Genomes from metagenomes: recovery and analysis of population genomes

Kate Ormerod
Australian Centre for Ecogenomics, School of Chemistry & Molecular Biosciences
Genomes from metagenomes

- Represent an consensus version of a particular species present within a dataset
Why assemble?

- We can pull functional and phylogenetic information from a metagenomic sample without any assembly – so why assemble?
  - Expanding the tree of life reference genomes
  - Gives context to the functional data – how much variation exists within a given family?
  - Functional assignment based on full length genes
  - Provides a basis for strain level comparisons
Generating population genomes

- **Experimental design**
  - Multiple samples
    - Different sites/biological entities
    - Time series
    - Different sample treatment

- **Initial de novo assembly**
  - Combined samples

- **Binning**
  - Based on coverage in individual samples and sequence composition
Experimental design

• How diverse is the environment you are sampling? How rare are the genomes you seek?
  – Assess using:
    • Test batch of shotgun samples
    • 16S rRNA profiling

• How easy is it to obtain samples?

• Is there likely to be host contamination?

• Depth can be achieved across multiple samples
Assembly & binning

• Assembly
  – Software
    • SPAdes
    • MetaVelvet
    • CLC Genomics Workbench
  – Read QC
    • eg Trimmomatic
  – Combined samples
    • All at once?
    • Subset?
  – Gap filling
    • eg ABySS
Assembly & binning

• Binning
  – **Software**
    • MetaBAT
    • CONCOCT
    • Canopy
    • GroopM
  – **Coverage/co-abundance**
    • Coverage within a single sample
    • Differential coverage
      – Coverage across different samples
  – **Sequence composition**
    • Tetranucleotide frequency patterns
      – Beware small contigs
Problems with population genomes

• **Completeness**
  – Insufficient coverage
  – Regions with different profile from the rest of the genome eg duplications, lateral transfer, transposable elements

• **Contamination**
  – Misassembly – chimeric contigs – lateral transfer, transposable elements
Assessing genome quality

- **CheckM**
  - Measures completeness and contamination
  - Lineage specific marker gene assessment

- **Read mapping**
  - Detection of structural variation, low/high coverage, different insert sizes

- **Reference genome coverage**

---

**CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes**

Donovan H. Parks, ¹ Michael Imelfort, ¹ Connor T. Skennerton, ¹ Philip Hugenholtz, ¹,² and Gene W. Tyson ¹,³

¹Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St. Lucia, QLD 4072, Queensland, Australia; ²Institute for Molecular Bioscience, The University of Queensland, St. Lucia, QLD 4072, Queensland, Australia; ³Advanced Water Management Centre, The University of Queensland, St. Lucia, QLD 4072, Queensland, Australia

Large-scale recovery of genomes from isolates, single cells, and metagenomic data has been made possible by advances in computational methods and substantial reductions in sequencing costs. Although this increasing breadth of draft genomes is providing key information regarding the evolutionary and functional diversity of microbial life, it has become impractical to finish all available reference genomes. Making robust biological inferences from draft genomes requires accurate estimates of their completeness and contamination. Current methods for assessing genome quality are ad hoc and generally make use of a limited number of marker genes conserved across all bacterial or archaeal genomes. Here we introduce CheckM, an automated method for assessing the quality of a genome using a broader set of marker genes specific to the position of a genome within a reference genome tree and information about the colocalization of these genes. We demonstrate the effectiveness of CheckM using synthetic data and a wide range of isolate, single-cell, and metagenome-derived genomes. CheckM is shown to provide accurate estimates of genome completeness and contamination and to outperform existing approaches. Using CheckM, we identify a diverse range of errors currently impacting publicly available isolate genomes and demonstrate that genomes obtained from single cells and metagenomic data vary substantially in quality. In order to facilitate the use of draft genomes, we propose an objective measure of genome quality that can be used to select genomes suitable for specific gene- and genome-centric analyses of microbial communities.

---

**Table 3. Controlled vocabulary of draft genome quality based on estimated genome completeness and contamination**

<table>
<thead>
<tr>
<th>Completeness</th>
<th>Classification</th>
<th>Contamination</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥90%</td>
<td>Near</td>
<td>≤5%</td>
<td>Low ³</td>
</tr>
<tr>
<td>≥70% to 90%</td>
<td>Substantial</td>
<td>5% to ≤10%</td>
<td>Medium</td>
</tr>
<tr>
<td>≥50% to 70%</td>
<td>Moderate</td>
<td>10% to ≤15%</td>
<td>High</td>
</tr>
<tr>
<td>&lt;50%</td>
<td>Partial</td>
<td>&gt;15%</td>
<td>Very high</td>
</tr>
</tbody>
</table>

³ Genomes estimated to have 0% contamination can be designated as having “no detectable contamination.”
Improving population genomes

- Reassembly of individual genome bins
  - Extract reads mapping to individual bins (BamM)

- Reference guided assembly

- Different binning tool and/or assembly tool
metaQUAST
Analysis of population genomes

• Major advantage = culture independent
  – Less biased view of the diversity of the chosen environment

