Microarray Data Analysis Using Partek® Genomic Suite

Xiaowen Wang
Field Application Specialist
Partek Inc.
Who is Partek?

- Founded in 1993
- Building tools for statistics & visualization
- Focused on genomics
- Thousands of customers worldwide
- Worldwide, world-class customer support
What is Partek® Genomics Suite™?

Desktop software on Windows, Linux, Mac
Two Modes of Operation

Workflow

- Select the Workflow for your assay
- Most common steps
- Can access menus between steps

Expert

- Use commands on top menu bar
- Right click row/column header
- Combine options on workflows
Gene Expression Workflow

• Import & normalization
• QA/QC – exploratory analysis
• Detect differential expression
• Visualization
• Biological interpretation

• Genomic integration
  • Copy Number integration
  • MicroRNA Integration
MicroRNA Expression Workflow

- Import & Normalization
- QA/QC – Exploratory Analysis
- Detect Differential Expression
- Integration with gene expression
  - Combine with targeted gene
  - Find overrepresented miRNA target sets
  - Correlate miRNA with mRNA
- Biological Interpretation
Exon Expression Workflow

- Import & normalization
- QA/QC – exploratory analysis
- Detect differential expression
  - Gene level
  - Exon level
- Detect alternative splicing
- Biological interpretation
- Genomic integration
Copy Number Workflow

- Import & create copy number
- QA/QC – exploratory analysis
- Detect amplification and deletion regions
  - HMM
  - Genomic segmentation
- Detect copy number changes among different population
- Find genes in aberration regions
- Genomic Integration
  - Integration with LOH
  - Integration with gene expression
LOH and Allele Specific Copy Number

- Import & create ASCN
- QA/QC – Sample QC and HWE
- Create LOH
- Detect Allelic imbalances
- Find genes in aberration regions
- Genomic Integration
  - Integration with copy number
  - Integration with gene expression
Association Workflow

- Import genotype calls
- QA/QC – Sample QC and HWE
- Generate sample IBS (Identity by State)
- Chi-square test
  - Allele
  - Genotype
  - Dominant/Recessive
- Linkage Disequilibrium
  - $D$
  - $D'$
  - $r^2$
- SNP Duo/Trio
  - Visualization
Tiling Data Analysis

- Can be used for ChIP-chip data or Methylation data
- Detect enriched regions
- Motif discovery
  - Discover de novo motif
  - Search for known motif
- Find Genes that overlap enriched regions
Methylation Data Analysis

- Illumina 27K and 450K chip
- Import directly from .idat
- Detect differential methylated loci
- Detect differential methylated CpG islands
- Biological Interpretation
Spreadsheet – main repository of the data
- No software imposed limitation on size of the data
- Hierarchical relationship
- Corresponding to a data file (text or binary formats)
- A project may contain many spreadsheets

Workflow – guide you through on data analysis
Import and Export

Import
• Text files (.txt, .csv)
• Vendor formats
  • Affymetrix, Illumina, Agilent, Nimblegen, Fluidigm, Nanostring, Taqman…
• GEO (Gene Expression Omnibus)
• .bam/sam, vcf/bcf, Flow’s output

Export:
• Text file
• Binary
• jpg tiff pdf svg ps
• Zipped project
Edit Spreadsheet

- Drag and drop to change column/row order
- Create unique list
- Clone spreadsheet
  - Make a copy of a spreadsheet
  - Make a subset of experiment
- Sort Rows: stable sorting
- Fill column: easily fill in the cells based on range
- Merge columns: combine values in multiple columns into one
- Split column: convert a text into separate columns
- more…
Transformation

- Verify and imput missing data
- Change text to upper/lower case
- Normalize to baseline
- LOESS: two color data normalization
- Adjust copy number: convert between CN & LR
- Transpose the spreadsheet
- Convert continuous column to categorical
Normalization and scaling

<table>
<thead>
<tr>
<th>Method of Scaling</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardize</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Shift Min</td>
<td>0.0</td>
<td>&lt;Grand Min</td>
</tr>
<tr>
<td>Shift Mean</td>
<td>0.0</td>
<td>&lt;Grand Mean</td>
</tr>
<tr>
<td>Mean Scale</td>
<td>1.0</td>
<td>&lt;Grand Mean</td>
</tr>
<tr>
<td>Linear Scale</td>
<td>Spec</td>
<td>min 0.0 max 1.0</td>
</tr>
</tbody>
</table>

Example: min 0.0 max 1.0: linearly scale min value to max value.
Valid Spec: min max mean mode median q1 q3 (1st/3rd quartile)

