micRobiome analysis

Kirsty Lee Garson
Katie Lennard
The microbiome is an essential part of human physiology

- more than half of the genes which constitute the human body are microbial

- this microbial gene pool is diverse and highly dynamic:
  - sensitive to changes in diet and behaviour
  - regular exchange of genes between microbes occurs
  - mutations occur more frequently than in the human genome
The microbiome is an essential part of human physiology

- the human microbiome plays a role in numerous physiological functions, including immunity and nutrition

- alterations in the composition of the human microbiome have been linked to a wide variety of diseases
Microbiome research is developing rapidly

- this is as result of recent advances in sequencing technology (next generation sequencing)

- there has been a corresponding surge in the development of analysis tools, many of which are implemented in R
Aim One: Derive sequence data from clinical samples

1. Collect samples
2. Extract DNA
3. Amplify and sequence a marker gene
4. Group sequences by 97% similarity
5. Assign taxonomy using reference databases
6. Align sequences and create a phylogeny

Sequence data:
- Group 1:  GATA[ACAGATGCAT
  GTATA[ACAGATGCAT
  GGATA[ACAGATGCAT
  GATCA[ACAGATGCAT
- Group 2:  ATAGA[ACAGATGCAT
  ATAGTATA[ACAGATGCAT
- Group 3:  TATGATA[ACAGACAT
  TATGTATA[ACAGACAT
  TAGG[ACAGACAT

Taxonomy:
- Group 1: Bacteria_Firmicutes...

Aim One: Derive sequence data from clinical samples

Kingdom                      Phylum                      Class                      Order                      Family                      Genus                      Species
"Bacteria"                  "Firmicutes"                "Bacilli"                   "Lactobacillales"            "Lactobacillaceae"           "Lactobacillus"             "inera"
Aim Two: Analyse the microbial composition of samples

What is the composition of these bacterial communities?
How does it differ between healthy individuals and those with disease?

(marker gene analysis)
2.1 Measure diversity within samples

\[ S_{\text{Chao1}} = S_{\text{obs}} + \frac{n_i^2}{2n_i^2} \]
2.2 Measure diversity across samples
2.2 Measure diversity across samples
2.2 Measure diversity across samples
2.3 Assess sample composition
2.4 Compare sample composition across groups

Before Circumcision

After Circumcision

Genus
- 1-68
- Anaerococcus
- Campylobacter
- Clostridium
- Corynebacterium
- Dialister
- Finegoldia
- Mobiluncus
- Peptococcus
- Peptoniphilus
- pH2
- Porphyromonas
- Prevotella
- Staphylococcus
- WAL_1855D
2.5 Compare the number of times microbes of interest occur across groups
Aim Three: Gain insight into the role these microbes play

Which genes are present?
What is the microbial community capable of?
*functional prediction, shotgun metagenomics*

Which genes are being expressed?
Which capabilities are currently in use?
*transcriptomics, proteomics*

Which metabolic processes are underway?
How are these capabilities being used to influence the human-microbial environment?
*metabolomics*
Cbio 16S analysis pipeline

Katie Lennard
Microbiome analysis workflow

Preprocess raw reads: (merge, filter, cluster OTUs, assign taxonomy)

Data preprocessing (UCT High Performance Cluster)

Import OTU table (and phylogenetic tree) into R (phyloseq package, import_biom)

Integrate microbiome and sample metadata -> phyloseq object

Import sample metadata (.txt file)

Exploratory analysis (barplots, heatmap, PCoA etc.) using customised functions (phyloseq, NMF)

Downstream statistical analyses: Differential abundance testing (metagenomeSeq); correlations analyses (corrplot, psych); unsupervised classification (fanny; pam; mclust); biomarker discovery (randomForest)

Functional characterisation: infer bacterial gene content from 16S data (PICRUSt, STAMP) – requires closed reference OTUs (map de novo IDs to Greengenes)
Microbiome analysis workflow

**Preprocess raw reads:** (merge, filter, cluster OTUs, assign taxonomy)

**Import OTU table (and phylogenetic tree) into R** *(phyloseq package, import_biom)*

**Import sample metadata (.txt file)**

**Import data into R**

**Integrate microbiome and sample metadata --> phyloseq object**

**Exploratory analysis** (barplots, heatmap, PCoA etc.) using customised functions *(phyloseq, NMF)*

**Downstream statistical analyses:**
- **Differential abundance testing** *(metagenomeSeq)*
- **correlations analyses** *(corrplot, psych)*
- **unsupervised classification** *(fanny; pam; mcclus)*
- **biomarker discovery** *(randomForest)*

