Addressing the challenges of Genomics Data Analysis in JMP Genomics

Dr. Valerie Nedbal
JMP Pharmaceutical Technical Manager

September 27th, 2010

Agenda

JMP Genomics

- Introduction
- Features and Benefits
- Live demonstration

  - Cross-referencing data set analysis

What is JMP Genomics?

JMP Genomics from SAS

- A solution aimed at analysis of high-throughput biological data
- All-in-one software for different data formats
  - Gene Expression
  - miRNA
  - Exon
  - SNP
  - Copy Number Variation
  - Methylation
  - Proteomics
  - Summarized Next Gen Sequencing Data
- Unique combination of JMP 8 and SAS 9.2

Highly Visual

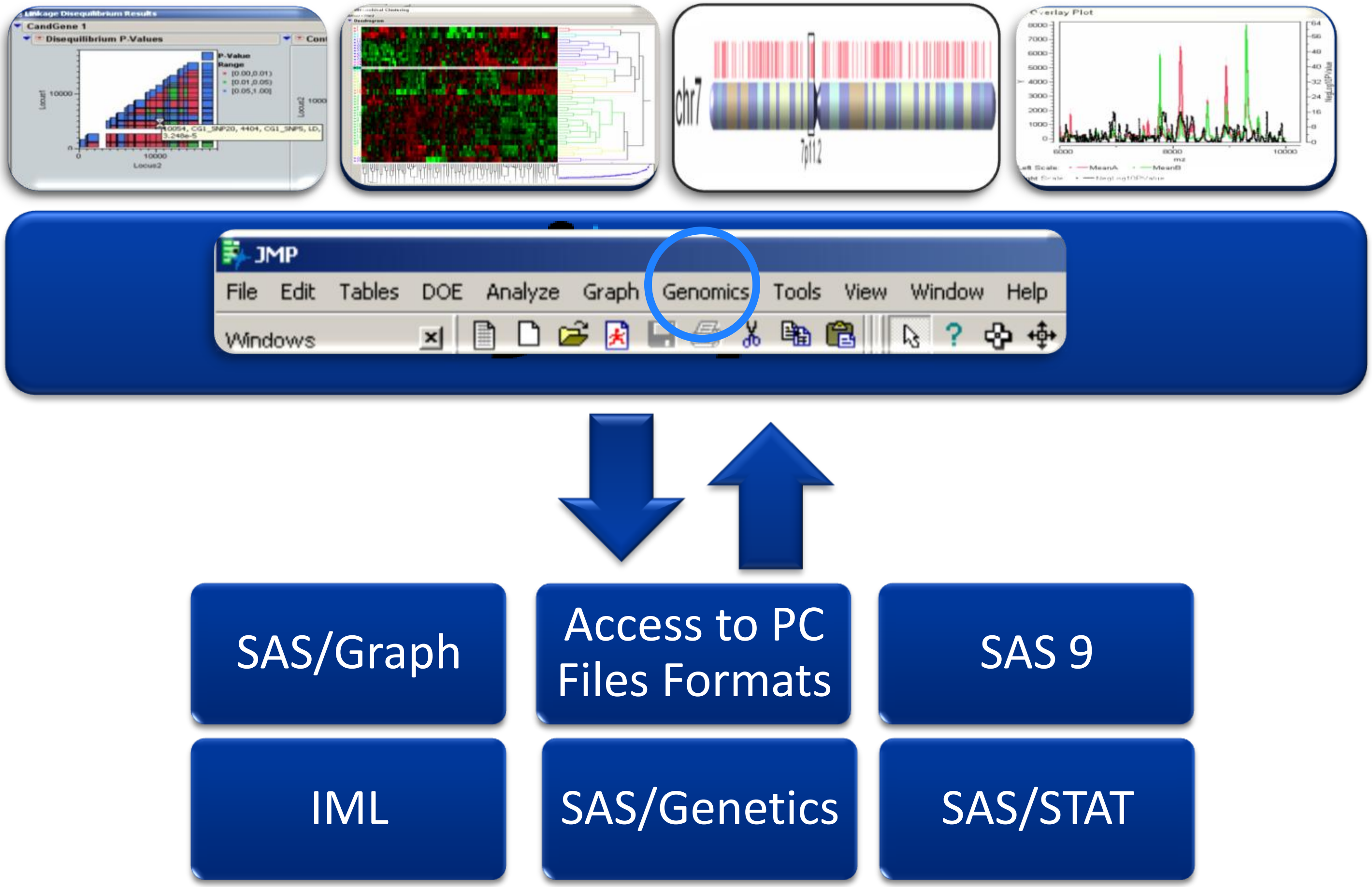
Interactive Graphics

Intuitive

Scalable

Validated Powerful Analytics

JMP Genomics Architecture



SAS/Graph

Access to PC Files Formats

SAS 9

IML

SAS/Genetics

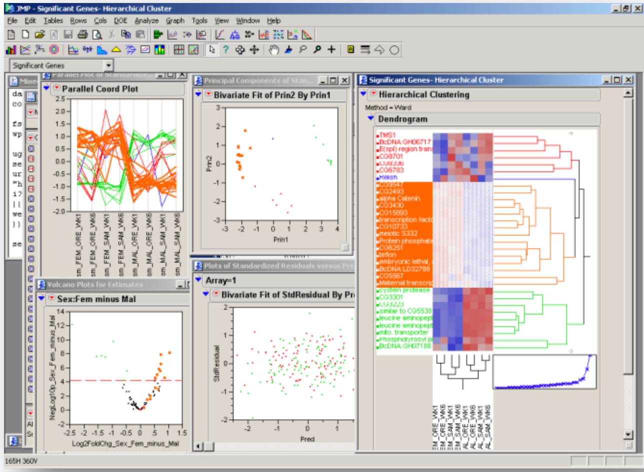
SAS/STAT

JMP Genomics Benefits

Large community users enables to analyse genomics data:

•Better

- •Highly visual, with interactive graphics linked to data
- •Based on proven and trusted analytics from SAS
- •Scalable support for large data sets
- •Open architecture - Extendable options for plug-ins



JMP - Significant Genes- Hierarchical Cluster
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help
Significant Genes
Parallel Coord Plot
Principal Components of Stan...
Bivariate Fit of Prin2 By Prin1
Prin2
Prin1
Significant Genes- Hierarchical Cluster
Hierarchical Clustering
Method = Ward
Dendrogram
Plots of Standardized Residuals versus Pre
Array=1
Bivariate Fit of StdResidual By Pr
StdResidual
Pred
Volcano Plots for Estimates
Sex:Fem minus Mal
NegLog10p_Sex_Fem_minus_Mal
Log2FoldChg_Sex_Fem_minus_Mal

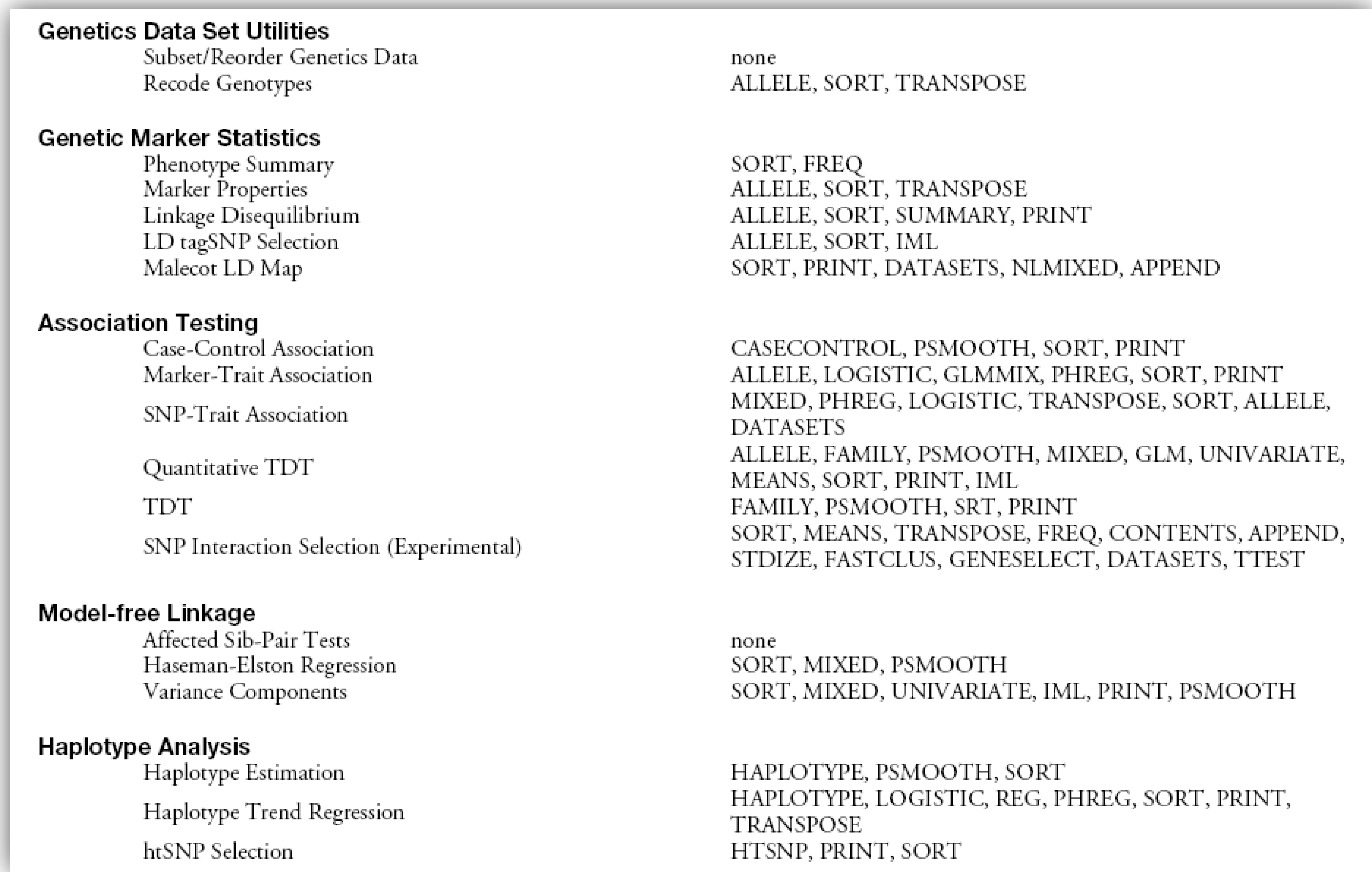
•Faster

- •All in one software cuts costs, reduces time wasted reformatting data for multiple packages.
- •Platform features support biologists and statisticians, enabling community wide genomics data analysis

•Easier

- •Easy to use, point and click menus and dialogs

List of Analytical Procedures in JMP Genomics using SAS Macro’s in the background

