Introducing Microbiome Bioinformatics

Part 3.

John Walshaw, GHFS, IFR

Recap: Aims

- Overview of types of **microbiome analysis**
  - with particular regard to **sequence informatics concepts**
- "Top down" – putting analysis tools and resources in context
- No highly detailed technicalities
  - No instructions on how to run particular programs
- Why you are using the bioinformatics approaches you use; pros, cons; alternatives

Series of talks

- At least 3 sessions to cover what I would like
- Beyond that – if there is demand –
  - can progress to more technical talks
  - especially about 16S analysis (probably)
  - increasingly metagenomics in GHFS research
- Informal and flexible
  - Please interrupt and ask questions
  - Suggestions for topics for further focus

Series of talks

- Part 1: 27/1/2017
  - "Biological and Experimental Stuff that a microbiome bioinformatician needs to know"
  - Overview of marker gene sequencing for community analysis
- Part 2: 10/2/2017
  - Overview of whole-metagenome sequencing
- Slideshows
  - http://ghfs1.ifr.ac.uk/ghfs/
  - (see posts of the above dates)
- Part 3: 24/2/2017
  - Focus on metatranscriptomics

Topics, top-down

Introduction

Types of environments studied

Variety of microbiomes, and domains of Life

Aims of studies:

"Communities"

Functions

Aims of DNA/ RNA sequencing approaches

"barcoding"- use of marker genes

"shotgun" metagenomics

metatranscriptomics

16S rRNA

Other markers

aims of whole-metagenome sequencing

Dealing with the data: Informatics concepts

16S rRNA

Communities in detail

"shotgun" meta-'omics

Later: more technical details

Metatranscriptomics informatics

There's more than one way to do it

Why Metatranscriptomics?

- Sample and sequence the RNA
  - To determine what is actually being expressed
- There may be metatranscriptome differences between subjects/disease states etc which have similar metagenomes
  - This has been demonstrated in some studies including in the human gut
- Discovery of new genes, thus far missed by metagenomics
  - How likely this is, depends on the microbiome
  - Amply demonstrated in some older ocean studies
  - ~ 90% of inferred ORFs *Gilbert et al.*(2008) *PLoS ONE* **3** (8) e3042
  - Less likely for the human gut prokaryote community
  - *What about the gut virome? Gut eukaryotes?....Discuss.*

Metatranscriptomics – the basics

- Might be done on its own, or applied alongside metagenomics
- Substantial quantities of RNA may be present
  - helps to inform metagenomics
- But skewed: >80% of total RNA is **rRNA**
- Around 15% is **tRNA**
- Usually no more than 5% is **mRNA** ; may be **considerably less**
- *e.g.* Westermann *et al.* (2012) *Nat. Rev. Microbiol.* **10** 618-30.
- Whether this matters depends on the aims
  - E.g. may need to enrich for mRNA
- As in normal transcriptomics, **mRNA** is itself skewed
  - Implications for sampling depth

Metatranscriptomics

- Comparison of two or more samples/environments:
  - 1. Biodiversity (*taxa*)
  - 2. Giant RNA-seq-type experiment (*genes*)
  - usually requires mRNA-enrichment
  - which is usually done experimentally
  - nowadays, best methods **remove 95-99% of rRNA**
    - E.g. Pérez-Pantoja & Tamames (2015) Prokaryotic Metatranscriptomics in *Hydrocarbon and Lipid Microbiology* pp69-98*,* Springer
    - Doing the sums implies that **> 15% of the total remaining RNA could be rRNA**
  - with *in silico* post-filtering to remove non-mRNA sequences which remain

Metatranscriptomics and databases

- Identifying (1) taxa and (2) genes
  - As with metagenomics, both aims rely on databases and reference sequences
  - To identify both genes and organisms of origin
- As with shotgun metagenomics:
  - genomic sequence databases
  - smaller, marker-gene databases
  - function-centric databases/protein sequence databases
  - possibly assemblies created from metagenomics
- Identified genes may be associated with **pathways**
  - E.g. KEGG pathways database

