Converting the N matrix to P such that:

$$P = \frac{N}{\sum_i \sum_j n_ij} \tag{20}$$

The *row masses* are represented by:

$$r_i = \sum_{j=1}^{J} p_ij \tag{21}$$

The *column masses* are represented by:

$$c_j = \sum_{i=1}^{I} p_ij \tag{22}$$

Correspondence Analysis II

For row and column masses, the diagonals are given by:

$$D_r = diag(r) \tag{23}$$

$$D_c = diag(c) \tag{24}$$

Distance between two rows i and i' is given by:

$$d^2(i,i') = \sum_{j=1}^{J} \frac{1}{c_j}\left(\frac{p_{ij}}{r_i} - \frac{n_{i'j}}{r'_i}\right)^2 \tag{25}$$

Euclidean distances weighted by the inverse of the corresponding frequency, hence *standardized* variance-wise.
Even if the rows i and i' are replaced by their sum of rows, then distances between columns would not change.
The inertia for i^{th} row profile is thus defined as:

CORRESPONDENCE ANALYSIS III

$$Rowinertia = Rowmass * Squareofdistancefromthecentroidoftherows \tag{26}$$

The underlying hypothesis for CA is that the rows and columns are independent. In a contingency table the theoretical value of a cell at (i, j) is given by, assuming the above hypothesis is true :

$$E_{i,j} = r_i * c_j \tag{27}$$

However the *observed* value at (i, j) is p_{ij}. Thus the Chi-square distance is alculated as :

$$\chi^2 = n \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \tag{28}$$

Consider the centroid z of the row vector points:

CORRESPONDENCE ANALYSIS IV

$$z = [c_1, c_2, ..., c_J] \tag{29}$$

The distance between any i^{th} row and it's centroid is given by, using the distance relation between rows from above:

$$d_{iz}^2 = \sum_{j=i}^{J} \frac{(\frac{p_{ij}}{r_i} - c_j)^2}{c_j} \tag{30}$$

which can be rewritten in terms of the centroid $\mu_{ij} = r_i c_j$ as:

$$d_{iz}^2 = \frac{1}{r_i} \sum_{j=i}^{J} \frac{(p_{ij} - \mu_{ij})^2}{\mu_{ij}} \tag{31}$$

Thus row inertia:

Correspondence Analysis V

$$r_i d_{iz}^2 = \sum_{j=i}^{J} \frac{(p_{ij} - \mu_{ij})^2}{\mu_{ij}} \tag{32}$$

The column inertia can be defined similarly.
Consider the residual matrix S:

$$S_{ij} = \left|\frac{p_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}}}\right| \tag{33}$$

In order to decompose S to lower dimensions consider SVD decomposition of S:

$$S = UD_\alpha V^T \tag{34}$$

where U,V are orthonormal $VV^T = 1$ and $UU^T = 1$ and D_α is a diagonal matrix with entries in descending order as λ_1, λ_2,....

Correspondence Analysis VI

The scores of the rows is then given by:

$$F = D_r^{\frac{-1}{2}} U D_\alpha \tag{35}$$

and the column scores are given by:

$$G = D_c^{\frac{-1}{2}} V D_\alpha \tag{36}$$

The dimension of these score matrices is $min(I-1, J-1)$ and essentially represent the *coordinates* of these row vectors in the higher-dimensional subspace.
Points in this space are so arranged that the euclidean distances between two points corresponds to the Chi-square distance in the original matrix.
In order to quantify the amount of inertia represented by this plot, we consider the following score:

Correspondence Analysis VII

$$\phi^2 = \sum_{i=1}^{I} r_i d_{iz}^2 \tag{37}$$

and the amount of inertia captured by he first two principal axes is given by:

$$\frac{\lambda_1^2 + \lambda_2^2}{\phi^2} \tag{38}$$

SVM I

Support Vector Machines are binary classifiers. Given a training set of (points,labels) (x_i, y_i) where $x_i \in \mathbf{R}$ and $y \in -1, 1]$. The idea is to search for a hyperplane that would separate the points with $y_i = 1$ from $y_i = -1$. There could be multiple hyperplanes like that, the focus is however only on the hyperplane that with maximum-margins(on both sides). Any such hyperplane satisfies:

$$w.x - b = 0 \tag{39}$$

If the data is linearly separable, two hyperplanes can be found :

$$w.x - b = 1 \tag{40}$$

$$w.x - b = -1 \tag{41}$$

SVM II

The distance between the two hyperplanes is $\frac{2}{||w||}$. Thus minimising $||w||$ would yield the required the hyperplane.
In order to prevent misclassification, the following constraints are required:

$$(w.x_i - b) \geq 1 \tag{42}$$

for x_i belonging to class 1 and

$$(w.x_i - b) \leq -1 \tag{43}$$

for x_i belonging to class -1 which can be combined as:

$$y_i(w.x_i - b) \geq 1 \tag{44}$$

and the objective function to be minimised under this constraint is : $||w||$