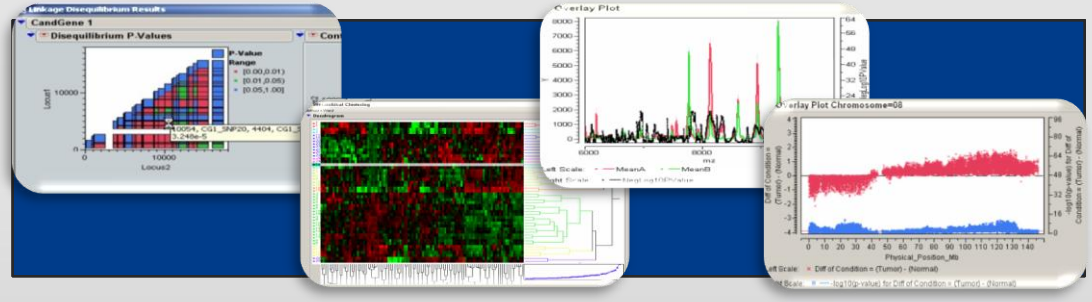


Genetics Data Set Utilities	
Subset/Reorder Genetics Data	none
Recode Genotypes	ALLELE, SORT, TRANSPOSE
Genetic Marker Statistics	
Phenotype Summary	SORT, FREQ
Marker Properties	ALLELE, SORT, TRANSPOSE
Linkage Disequilibrium	ALLELE, SORT, SUMMARY, PRINT
LD tagSNP Selection	ALLELE, SORT, IML
Malecot LD Map	SORT, PRINT, DATASETS, NLMIXED, APPEND
Association Testing	
Case-Control Association	CASECONTROL, PSMOOTH, SORT, PRINT
Marker-Trait Association	ALLELE, LOGISTIC, GLMMIX, PHREG, SORT, PRINT
SNP-Trait Association	MIXED, PHREG, LOGISTIC, TRANSPOSE, SORT, ALLELE, DATASETS
Quantitative TDT	ALLELE, FAMILY, PSMOOTH, MIXED, GLM, UNIVARIATE, MEANS, SORT, PRINT, IML
TDT	FAMILY, PSMOOTH, SRT, PRINT
SNP Interaction Selection (Experimental)	SORT, MEANS, TRANSPOSE, FREQ, CONTENTS, APPEND, STDIZE, FASTCLUS, GENESELECT, DATASETS, TTEST
Model-free Linkage	
Affected Sib-Pair Tests	none
Haseman-Elston Regression	SORT, MIXED, PSMOOTH
Variance Components	SORT, MIXED, UNIVARIATE, IML, PRINT, PSMOOTH
Haplotype Analysis	
Haplotype Estimation	HAPLOTYPE, PSMOOTH, SORT
Haplotype Trend Regression	HAPLOTYPE, LOGISTIC, REG, PHREG, SORT, PRINT, TRANSPOSE
htSNP Selection	HTSNP, PRINT, SORT

JMP Genomics Platform

Consumers (Biologists)

Knowledge Deployment



JMP G Workflows

Creators (Statisticians)

Knowledge Generation

Statistician Tools

SAS

R

Other

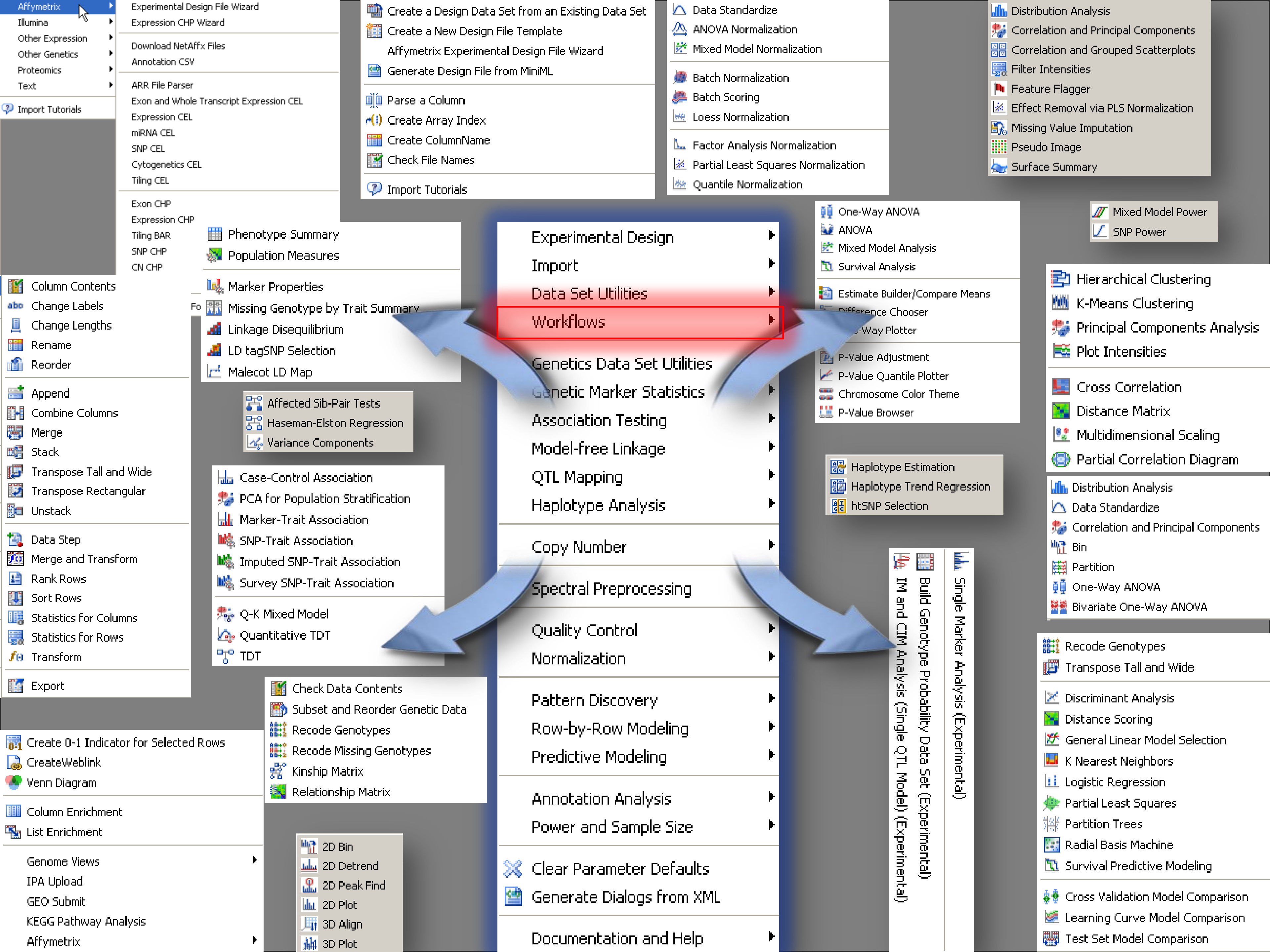
Data Layer

Data Gathering

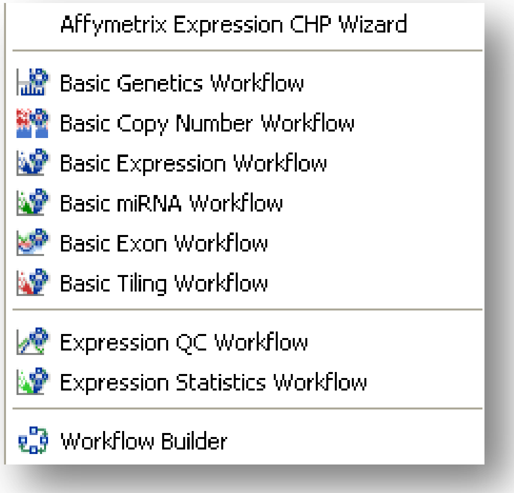
Affymetrix
Illumina
Xls, csv, txt

Agilent
HapMap
Pedigree

ABI Analyst
Nexus
...


Affymetrix
Illumina
Other Expression
Other Genetics
Proteomics
Text
Import Tutorials
Experimental Design File Wizard
Expression CHP Wizard
Download NetAffx Files
Annotation CSV
ARR File Parser
Exon and Whole Transcript Expression CEL
Expression CEL
miRNA CEL
SNP CEL
Cytogenetics CEL
Tiling CEL
Exon CHP
Expression CHP
Tiling BAR
SNP CHP
CN CHP
Create a Design Data Set from an Existing Data Set
Create a New Design File Template
Affymetrix Experimental Design File Wizard
Generate Design File from MiniML
Parse a Column
Create Array Index
Create ColumnName
Check File Names
Import Tutorials
Data Standardize
ANOVA Normalization
Mixed Model Normalization
Batch Normalization
Batch Scoring
Loess Normalization
Factor Analysis Normalization
Partial Least Squares Normalization
Quantile Normalization
Distribution Analysis
Correlation and Principal Components
Correlation and Grouped Scatterplots
Filter Intensities
Feature Flagger
Effect Removal via PLS Normalization
Missing Value Imputation
Pseudo Image
Surface Summary
Phenotype Summary
Population Measures
Marker Properties
Missing Genotype by Trait Summary
Linkage Disequilibrium
LD tagSNP Selection
Malecot LD Map
Experimental Design
Import
Data Set Utilities
Workflows
Genetics Data Set Utilities
Genetic Marker Statistics
Association Testing
Model-free Linkage
QTL Mapping
Haplotype Analysis
Copy Number
Spectral Preprocessing
Quality Control
Normalization
Pattern Discovery
Row-by-Row Modeling
Predictive Modeling
Annotation Analysis
Power and Sample Size
Clear Parameter Defaults
Generate Dialogs from XML
Documentation and Help
One-Way ANOVA
ANOVA
Mixed Model Analysis
Survival Analysis
Estimate Builder/Compare Means
Difference Chooser
-Way Plotter
P-Value Adjustment
P-Value Quantile Plotter
Chromosome Color Theme
P-Value Browser
Mixed Model Power
SNP Power
Hierarchical Clustering
K-Means Clustering
Principal Components Analysis
Plot Intensities
Cross Correlation
Distance Matrix
Multidimensional Scaling
Partial Correlation Diagram
Column Contents
Change Labels
Change Lengths
Rename
Reorder
Append
Combine Columns
Merge
Stack
Transpose Tall and Wide
Transpose Rectangular
Unstack
Data Step
Merge and Transform
Rank Rows
Sort Rows
Statistics for Columns
Statistics for Rows
Transform
Export
Affected Sib-Pair Tests
Haseman-Elston Regression
Variance Components
Case-Control Association
PCA for Population Stratification
Marker-Trait Association
SNP-Trait Association
Imputed SNP-Trait Association
Survey SNP-Trait Association
Q-K Mixed Model
Quantitative TDT
TDT
Haplotype Estimation
Haplotype Trend Regression
htSNP Selection
Distribution Analysis
Data Standardize
Correlation and Principal Components
Bin
Partition
One-Way ANOVA
Bivariate One-Way ANOVA
Single Marker Analysis (Experimental)
Build Genotype Probability Data Set (Experimental)
IM and CIM Analysis (Single QTL Model) (Experimental)
Recode Genotypes
Transpose Tall and Wide
Discriminant Analysis
Distance Scoring
General Linear Model Selection
K Nearest Neighbors
Logistic Regression
Partial Least Squares
Partition Trees
Radial Basis Machine
Survival Predictive Modeling
Cross Validation Model Comparison
Learning Curve Model Comparison
Test Set Model Comparison
Create 0-1 Indicator for Selected Rows
CreateWeblink
Venn Diagram
Column Enrichment
List Enrichment
Genome Views
IPA Upload
GEO Submit
KEGG Pathway Analysis
Affymetrix
Check Data Contents
Subset and Reorder Genetic Data
Recode Genotypes
Recode Missing Genotypes
Kinship Matrix
Relationship Matrix
2D Bin
2D Detrend
2D Peak Find
2D Plot
3D Align
3D Plot
Fo

