Statistical challenges and new data analysis methods for microbiome data integration

Kim-Anh Lê Cao

NHMRC Career Development Fellow
Centre for Systems Genomics
School of Mathematics and Statistics
The University of Melbourne
Microbiome is a complex organism

- Uncover the role of microbiome, as well as its composition
- Technological advances: culture-independent & NGS
- Not only about cataloging organisms, but biological functions that affect host and participate in disease processes
- Require advanced bioinformatics and computational statistics
We need a holistic view
We need cutting-edge statistical methods to study the microbiome as a whole

- Many organisms are co-dependent on each other within an ecosystem
  - limited insight when considered individually
- Microbiome has a major impact on human health and disease treatment outcomes
  - Understand microbiome - host interactions
We need cutting-edge statistical methods to study the microbiome as a whole

- Many organisms are co-dependent on each other within an ecosystem
  \[\Rightarrow\text{limited insight when considered individually}\]
- Microbiome has a major impact on human health and disease treatment outcomes
  \[\Rightarrow\text{Understand microbiome - host interactions}\]

**Multivariate statistical methods for microbiome data analysis:**

- Consider all bacterial taxa together
- Reduce data dimension and enable data visualisation (e.g. PCA, PCoA)
- Enable the integration of multiple sources of biological data
We need open-source and user-friendly tools

The **Omics** R toolkit (since 2009)

- French’Oz team:
  4 core, 1 developer, students, collaborators
- Today: 17 novel multivariate methods
- 21K downloads in 2016
- 14 multi-day workshops since 2014 (FR, AUS, NZ)

Our research program focuses on the development of multivariate statistical methodologies, their applications in areas informed by biology, and the training of the new generation of computational biologists.

www.mixOmics.org
Microbial count data

Example of data from 16S rRNA sequencing after taxonomic classification:

<table>
<thead>
<tr>
<th>Betaproteobacteria</th>
<th>Alphaproteobacteria</th>
<th>Actinobacteria</th>
<th>Clostridia</th>
<th>Bacteroidia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feces659</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>98</td>
</tr>
<tr>
<td>Feces309</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mouth599</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Mouth386</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Feces32</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>Plaque240</td>
<td>24</td>
<td>0</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>Plaque244</td>
<td>230</td>
<td>0</td>
<td>153</td>
<td>0</td>
</tr>
<tr>
<td>Plaque235</td>
<td>143</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Plaque245</td>
<td>128</td>
<td>0</td>
<td>102</td>
<td>0</td>
</tr>
<tr>
<td>Plaque246</td>
<td>42</td>
<td>0</td>
<td>7</td>
<td>0</td>
</tr>
</tbody>
</table>

→ we want to identify bacterial taxa that explain differences between sample groups (beta diversity analyses)
Data are sparse with skewed distribution

Many microbial organisms are absent from a large number of samples.
Data are sparse with skewed distribution

Many microbial organisms are absent from a large number of samples

\[ \Rightarrow \text{'zero-inflated' distribution} \]
Data with high variability

<table>
<thead>
<tr>
<th>Variability</th>
<th>Data challenges</th>
<th>Proposed solutions</th>
<th>Multivariate statistical methods</th>
<th>Conclusions</th>
</tr>
</thead>
</table>

Data with high variability need to be accounted for in the statistical analysis.
Variability also due to varying sequencing depth per sample
Variability also due to varying sequencing depth per sample

Pragmatic solution: transform count data into relative proportions

Relative proportion data are compositional

Relative proportion data ‘live’ in a constraint space (a ‘simplex’)

Image adapted from http://qedinsight.com
Compositional data: consequences

- Bacterial counts are relative to each other ➞ spurious correlation (Pearson in 1897!)

- Compositional data are bounded ➞ data distribution is not normal ➞ most statistical methods assume non-bounded data!
Compositional data: consequences

- Bacterial counts are relative to each other \(\leadsto\) spurious correlation (Pearson in 1897!)

- Compositional data are bounded
  \(\leadsto\) data distribution is not normal
  \(\leadsto\) most statistical methods assume non-bounded data!

‘History has proved him [Pearson] correct: over the succeeding years and indeed right up to the present day, there has been no other form of data analysis where more confusion has reigned and where more improper and inadequate statistical methods have been applied’ (Aitchison, 2011).

\(\leadsto\) potentially spurious statistical results!

Batch effects: systematic sources of variation that may outweigh biological variation

e.g. Sequencing runs (left), batches of expt (right), cage, country, differences in protocols or laboratories
Sparse data

Data are sparse and skewed

- Solution 1: apply or develop non parametric methods:
  → Univariate methods limited by small sample sizes
  → Multivariate methods (focus of this talk)

- Solution 2: develop statistical models to handle zero-inflated distributions (univariate)
  → Generalised Linear Model not appropriate
  → Zero Inflated Poisson, Zero Inflated Negative Binomial models and variants*

Review: Xu et al. (2015), Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data, *PLOS ONE* 10(7)
Data are compositional

- Pragmatic solution 1: log ratio transformation
  → Project data into a domain of real numbers (this talk)

- Solution 2: novel univariate statistical models to respect the intrinsic data characteristics
  → Constraint regression models*

Data contain batch effects (sometimes)

- Solution 1: **correct for** batch effect **prior** to statistical analysis
  - Batch Mean Centering, Combat* (originally developed for microarray data), **may also remove biological variation**!

