The novel microbial proteome associated with methane production in cattle as defined by metagenomic sequencing

Mick Watson
Edinburgh Genomics &
The Roslin Institute
University of Edinburgh

Edinburgh Genomics
- Genomics facility based at the University of Edinburgh
- Available for collaborations on an academic, non-profit basis
- Formed from merger of
  - ARK-Genomics
  - The GenePool
- Funded by three major bio UK research councils
- A range of technologies and expertise available

http://genomics.ed.ac.uk

FOOD SECURITY

Challenges in food security
- The World’s food system doesn’t work:
  - 1.5bn overweight, 500m obese
  - 925m experience hunger, +1bn “hidden hunger”
- Moving forward, there are a number of key pressures:

Over the next 50 years, the world’s farmers and ranchers will be called upon to produce more food than has been produced in the past 10,000 years combined, and to do so in environmentally sustainable ways.

Jacques Diouf, FAO Director General, 2007

1. WHO [http://www.who.int/mediacentre/factsheets/fs311/en/]
2. Hunger report "Nutrition and agriculture challenges for human nutrition sustainability"
In order to survive, humans are changing our planet - literally

Three gorges dam
- The three gorges dam in China is so large we have changed the Earth’s spin and shape
- 410 miles in length and 0.70 miles in width (average)
- Surface area of the reservoir is 1045 km²
- Contains about 9.43 cubic miles of water.
- That water will weigh more than 42 billion tons
- A shift of such a mass predicted to increase the length of day by 0.06 microseconds
- Makes the Earth slightly more round in the middle and flat on the top.
- Shifted the pole position by about two centimetres (0.8 inch)

Should we be worried?
- Yes and no! Humans have been improving food for many centuries
- Note: no information for beef cattle and sheep

Feeding our expanding species is a major problem
The Roslin Institute: Aims

- Enhance animal health and welfare through knowledge of animal genetics
- Enhance sustainability and productivity of livestock systems
- Enhance food safety by understanding host-pathogen interactions
- Enhance human health through an understanding of disease
- Identify new and emerging zoonoses
- Enhance quality of life for farmed animals

Highlights


RUMEN GUT METAGENOMICS

Prevailing theory of the individual

- An individual consists of at least 10x as many bacterial cells as host cells
- Each individual is a “supra-organism”
  - a composite of host and microbial cells contribute the functions necessary for the individual to survive
- The genetic landscape of any individual is a composite of the host genome and the genomes of the millions of microbial symbionts that live on and within that individual
- It is clearly important to take a holistic view when examining any animal phenotype

Focus

- Move from discovery science to applied science
- “What’s there?” → “What can we do with it?”

METAGENOMIC SEQUENCING OF RUMEN GUTS

Why are we studying it?

- Energy from food
  “Our results indicate that the obese microbiome has an increased capacity to harvest energy from the diet. Furthermore, this trait is transmissible: colonization of germ-free mice with an ‘obese microbiota’ results in a significantly greater increase in total body fat than colonization with a ‘lean microbiota.’”
- Novel enzyme discovery
  “An initial assembly of the metagenomic sequence resulted in 179,092 scaffolds... Only 47 (0.03%) of the assembled scaffolds showed high levels of similarity to previously sequenced genomes available in GenBank. These results suggest that the vast majority of the assembled scaffolds represent segments of hitherto uncharacterized microbial genomes.”
- Methane Emissions
  Globally, ruminant livestock produce about 80 million metric tons of methane annually, accounting for about 20% of global methane emissions from human-related activities.
METHANE PRODUCTION IN CATTLE

Methane production

- Methane is a natural product of anaerobic microbial fermentation
  - Rumen is anaerobic
- Methane is a greenhouse gas (GHG) with a global warming potential 25-fold that of carbon dioxide (IPCC 2006).
- Ruminants are the major producers of methane emissions from anthropogenic activities,
  - accounting for 37% of total GHG from agriculture in the UK
- Methane emissions from cattle are entirely microbial in origin

Related to abundance?

- Hypothesis: Methanogenic activity is proportional to the abundance of methanogenic archaea
  - Many have tried and failed
  - No relationship between metagenomic abundance
  - Clear relationship between metatranscriptomic abundance

Our data set

- Steers chosen from longitudinal study
- Chose high and low methane emitters matched for breed and diet
- Submitted for metagenomic sequencing
- Approx. 11Gb per sample

Relationship to archaeal abundance

- Mapped metagenomic reads to Greengenes database
- Recorded all hits in database that are as good as best hit
- Calculated lowest common taxon (in this case, Kingdom)
- Matched for breed and diet, high methane correlates with high archaeal abundance
- qPCR confirms this

Relationship to enzyme abundance

- Mapped metagenomic reads to KEGG
- Matched for breed and diet, the abundance of several enzymes is associated with methane production
Relationship to enzyme abundance
• Mapped metagenomic reads to KEGG
• Matched for breed and diet, the abundance of several enzymes is associated with methane production

Methane pathway
• Fig on left is from Shi et al Genome Research 24(9):1517-25
• Fig on right is same enzymes in our data set, matched for breed and diet