Workflows for ease of use



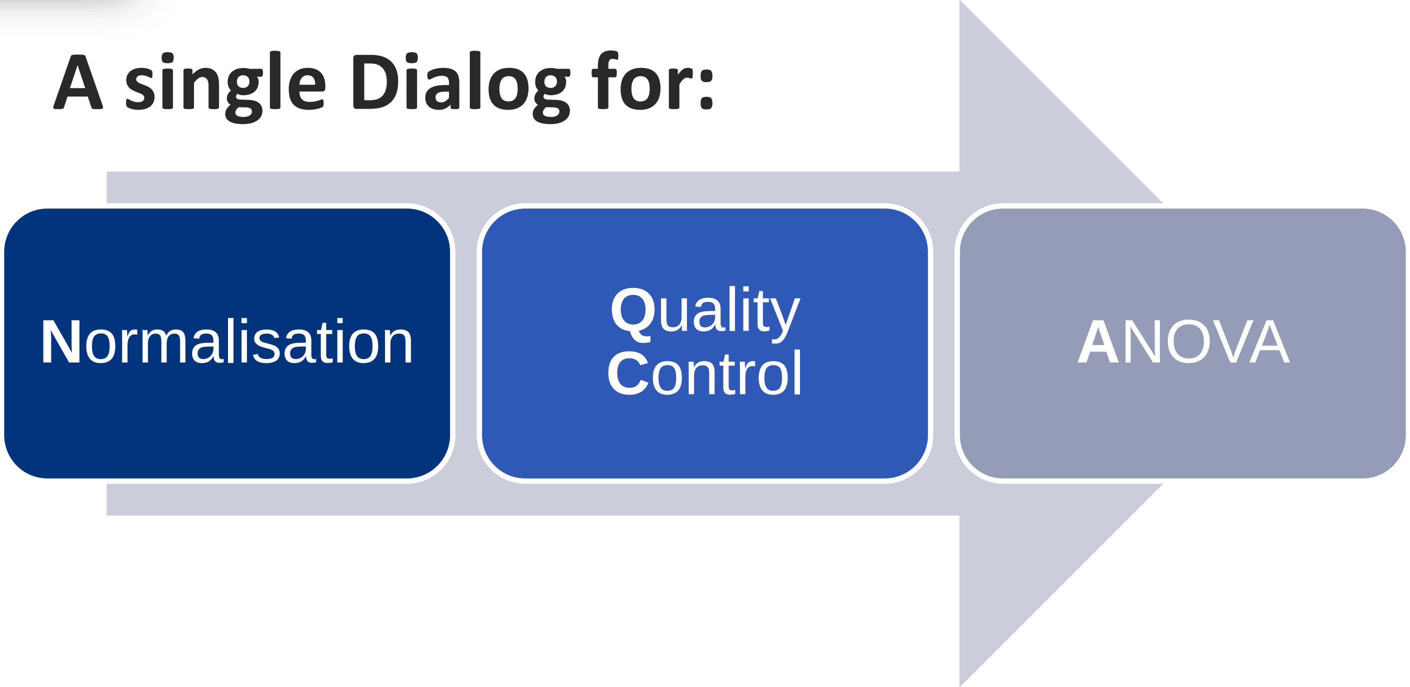
Affymetrix Expression CHP Wizard

- Basic Genetics Workflow
- Basic Copy Number Workflow
- Basic Expression Workflow
- Basic miRNA Workflow
- Basic Exon Workflow
- Basic Tiling Workflow

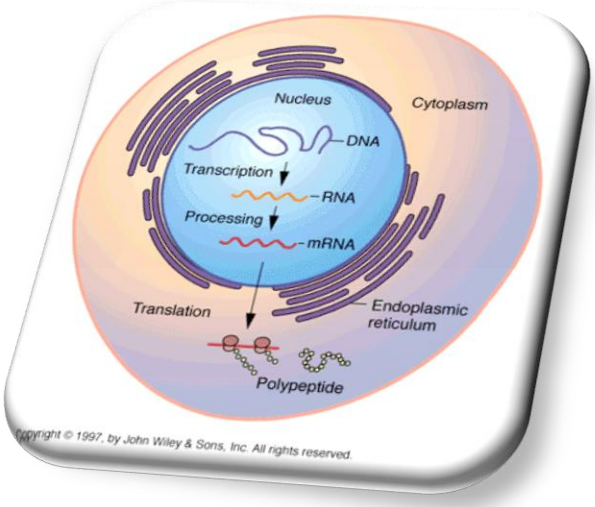
- Expression QC Workflow
- Expression Statistics Workflow

- Workflow Builder

A single Dialog for:

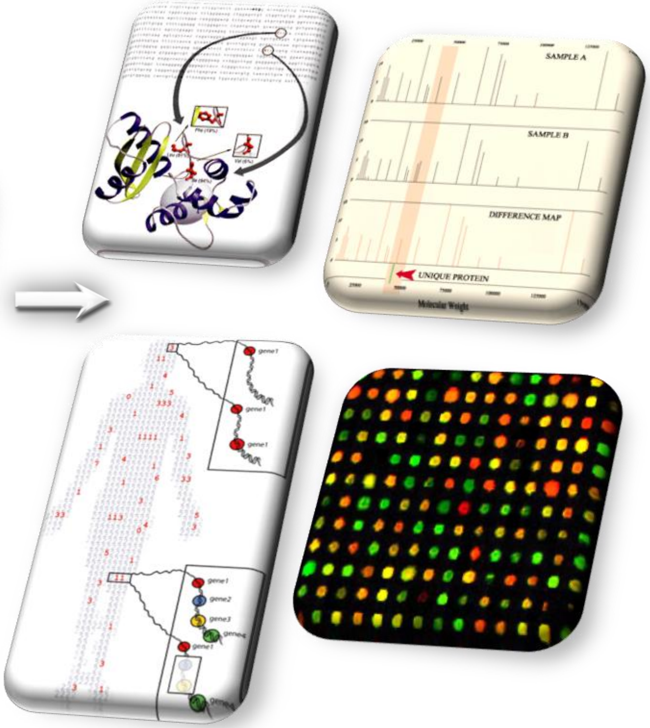


JMP Genomics assess SNP, Gene Expression, Alternative Splicing, Epigenetics, Gene Copy Number and Protein Sequence Variation



Copyright © 1997, by John Wiley & Sons, Inc. All rights reserved.

Central Dogma



Biomarker Discovery

Disease

Traits

Adverse Events

Experimental Design

Epigenetic signature of breast cancer and its association with gene expression and copy number

- Crossreferencing data sets generated from multiple whole-genome platforms
  - Simultaneous highresolution, whole-genome analyses using Affymetrix gene expression (U133), promoter (1.0R) and SNP/CNV (SNP 6.0) microarray platforms to correlate epigenetic (DNA methylation), gene expression and combination single nucleotide polymorphism / copy number variant (SNP 6.0) microarrays
  - GSE 15619 (July 2008)

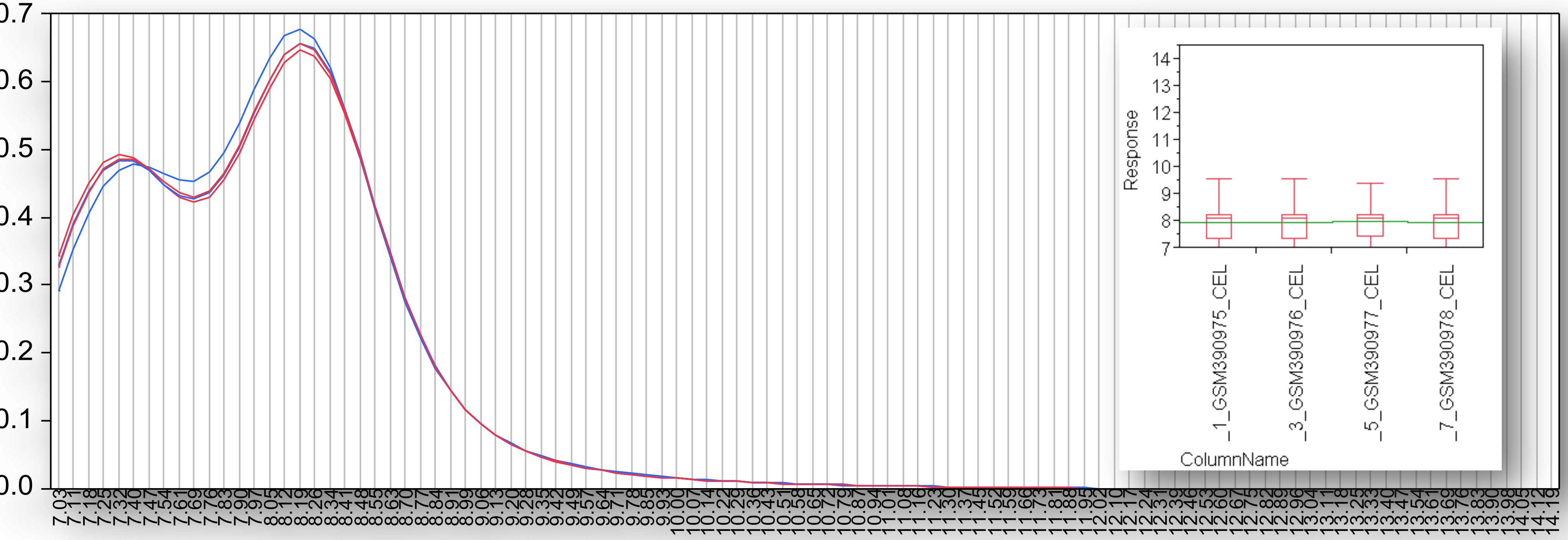
Epigenetic signature of breast cancer and its association with gene expression and copy number

- Comparison of 2 Breast Cancer cell lines:
  - 468GFP: Parental cell line
  - 468GFP – LN: Highly Metastatic cell line
- Copy Number Variation Data
  - DNA was compared of 2 biological replicates of a highly metastatic breast cancer cell line (468GFP-LN) to 2 biological replicates from the parental cell line, 468GFP
- Expression Data
  - Expression was compared of 3 biological replicates of a highly metastasic cancer cell line MDA-MB-468GFP-LN to 3 biological replicates of a control group MDA-MB-468GFP
- Epigenetic Mapping
  - DNA derived from 3 biological replicates of a highly metastatic (via Lymph Nodes) Breast cancer cell line (468GFP-LN) was compared to 3 biological replicates of DNA prepared from the parental cell line, 468GFP

