Sample Size Does Matter: Scaling Up Analysis in Galaxy with Metagenomics

Daniel Blankenberg
dan@bx.psu.edu
Post Doctoral Research Associate
Galaxy Team / Penn State
Metagenomics Powerhouse

Sarah J. Craig, PhD
sarah@bx.psu.edu

Center for Medical Genomics, Penn State University, University Park
Hershey Medical School, Pennsylvania, USA
Department of Biology, Penn State University, University Park, Pennsylvania

~500 Samples: Buccal & Stool from Mother-Child Pairs
Overview

- Brief Introduction to Metagenomics
- Doing It
- Challenges, Solutions, and Future Work
Overview

- Brief Introduction to Metagenomics
- Doing It
- Challenges, Solutions, and Future Work
Metagenomics

- Study of genetic material recovered directly from environmental samples
  - isolation and lab cultivation not required
- High-throughput sequencing
  - 16S rRNA targeted
  - Whole-genome shotgun

Metagenomics

Clinical or environmental sample containing microbes

Extraction of genomic DNA

What microbes are present? or What are the microbes doing?

Amplicon metagenomic sequencing (16S rRNA gene: bacteria)

Amplification and sequencing of heterogeneous mixture of 16S rRNA genes

Bioinformatics analysis
- Alignment
- Classification
- Operational taxonomic unit (OTU) analysis
- Phylogenetic analysis

Whole-genome shotgun metagenomic sequencing

Fragmenting of DNA, construction of paired-end libraries, sequencing of heterogeneous mixture of DNA

Bioinformatics analysis
- Functional assignment of reads
- Identification of enriched metabolic pathways and gene functions
- Comparative metagenomics

16S rRNA


Overview

• Brief Introduction to Metagenomics
• **Doing It**
• Challenges, Solutions, and Future Work
Doing It

- QC and Preparation
- Classify Reads
- Normalize
- alpha diversity
- beta diversity
- Downstream Analysis and Visualization
Doing It

- QC and Preparation
- Classify Reads
- Normalize
- alpha diversity
- beta diversity
- Downstream Analysis and Visualization
Importing FTP uploaded files into Galaxy

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at 54.211.185.247 using your Galaxy credentials (email address and password).

Available files:

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Created</th>
</tr>
</thead>
<tbody>
<tr>
<td>F3D0_S188_L001_R1_001.fastq</td>
<td>4.2 MB</td>
<td>05/25/2016 11:01:21 PM</td>
</tr>
<tr>
<td>F3D0_S188_L001_R2_001.fastq</td>
<td>4.2 MB</td>
<td>05/25/2016 11:01:22 PM</td>
</tr>
<tr>
<td>F3D0_S141_S207_L001_R1_001.fastq</td>
<td>3.2 MB</td>
<td>05/25/2016 11:01:23 PM</td>
</tr>
<tr>
<td>F3D0_S141_S207_L001_R2_001.fastq</td>
<td>3.2 MB</td>
<td>05/25/2016 11:01:24 PM</td>
</tr>
</tbody>
</table>

Choose FTP file

Start
Input: a List of paired sequencing Reads
Outputs: QC’d Joined paired reads, filtered for chimeras (FASTA)
Quality Reports (html and single summary table)
Create Workflow for Easy-to-Read FastQC Report

Use as a SubWorkflow
Treat an entire workflow as a single Tool Module
Single Summary FastQC Report
Doing It

- QC and Preparation
- **Classify Reads**
- Normalize
- alpha diversity
- beta diversity
- Downstream Analysis and Visualization
Classify Reads

• “Binning”
  • "best effort" to identify reads or contigs with certain groups of organisms designated as OTUs
• Operational Taxonomic Units (OTUs)
  • operational definition of a species or group of species
• Algorithms
  • Taxonomy Dependent
    • alignment to known sequences and species
    • misses reads that are absent in database
    • obtain estimates of the profile/abundance of ‘known’ taxonomic groups
  • Taxonomy Independent
    • group/bin reads in a given dataset based on their mutual similarity
    • considers content of reads only
    • no database
Kraken Classifier

