Computational techniques for understanding host transcriptional responses to TB

Fergal Duffy
TB Bioinformatics Working Group

12 November 2019
Talk Overview

• Host transcriptional responses to TB
  1. Sources of open data
  2. Pre-processing transcriptional profiles
  3. Cross-study comparisons
     • Case study: Development of ACS-CoR
  4. Cross-species comparisons
     • Case study: Mouse ULD signature
A typical workflow

1. Acquire data
   - Yours or public data
   - RNAseq or microarray

2. Preprocess
   - Normalization

3. Downstream analyses
   - Differential expression
   - Cross study comparisons
   - Predictive models
Understanding host transcriptional responses

• Most common: systematic responses in blood
  – Either whole blood or PBMCs
  – Using microarrays or RNAseq or ...
  – 12,000 – 40,000 “measurements”
    • Can be array probes/gene counts/splice junctions
Why blood?

- Local responses determine granuloma outcome
- But we can still learn from systemic responses
- Blood is
  - Accessible
  - Many tools to analyze
  - Suitable for use in clinical diagnostic

Lin et al, 2014
Many blood transcriptional signatures to diagnose TB
Many blood transcriptional signatures to diagnose TB

- Table lists human whole-blood microarray study cohorts available in GEO as of 2016
- Total of ~2,500 transcriptional profiles including TB, LTB and related diseases and healthy individuals
- These have been the basis of many TB diagnostic signatures
- Meta-analysis led to 3-gene signature
  - GB5, DUSP3, KLF2

<table>
<thead>
<tr>
<th>Year</th>
<th>Reference</th>
<th>Platform</th>
<th>Use</th>
<th>Country</th>
<th>Age</th>
<th>HIV status</th>
<th>Active tuberculosis controls or smear</th>
<th>Healthy controls</th>
<th>Latent tuberculosis</th>
<th>Other disease</th>
<th>Active tuberculosis</th>
<th>Treatment</th>
<th>Total</th>
<th>Miscellaneous</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>Sweeney et al, 2016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sweeney et al, 2016
Open TB transcriptional data on GEO

(1) Build your query

(2) Look for result Series
Open TB transcriptional data on SRA/BioBProject

Open TB transcriptional data on ArrayExpress

https://www.ebi.ac.uk/arrayexpress/

Search with Booleans (‘OR’ ‘AND’) and filter to find human array assays.
Microarray basics

- Labelled fluorescent transcripts hybridized to probes
- Chip scanned to image, continuous intensity values represent gene expression levels
Microarray normalization

- Background Correction
- Log transform
  - Results in approximately normally distributed expression values
- Quantile Normalization between arrays
- R limma package
RNAseq basics

- Count number of reads mapping to areas of interest
  - Different statistical approaches to count-based vs continuous measurements (e.g. microarrays)
- Also consider sequencing depth
RNAseq normalization

- Estimate mean and variance of counts and use to correct bias in low counts
  - Negative binomial
  - Poisson

- Correct for sequencing depth
  - Counts per million

- Log transform

- R limma/DESeq packages
Differential Expression Analysis

- **Does my gene of interest differ across **two (or more) conditions** in one study?**
  - Same approach for microarrays/RNAseq
  - t-test, linear models, ANOVA
  - Correct for multiple test: false discovery approach
Challenges for signature development

• How do we design a statistically interpretable, robust and universal signature?
  – n (samples) << p (genes)

• How do we deal with technical variation/batch effects between studies?

• How do we translate to new sequencing technologies?
Case Study: ACS-CoR
Case Study: ACS-CoR

Zak et al, 2016

• Cohort: Latent TB+ South African Adolescents
  – Recruited and prospectively followed for 2 years
  – 46 individuals developed active TB (Progressors) matched to 107 Control individuals
  – Whole blood RNAseq at enrollment
    • optionally at 180 days, 360 days, 540 days post enrollment

• Goal: Develop a transcriptional signature of TB risk
  – Look for transcriptional responses specific to Progressors prior to disease onset
ACS-CoR signature discovery workflow

- Derived & cross-validated a TB risk signature (RNAseq) on LTB+ adolescents
  - ACS 3:1 training: test split
- Translated to qPCR and validated on a new cohort of TB household contacts
ACS-CoR: Training a Pairwise SVM

Advantages
- Interpretable
- Mathematically straightforward
- Visualizable
- Voting is robust to missing genes