What’s in there?
• Assembled all 8 metagenomes with MetaVelvet
  – Predicted genes with Prokka
  – Annotated using Pfam domains
• 1.5 million gene/protein predictions
• Less than half have any known domain
• From 44 KEGG enzymes
  – 7021 representatives in our data
  – 5942 unique protein sequences
• Only 29 have exact match in NR
• Only 60 are 100% conserved
• At 90% identity, 807 / 5942 have hit

Rumen metagenomics
• It is possible to assemble contigs from deep sequencing of rumen microbiomes
• It is possible to extract novel proteins/enzymes and predict domains/functions
• About half of the predicted proteins have a domain...
  ... but this includes domains of unknown function
• The vast majority of the genomic landscape is novel – most contigs don’t hit anything known
• In animals matched for breed and diet, methane emissions correlate with (i) abundance of archaea and (ii) abundance of enzymes, many of which are involved in methane production

Metagenomic Assembly

<table>
<thead>
<tr>
<th>Run</th>
<th>Read Pairs</th>
<th>Read Length</th>
<th>GigaBases</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRR094437</td>
<td>217,692,239</td>
<td>2x101</td>
<td>87.9</td>
</tr>
<tr>
<td>SRR094926</td>
<td>248,205,313</td>
<td>2x101</td>
<td>100.3</td>
</tr>
<tr>
<td>SRR094166</td>
<td>4,882,637</td>
<td>2x125</td>
<td>2.4</td>
</tr>
<tr>
<td>SRR094403</td>
<td>4,766,838</td>
<td>2x125</td>
<td>2.4</td>
</tr>
<tr>
<td>SRR094405</td>
<td>4,805,576</td>
<td>2x125</td>
<td>2.4</td>
</tr>
<tr>
<td>SRR094415</td>
<td>5,667,078</td>
<td>2x125</td>
<td>2.8</td>
</tr>
<tr>
<td>SRR094416</td>
<td>4,776,805</td>
<td>2x125</td>
<td>2.4</td>
</tr>
<tr>
<td>SRR094417</td>
<td>4,960,626</td>
<td>2x125</td>
<td>2.5</td>
</tr>
<tr>
<td>SRR094418</td>
<td>4,715,030</td>
<td>2x125</td>
<td>2.4</td>
</tr>
</tbody>
</table>
Notes:
MetaVelvet partitions graph on coverage and connectivity
IDBA-UD designed for uneven coverage (replaces Meta-IDBA)
MetAMOS is a pipeline, can use SOAPdenovo or Meta-IDBA (we used the latter)
RayMeta is a scalable, distributed assembler
SOAPdenovo included as memory efficient and has been used historically

Immediate problems
• SGE cluster, 24 cores, 512Gb RAM
• Focused on “deep” dataset
  – 931M reads, 188Gbases of sequence
• Immediate problem!
  – Every single assembler exceeded the 512Gb RAM
  – Some used up 1024Gb+ (i.e. 512Gb in swap)
• Oops
  – How to annoy an entire institute……
• Focused on smaller dataset
  – 34M reads, 17.1Gbases of sequence
  – Trimmed to Q30 before assembly
  – Used K of 51 for consistency

How much assembled?
• What happens when we map (using BWA) the reads back to the assembly?

Scaffolds using pairs
• We didn’t use 3Kb or 5Kb jumping libraries
• Used internal scaffolder, or SSPACE, or both

Complete genes
Repeat regions in metagenomes are orthologous genes!
• Good news for IDBA-UD, bad news for metagenomics
• Hess et al (using Velvet) reported 55% complete genes
Run time

<table>
<thead>
<tr>
<th>Assembler</th>
<th>Total time</th>
<th>Individual components</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROOPManor</td>
<td>29 mins</td>
<td>velvatrol (9h, 25m) + velvatrol (9h, 25m)</td>
</tr>
<tr>
<td>MetaRoof</td>
<td>30 mins</td>
<td>velvatrol (9h, 25m) + velvatrol (9h, 25m)</td>
</tr>
<tr>
<td>RAPManor</td>
<td>30 mins</td>
<td>velvatrol (9h, 25m) + velvatrol (9h, 25m)</td>
</tr>
<tr>
<td>IDBA-UD</td>
<td>35 mins</td>
<td>velvatrol (9h, 25m) + velvatrol (9h, 25m)</td>
</tr>
</tbody>
</table>

Conclusions

- Metagenomic assembly remains an unsolved problem
- Data are large and complex and need RAM++
- Assemblies fail in the middle of genes
  - Exactly where you don’t want them to
- Assemblers vary in their resource requirements and results
- Our data suggest IDBA-UD is the leading metagenomic assembler (on this data set)

Acknowledgements

**Funders:** BBSRC, Roslin Foundation, TSB

**People:** Christelle Robert, Pablo Fuentes-Utrilla, Richard Talbot, Edinburgh Genomics, David Hume, John Wallace, Rainer Roehe, Rodrigo Bacigalupe

**Edinburgh Genomics:** [http://genomics.ed.ac.uk](http://genomics.ed.ac.uk)

**The Roslin Institute:** [http://www.roslin.ed.ac.uk](http://www.roslin.ed.ac.uk)

Follow me:

Twitter: @BioMickWatson
Blog: biomickwatson.wordpress.com