Remember, **functions** tend to be more conserved than **phylotypes**

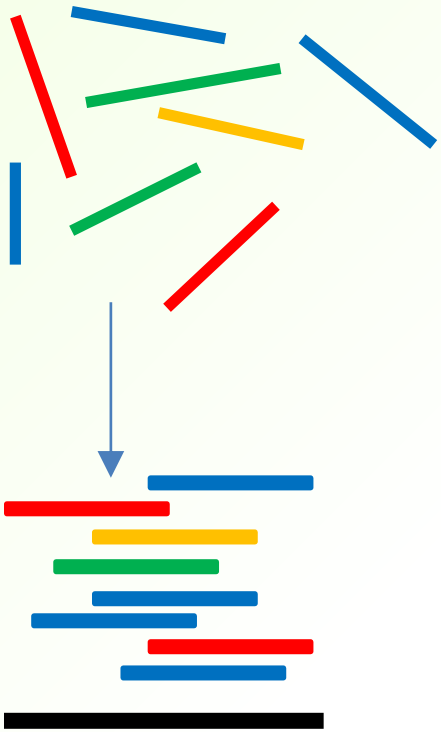
Metatranscriptomics: amplification

- As with metagenomics, amplification may be necessary
  - In metatranscriptomics, often to enrich for mRNA
- By various methods
- RNA linear amplification
  - First step, polyadenylate the RNA; → 1-stranded cDNA → 2-stranded
  - Get (cDNA sequences of) transcript **and** reverse complement, indistinguishably
- MDA (multiple displacement amplification; see part 2)
  - May be biased in favour of low GC-content genomes, but is partly dependent on protocol
  - (also produces cDNA of course)
- Strand-specific methods: uses dUTP markers to distinguish the first cDNA strand from the second
  - sequence-database matching methods perform matches equally capable with forward- or reverse-complement sequences
  - but knowing which is the forward strand potentially aids in resolving whether some poorer matches are 'real' or not
  - But this is also another source of bias

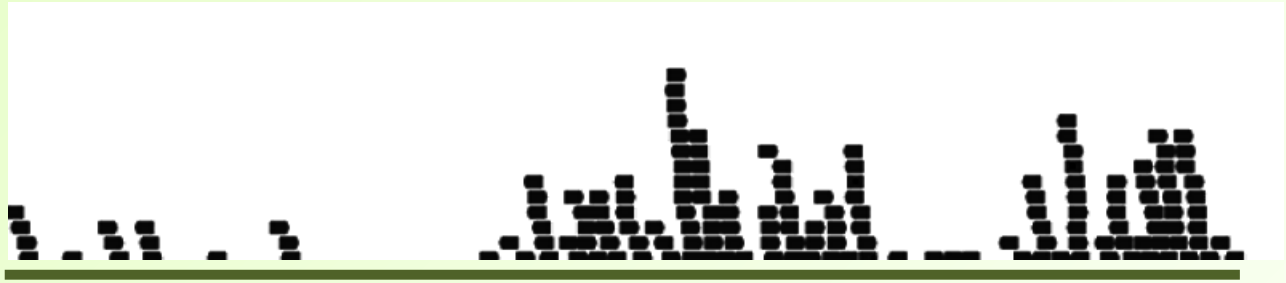
- Strand-specific cDNA synthesis by DUTP marking
  - "this procedure has been known **to introduce 1-2% *Escherichia coli* genomic DNA into the final cDNA library** (a result of *E. coli*-derived DNA polymerase I and ligase being used in the cDNA generation steps).Including versus excluding *E. coli* sequences in downstream bioinformatic analyses **did not affect the conclusions** of this work."
  - *(my emphases)*
  - Franzosa *et al.* (2014), Relating the metatranscriptome and metagenome of the human gut, *Proc. Natl. Acad. Sci. U. S. A.* **111** E2329-38

Metatranscriptomics : “assembly”?

- Rescuing of complete, single-species transcripts (“assembly”) may be very challenging
  - And unnecessary
- As with metagenomics, clustering of very similar sequences (in this case transcript fragments) would be the general case
- As with metagenomics, attempts at “assembly” (clustering) **may not be necessary**
  - Depends on the approach; **some are read-by-read**
  - But longer sequences improve database matching
  - Functional annotation rate increased up to 6-fold, depending on assembly length (Celaj *et al.* (2014) *Microbiome* **2** : 39
  - Collapsing multiple → single sequences: frequency issues
- Some very useful informatics methods are completely identical whether applied to metagenomics or metatranscriptomics



Metatranscriptomics : mapping reads?