- assign taxonomic labels to short DNA sequences
- exact alignment of k-mers

Input: a List of sequencing Reads (FASTA or FASTQ)
Outputs: Summary Table of counts (filtered or non-filtered joined)
## Abundance Counts

<table>
<thead>
<tr>
<th>Analysis Data</th>
<th>Workflow</th>
<th>Shared Data</th>
<th>Visualization</th>
<th>Admin</th>
<th>Help</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Image of Galaxy/ Dan's Brubeck Playground" /></td>
<td><a href="link">Link</a></td>
<td><a href="link">Link</a></td>
<td><a href="link">Link</a></td>
<td><a href="link">Link</a></td>
<td><a href="link">Link</a></td>
<td><a href="link">Link</a></td>
</tr>
</tbody>
</table>

### Tools

- **Analysis**
  - Numerical
  - Graph/Display Data
  - Region
  - Metagenomic analysis
  - VCF Tools

- **Data**
  - Encode
  - Filter and Sort
  - Join
  - Subtract and Group

- **Files**
  - Convert Formats
  - Extract Features
  - Fetch Sequences
  - Fetch Alignments
  - Get Genomic Features
  - Operate on Genomic Intervals

- **Statistics**
  - Wavelet Analysis
  - Graph Display Data
  - Regional Variation
  - Multiple Analysis
  - Evolution
  - MultiT"Algorithms

- **Python**
  - Multiple Alignments
  - Metagenomic analysis

### Metagenomic analysis

| FNA A | FNA B | FNA C | FNA D | FNA E | FNA F | FNA G | FNA H | FNA I | FNA J | FNA K | FNA L | FNA M | FNA N | FNA O | FNA P | FNA Q | FNA R | FNA S | FNA T | FNA U |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.123 | 0.456 | 0.789 | 0.123 | 0.456 | 0.789 | 0.123 | 0.456 | 0.789 | 0.123 | 0.456 | 0.789 | 0.123 | 0.456 | 0.789 | 0.123 | 0.456 | 0.789 | 0.123 | 0.456 | 0.789 | 0.123 |

---

*Note: The table above represents simulated data for demonstration purposes.*
Doing It

- QC and Preparation
- Classify Reads
- **Normalize**
- alpha diversity
- beta diversity
- Downstream Analysis and Visualization
Normalizing

- Rarefy
- Total sum scaling
- Cumulative Sum Scaling
- DeSeq2 process
- ...