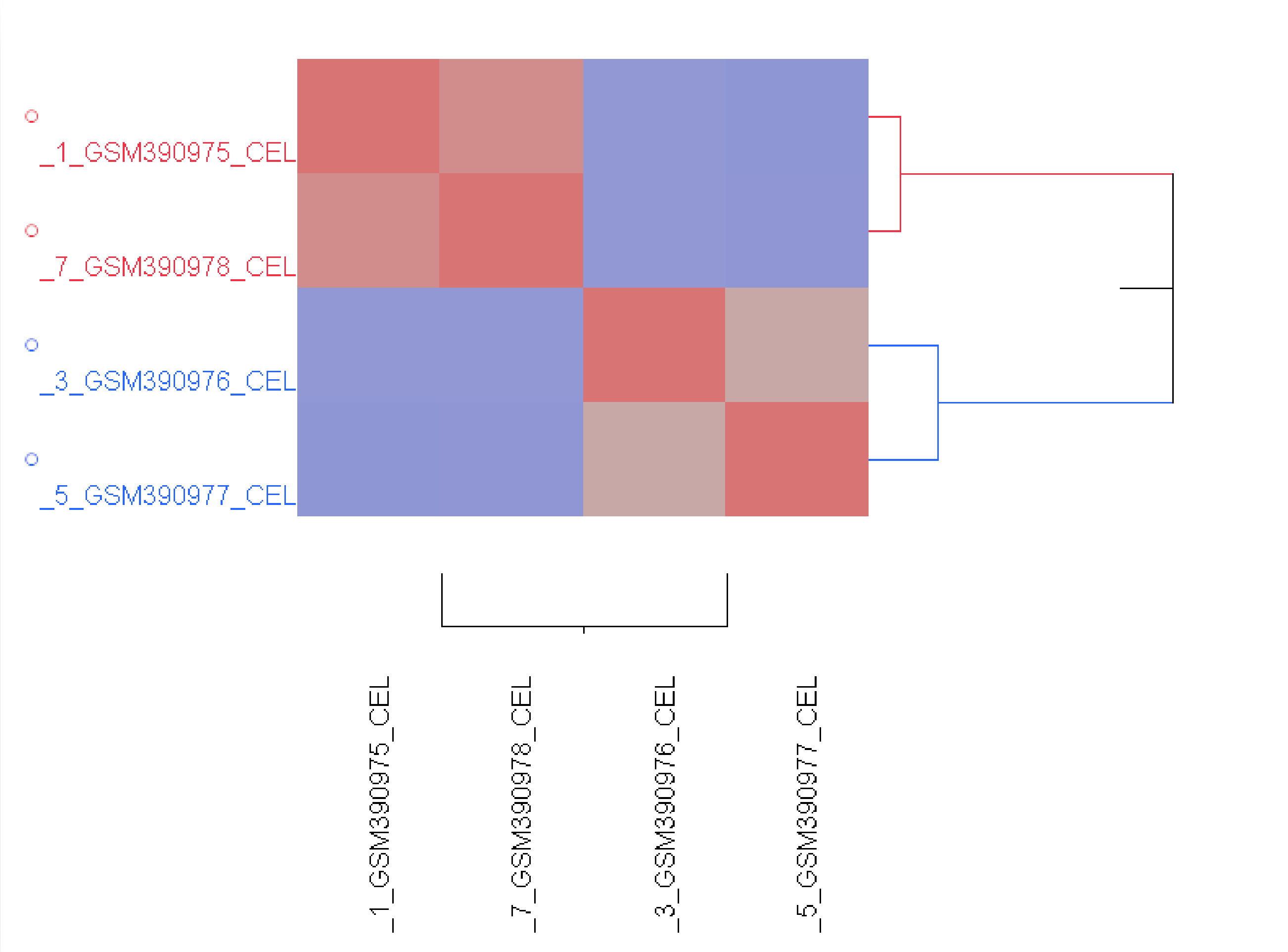
First data set: Copy Number Variation

- Step 1: Quality Control Checks
- Step 2: ANOVA to find out copy number significant differences
- Step 3: Partition analysis to define break positions
- Step 4: Gene Mapping

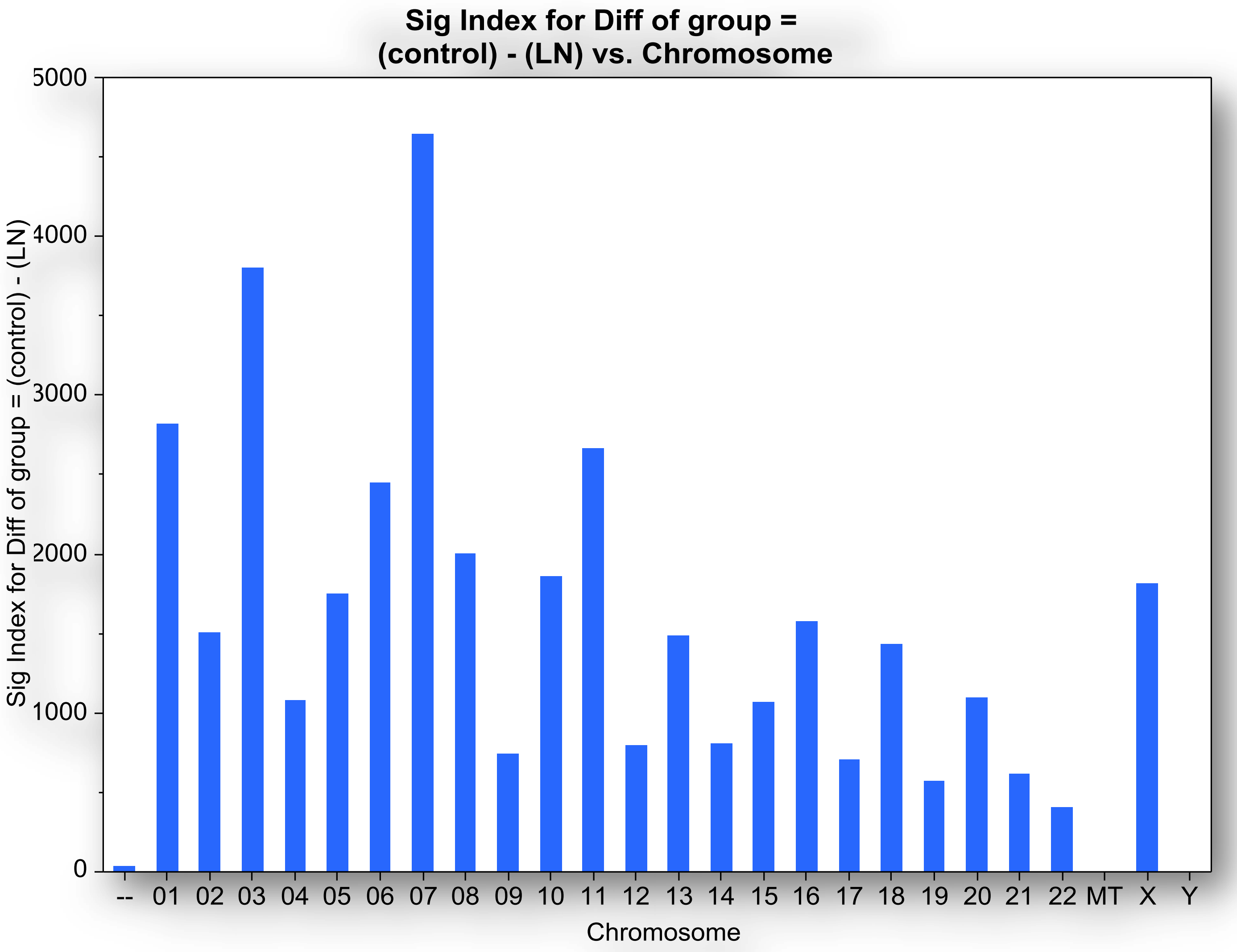
Copy Number Variation – Distribution Analysis



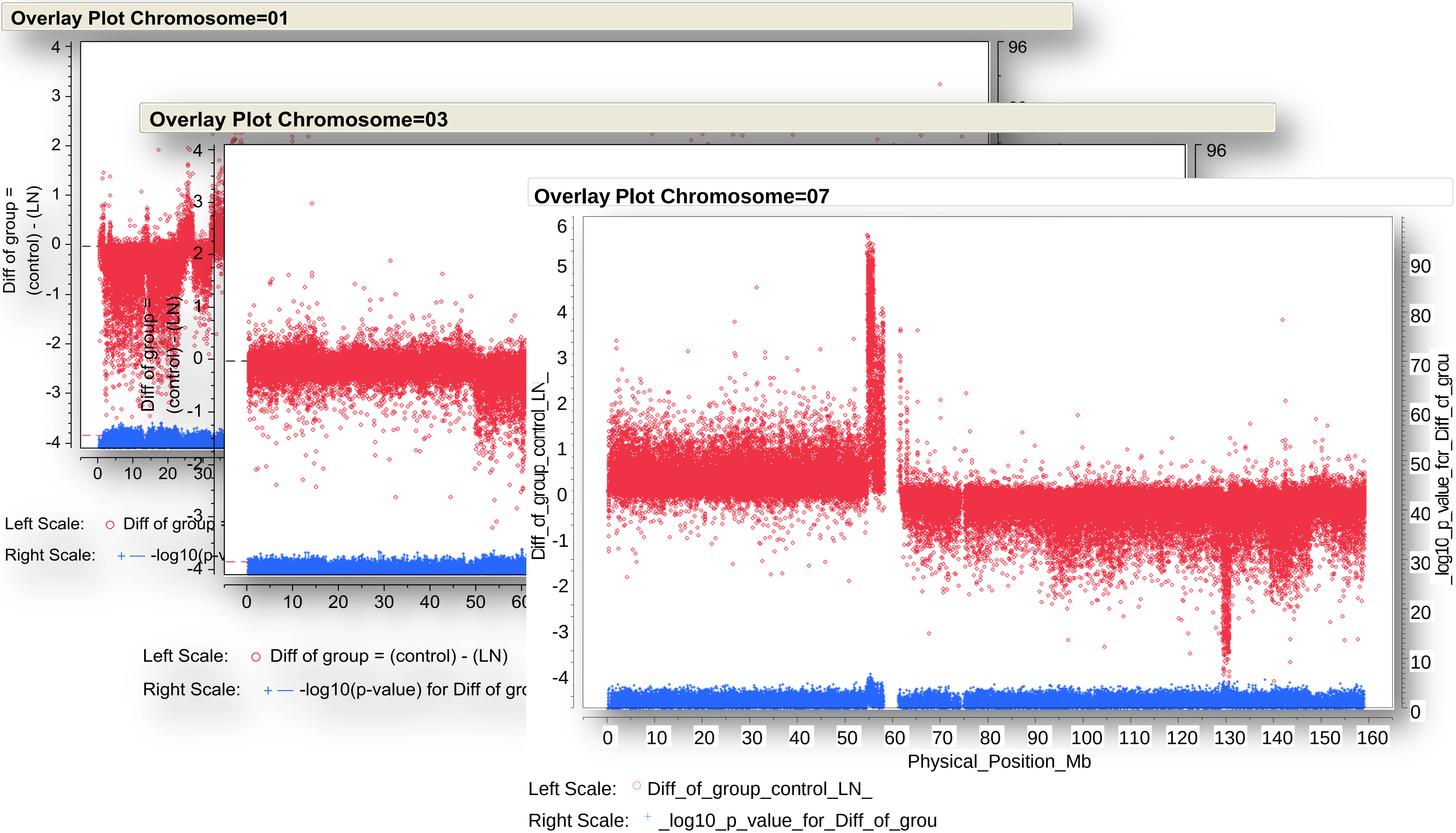
Copy Number Variation – Hierarchical Cluster Tree on Correlation



