Use Case 16: Differential Expression Analysis of Two Breast Cancer Cell Lines Using Cuffdiff

Epigenome Informatics Workshop
Bioinformatics Research Laboratory
mRNA-Seq

- Read mapping
  - tophat
- Gene expression estimation
  - cufflinks
  - Confidence intervals
- Gene expression changes
  - Sample groups
  - cuffdiff
Use case for RNA-Seq tools

• 2 breast cancer cell lines

  – Joe Gray 51 breast cancer cell lines panel
Evaluate Gene Expression Differences

<table>
<thead>
<tr>
<th>Sample</th>
<th>Luminal/Basal</th>
<th>ER status</th>
<th>PR status</th>
<th>Her2/ERBB2 status</th>
</tr>
</thead>
<tbody>
<tr>
<td>BT474</td>
<td>Luminal</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>HCC1143</td>
<td>BasalA</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Exercise plan

• Evaluate differences between sample groups
  – *Luminal vs basal breast cancer*
  – *Cuffdiff tool*
  – Gene enrichment via GSEA/MSigDB

• Process subset of one of the cell lines
  – BT474
Evaluate Gene Expression Differences

• Goal
  – Gene expression comparison
    • Luminal vs Basal

• Tasks – Completed by BRL staff
  – RNA-Seq received from Joe Gray and uploaded to the Genboree Workbench
  – Map reads using Tophat (via Genboree Workbench)
    • File locations (output from Tophat job)
      – Epigenome Toolset Demo Input Data -> Databases -> RNA-Seq Tool Demo -> Files
        » TopHat-BT474_accepted_hits.bam (Luminal)
        » TopHat-HCC1143_accepted_hits.bam (Basal)
    • Job metrics
      – 4-10 hours
      – 8 cores

• Tasks – To do for workshop attendees
  – Run cuffdiff to find significant changes in transcript expression, splicing, and promoter use
Create an Entity List for BT474

Drag TopHat output “BT474_accepted_hits.bam” into Input Data
Choose a Destination for the Entity List

Drag the destination database into Output Targets (this tells Genboree where to deposit the Entity List)
Invoke the “Create File Entity List” Tool

Invoke “Create File Entity List”, which will present the dialogue box on the next slide.
Give Your First Entity List a Meaningful Name

Name your entity list, and then click “submit”
The entity list will be created immediately and will appear in your destination database under “Lists and Selections” → “Lists of Files” (next slide)
The entity list is located in your destination database under “Lists and Selections” → “Lists of Files”. Note the new icon.
Create an Entity List for HCC1143
(execute same steps as for BT474)
Create the HCC1143 Entity List

- Open Epigenome Toolset Demo Input Data → RNA-Seq Tool Demo → Files and drag HCC1143_accepted_hits.bam into Input Data
- Drag your destination database (GenboreeUser_database) into Output Targets
- Invoke “Create File Entity List”
Give Your Second Entity List a Meaningful Name

Name your entity list, and then click “submit”
The entity list will be created immediately and will appear in your destination database under “Lists and Selections” → “Lists of Files” (next slide)
Run Cuffdiff to Find Expression Differences

- Drag the entity lists into Input Data
- Drag the destination database into Output Targets
- Invoke Transcriptome → Analyze RNA-Seq Data → Detect... by Cuffdiff

Note: Cuffdiff results are deposited in GenboreeUser_database → Files → Cuffdiff
Name The Two Samples Before Submitting
Message Denoting Successful Cuffdiff Submission

Detect Transcription Changes by Cuffdiff

Job Id: wbJob-cuffdiff-IwKv5s-2270

Your job has been successfully submitted. You will be notified by email when your job has completed.

If you have questions, please contact genboree_admin@genboree.org for assistance.
Hello Genboree User,

Your job completed successfully.

Job Summary:
JobID - wbJob-cuffdiff-lwKv5s-2270

Additional Info:
Database: 'GenboreeUser_database'
Group: 'GenboreeUser_group'

You can download result files from the 'Cuffdiff-2013-2-19-15:59:47' folder under the 'Cuffdiff' directory.
Note that files without data are stored under the 'raw' folder.