<table>
<thead>
<tr>
<th>Method of Normalization</th>
<th>Base</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log (x+offset)</td>
<td>2.0</td>
</tr>
<tr>
<td>Power (x^y)</td>
<td>2.0</td>
</tr>
<tr>
<td>Square Root</td>
<td></td>
</tr>
<tr>
<td>Generalized Log</td>
<td></td>
</tr>
<tr>
<td>Variance Stabilization</td>
<td>Configure...</td>
</tr>
<tr>
<td>Quantile Normalization</td>
<td></td>
</tr>
<tr>
<td>Normalize To Control</td>
<td></td>
</tr>
<tr>
<td>Box-Cox Auto</td>
<td></td>
</tr>
<tr>
<td>Set Values to At Least</td>
<td>0.001</td>
</tr>
<tr>
<td>Ratio to Fold Change</td>
<td></td>
</tr>
<tr>
<td>Sine</td>
<td></td>
</tr>
<tr>
<td>Logit</td>
<td></td>
</tr>
<tr>
<td>Probit</td>
<td></td>
</tr>
</tbody>
</table>

Values:
- Log (x+offset)
- Power (x^y)
- Square Root
- Generalized Log
- Variance Stabilization
- Quantile Normalization
- Normalize To Control
- Box-Cox Auto
- Set Values to At Least
- Ratio to Fold Change
- Sine
- Logit
- Probit

Save Reference File:

Invert (1/x)

CUMulative Distribution Function
Data Filtering

- Interactive filter
  - Helpful on visually check sample labels
  - Easily filter based on group or range
- Row/Column filter manager
- Filter based on a list
- Filter based on genomic location
- Randomize and re sampling the data
Exploratory Analysis

- Dimension reduction
  - Principal components analysis
  - Multidimensional scaling
- Clustering analysis
  - Hierarchical clustering
  - K-means clustering
  - Self organizing map
- Sort rows by prototype
  - Time series data analysis
Inferential Statistics

- Parametric
  - t-Test, z-Test, ANOVA, Welch’s ANOVA, Pearson Correlation
- Non-parametric
  - Mann-Whitney, Kruskal-Wallis, Friedman, Chi-square, Rank correlation
- Others
  - Power Analysis
  - Survival Analysis
    - Cox regression
    - Kaplan-Meier Curve
  - Multiple test corrections
  - Descriptive statistics

\[ t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1-\bar{x}_2}} \sim \frac{\text{signal}}{\text{noise}} \]
Analysis of Variance

• No limitation on number of factors
• Factors can be categorical and/or numerical
• Automatically detect nested/crossed relationship
• Handle balanced or unbalanced experiment design
• Perform mixed model when random effect included
• It is also applied in:
  • Remove batch effect (batch remover)
  • Detect alternative splicing (alt-splicing ANOVA)
  • Detect differential expressed functional group (GO ANOVA)
  • Integrate with copy number analysis (Copy ANOVA)
Prediction Modeling

- Variable Selection
  - ANOVA
  - Forward selection
  - Backward elimination
  - Genetic algorithm
- Classification
  - K nearest neighbor
  - Nearest centroid
  - Discriminate analysis
  - Support vector machine
  - Partial least squares
- Cross validation
Visualization

- Genome Browser
- Heat map
- Dot plot
- Scatter plots
- Profile
- Venn Diagram
- Histogram
- Bar charts, Pie Charts
- Volcano Plots
- MA plots
- …much more
Integration Approaches

- Combine data from different experiments
- Correlate data from different experiments/assays/platforms
- Visualize data from different experiments/assays/platforms
- Venn Diagram—5-way
- Various “Tools” commands
Biological Interpretation

- Biological relevance is not usually found in only a single gene

- Database:
  - Gene Ontology,
  - KEGG Pathways
  - Custom annotation
    - GMT, GAF, text file

- Method:
  - Enrichment: test if lead genes are over-represented in any pathway
  - Pathway ANOVA: detect differentially expression pathway
What is Partek® Pathway™?

• Seamlessly integrated with PGS
• Find enriched pathway
• Detect differentially expressed pathway
• Visualize gene relationships
• Color the genes based on stat results
• Search for specific pathway and gene
• Support all species in KEGG database

Partek Pathway
Extends Biological Interpretation
Pathway Enrichment vs. Pathway ANOVA

• Breast cancer cell lines

• 2 factors
  1. Tumor status, Normal vs. Tumor
  2. ER status (+ and -)
Pathway Enrichment Result

- Only 3 significant pathways with respect to ER status
- Link to breast cancer is very unclear…
Pathway ANOVA Result

