Meta²genomics

CNB/CSIC
Summary

- The need for meta-metagenomics
- The micro-bee
- Accuracy of metagenomics
- When is enough enough?
- Speeding up the process
- Comparing studies
Preamble

- Due to time constraints this is only an overview
- All the major points have been addressed
- Only some illustrative data will be provided
- A full description of all this work is being submitted for publication
The need for meta-metagenomics
Common trends

- There is a need to identify common trends across metagenomic studies
  - Economy
    - Do not repeat studies
  - Practical
    - Full reproducibility is rarely achievable (if ever)
Example: Maize rhizosphere

- We conducted studies at different locations, over different yearly cultivation cycles.
- Each study considered different conditions
  - Different times
  - Different location
  - Different maize cultivars
  - Different treatments
- Goal: identify cumulative effect of herbicides.
  - Each study led naturally to the next analysis
A bit of history

- Started with cultivable bacteria
- Moved to metagenomics using 16S-V6 (short read lengths)
- Test normal maize
- Test cotton
- Test herbicide resistant maize
- Test and compare additional herbicides
- Test herbicide combinations...
- Each step must build on previous experience
Scientific limitations

- One cannot justify a new experiment before finishing the previous ones.
- But then it must be done next year (with different climate).
- If cumulative effects are expected, then it must also be done on a new, virgin soil.
- As years and locations change, so do environmental conditions.
The trivial approach

- A possible solution
  - Repeat the experiment (e.g. include previous treatments) in all subsequent instances
  - Replicate the experiment on different soils at the same time
  - Replicate the experiment at different times

- Problems
  - Must use the same technology
  - Must repeat work already done
  - Must waste a lot of money
The not-so-trivial approach

- Try to reuse as much information as possible
  - Some experiments will need to be repeated in all cases (e.g. control)
  - Consider the possible impact of experimental conditions
    - Time
    - Location
    - Methods
    - Treatment
    - Etc...
  - Analyze heterogeneous data
The micro-bee
Bees

- Produce honey
- Pollinate plants
  - 60-80% of the world flowering plants and 35% of crop production depend on animal pollination
- Are terribly sensitive to pollution
  - Air pollution
  - Light pollution
  - Cell-phone radiation
  - Pesticide misuse
  - Global warming
The micro-bee

- **Framework:**
  - CBRN P35 EU-Africa cooperation project.
- **Goal:**
  - find an easy way to identify soil/water contamination
- **Question:**
  - is there a microbe species (or higher taxa) that can identify contamination?
- **Premises:**
  - Previous meta-genomic studies show that some phylogenetic groups tend to be consistently affected
The trivial approach

- Conduct experiments on as many locations as possible
- Repeat several years (to correct for climate changes)
- Test as many contaminants as possible
- Impoverish your funding agency
The not-so-trivial approach

- Collect as many previous studies as possible
- Compare them
- Identify a species -or taxonomic group- that is consistently affected by aggressive treatments
- Develop a simple test for changes in the micro-bee population.
Data sources

- Heterogeneous data from different experiments and authors
  - Pesticide treatments
  - Grassland soils
  - Maize cultures
  - Cotton cultures
  - Etc...

- Retrieved from SRA

- Original analyses must be replicated
  - At least to the extent required by our goal
Measuring accuracy
The problem

- Taxonomy assignment is based on similarity
  - Different species differ in ~3%
  - 97% similarity → same species
- Knowledge limits
  - Not all bacterial sequences are known
- Practical limits
  - Some species are known to be indistinguishable by some methods
- how many species can we identify?
Measuring accuracy

- Cluster all sequences known at 97% similarity
  - Clusters gives the maximum number of groups that can be unequivocally identified
  - Singleton clusters give the maximum number of species that can be identified
- Must be checked for each method
  - Reference sequence
  - Clustering/identification method (blast, uclust, RDP, Rtax, etc...)
  - Etc...
Similarity classification

- VAMPS 16S rRNA hyper-variable regions 97% (subset)

<table>
<thead>
<tr>
<th>Region</th>
<th>N seqs</th>
<th>Avg. Len.</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>V3</td>
<td>118982</td>
<td>76</td>
<td>34951</td>
</tr>
<tr>
<td>V3V5</td>
<td>203487</td>
<td>362</td>
<td>34700</td>
</tr>
<tr>
<td>SSU</td>
<td>401607</td>
<td>900</td>
<td>24276</td>
</tr>
</tbody>
</table>

NOTES:
SSU includes non-hyper-variable regions
More sequences or more length do not imply greater power
What if I do not use similarity?

Blast 97% LCA

RDP

RTax
Do with less—so they'll have enough!

RATIONING GIVES YOU YOUR FAIR SHARE

When is enough enough?
Identifying genetic biodiversity

- Saturating OTUS requires ~400,000 reads
- Saturating CHAO1/ACE requires ~40,000
- We need to know the shape of the distribution
Adjusting curves

- Most current methods use a standard curve (e.g. lognormal log mean=1, log sd=1)
- Does this reflect reality?

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Log mean</th>
<th>Log SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>FMG1 (Nacke et al.)</td>
<td>1.08</td>
<td>1.15</td>
</tr>
<tr>
<td>UPG1</td>
<td>1.34</td>
<td>0.78</td>
</tr>
<tr>
<td>UPG3</td>
<td>0.94</td>
<td>1.18</td>
</tr>
<tr>
<td>PriestPot (Quince et al.)</td>
<td>0.93</td>
<td>1.39</td>
</tr>
<tr>
<td>r143_s2 (Huse et al.)</td>
<td>1.411</td>
<td>1.94</td>
</tr>
<tr>
<td>Zaragoza Avg (Valverde et al)</td>
<td>1.77</td>
<td>1.71</td>
</tr>
<tr>
<td>ZC1</td>
<td>1.30</td>
<td>1.31</td>
</tr>
<tr>
<td>ZC2</td>
<td>1.85</td>
<td>1.61</td>
</tr>
<tr>
<td>ZG1</td>
<td>2.14</td>
<td>1.36</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Speeding up
Test and compare alternatives

- Taxonomical classification
  - BLAT / BOWTIE
  - Similarity algorithms
  - RDP
  - Rtax

- Select appropriate sample size
  - Compare with saturated studies
    - Illumina
  - Consider curve fitting: rely on preliminary studies
  - Allow for experimental error
Comparing experiments
The problem

- Taxonomical comparisons are hard
  - Huge amounts of categorical data
  - Many non-shared groups
  - Various hierarchical levels
- We need a systematic approach to compare taxonomic hierarchies
  - How similar are two populations?
  - Are cladistic differences significant?
TaxFrac

- A novel approach to taxonomic comparison using full-knowledge
  - Consider all cladistic levels
  - Define a comparison metric
  - Define a statistical validation method

- Answer the question
  - “how similar are two populations?”
Item-level validation

- Two basic questions:
  - How similar are two populations?
  - Are differences significant?

- Road blocks:
  - How variable are specific sub-populations?
  - Dealing with undetectable sub-populations?

- Approaches
  - Subsampling (good for a single experiment)
  - Compare many studies (required for cross-experimental comparison)
  - Ignore method-specific discrepancies
So, what?

- The more data we collect the better

- Metagenomics is still young

- Probably any conclusion we make now will need to be reviewed in the future

- But we can start to consider it right now.
Thanks

- To all of you
- To the organizers
- To our sponsors
  - EU COST: SEQAHEAD
  - CYTED: FreeBIT
  - EU CBRN: P35
  - CSIC, Spanish Government

jrvalverde@cnb.csic.es