- The Genboree Team
Highlight the file of interest, and then click on the link to download.
Example Data Returned by Cuffdiff

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>test_id</td>
<td>gene_Name</td>
<td>gene_id</td>
<td>locus</td>
<td>sample_1</td>
<td>sample_2</td>
</tr>
<tr>
<td>2</td>
<td>NM_000014</td>
<td>A2M</td>
<td>NM_000014</td>
<td>chr12:9217772-9268558</td>
<td>Basal_Br_Cancer_HCC1143</td>
<td>Lumina_Br_Cancer_BT474</td>
</tr>
<tr>
<td>54</td>
<td>NM_000067</td>
<td>CA2</td>
<td>NM_000067</td>
<td>chr8:86376130-86393721</td>
<td>Basal_Br_Cancer_HCC1143</td>
<td>Lumina_Br_Cancer_BT474</td>
</tr>
<tr>
<td>191</td>
<td>NM_000206</td>
<td>IL2R</td>
<td>NM_000206</td>
<td>chrX:70327253-70314811</td>
<td>Basal_Br_Cancer_HCC1143</td>
<td>Lumina_Br_Cancer_BT474</td>
</tr>
<tr>
<td>189</td>
<td>NM_000214</td>
<td>JAG1</td>
<td>NM_000214</td>
<td>chr20:10618331-10654694</td>
<td>Basal_Br_Cancer_HCC1143</td>
<td>Lumina_Br_Cancer_BT474</td>
</tr>
<tr>
<td>401</td>
<td>NM_000422</td>
<td>KRT17</td>
<td>NM_000422</td>
<td>chr17:39775691-39780682</td>
<td>Basal_Br_Cancer_HCC1143</td>
<td>Lumina_Br_Cancer_BT474</td>
</tr>
<tr>
<td>457</td>
<td>NM_000486</td>
<td>AQP2</td>
<td>NM_000486</td>
<td>chr12:50344523-50352664</td>
<td>Basal_Br_Cancer_HCC1143</td>
<td>Lumina_Br_Cancer_BT474</td>
</tr>
<tr>
<td>563</td>
<td>NM_000598</td>
<td>IGFBP3</td>
<td>NM_000598</td>
<td>chr7:45951843-45960871</td>
<td>Basal_Br_Cancer_HCC1143</td>
<td>Lumina_Br_Cancer_BT474</td>
</tr>
<tr>
<td>579</td>
<td>NM_000615</td>
<td>NCAM1</td>
<td>NM_000615</td>
<td>chr11:112831968-113149158</td>
<td>Basal_Br_Cancer_HCC1143</td>
<td>Lumina_Br_Cancer_BT474</td>
</tr>
<tr>
<td>673</td>
<td>NM_000660</td>
<td>TGFBI</td>
<td>NM_000660</td>
<td>chr19:41836811-41859831</td>
<td>Basal_Br_Cancer_HCC1143</td>
<td>Lumina_Br_Cancer_BT474</td>
</tr>
<tr>
<td>805</td>
<td>NM_000864</td>
<td>HTR1D</td>
<td>NM_000864</td>
<td>chr1:23518387-23521222</td>
<td>Basal_Br_Cancer_HCC1143</td>
<td>Lumina_Br_Cancer_BT474</td>
</tr>
<tr>
<td>953</td>
<td>NM_000916</td>
<td>OXTR</td>
<td>NM_000916</td>
<td>chr3:8792094-8811300</td>
<td>Basal_Br_Cancer_HCC1143</td>
<td>Lumina_Br_Cancer_BT474</td>
</tr>
<tr>
<td>984</td>
<td>NM_001001389</td>
<td>CD44</td>
<td>NM_001001389</td>
<td>chr11:35160516-35253949</td>
<td>Basal_Br_Cancer_HCC1143</td>
<td>Lumina_Br_Cancer_BT474</td>
</tr>
</tbody>
</table>

Note: Sheet sorted by the ‘significant’ column
Further Analysis Using GSEA/MSigDB

• Evaluate differences between sample groups
  – Luminal vs basal breast cancer
  – Cuffdiff tool
  – *Gene enrichment via GSEA/MSigDB*
GSEA/MSigDB

• **Gene Set Enrichment Analysis**

• **Molecular Signatures Database**

• Exposed as a web service
Register with MSigDB

http://www.broadinstitute.org/gsea/login.jsp

Click here to register to view the MSigDB gene sets and/or download the GSEA software. This helps us track and better serve our user community.