Copy Number Variation - ANOVA



Copy Number Variation – Chromosomal Position Plot

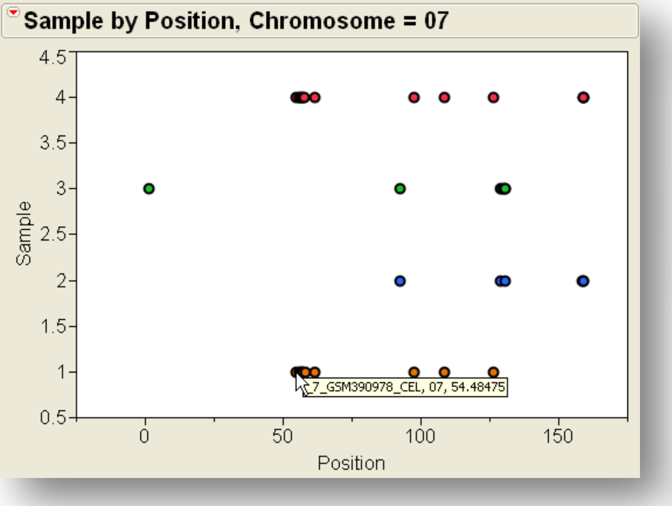


Overlay Plot Chromosome=01

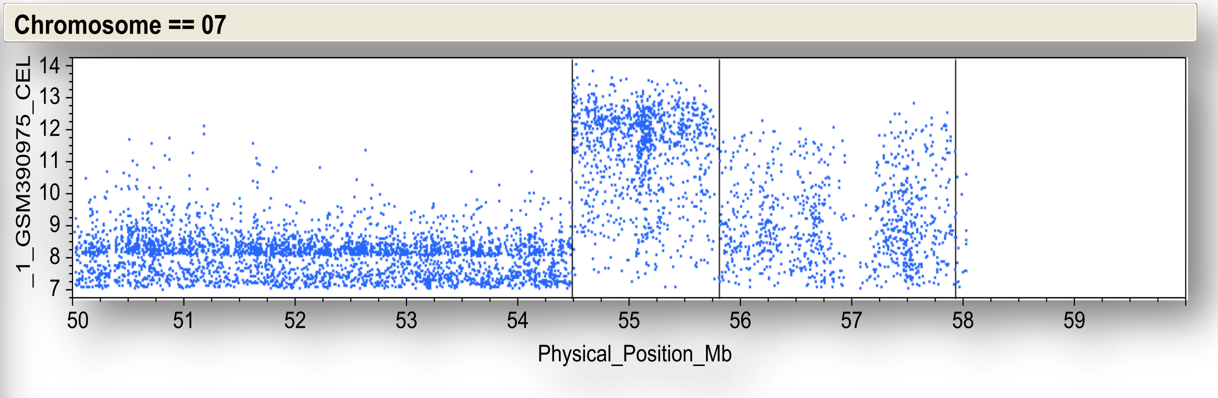
Overlay Plot Chromosome=03

Overlay Plot Chromosome=07

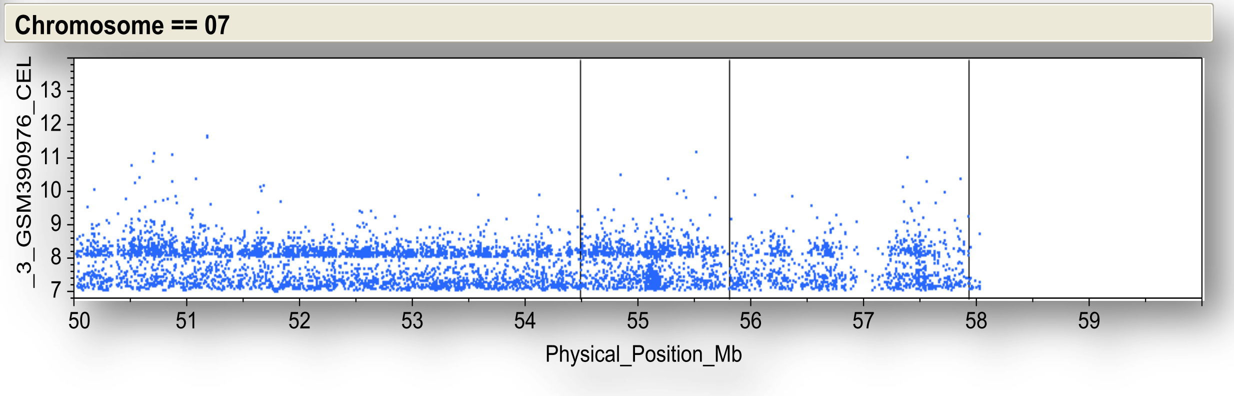
Partition Analysis: Break Positions on Chr.7



Sample by Position, Chromosome = 07

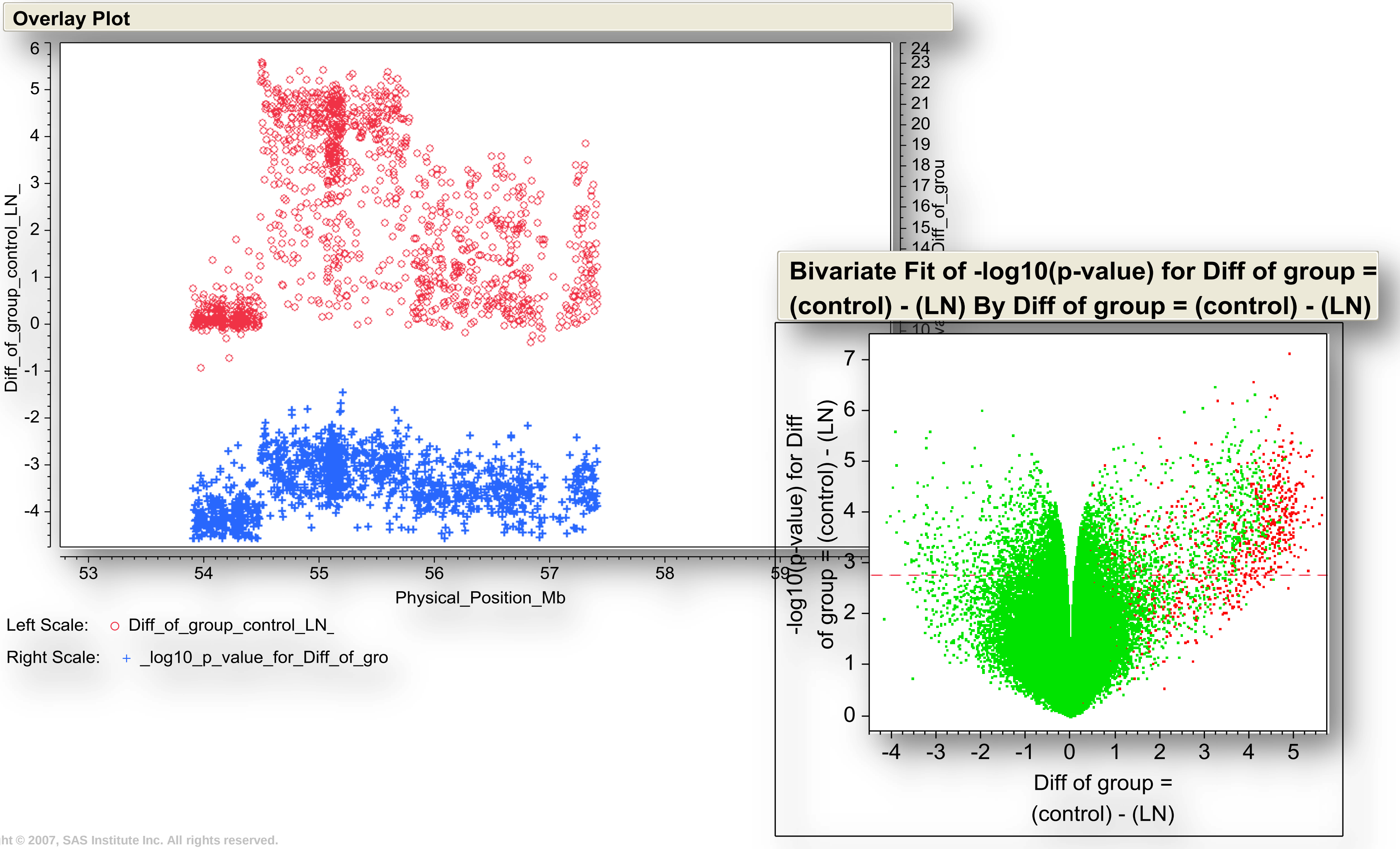


Chromosome == 07



Chromosome == 07

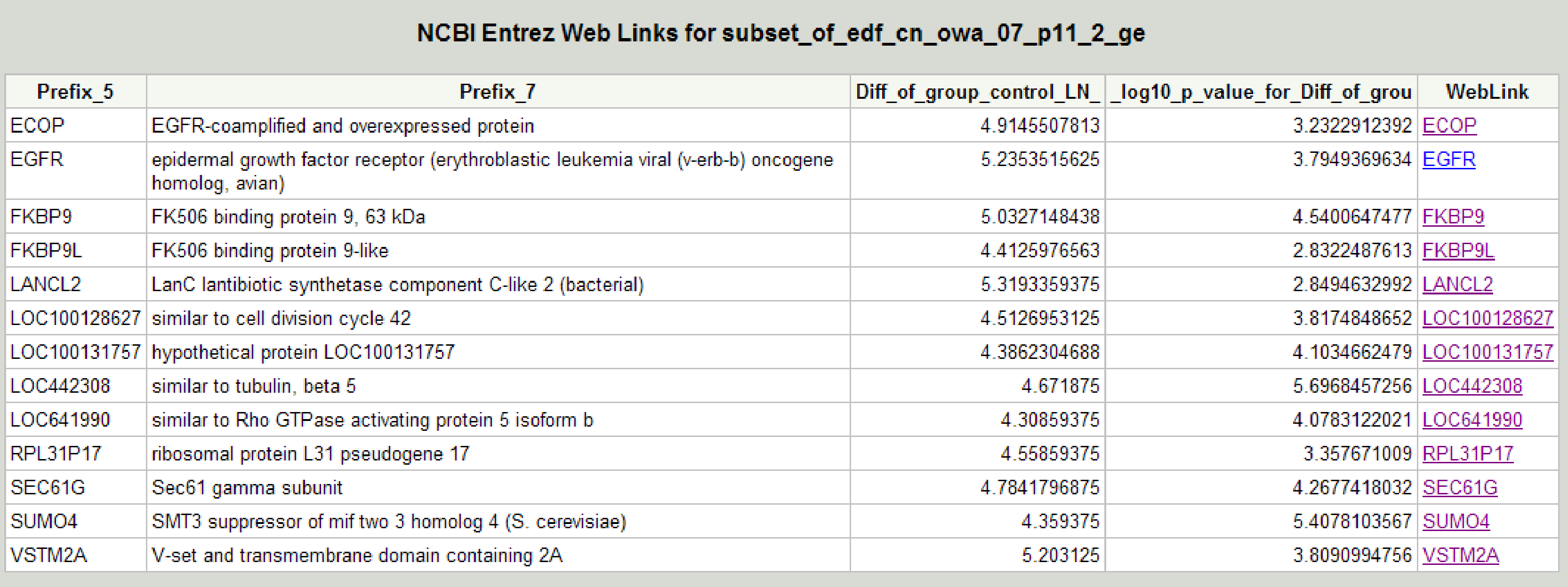
Copy Number Variation – Chr. 7 – Cytoband p11.2



Overlay Plot

Bivariate Fit of -log10(p-value) for Diff of group = (control) - (LN) By Diff of group = (control) - (LN)

Copy Number Variation – Gene Mapping on Chr. 7 – p11.2



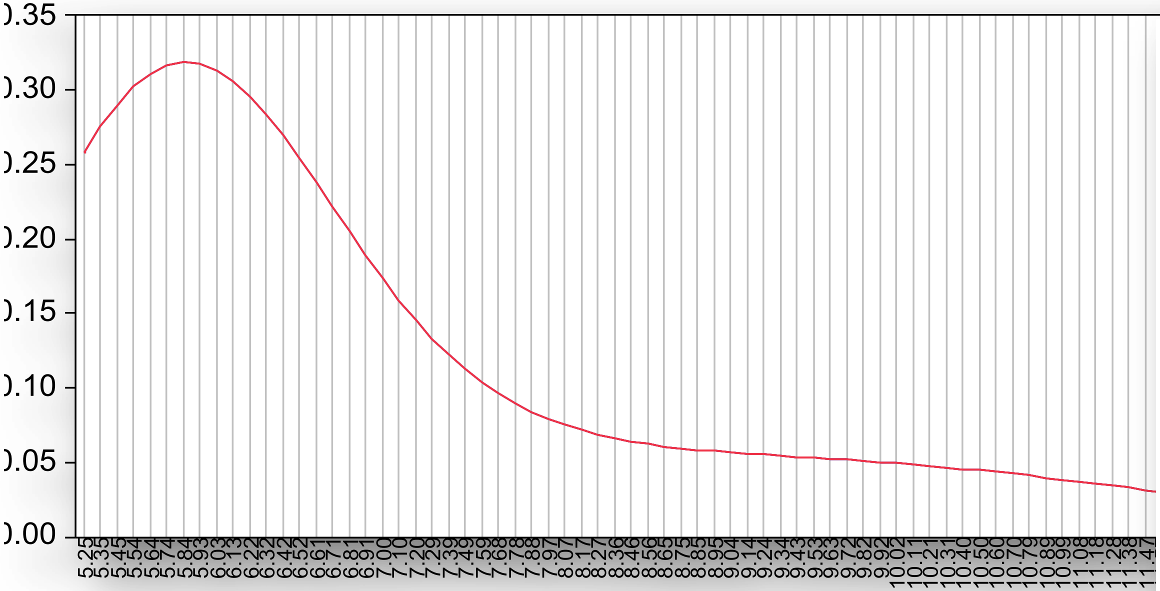
NCBI Entrez Web Links for subset_of_edf_cn_owa_07_p11_2_ge