**Functional characterisation:** infer bacterial gene content from 16S data *(PICRUSt, STAMP)* – requires closed reference OTUs (map de novo IDs to Greengenes)

```r
str(M)
```

Formal class 'phyloseq' [package "phyloseq"] with 5 slots
  .@ otu_table: Formal class 'otu_table' [package "phyloseq"]
  . . . . . .- attr(*, "dimnames")=List of 2
  . . . . . . . : chr [1:510] "OTU_7" "OTU_294" "OTU_214" "OTU_295"
  . . . . . . . : chr [1:182] "SW16" "SW17" "SW18" "SW23" ..
  .@ taxa_are_rows: logi TRUE
  .@ tax_table: Formal class 'taxonomyTable' [package "phylos"
  . . . . .@ .Data : chr [1:510, 1:7] "Bacteria" "Bacteria" "Bc"
  . . . . . .- attr(*, "dimnames")=List of 2
  . . . . . . . : chr [1:510] "OTU_7" "OTU_294" "OTU_214" ..
  . . . . . . . : chr [1:7] "Kingdom" "Phylum" "Class" "Order"
  .@ sam_data : 'data.frame': 182 obs. of 148 variables:
  . . . . .@ .Data : List of 148
  . . . . . . .: int [1:182] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  . . . . . . .: Factor w/ 2 levels "0",...: 1 1 1 1 1 1 1 1
  . . . . . . .: Factor w/ 2 levels "0",...: 1 1 1 1 1...
Microbiome analysis workflow

Preprocess raw reads: (merge, filter, cluster OTUs, assign taxonomy)

Import OTU table and phylogenetic tree into R (phyloseq package, import_biom)

Integrate microbiome and sample metadata -> phyloseq object

Exploratory analysis (barplots, heatmap, PCoA etc.) using customised functions (phyloseq, NMF)

Exploratory

Downstream statistical analyses: Differential abundance testing (metagenomeSeq); correlations analyses (corrlplot, psych); unsupervised classification (fanny; pam; mclust); biomarker discovery (randomForest)

Functional characterisation: infer bacterial gene content from 16S data (PICRUSt, STAMP) – requires closed reference OTUs (map de novo IDs to Greengenes)

Genus-level abundance by BV all samples

Summary barplots

Genus
- Clostridium
- Dialister
- Finegoldia
- Lactobacillus
- Megasphaera
- Prevotella
- Shuttleworthia
- Sneathia

Percentage of Sequences
Microbiome analysis workflow

Preprocess raw reads: (merge, filter, cluster OTUs, assign taxonomy)

OTU table (.biom format) & phylogenetic tree

Import OTU table (and phylogenetic tree) into R (phylolseq package, import_biom)

Import sample metadata (.txt file)

Integrate microbiome and sample metadata -> phyloseq object

Exploratory analysis (barplots, heatmap, PCoA etc.) using customised functions (phylolseq, NMF)

Downstream statistical analyses: Differential abundance testing (metagenomeSeq); correlated analyses (corrplot, psych); unsupervised classification (fanny; pam; mclust); biomarker discovery (randomForest)

Beta diversity: NMDS/PCoA

NMDS of 16S microbiome, Bray-Curtis distance, k=2

- Inflammation
  - High
  - Low
  - Med

- BV
  - 0
  - 1
  - Intermediate
Microbiome analysis workflow

Preprocess raw reads: (merge, filter, cluster OTUs, assign taxonomy)

OTU table (.biom format) & phylogenetic tree

Import OTU table (and phylogenetic tree) into R (phyloseq package, import_biom)

Import sample metadata (.txt file)

Integrate microbiome and sample metadata --> phyloseq object

Exploratory analysis (barplots, heatmap, PCoA etc.) using customised functions (phyloseq, NMF)

Functional characterisation: infer bacterial gene content from 16S data (PICRUSt, STAMP) – requires closed reference OTUs (map de novo IDs to Greengenes)

Downstream statistical analyses: Differential abundance testing (metagenomeSeq); correlations analyses (corplot, psych); unsupervised classification (fanny; pam; mclust); biomarker discovery (randomForest)

Annotated heatmaps
Microbiome analysis workflow

Preprocess raw reads: (merge, filter, cluster OTUs, assign taxonomy)

OTU table (.biom format) & phylogenetic tree

Differential abundance testing

Import OTU table (and metadata if available)