If you have already registered for GSEA or MSigDB please enter your registration email address below.

Items marked with * are required.

Email: *

[login button]
Register with MSigDB

GSEA/MSigDB Registration and License Agreement

 Instructions to obtain GSEA software and/or MSigDB gene sets. Please Read carefully.

1. Fill in the form below.
2. The software and gene sets are freely available to individuals in academic and private institutions. There are no licensing fees.
3. Source code is freely available.
4. Read the license agreement and make sure you agree with the terms of the agreement.
   If so, click the 'I Agree' button at the end of the form and you will be transferred to the GSEA download page.

Items marked with * are required.

Name: *

Email: *
(You will receive a registration notification email.)

Organization: *

Country: *
select a country

Join mailing list:
✓ notify me of GSEA updates
(You will receive a confirmation email. Reply to join the list.)

Comments:

GSEA and MSigDB license agreements:

**** GSEA/MSigDB LICENSE AGREEMENT ****

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
SINGLE USER LICENSE AGREEMENT FOR INTERNAL RESEARCH PURPOSES ONLY

This Agreement is made between Massachusetts Institute of Technology with a principal address at 77 Massachusetts Avenue, Cambridge, MA 02139 ("MIT") and the subscriber above ("LICENSEE"), and is effective at the date the downloading is completed and proper registration/licensure
Login to MSigDB

http://www.broadinstitute.org/gsea/login.jsp

Next:

• Click ‘Explore the Molecular Signatures Database (MSigDB)’
  • http://www.broadinstitute.org/gsea/msigdb/index.jsp
• Click ‘Investigate Gene Sets’
  • http://www.broadinstitute.org/gsea/msigdb/annotate.jsp
Use MSigDB

Investigate Gene Sets

Gain further insight into the biology behind a gene set by using the following tools:

- **compute overlaps** with other gene sets in MSigDB (more...)
- **display the gene set expression profile** based on a selected compendium of expression data (more...)
- **categorize** members of the gene set by gene families (more...)

Gene Identifiers

- C1: positional gene sets
- C2: curated gene sets
- CGP: chemical and genetic perturbations
- CP: canonical pathways
- CP: BIOCARTA: BioCarta gene sets
- CP: KEGG: KEGG gene sets
- CP: REACTOME: Reactome gene sets
- C3: motif gene sets
- MIR: microRNA targets
- TFT: transcription factor targets
- C4: computational gene sets
- CGN: cancer gene neighborhoods
- CM: cancer modules
- C5: GO gene sets
- BP: GO biological process
- CC: GO cellular component
- MF: GO molecular function

Gene Identifier Platform

[GENE SYMBOL]
Gene expression differences

Filter by “significant” column

Copy “official” gene symbol
Use MSigDB

- **Gene Identifiers**
  - CA2, HMBS, ITGA6, JAG1, LAMA3, NF1, KRT17, NOTCH3, APRT, F12, TGFBR1, ADORA2A, ALDH1A3, TSP1, CRAT, GSTT1, HTR1D, OXTR, CD44, CD44, SMAD5, MTUS1, ATP5C1, ZNF787, PTHR1, ARHGEF35, MAGED1, RHBD1, LSS

- **Compute Overlaps**
  - Select gene sets
    - C1: positional gene sets
    - C2: curated gene sets
    - CGP: chemical and genetic perturbations
    - CP: canonical pathways
      - CP:BIOCARTA: BioCarta gene sets
      - CP:KEGG: KEGG gene sets
      - CP:REACTOME: Reactome gene sets
    - C3: motif gene sets
    - MIR: microRNA targets
    - TFT: transcription factor targets
    - C4: computational gene sets
      - CGN: cancer gene neighborhoods
      - CM: cancer modules
    - C5: GO gene sets
      - BP: GO biological process
      - CC: GO cellular component
      - MF: GO molecular function

- **Number of gene sets returned**: 30
- **Diff expressed genes**

- **Gene Identifier Platform**
  - GENE SYMBOL
Use MSigDB

Compute Overlaps for Selected Genes

Converted 701 submitted identifiers into 599 gene symbols. [click here for details.]