ACS-CoR is based on pairs of splice junctions
1. Pick differentially expressed splice junctions
2. Make junction pairs
3. Fit linear decision boundary (Support vector machine: SVM)
4. Select best pairs
5. All pairs ‘vote’ to classify new samples
RNAsseq enables splice junction counting

Lines represent junction-spanning reads
ACS-CoR: Training a Pairwise SVM

- **5X Cross-validation:**
  - Randomly split your data into 5
  - Train on 4x, predict on 1x
  - Unbiased estimate of prediction performance

- Also Internal cross validation was used to select robust gene pairs
ACS –CoR: Housekeeping normalization

Reference junctions used for normalization of the signature

Expression relative to a panel of constitutively expressed housekeeping junctions = more robust to technical variability and facilitates translation to PCR

Expression = \log(Junction_{\text{normCPM}}) - \log(\text{mean}(\text{refJunction}_{\text{normCPM}}))
Splice junctions, genes and chromosomal locations that comprise the signature

51 Junctions
16 Genes
Visualizing the ACS-CoR signature

**Single Progressor vs Control**
Green lines indicate pair votes for ‘Control’, red for ‘Progressor’

**All Training samples**
Columns represent splice junctions
Rows represent mean expression +/- SD
Quantifying ACS-CoR performance

### Example Confusion matrix: ACS test set

<table>
<thead>
<tr>
<th>True Progressors</th>
<th>True Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Progressors</td>
<td>6 (TP)</td>
</tr>
<tr>
<td>Predicted Controls</td>
<td>9 (FN)</td>
</tr>
</tbody>
</table>

- **Sensitivity:** Progressors predicted to be progressors (TPR)
  - \( \frac{TP}{TP + FN} = \frac{6}{6 + 9} = 40\% \)

- **Specificity:** Controls predicted to be controls (TNR)
  - \( \frac{TN}{TN + FN} = \frac{70}{70 + 7} = 91\% \)
Quantifying ACS-CoR performance

- ACS-CoR provides a score, so have to pick a prediction threshold

- Tradeoff between sensitivity and specificity
  - High threshold, high specificity, low sensitivity
  - And vice-versa
Quantifying ACS-CoR performance

- ROC curves plot sensitivity vs specificity over all threshold values
- Shown: Cross validation ACS-CoR scores on the training set

- Area under the curve represents model performance on ACS test set at all thresholds
  - 1 = perfect classifier
  - 0.5 = random guessing
Translating ACS-CoR from RNAseq to qPCR

• Motivation:
  – qPCR is cheap and accurate, suitable for low cost screening test
  – Enabled validation on a new cohort with a locked-down signature without further expensive RNAseq
Quantifying ACS-CoR performance

- ACS-CoR showed very similar performance on the independent test set for RNAseq and qPCR.
Translating ACS-CoR from RNAseq to qPCR

- Splice junctions were translated directly to PCR primers (control and signature)
- All ACS cDNA was measured again via qPCR
- Individual SVM thresholds were rederived for each junction pair on the ACS training set
- GC6 validation set expression measured via PCR and ACS-CoR scores calculated
qPCR ACS-CoR validates on GC6 HHCs

- GC6-74 was a study of individuals after exposure to TB
- Like ACS, individuals enrolled were healthy, small number of Progressors
ACS-CoR was also translated to microarray

- ACS-CoR splice junctions pairs were collapsed into gene pairs
- All array probes mapping to these genes were identified and all possible probe pairs constructed and new SVM boundaries were fit for each resulting pair
  - Based on discriminating Berry et al, 2010 active TB vs latent TB samples
- Reparameterized array-based ACS-CoR used to classify existing microarray cohorts
ACS-CoR was also translated to microarray

- ACS-CoR discriminates
  - Active TB vs latent TB
  - Active TB vs other inflammatory/pulmonary diseases
  - In 5 existing cohorts
ACS-CoR Summary

- ACS-CoR signature predicted TB risk
- Visualizable and interpretable
- Validated in independent test sets
- Translated to qPCR and microarray
- These didn’t happen by accident!
Conclusions

• There is a large amount of freely available TB host transcriptional data

• TB Host transcriptional responses are shared across species

• Use this data to support your own hypotheses!
Acknowledgements

ULD mouse Infection
Courtney Plumlee
Kevin Urdahl

ACS & GC6 Signatures
Dan Zak
Ethan Thompson
ACS and GC6 study teams

Aitchison Lab
John Aitchison

RhCMV NHP data
Scott Hansen
Louis Picker