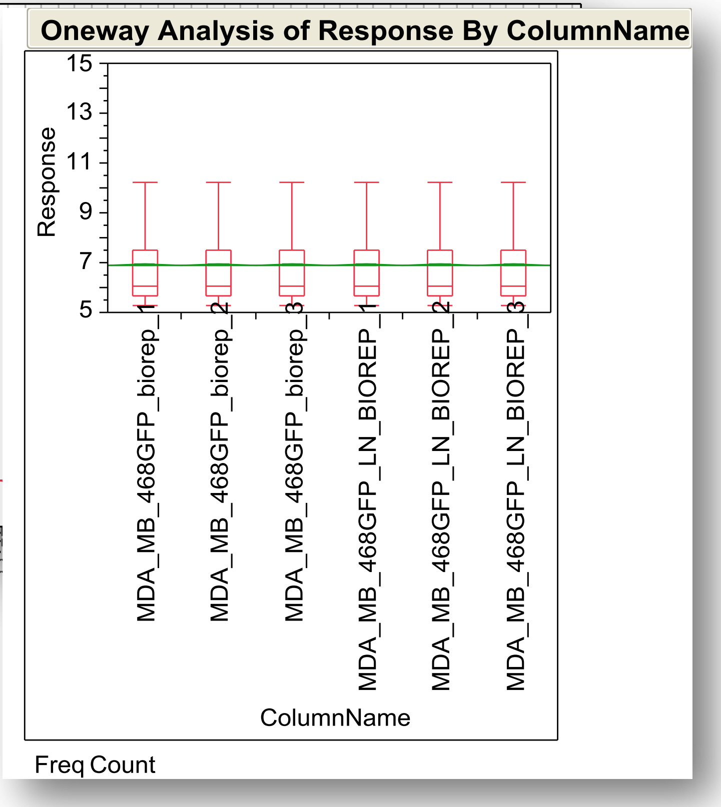
Prefix_5	Prefix_7	Diff_of_group_control_LN_	_log10_p_value_for_Diff_of_grou	WebLink
ECOP	EGFR-coamplified and overexpressed protein	4.9145507813	3.2322912392	ECOP
EGFR	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	5.2353515625	3.7949369634	EGFR
FKBP9	FK506 binding protein 9, 63 kDa	5.0327148438	4.5400647477	FKBP9
FKBP9L	FK506 binding protein 9-like	4.4125976563	2.8322487613	FKBP9L
LANCL2	LanC lantibiotic synthetase component C-like 2 (bacterial)	5.3193359375	2.8494632992	LANCL2
LOC100128627	similar to cell division cycle 42	4.5126953125	3.8174848652	LOC100128627
LOC100131757	hypothetical protein LOC100131757	4.3862304688	4.1034662479	LOC100131757
LOC442308	similar to tubulin, beta 5	4.671875	5.6968457256	LOC442308
LOC641990	similar to Rho GTPase activating protein 5 isoform b	4.30859375	4.0783122021	LOC641990
RPL31P17	ribosomal protein L31 pseudogene 17	4.55859375	3.357671009	RPL31P17
SEC61G	Sec61 gamma subunit	4.7841796875	4.2677418032	SEC61G
SUMO4	SMT3 suppressor of mif two 3 homolog 4 (S. cerevisiae)	4.359375	5.4078103567	SUMO4
VSTM2A	V-set and transmembrane domain containing 2A	5.203125	3.8090994756	VSTM2A

Second data set: Gene Expression Data

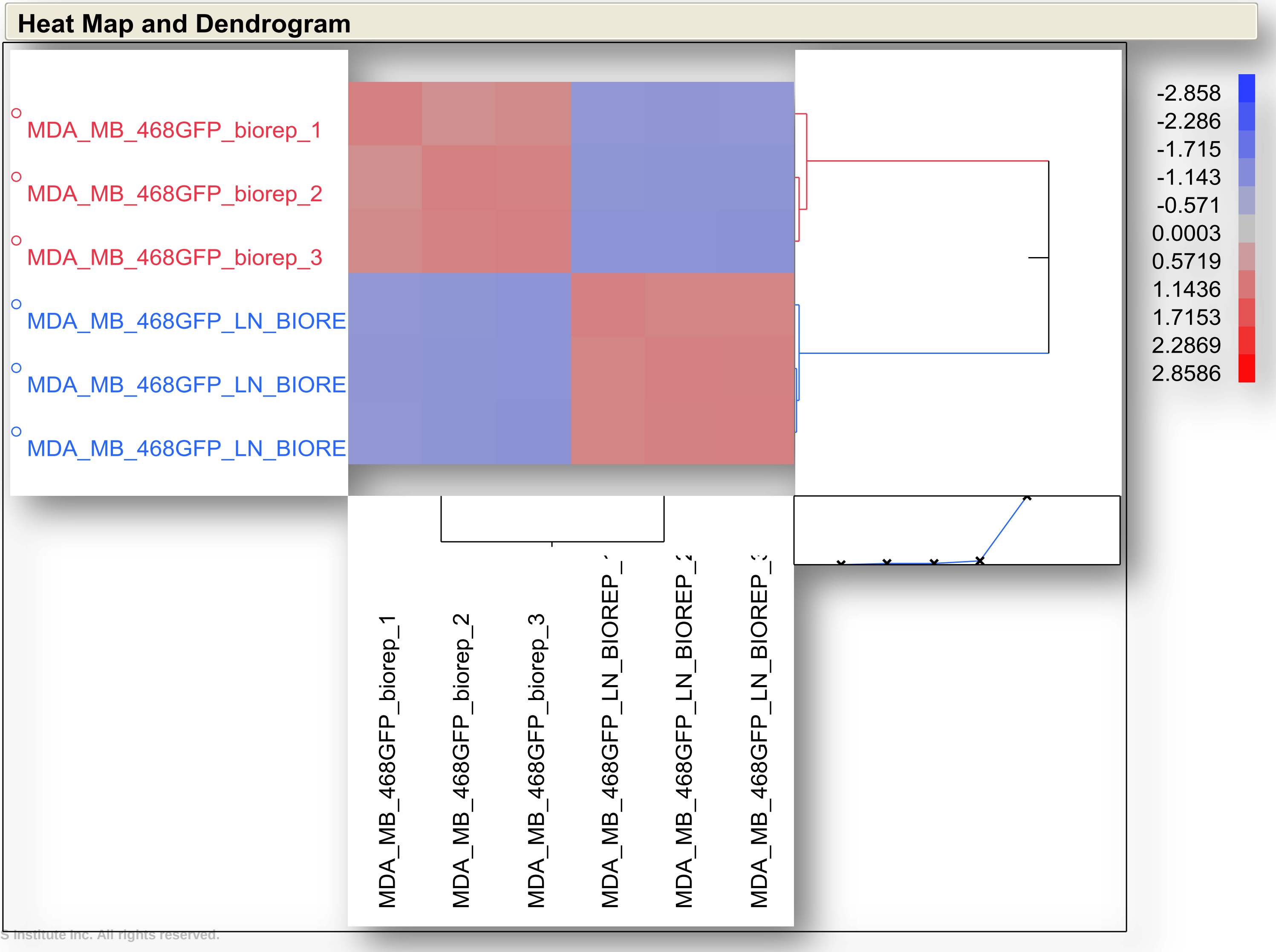
- Step 1: Quality Control Checks
- Step 2: ANOVA to find out significant differencially expression levels mapping the chromosome cytoband of interest
- Step 3: Pearson Correlation between Copy Number Variation and Gene Expression Data

Expression Data – Distribution Analysis

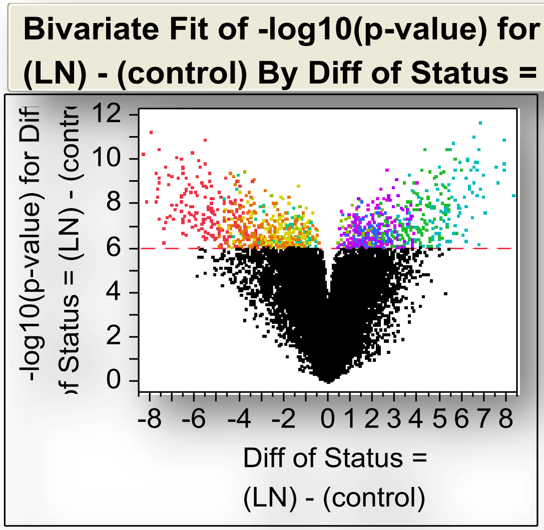



Oneway Analysis of Response By ColumnName

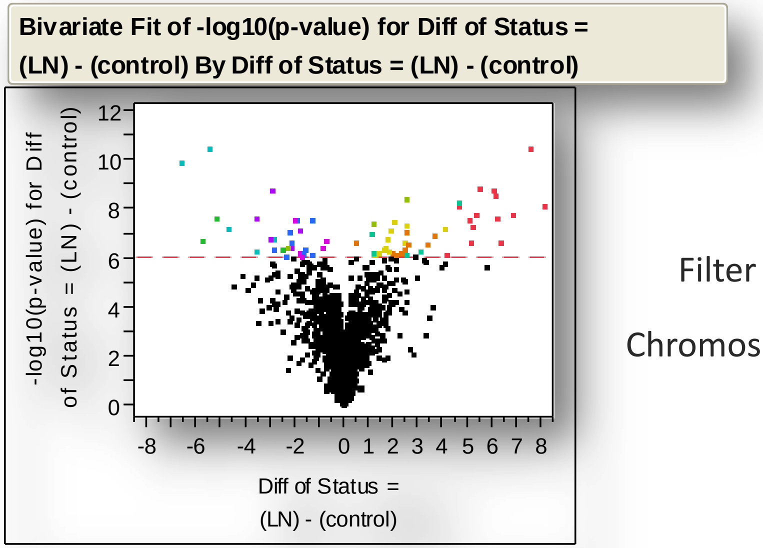
Expression Data – Hierarchical Cluster Tree on Correlation



Expression Data - ANOVA

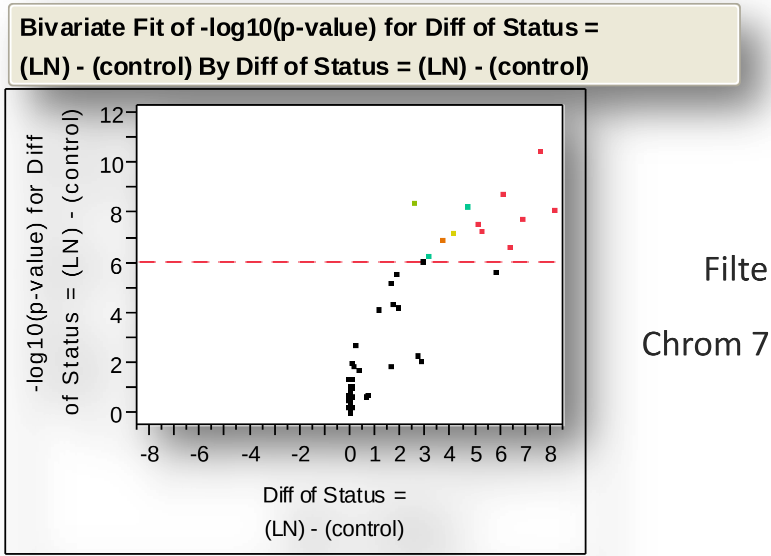

Bivariate Fit of -log10(p-value) for Diff of Status = (LN) - (control) By Diff of Status = (LN) - (control)

Overall


Bivariate Fit of -log10(p-value) for Diff of Status = (LN) - (control) By Diff of Status = (LN) - (control)