- No pre-filtering
- No information loss
- Many more pathways

<table>
<thead>
<tr>
<th>Pathway Name</th>
<th>Database</th>
<th># Pts</th>
<th>Pathway ID</th>
<th>p-value(FR Status)</th>
<th>FoldChange(FR vs. ER)</th>
<th>FoldChange(FR+ vs. ER-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Toll-like receptor signaling pathway</td>
<td>kegg</td>
<td>57</td>
<td>kegg_pathway_149</td>
<td>2.16305e-007</td>
<td>-1.66921</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>2. Fc epsilon RT signaling pathway</td>
<td>kegg</td>
<td>45</td>
<td>kegg_pathway_158</td>
<td>3.51312e-007</td>
<td>-1.30689</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>3. Epithelial cell signaling in Helicobacter pylori</td>
<td>kegg</td>
<td>55</td>
<td>kegg_pathway_210</td>
<td>2.14334e-006</td>
<td>1.34046</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>4. RIG-I-like receptor signaling pathway</td>
<td>kegg</td>
<td>44</td>
<td>kegg_pathway_151</td>
<td>2.79357e-006</td>
<td>1.34439</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>5. Shigellosis</td>
<td>kegg</td>
<td>52</td>
<td>kegg_pathway_212</td>
<td>3.15081e-006</td>
<td>1.18379</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>6. Acherens junction</td>
<td>kegg</td>
<td>60</td>
<td>kegg_pathway_143</td>
<td>3.372e-006</td>
<td>1.23759</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>7. Dorso-ventral axis formation</td>
<td>kegg</td>
<td>17</td>
<td>kegg_pathway_133</td>
<td>4.7785e-006</td>
<td>1.48635</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>8. Circadian rhythm - mammal</td>
<td>kegg</td>
<td>14</td>
<td>kegg_pathway_162</td>
<td>7.07725e-006</td>
<td>1.32232</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>9. Pertussis</td>
<td>kegg</td>
<td>45</td>
<td>kegg_pathway_214</td>
<td>1.22237e-005</td>
<td>-2.18475</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>10. Galactose metabolism</td>
<td>kegg</td>
<td>19</td>
<td>kegg_pathway_6</td>
<td>1.28657e-005</td>
<td>1.57259</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>11. Small cell lung cancer</td>
<td>kegg</td>
<td>69</td>
<td>kegg_pathway_243</td>
<td>1.39088e-005</td>
<td>-1.43267</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>12. Jak-STAT signaling pathway</td>
<td>kegg</td>
<td>66</td>
<td>kegg_pathway_153</td>
<td>1.6696e-005</td>
<td>1.7822</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>13. Folate biosynthesis</td>
<td>kegg</td>
<td>9</td>
<td>kegg_pathway_79</td>
<td>2.11929e-005</td>
<td>-2.92003</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>14. Renal cell carcinoma</td>
<td>kegg</td>
<td>60</td>
<td>kegg_pathway_232</td>
<td>2.31282e-005</td>
<td>1.19673</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>15. Apoptosis</td>
<td>kegg</td>
<td>63</td>
<td>kegg_pathway_129</td>
<td>2.52002e-005</td>
<td>-1.36823</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>16. Long-term depression</td>
<td>kegg</td>
<td>32</td>
<td>kegg_pathway_172</td>
<td>2.97473e-005</td>
<td>-1.14138</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>17. Pathways in cancer</td>
<td>kegg</td>
<td>223</td>
<td>kegg_pathway_229</td>
<td>3.29722e-005</td>
<td>1.36491</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>18. Toxoplasmosis</td>
<td>kegg</td>
<td>80</td>
<td>kegg_pathway_220</td>
<td>3.3033e-005</td>
<td>1.43154</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>19. Adipocytokine signaling pathway</td>
<td>kegg</td>
<td>42</td>
<td>kegg_pathway_181</td>
<td>3.59866e-005</td>
<td>-1.3189</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>20. Glycosaminoglycan biosynthesis - chondroitin</td>
<td>kegg</td>
<td>15</td>
<td>kegg_pathway_52</td>
<td>4.39343e-005</td>
<td>1.57672</td>
<td>ER+ down vs ER-</td>
</tr>
<tr>
<td>21. T cell receptor signaling pathway</td>
<td>kegg</td>
<td>67</td>
<td>kegg_pathway_156</td>
<td>4.71752e-005</td>
<td>1.81119</td>
<td>ER+ down vs ER-</td>
</tr>
</tbody>
</table>
Pathway ANOVA finds important results

- **ERBB2** encodes Her2
- Linked to breast cancer
- Filtered out by Enrichment
What is disruption?

Disruption is something you can ONLY see with Partek

- Sometimes we are not expecting an entire pathway to change
- In this case we are interested in “disruption” of the pathway
- These changes can cancel each other out and remain hidden
Pathway ANOVA
Show my favorite pathway

Gene name, pathway name, etc.
Partek is Your Partner

Free Software Trials
• partek.com/trial
• Email: licensing@partek.com

Self-learning
• Help > On-line tutorials
• Recorded webinars

Regional Technical Support
• Email: support@partek.com
• Phone: +1-314-878-2329