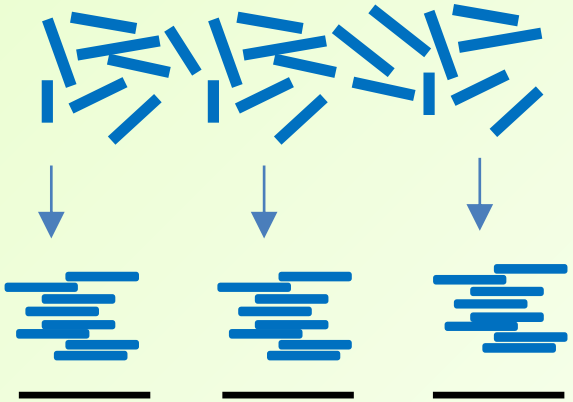


- Some fundamental differences compared to real RNA-seq:
  - The transcripts arise from (possibly very many) different genomes
  - You may not have reference genomes for all of these
  - Even if you do, it may not always be possible to determine exactly which one is 'correct'
    - **This may not matter**, depending on the aims
    - Especially for **functional identification**, "chimaeric" mappings may not matter
  - The more rigorous approaches to **quantification** in RNA-seq, involving a single known reference genome, cannot be necessarily be assumed to be appropriate for your metatranscriptomics data
    - Some genomes' transcripts may have been only very sparsely sampled

Metatranscriptomics and metagenomics in tandem

Metagenomic reads

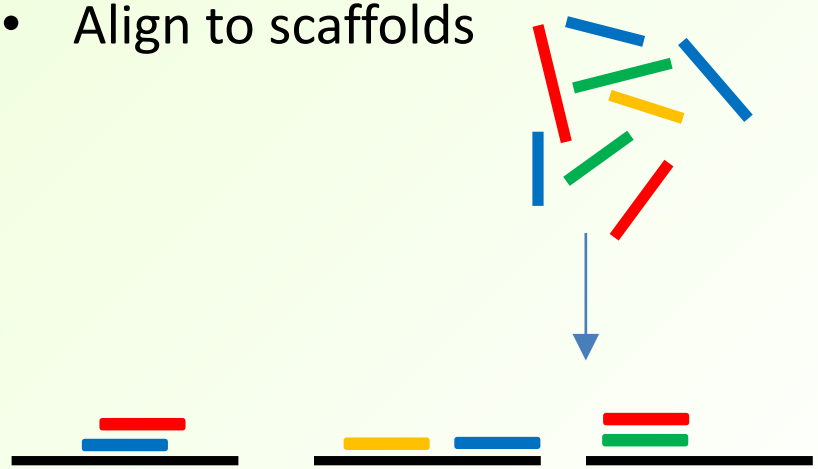
- Assemble reads into longer "scaffolds"
- Likely to be chimaeric



- Identify possible coding regions

Metatranscriptomic reads

- Align to scaffolds



- e.g. Durbán *et al.* (2013) Instability of the faecal microbiota in diarrhoea-predominant irritable bowel syndrome *FEMS Microbiol. Ecol.* **86** 581-9

Which approaches are actually used?

- Review of 27 metatranscriptomics studies published between 2013-2015
  - only 4 are human microbiome studies, of which 2 concern GIT
- More than half involved no assembly or mapping; 4 studies employed both
- Pérez-Pantoja & Tamames (2015) Prokaryotic Metatranscriptomics in *Hydrocarbon and Lipid Microbiology* pp69-98, Springer

Metatranscriptomics and 16S sequences

- If you sample/sequence "all" the metatranscriptome
  - you get mostly rRNA
  - cells make loads of ribosomes!
- **In principle, this is ideal for community analysis**
  - In a very similar manner to 16S amplicons
  - But assays "*who is transcribing*" more than "*who is there*"
- The (relatively small) amount of mRNA can be used simultaneously for functional studies
- **In practice**, a metatranscriptomics study is likely to target a particular aspect such as expression of protein-coding genes
  - So would be experimentally **enriched for mRNA**
  - Taxonomic/phylotypic identification (community analysis) might be done in a parallel sequencing experiment (**e.g. by 16S amplicons**)

Example: metatranscriptomics alone for community analysis

- Turner *et al.* (2009) – soil/rhizosphere environment
  - Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere microbiome of plants *ISME J.* **7** 2248-58
- RNA: mostly ribosomal
  - both small and large subunits
  - both prokaryote and eukaryote
  - used for community analysis
  - a small proportion was mRNA

for Fig 3a from this paper, please see original at http://www.nature.com/ismej/journal/v7/n12/fig_tab/ismej2013119f3.html#figure-title