• Major challenge = culture independent
  – Therefore functional information may not be available
  – Describing new genus, family, phylum?
Finding the story in your genomes

• Inferring core metabolism
  – Energy
  – Food

• Environment specific points of interest
  – Antibiotic resistance
  – Virulence factors

• Comparisons to phylogenetic and environmental neighbours
Analysis of population genomes: an example

Genome annotation

- Annotation
  - Prokka
  - NCBI
- Annotation + analysis
  - RAST
  - IMG
  - KBase
<table>
<thead>
<tr>
<th>Function</th>
<th>Subsystem</th>
<th>Primary class</th>
<th>Secondary class</th>
<th>Totals</th>
<th>Families</th>
<th>Family genes</th>
<th>Family genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta-lactamase (EC 3.5.2.6)</td>
<td>Beta-lactamase</td>
<td>Virulence</td>
<td>Resistance to antibiotics and toxic compounds</td>
<td>Families: 2</td>
<td>Genes: 4</td>
<td>Genomes: 4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>H3.CDS.4062</td>
<td>M2.CDS.1759</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metal-dependent hydrolases of the beta-lactamase superfamily 1</td>
<td>Beta-lactamase</td>
<td>Virulence</td>
<td>Resistance to antibiotics and toxic compounds</td>
<td>Families: 1</td>
<td>Genes: 2</td>
<td>Genomes: 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>H7.CDS.3292</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zn-dependent hydrolase (beta-lactamase superfamily)</td>
<td>Hypothetical</td>
<td>Experimental</td>
<td></td>
<td>Families: 1</td>
<td>Genes: 1</td>
<td>Genomes: 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DUF454</td>
<td>Subsystems</td>
<td></td>
<td>H3.CDS.4657</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beta-lactamase (Cephalosporinase) (EC 3.5.2.6)</td>
<td>Beta-lactamase</td>
<td>Virulence</td>
<td>Resistance to antibiotics and toxic compounds</td>
<td>Families: 1</td>
<td>Genes: 1</td>
<td>Genomes: 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>M2.CDS.3334</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metallo-beta-lactamase superfamily domain protein in prophage</td>
<td>Hypothetical</td>
<td>Experimental</td>
<td></td>
<td>Families: 1</td>
<td>Genes: 1</td>
<td>Genomes: 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DUF454</td>
<td>Subsystems</td>
<td></td>
<td>M2.CDS.3735</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Showing 1 to 5 of 5 entries (filtered from 1,998 total entries)
Functional categorisation

• Databases assign function to protein annotations using homology
  – CAZy: carbohydrate active enzymes
  – KEGG
  – COG
  – EggNOG

• Differentiation between:
  – Family members
  – Phylogenetic neighbours
  – Environmental neighbours
CAZy: carbohydrate active enzymes
KEGG & COG: intrafamily comparisons
KEGG & COG: interfamily comparisons
CAZy: niche companion comparison
Database bias
From annotations to pathways
From annotations to pathways

- **KEGG**
  - 490 reference pathways

- **BioCyc**
  - MetaCyc
    - 2,453 pathways
    - 2,063 organisms
  - Pathway Tools

- **RAST**
  - ModelSEED

- **IMG**

- **Kbase**
Pathway Tools

Pathway: xylose degradation I

**Pathway Summary from MetaCyc:**
- Xylose, which can serve as a total source of carbon and energy for *Escherichia coli* K-12 substract, MG1655, enters the cell either through a low-affinity, proton-motive force-driven or a high-affinity, ATP-driven (ABC) transport system, so it is not phosphorylated during entry. Once inside the cell, an isomerase converts it to D-xyulose and subsequently a kinase converts it to D-xyulose 5-phosphate, an intermediate of the pentose phosphate pathway. Hence it flows through the pathways of central metabolism to satisfy the cell’s need preceptor precursor metabolites, reducing power, and metabolic energy.

**Pathway Evidence Glyph:**
- Key to pathway glyph edge colors:
  - Red: Enzyme is present in this organism
  - Green: Enzyme catalyzing this reaction is present in this organism

**Promoters:**
- Mayer, C. and W. Boes. Hexose/Pentose and Hexitol/Pentitol Metabolism. EcoSal Module 3.4.1 [ECOSAL]

**Credits:**
- Created in MetaCyc 10-Sep-1994 by Riley M. Marine Biological Laboratory
- Revised in MetaCyc 16-Jun-2006 by Ingraham JL, UC Davis
- Updated in MetaCyc 09-Dec-2015 by Please Edit This
## Cross-Species Comparison: xylose degradation

Note: In addition to reflecting differences in biology of different organisms, these statistics will reflect differences in the levels of curation, data availability, and completeness of the PGDBs for these organisms.

