MetaBAT: An Efficient Tool for Accurately Reconstructing Single Genomes from Complex Microbial Communities

Don Kang
Metagenomics at Gene or Pathway Levels

20,000 New Cellulase Genes

Expression of Methanogenesis Pathways

Wasn’t me!
But, Genomes would be Better!

• Able to get a full picture of metabolic capacity of an individual member of the community
• Study genome dynamics of individual members
  – Genome-wide sweep, gene gain/loss analysis
• Understanding inter-species interaction
How can we construct single genomes from metagenomic data?
Genome Reconstruction from Metagenomic Data

Metagenome → Shotgun Sequencing → Reads → Initial Assembly → Contigs → Metagenome
Genome Reconstruction from Metagenomic Data

Metagenome → Shotgun Sequencing → Reads → Initial Assembly → Contigs → “Binning”
Existing Binning Methods

• **Reference Based Binning**
  – Phylogeny based

• **De novo Binning**
  – Sequence composition
  – Abundance
  – Both

  ➢ Inaccurate for complex metagenomes
  ➢ Manual
  ➢ Not scalable for many samples
Co-Abundance (coverage covariance) Binning
Abundance (Coverage) Binning

Metagenome → Shotgun Sequencing → Reads → Initial Assembly → Contigs

Ideally, contigs from the same genome should have the same coverage.

But, single abundance cannot differentiate multiple genomes of similar abundance?
Co-abundance Binning

Multiple samples (libraries) help to differentiate the similar abundance in single sample (library).
Design Goals for Binning Software

• Automated Unsupervised Co-abundance Binning
  – Integration of tetranucleotide frequency (TNF) and (or) abundance (ABD) as features
  – Handling of multiple ABDs from samples
• Highly Efficient
  – A couple of hours to bin millions of contigs having thousands of samples
  –Runnable in a single node (<20G memory)
• Reproducible and Reliable
  – Robust to noise in contigs or samples
  – Designed to have high specificity than sensitivity
• Flexible
  – Handle any number of samples
  – Adjustable parameter setting to change sensitivity and specificity
• Simple
  – Easy to run and fully automated
Run MetaBAT!

runMetaBat.sh assembly.fasta *.bam
Benchmarks of Automated Metagenome Binners With A Medium Sized Data Set

- 5 binning methods
- 264 human gut metagenomic samples (ERP000108)
  - Assembled into 200K contigs
  - Used a method (CheckM) to estimate completeness and precision based on single copy genes
The Contestants

- **MetaBAT**
  - Sequence composition (TNF) + Co-abundance
- **CONCOCT**
  - Sequence composition + Co-abundance
- **GroopM**
  - Sequence composition + Co-abundance
  - Optional manual steps
- **MaxBin**
  - Sequence composition + Abundance
- **Canopy**
  - General purpose clustering algorithm
  - Co-abundance only
MetaBAT found the most genomes

For details, refer to https://bitbucket.org/berkeleylab/metabat/wiki/Benchmark_MetaHIT
MetaBAT runs very efficiently

<table>
<thead>
<tr>
<th></th>
<th>MetaBAT</th>
<th>Canopy</th>
<th>CONCOCT</th>
<th>MaxBin</th>
<th>GroopM**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Bins Identified (&gt;200kb)</td>
<td>234</td>
<td>223</td>
<td>260</td>
<td>168</td>
<td>335</td>
</tr>
<tr>
<td>Number of Genomes Detected (Precision &gt; .9 &amp; Recall &gt; .3)</td>
<td>130</td>
<td>96</td>
<td>64</td>
<td>39</td>
<td>28</td>
</tr>
<tr>
<td>Wall Time (16 cores; 32 hyper-threads)</td>
<td>00:03:36</td>
<td>00:02:31*</td>
<td>82:19:53</td>
<td>06:49:39</td>
<td>12:19:12</td>
</tr>
<tr>
<td>Peak Memory Usage (for binning step)</td>
<td>3.0G</td>
<td>1.6G*</td>
<td>7G</td>
<td>5.8G</td>
<td>6.3G</td>
</tr>
</tbody>
</table>

*Canopy only use abundance table as input, so it should have taken more time and memory to read and write sequence data like the others

**Manual steps were not used

For details, refer to [https://bitbucket.org/berkeleylab/metabat/wiki/Benchmark_MetaHIT](https://bitbucket.org/berkeleylab/metabat/wiki/Benchmark_MetaHIT)
Binners complement each other

For details, refer to [https://bitbucket.org/berkeleylab/metabat/wiki/Benchmark_MetaHIT](https://bitbucket.org/berkeleylab/metabat/wiki/Benchmark_MetaHIT)

130/144 (90%)
Can MetaBAT Scale to Huge Data Set?

- 1704 human gut metagenomic samples (ERP002061)
- >1M contigs over 1kb
- Only MetaBAT and Canopy was able to handle the amount of data
- 3 hours in a single node (with 32 threads using 17G memory)
- MetaBAT produced 790 (out of 1634) genome bins with >30% completeness and <5% contamination
- Using genome bins as seeds, we recruited & reassembled reads to improve the quality of bins.
The Quality of Genome Bins Approximates High Quality Draft Genomes

MetaBAT  MGS
+ Reassembly  Draft Genomes

342/373 (92%)
Acknowledgement

Zhong Wang  Rob Egan  Jeff Froula