Filter on

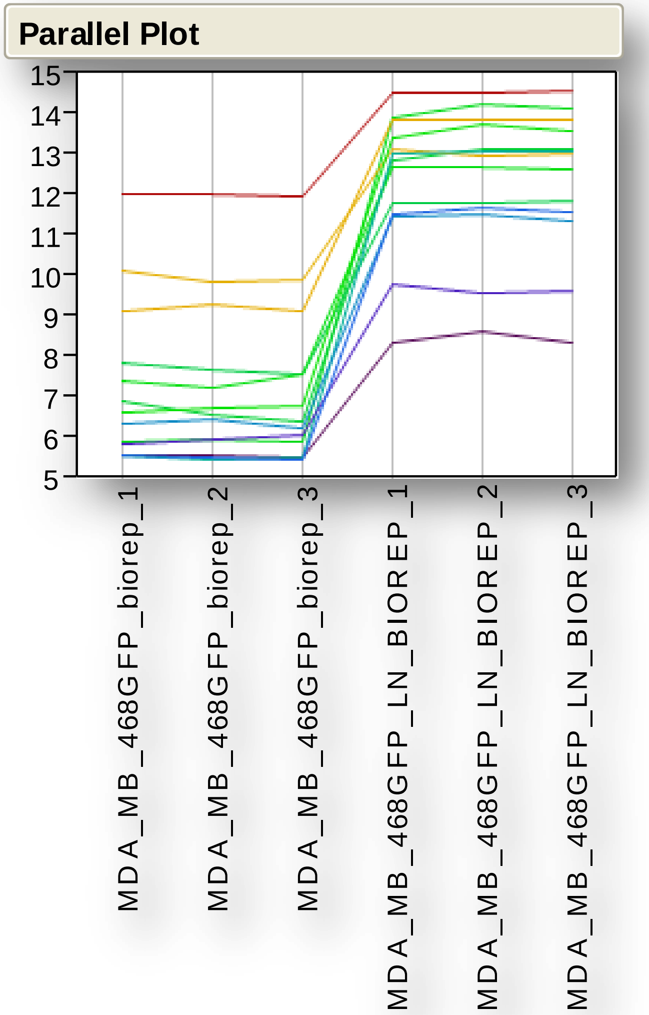
Chromosome 7


Bivariate Fit of -log10(p-value) for Diff of Status = (LN) - (control) By Diff of Status = (LN) - (control)

Filter on

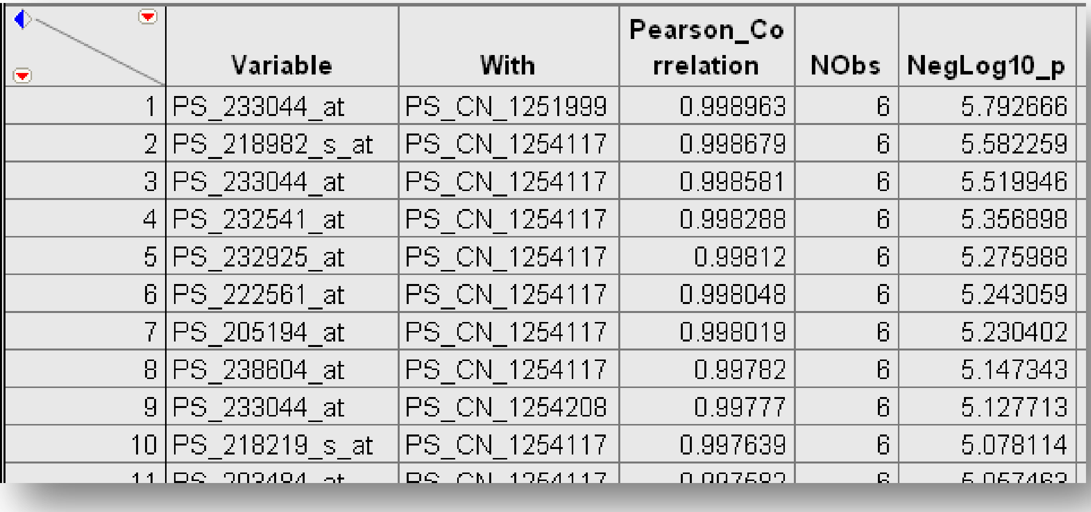
Chrom 7 – p11.2

Expression Data – Plot Intensities

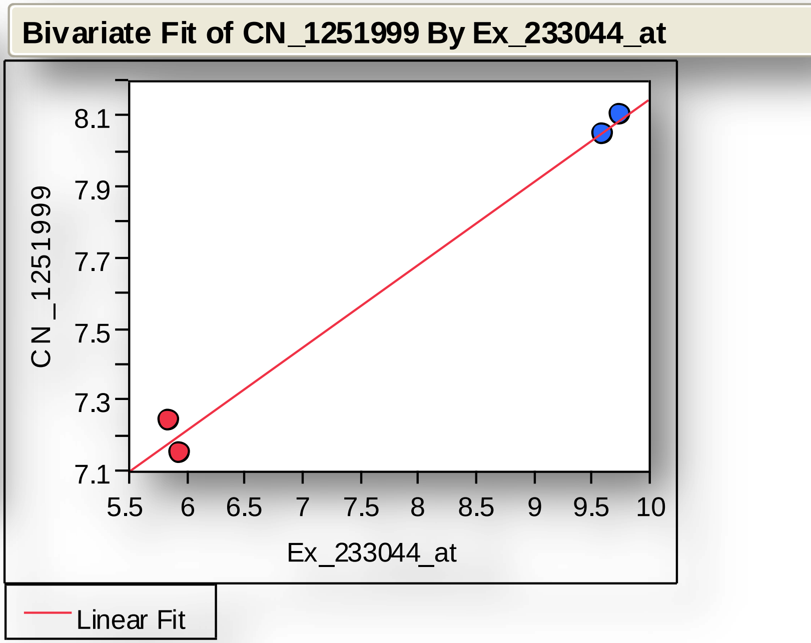


- Plot intensities levels of Chrom 7 p11.2
- There is a perfect correlation with the copy number variation outcome

Pearson Correlation of CNV and Expression Data



	Variable	With	Pearson_Correlation	NObs	NegLog10_p
1	PS_233044_at	PS_CN_1251999	0.998963	6	5.792666
2	PS_218982_s_at	PS_CN_1254117	0.998679	6	5.582259
3	PS_233044_at	PS_CN_1254117	0.998581	6	5.519946
4	PS_232541_at	PS_CN_1254117	0.998288	6	5.356898
5	PS_232925_at	PS_CN_1254117	0.99812	6	5.275988
6	PS_222561_at	PS_CN_1254117	0.998048	6	5.243059
7	PS_205194_at	PS_CN_1254117	0.998019	6	5.230402
8	PS_238604_at	PS_CN_1254117	0.99782	6	5.147343
9	PS_233044_at	PS_CN_1254208	0.99777	6	5.127713
10	PS_218219_s_at	PS_CN_1254117	0.997639	6	5.078114
11	PS_203494_at	PS_CN_1254117	0.997582	6	5.057463



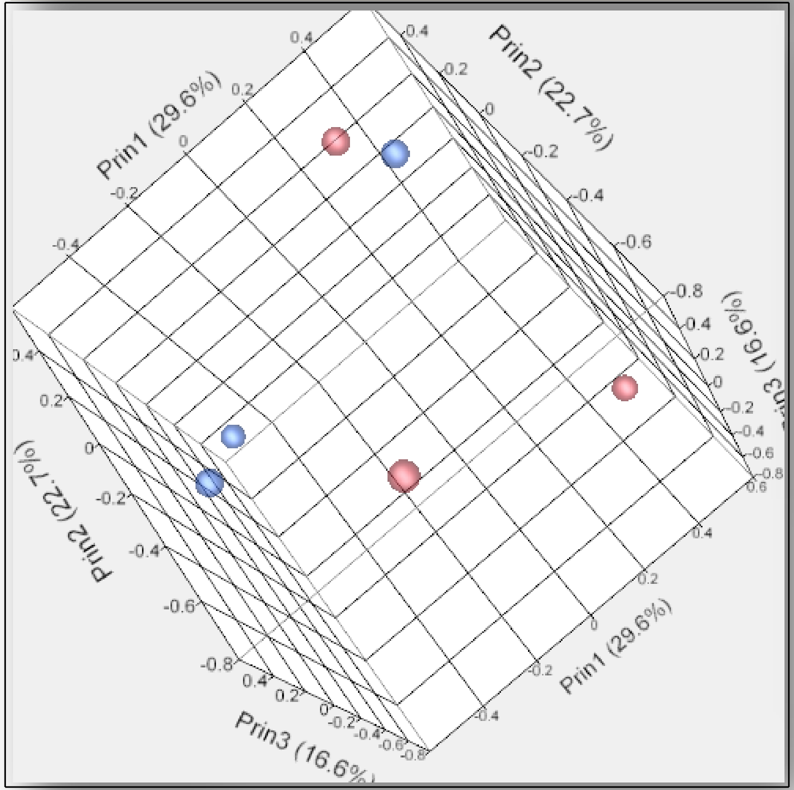
Bivariate Fit of CN_1251999 By Ex_233044_at

Third data set: Methylation Pattern

- Step 1: Quality Control Checks
- Step 2: ANOVA to find out significant methylation differences
- Step 3: Methylation Mapping