<table>
<thead>
<tr>
<th>Collections</th>
<th># Overlaps Shown</th>
<th># Gene Sets in Collections</th>
<th># Genes in Comparison (n)</th>
<th># Genes in Collections (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2, C4, C5</td>
<td>10</td>
<td>5607</td>
<td>599</td>
<td>22684</td>
</tr>
</tbody>
</table>

Click the gene set name to see the gene set page. Click the number of genes [in brackets] to download the list of genes.

Color bar shading from light green to black, where lighter colors indicate more significant p values (< 0.05) and black indicates less significant p values (>= 0.05).

Export: Excel

<table>
<thead>
<tr>
<th>Gene Set Name [ # Genes (K) ]</th>
<th>Description</th>
<th># Genes in Overlap (k)</th>
<th>k/K</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUYTTEN_NIPP1_TARGETS_DN [777]</td>
<td>Genes down-regulated in PC3 cells (prostate cancer) after knockdown of NIPP1 [Gene ID=5511] by RNAi.</td>
<td>67</td>
<td></td>
<td>0 e-0</td>
</tr>
<tr>
<td>SMID_BREAST_CANCER_BASAL_DN [713]</td>
<td>Genes down-regulated in basal subtype of breast cancer samples.</td>
<td>65</td>
<td></td>
<td>0 e-0</td>
</tr>
<tr>
<td>SMID_BREAST_CANCER_LUMINAL_B_DN [599]</td>
<td>Genes down-regulated in the luminal B subtype of breast cancer.</td>
<td>63</td>
<td></td>
<td>0 e-0</td>
</tr>
<tr>
<td>CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_Bмагази [462]</td>
<td>The ‘group 5 set’ of genes associated with acquired endocrine therapy resistance in breast tumors expressing ESR1 but not ERBB2 [Gene ID=2099, 2004].</td>
<td>54</td>
<td></td>
<td>0 e-0</td>
</tr>
<tr>
<td>SMID_BREAST_CANCER_BASAL_UP [675]</td>
<td>Genes up-regulated in basal subtype of breast cancer samples.</td>
<td>78</td>
<td></td>
<td>0 e-0</td>
</tr>
<tr>
<td>CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN [456]</td>
<td>Genes down-regulated in luminal-like breast cancer cell lines compared to the basal-like ones.</td>
<td>59</td>
<td></td>
<td>0 e-0</td>
</tr>
</tbody>
</table>
Use MSigDB

<table>
<thead>
<tr>
<th>Gene Set Name</th>
<th>Description</th>
<th>Count</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMID_BREAST_CANCER_BASAL_DN [713]</td>
<td>Genes down-regulated in basal subtype of breast cancer samples.</td>
<td>65</td>
<td>$0,e^0$</td>
</tr>
<tr>
<td>SMID_BREAST_CANCER_LUMINAL_B_DN [599]</td>
<td>Genes down-regulated in the luminal B subtype of breast cancer.</td>
<td>63</td>
<td>$0,e^0$</td>
</tr>
<tr>
<td>SMID_BREAST_CANCER_BASAL_UP [676]</td>
<td>Genes up-regulated in basal subtype of breast cancer samples.</td>
<td>78</td>
<td>$0,e^0$</td>
</tr>
<tr>
<td>CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_SAL_DN [456]</td>
<td>Genes down-regulated in luminal-like breast cancer cell lines compared to the basal-like ones.</td>
<td>59</td>
<td>$0,e^0$</td>
</tr>
</tbody>
</table>

Enrichments for gene sets differentiating luminal vs basal breast cancer cells
Acknowledgments

• BRL
  – Aleksandar Milosavljevic
  – Cristian Coarfa, Alan R Harris
• BRL core
  – Matt Roth, Kevin Riehle
• Genboree programming team
  – Andrew Jackson, Sameer Paithankar, Sriram Raghuram
• Former BRL members
  – Chia-Chin Wu, Arpit Tandon
• Peak Calling
  – Anshul Kundaje
  – Bob Thurman (UW)
  – Noam Shoresh (BI)
  – Martin Hirst (BCGSC)
  – Lee Daniels (NIH)
  – Wei Li (BCM)
• RNA-Seq
  – Adrian Lee, Joe Gray
  – Laising Yen, Kalpana Kannan
  – Sean McGuire, Christine Pichot