<table>
<thead>
<tr>
<th>OTU</th>
<th>percent_positive_group</th>
<th>+samples in group</th>
<th>-samples in group</th>
<th>mean_positive_group</th>
<th>mean_positive_group</th>
<th>oddsRatio</th>
<th>lower</th>
<th>upper</th>
<th>fisherP</th>
<th>fisherAdjP</th>
<th>coeff</th>
<th>pvalues</th>
<th>adPvalues</th>
<th>Kingdom</th>
</tr>
</thead>
<tbody>
<tr>
<td>OTU_254</td>
<td>100</td>
<td>100</td>
<td>89</td>
<td>72</td>
<td>9139</td>
<td>2402</td>
<td>0</td>
<td>1</td>
<td>4.3814786</td>
<td>3.50E-08</td>
<td>1.18E-07</td>
<td>Bacteria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTU_59</td>
<td>97.8</td>
<td>97.8</td>
<td>87</td>
<td>67</td>
<td>4343</td>
<td>1414</td>
<td>3.2233077</td>
<td>0.50865097</td>
<td>34.8420058</td>
<td>2.4392345</td>
<td>3.9872103</td>
<td>4.3369</td>
<td>3.66E-08</td>
<td>1.23E-07</td>
</tr>
<tr>
<td>OTU_27</td>
<td>97.8</td>
<td>97.8</td>
<td>87</td>
<td>66</td>
<td>3220</td>
<td>968</td>
<td>3.91224426</td>
<td>0.6737754</td>
<td>40.9629286</td>
<td>1.4084517</td>
<td>0.2565461</td>
<td>-4.3022808</td>
<td>2.81E-08</td>
<td>9.67E-08</td>
</tr>
<tr>
<td>OTU_314</td>
<td>19.1</td>
<td>4.2</td>
<td>17</td>
<td>3</td>
<td>67</td>
<td>46</td>
<td>5.38194296</td>
<td>1.46306825</td>
<td>29.93353159</td>
<td>0.00398918</td>
<td>0.016676011</td>
<td>4.2805204</td>
<td>5.79E-07</td>
<td>1.96E-07</td>
</tr>
<tr>
<td>OTU_417</td>
<td>87.6</td>
<td>81.9</td>
<td>78</td>
<td>59</td>
<td>1494</td>
<td>639</td>
<td>1.5790049</td>
<td>0.96930478</td>
<td>4.14219894</td>
<td>0.37566612</td>
<td>0.54121822</td>
<td>-3.997862</td>
<td>5.41E-08</td>
<td>1.75E-07</td>
</tr>
<tr>
<td>OTU_9</td>
<td>93.3</td>
<td>90.3</td>
<td>83</td>
<td>65</td>
<td>1266</td>
<td>456</td>
<td>1.45599802</td>
<td>0.40568486</td>
<td>5.61223037</td>
<td>0.56702208</td>
<td>0.714024529</td>
<td>-3.8579988</td>
<td>6.55E-08</td>
<td>2.07E-07</td>
</tr>
<tr>
<td>OTU_148</td>
<td>82</td>
<td>69.9</td>
<td>73</td>
<td>48</td>
<td>1064</td>
<td>119</td>
<td>2.53413901</td>
<td>1.18007381</td>
<td>5.71388377</td>
<td>0.01190991</td>
<td>0.03578463</td>
<td>-3.255635</td>
<td>1.22E-07</td>
<td>3.51E-07</td>
</tr>
<tr>
<td>OTU_7</td>
<td>76.4</td>
<td>47.7</td>
<td>68</td>
<td>34</td>
<td>389</td>
<td>219</td>
<td>3.58801858</td>
<td>1.73818871</td>
<td>7.51348729</td>
<td>0.00015486</td>
<td>0.000114286</td>
<td>-3.0135061</td>
<td>1.98E-08</td>
<td>4.97E-08</td>
</tr>
<tr>
<td>OTU_255</td>
<td>84.3</td>
<td>72.2</td>
<td>75</td>
<td>52</td>
<td>1549</td>
<td>141</td>
<td>2.05102697</td>
<td>0.89323826</td>
<td>4.82550329</td>
<td>0.08049993</td>
<td>0.156890011</td>
<td>-2.9487617</td>
<td>1.12E-05</td>
<td>2.56E-05</td>
</tr>
<tr>
<td>OTU_658</td>
<td>93.3</td>
<td>87.5</td>
<td>83</td>
<td>63</td>
<td>420</td>
<td>146</td>
<td>1.96779079</td>
<td>0.59001023</td>
<td>7.08991045</td>
<td>0.27742493</td>
<td>0.43940622</td>
<td>-2.8897097</td>
<td>1.89E-09</td>
<td>8.35E-09</td>
</tr>
<tr>
<td>OTU_178</td>
<td>100</td>
<td>100</td>
<td>89</td>
<td>72</td>
<td>10637</td>
<td>3587</td>
<td>0</td>
<td>1</td>
<td>2.8715645</td>
<td>4.26E-08</td>
<td>1.04E-05</td>
<td>Bacteria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTU_16</td>
<td>100</td>
<td>100</td>
<td>89</td>
<td>72</td>
<td>7149</td>
<td>2672</td>
<td>0</td>
<td>1</td>
<td>2.6568928</td>
<td>2.35E-05</td>
<td>5.19E-05</td>
<td>Bacteria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTU_307</td>
<td>100</td>
<td>98.6</td>
<td>89</td>
<td>71</td>
<td>4700</td>
<td>1956</td>
<td>Inf</td>
<td>0.003169487</td>
<td>0.44720497</td>
<td>0.61808817</td>
<td>-2.5817595</td>
<td>4.50E-05</td>
<td>9.62E-05</td>
<td>Bacteria</td>
</tr>
</tbody>
</table>