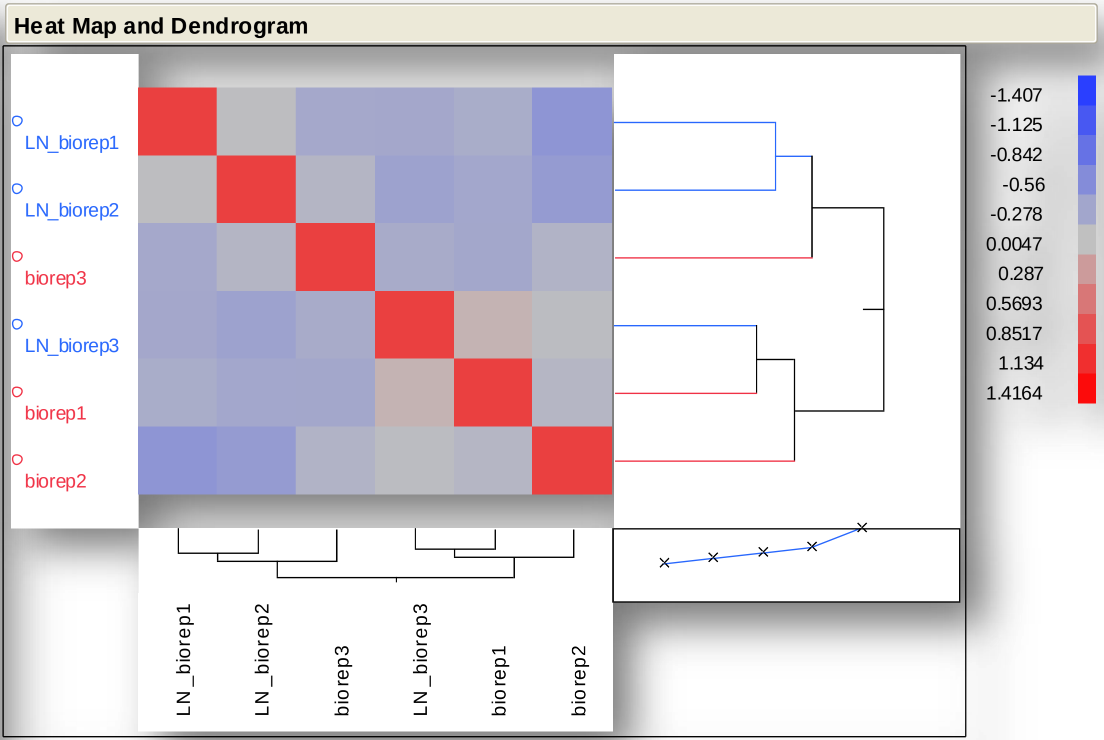
Epigenetics: Methylation Profiling

Scatterplot 3D

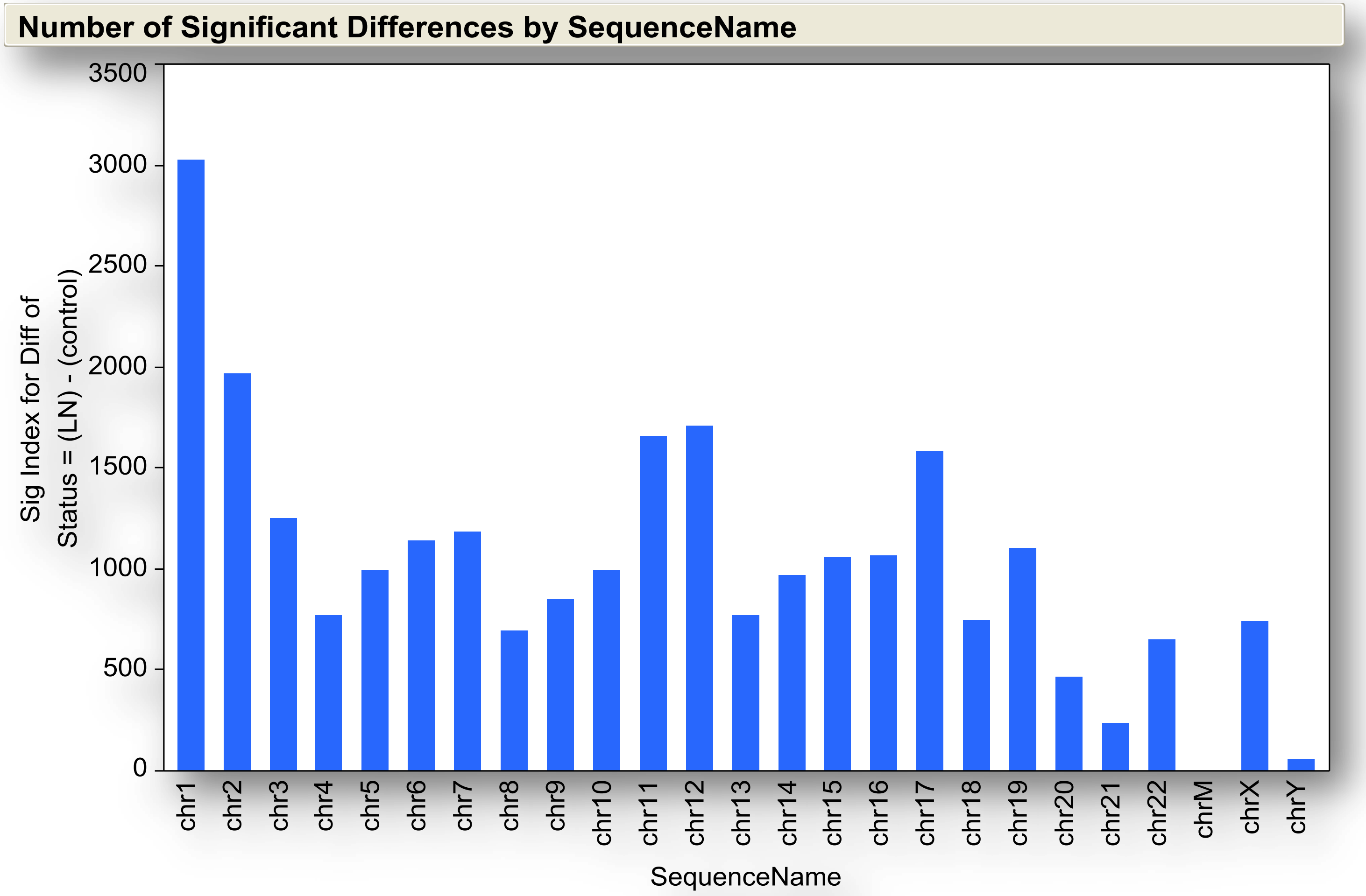


Data Columns Prin1 (29.6%) Prin2 (22.7%) Prin3 (16.6%)

Heat Map and Dendrogram

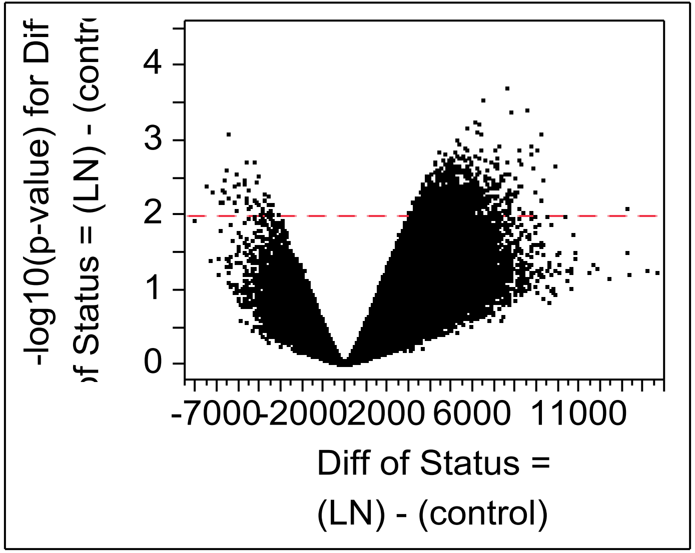


Epigenetics: ANOVA analysis on Methylation Profiling

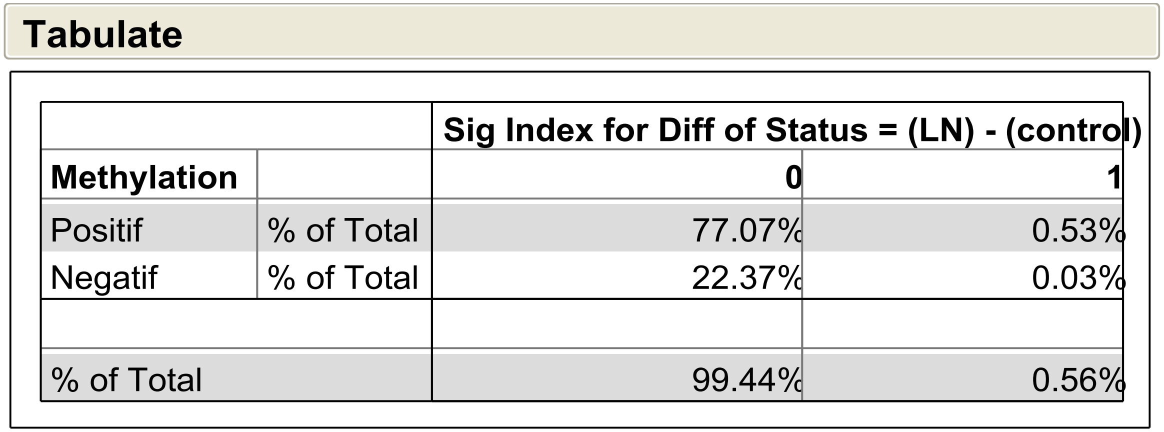


Epigenetics: ANOVA analysis on Methylation Profiling

Bivariate Fit of -log10(p-value) for Diff of Status = (LN) - (control) By Diff of Status = (LN) - (control)



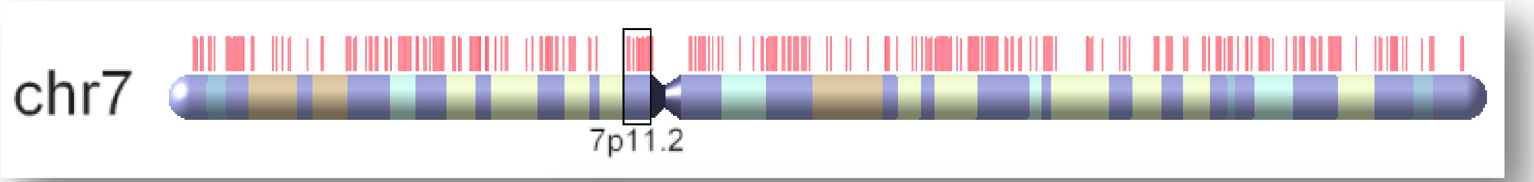
Tabulate

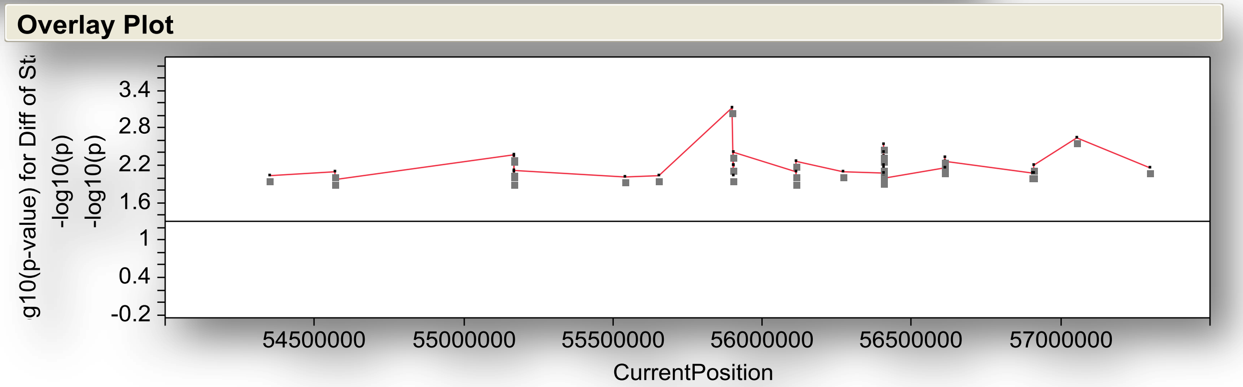


		Sig Index for Diff of Status = (LN) - (control)	
Methylation		**0**	**1**
Positif	% of Total	77.07%	0.53%
Negatif	% of Total	22.37%	0.03%
% of Total		99.44%	0.56%

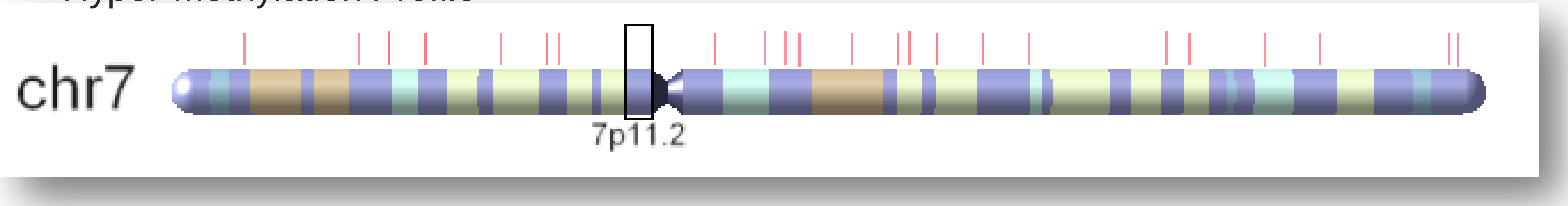
Epigenetics: Methylation Profiling of Chr7

Hypo-Methylation Profile





Hyper-Methylation Profile



Conclusion

- We have demonstrated how the cross-correlation tool in JMP Genomics simplifies the task of finding regions of correlation between SNP intensity, expression levels and methylation patterns.
- However, cross-correlation analysis is highly flexible and may be used for paired analysis of many other data types. For example, quantitative measures of expression or protein amounts may be paired combination with miRNA data to look for potential regulatory interactions.

Any Further Information …

valerie.nedbal@eur.sas.com

Or go to

www.jmp.com/software/genomics