Solution 2: **account for** batch effect **directly into** the statistical model
  - **univariate** methods: include batch information in the regression model
  - **multivariate** methods: preliminary results in this talk

*Johnson et al (2007), Adjusting batch effects in microarray expression data using empirical Bayes methods, *Biostatistics* 8(1)*
Our multivariate methods for microbiome data analysis

Methods implemented in mixOmics:

- **Identification of microbial signatures** to characterise a phenotype
- **Integration** with other high-throughput 'omics data for a better understanding of host - microbiome interactions
- **Longitudinal analyses** to understand disease progression


A multivariate method to account for batch effects

Eva Wang, Helen Benham, Ranjeny Thomas, UQDI

Classical discriminant analysis to
discriminate 28 healthy controls vs
86 rheumatoid arthritis patients

New method accounts for batch
effects and identifies a microbial
signature

RA study: Oral microbiome

unpublished preliminary results
A method to identify a multivariate microbial signature

A predictive statistical model based on microbial signature

Vanessa Lakis, Helen Benham, Ranjeny Thomas, UQDI

Method identifies a microbial signature:

- 94% accuracy to predict healthy controls vs rheumatoid arthritis
- First Degree Relatives predicted as likely-HC (29) or likely-RA (61)

RA study: Faecal microbiome

unpublished preliminary results
A predictive statistical model based on microbial signature

Vanessa Lakis, Helen Benham, Ranjeny Thomas, UQDI

Microbial signature discriminates HC vs RA

Prediction of FDR as a risk score from likely-HC to likely-RA

RA study: Faecal microbiome

unpublished preliminary results
A multivariate method to integrate ‘omics data

Amrit Singh (UBC), Florian Rohart (UQDI)

14 asthmatic individuals undergoing allergen inhalation challenge

Leukocyte gene expression and plasma metabolite abundance reduced to pathways using eigengene summarisation (Langfelder et al. 2008).

DIABLO seeks for subsets of ‘omics variables maximally correlated.


Kim-Anh Lê Cao
May 19-20 2017
207 Falk Foundation Gut Microbiome Symposium
**A multivariate method to integrate ‘omics data**

**Amrit Singh (UBC), Florian Rohart (UQDI)**

Our multi-‘omics signature suggests mechanistic link with response to allergen challenge across different biological layers

- **Cell types**: eosinophils and basophils (hallmarks of allergic asthma)

- **Gene pathway**: Valine, leucine and isoleucine biosynthesis (↑ post challenge)

- **Metabolite pathway**: Valine, leucine and isoleucine metabolism (↑ post challenge)

**Correlation btw cells, genes/metab pathways**

Integrate longitudinal microbial profiles with islet autoantibody markers

DIABIMMUNE T1D study: 33 genetically predisposed infants, faecal samples from 50 to 1,500 days to elucidate the role of microbial and host inflammatory activity in T1D (Kostic et al. 2015, *Cell host & microbe*).

E006574: seroconvert at 532Days, T1D at 1339Days

E022137: seroconvert at 562Days

Methods from:

unpublished preliminary results...
Integrate longitudinal microbial profiles with islet autoantibody markers

DIABIMMUNE T1D study: 33 genetically predisposed infants, faecal samples from 50 to 1,500 days to elucidate the role of microbial and host inflammatory activity in T1D (Kostic et al. 2015, *Cell host & microbe*)

![Time vs Intensity graph]

**Microbial signature:**
- **Blautia, Coprococcus and Streptococcus** positively correlated with ↑ insulin autoantibody IAA
- **Bacteroides**, members of Lachnospiraceae and Veillonellaceae, *Faecalibacterium (prausnitzii)* negatively correlated with IAA.

E006574: seroconvert at 532Days, T1D at 1339Days

Methods from:

**unpublished preliminary results**
Statistics in the microbiome era

- After some initial stalling, computational field is moving at a (too?) fast pace, yet no consensus (normalisation, analysis)
- Rigorous statistics needed to address the data challenges and ensure reliable results
- Flexibility of multivariate methods and novel developments:
  - data exploration; classification; integration of multiple data sets; biomarker identification
  - provide a deeper understanding of the microbiome as a biological system

www.mixOmics.org  @mixOmics_team

mixomics@math.univ-toulouse.fr  kimanh.lecao@unimelb.edu.au
Acknowledgements

mixOmics team

France: Sébastien Déjean, François Bartolo
Australia: K-A Lê Cao, Florian Rohart, Benoît Gautier
Many mixOmics users

Lê Cao past lab at UQ
Staff: Florian Rohart, Nicholas Matigian, Anne Bernard, Benoît Gautier
PhD students: Ralph Patrick, Chao Liu, Amrit Singh, Jasmin Straube, Aimee Hanson
Masters/Hons: Eva Wang, Vanessa Lakis, Zoe Welham, Nick d’Arcy, Thom Cuddihy

My dear microbiome collaborators

Thomas lab: Helen Benham, Linda Rehaume, Ranjeny Thomas (UQ)
Brown lab: Mary-Ellen Costello, Matt Brown (QUT)
Hugenholtz lab (UQ)

We're Hiring two senior postdocs in computational biostatistics, University of Melbourne

Kim-Anh Lê Cao
207 Falk Foundation Gut Microbiome Symposium