Downstream statistical analyses: Differential abundance testing (metagenomeSeq); correlations analyses (corrplot, psych); unsupervised classification (fanny; pam; mclust); biomarker discovery (randomForest)

Functional characterisation: infer bacterial gene content from 16S data (PICRUSt, STAMP) – requires closed reference OTUs (map de novo IDs to Greengenes)

Downstream analyses
Microbiome analysis workflow

Preprocess raw reads: (merge, filter, cluster OTUs, assign taxonomy)

OTU table (biom format) & phylogenetic tree

Import OTU table (and phylogenetic tree) into R (phyloseq package, import_biom)

Import sample metadata (.txt file)

Integrate microbiome and sample metadata --> phyloseq object

Exploratory analysis (barplots, heatmap, PCoA etc.) using customised functions (phyloseq, NMF)

Downstream statistical analyses: Differential abundance testing (metagenomeSeq); correlations analyses (corrplot, psych); unsupervised classification (fanny; pam; mclust); biomarker discovery (randomForest)

Functional characterisation: infer bacterial gene content from 16S data (PICRUSt, STAMP) – requires closed reference OTUs (map de novo IDs to Greengenes)

Downstream analyses
Microbiome analysis workflow

Preprocess raw reads: (merge, filter, cluster OTUs, assign taxonomy)

Import OTU table (and phylogenetic tree) into R (phyloseq package, import_biom)

Integrate microbiome and sample metadata --> phyloseq object

OTU table (.biom format) & phylogenetic tree

Import sample metadata (.txt file)

Exploratory analysis (barplots, heatmap, PCoA etc.) using customised functions (phyloseq, NMF)

Downstream analyses: Differential abundance testing (metagenomeSeq); correlations analyses (corrplot, psych); unsupervised classification (fanny; pam; mclust); biomarker discovery (randomForest)

Functional characterisation: infer bacterial gene content from 16S data (PICRUSt, STAMP) – requires closed reference OTUs (map de novo IDs to Greengenes)

Downstream analyses

unsupervised classification

NMDS of PICRUSt metagenome, k=3 fuzzy clusters

Fuzzy classes

- A
- B
- C
- no.cluster

Nugent score

0.0  2.5  5.0  7.5  10.0

NMDS1

NMDS2
Microbiome analysis workflow

Preprocess raw reads: (merge, filter, cluster OTUs, assign taxonomy)

Import OTU table (and phylogenetic tree) into R (phyloseq package, import_biom)

Integrate microbiome and sample metadata -> phyloseq object

OTU table (.biom format) & phylogenetic tree

Import sample metadata (.txt file)

Exploratory analysis (barplots, heatmap, PCoA etc.) using customised functions (phyloseq, NMF)

Biomarker discovery: random forests

Downstream analyses

ROC Curve taxa Predictors

AUC = 0.96886
PPV = 0.90909
NPV = 0.96

Downstream statistical analyses: Differential abundance testing (metagenomeSeq); correlations analyses (corrplot, psych); unsupervised classification (fanny; pam; mclust); biomarker discovery (randomForest)

Functional characterisation: infer bacterial gene content from 16S data (PICRUSt, STAMP) – requires closed reference OTUs (map de novo IDs to Greengenes)