Metatranscriptomics and some examples of the “compar-ome”

	Metagenomic reads	Metatranscriptomic reads	16S amplicon reads
Taxonomic abundances			
Gene (functional) abundances			
Variation between samples: taxa			
Variation between samples: functions			

Example: metatranscriptomics in tandem with 16S sequencing

- Poretsky *et al.* (2009) – ocean environment
  - Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre *Environ. Microbiol.* **11** (6) 1358-75
- RNA: applied two rounds of **mRNA-enrichment/rRNA-depletion** using different methods
  - 37% of remaining RNA was identified as rRNA, by comparison with RDP database
- 16S amplicons: very long – **used <u>Sanger</u> sequencing**
- Also performed *cell counts* for some organisms

Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre

[Figure 3 of Poretsky *et al.* (2009)
Environmental Microbiology
11:1358-1375
(see link below)]

Environmental Microbiology
Volume 11, Issue 6, pages 1358-1375, 4 FEB 2009 DOI: 10.1111/j.1462-2920.2008.01863.x
http://onlinelibrary.wiley.com/doi/10.1111/j.1462-2920.2008.01863.x/full#f3

Example: metatranscriptomics in tandem with metagenomics

- *Franzosa et al.* (2014) Relating the metatranscriptome and metagenome of the human gut *Proc. Natl. Acad. Sci. U. S. A.* **111** E2329-38
- Three informatics aspects to focus on here:
  1. Compare transcript abundance to abundance of their corresponding genes
  2. Compare variation between subjects of transcript abundances to gene abundances
  3. Granularity of mapped pathways

Example: metatranscriptomics in tandem with metagenomics

1. Compare transcript abundance to abundance of their corresponding genes [*Franzosa et al.* (2014)]

– About **40%** of the transcripts with a small or no fold-change c.f. gene abundance

– About **20% have a fold-change of > 10** (up or down)

– Functional families (and taxonomic groups) can be associated with these

  - E.g. most of the most strongly 'overexpressed' genes encode ribosomal proteins

Functional diversity at the transcriptional level suggests a pattern of subject-specific metagenome regulation.

[Fig. 5 from Franzosa *et al.* (2014) : see below for reference
http://www.pnas.org/content/111/22/E2329]

Eric A. Franzosa et al. PNAS 2014;111:E2329-E2338

3. Granularity of mapped pathways [*Franzosa et al.* (2014)]

- One pathway with ‘overexpressed’ genes: TCA cycle
- But only one part of it
- (high-level aerobic metabolism unlikely)
- http://www.genome.jp/kegg/pathway/map/map00020.html

Example: metatranscriptomics in tandem with metagenomics (ii)

- A further consideration:

4. How well conserved is apparent *function* between samples
   - Compared to conservation of taxonomic groups?

- Example:
  - Durbán *et al.* (2013) Instability of the faecal microbiota in diarrhoea-predominant irritable bowel syndrome *FEMS Microbiol. Ecol.* **86** 581-9

[for Fig 2 of
Durbán *et al.* (2013),
Refer to URL below]

[for Fig 3 of
Durbán *et al.* (2013),
Refer to URL below]

From Durbán *et al.* (2013)
FEMS Microbiol. Ecol. **86** 581-9
http://dx.doi.org/10.1111/1574-6941.12184

What sequence-based meta-'omics do

'Omics	Community analysis? (who is in there?)	Functional analysis? (what are they doing?)	Assembly of whole or partial genomes ?
16S/18S amplicon sequencing Targeted amplicons, usually **segments of**: • 16S rRNA genes (prokaryotes) • 18S rRNA or genes ITS (eukaryotes esp. fungi)	yes	No (not directly)	no
Shotgun Metagenomics	yes	yes	yes (to some extent)
Metatranscriptomics (community RNA-Seq)	yes	yes	no

Topics, top-down

Introduction

Types of environments studied

Variety of microbiomes, and domains of Life

Aims of studies:

"Communities"

Functions

Aims of DNA/ RNA sequencing approaches

"barcoding"- use of marker genes

"shotgun" metagenomics

metatranscriptomics

16S rRNA

Other markers

aims of whole-metagenome sequencing

Dealing with the data: Informatics concepts

16S rRNA

Communities in detail

"shotgun" meta-'omics

Later: more technical details