This table compares a single pathway across the requested organism databases. The Evidence Glyph is a cartoon of the pathway diagram showing which steps have identified enzymes, which are pathway holes, and which steps are unique to this pathway (a color key is included at the bottom of this page). For each reaction step, we list the identified enzymes and genes. A pathway may not be present in an organism database even if enzymes have been identified for one or more of its reactions, and we indicate if this is the case. Finally, for organism databases in which operons exist (i.e. prokaryotes) and have been defined, we include diagrams showing all the operons that contain the genes for this pathway. Genes that are actually involved in the pathway are colored dark purple, whereas other genes in the same operon which are not involved in the pathway are shaded a lighter color. Clicking on an organism name will take you to the display page for the pathway in that organism database.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Evidence Glyph</th>
<th>Enzymes and Genes for xylose degradation</th>
<th>Operons</th>
</tr>
</thead>
<tbody>
<tr>
<td>B. intestinohominis YIT 11860</td>
<td>![Evidence Glyph]</td>
<td><em>This pathway is not marked as present in this organism.</em> No enzymes or genes have been identified for this pathway</td>
<td></td>
</tr>
<tr>
<td>B. viscercola DSM 18177</td>
<td>![Evidence Glyph]</td>
<td><em>This pathway is not marked as present in this organism.</em> No enzymes or genes have been identified for this pathway</td>
<td></td>
</tr>
<tr>
<td>C. fastidiosus NSB1</td>
<td>![Evidence Glyph]</td>
<td>EC 5.3.1.5 <em>xylose isomerase; NSB1T_RS21625</em> EC 2.7.1.1<em>None</em></td>
<td></td>
</tr>
<tr>
<td>E. coli K-12 substr. MG1655</td>
<td>![Evidence Glyph]</td>
<td>EC 5.3.1.5 <em>xylose isomerase; xylA</em> EC 2.7.1.1<em>Xyulokinase; xylB</em></td>
<td></td>
</tr>
<tr>
<td>H. M2</td>
<td>![Evidence Glyph]</td>
<td><em>This pathway is not marked as present in this organism.</em> No enzymes or genes have been identified for this pathway</td>
<td></td>
</tr>
<tr>
<td>H. GP1</td>
<td>![Evidence Glyph]</td>
<td>EC 5.3.1.5 <em>Xylose isomerase; xylA_1</em> EC 2.7.1.1<em>Xyulokinase; xylB_2</em> Xyulokinase; xylB_1</td>
<td></td>
</tr>
<tr>
<td>H. GP2</td>
<td>![Evidence Glyph]</td>
<td><em>This pathway is not marked as present in this organism.</em> No enzymes or genes have been identified for this pathway</td>
<td></td>
</tr>
<tr>
<td>H. GP3</td>
<td>![Evidence Glyph]</td>
<td>EC 5.3.1.5 <em>Xylose isomerase; xylA_3</em> EC 2.7.1.1<em>Xyulokinase; xylB</em></td>
<td></td>
</tr>
<tr>
<td>H. GP4</td>
<td>![Evidence Glyph]</td>
<td>EC 5.3.1.5 <em>None</em> EC 2.7.1.1<em>Xyulokinase; xylB</em></td>
<td></td>
</tr>
<tr>
<td>H. H1</td>
<td>![Evidence Glyph]</td>
<td><em>This pathway is not marked as present in this organism.</em> No enzymes or genes have been identified for this pathway</td>
<td></td>
</tr>
<tr>
<td>H. H10</td>
<td>![Evidence Glyph]</td>
<td><em>This pathway is not marked as present in this organism.</em> No enzymes or genes have been identified for this pathway</td>
<td></td>
</tr>
</tbody>
</table>
Pathway Tools
Metabolic gap filling

• Are there missing enzymes within core pathways?
  – Are these truly missing or an assembly artefact?
  – Is there something missing suggestive of reliance on other members of the community?

• How does your prediction compare to other, possibly cultured, species from similar environment?
  – Are there known culturing requirements for related species?
Metabolic overview

Vitamin production: thiamine, riboflavin, niacin, pantothenate, biotin, folate

Oxidative stress: superoxide dismutase, catalase, peroxiredoxins, thioredoxins, ferritin

Amino acid synthesis:
- Asp, Gin, Glu, Gly, Met, Val, Iso, Leu
- Arg, Asn, Cys, His, Ser, Pro, Thr, Phe, Tyr
- Ala, Lys, Tyr
Summary

Plan
- Test your environment before committing
- Depth can be achieved across multiple samples

Assemble
- Combine all samples or subsets?

Bin
- More samples means improved differential coverage binning
- Try multiple binning programs

Annotate
- Databases may not capture everything - make use of multiple databases

Infer
- Core metabolism can be inferred from reference pathways, phylogenetic neighbours and environmental neighbours

Useful references


For the utility of different extraction methods:
Acknowledgements

ACE
Phil Hugenholtz
Gene Tyson
David Wood
Nancy Lachner
Joshua Daly
Nicola Angel
Serene Low

TRI, University of Queensland
Mark Morrison

University of Newcastle
Phil Hansbro
Shaan Gellatly

IMB, University of Queensland
Matt Cooper

AIBN, University of Queensland
Lars Nielsen
Cristiana Dal’Molin
Robin Palfreyman

QFAB
Jeremy Parsons