• Have we sequenced enough?
• Normalize counts across samples
Vegan Tool Suite - Rarefaction Plot
<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;x&quot;</td>
<td>0.0439628348495116</td>
</tr>
<tr>
<td>&quot;X102C1Buccal_S13_L001_001_2&quot;</td>
<td>0.0436719139786513</td>
</tr>
<tr>
<td>&quot;X102C1Stool_S1_L001_001_2&quot;</td>
<td>0.0441353284015752</td>
</tr>
<tr>
<td>&quot;X102M.buccal_2&quot;</td>
<td>0.0412752117548852</td>
</tr>
<tr>
<td>&quot;X103C1.buccal_S26_L001_001_2&quot;</td>
<td>0.0375047214380159</td>
</tr>
<tr>
<td>&quot;X103C1.stool_S15_L001_001_2&quot;</td>
<td>0.0468916650084611</td>
</tr>
<tr>
<td>&quot;X103M.buccal_S37_L001_001_2&quot;</td>
<td>0.0355364524551467</td>
</tr>
<tr>
<td>&quot;X105C1Buccal_S14_L001_001_2&quot;</td>
<td>0.0475142673724111</td>
</tr>
<tr>
<td>&quot;X105C1Stool_S2_L001_001_2&quot;</td>
<td>0.0470150093788334</td>
</tr>
<tr>
<td>&quot;X105M.buccal_S1_L001_001_2&quot;</td>
<td>0.046748258733604</td>
</tr>
<tr>
<td>&quot;X106C1Buccal_S15_L001_001_2&quot;</td>
<td>0.0534846235423753</td>
</tr>
<tr>
<td>&quot;X106C1Stool_S3_L001_001_2&quot;</td>
<td>0.040421620256698</td>
</tr>
<tr>
<td>&quot;X107C1Buccal_S16_L001_001_2&quot;</td>
<td>0.0471970891841178</td>
</tr>
<tr>
<td>&quot;X107C1Stool_S4_L001_001_2&quot;</td>
<td>0.0533971575341763</td>
</tr>
<tr>
<td>&quot;X107M.buccal_S3_L001_001_2&quot;</td>
<td>0.0472890304994955</td>
</tr>
<tr>
<td>&quot;X108C1Buccal_S17_L001_001_2&quot;</td>
<td>0.0481609441445347</td>
</tr>
<tr>
<td>&quot;X108C1Stool_S5_L001_001_2&quot;</td>
<td>0.045170380186428</td>
</tr>
<tr>
<td>&quot;X108M.buccal_S4_L001_001_2&quot;</td>
<td>0.0378253483991975</td>
</tr>
<tr>
<td>&quot;X109C1Buccal_S18_L001_001_2&quot;</td>
<td>0.0310756229168307</td>
</tr>
<tr>
<td>&quot;X109C1Stool_S6_L001_001_2&quot;</td>
<td>0.0364931311092052</td>
</tr>
<tr>
<td>&quot;X109M.buccal_S5_L001_001_2&quot;</td>
<td>0.0375674715083014</td>
</tr>
<tr>
<td>&quot;X110C1Buccal_S19_L001_001_2&quot;</td>
<td>0.035845121476847</td>
</tr>
<tr>
<td>&quot;X110C1Stool_S7_L001_001_2&quot;</td>
<td>0.0492491404316483</td>
</tr>
<tr>
<td>&quot;X110M.buccal_S6_L001_001_2&quot;</td>
<td>0.039821866382552</td>
</tr>
<tr>
<td>&quot;X112C1.buccal_S27_L001_001_2&quot;</td>
<td>0.0325941364457868</td>
</tr>
<tr>
<td>&quot;X112C1.stool_S16_L001_001_2&quot;</td>
<td>0.0524261115683378</td>
</tr>
<tr>
<td>&quot;X112M.buccal_S38_L001_001_2&quot;</td>
<td>0.0359530450432949</td>
</tr>
<tr>
<td>&quot;X113C1.buccal_S28_L001_001_2&quot;</td>
<td>0.0360305077235005</td>
</tr>
<tr>
<td>&quot;X113C1.stool_S17_L001_001_2&quot;</td>
<td>0.0517149011158896</td>
</tr>
</tbody>
</table>
Doing It

- QC and Preparation
- Classify Reads
- Normalize
- **alpha diversity**
- beta diversity
- Downstream Analysis and Visualization
\(\alpha\)-diversity

- biodiversity in a defined habitat
- ecosystem
- sample
- Community Richness
- how many organisms are really there
Alpha Diversity

Galaxy interface showing the Vegan Diversity Index tool with a dataset for Vegan Rarefaction on data 500 (Random rarefied community matrix). The tool calculates diversity indices using vegan and selected methods.
Doing It

- QC and Preparation
- Classify Reads
- Normalize
- alpha diversity
  - beta diversity
- Downstream Analysis and Visualization
• compares species diversity between habitats
• samples
• How similar are two samples?
### Beta Diversity

#### Vegan Beta Diversity curve and statistics (Galaxy Version 0.0.2)

**File with abundance values for community**

- Column: 1
- Species, phylum, etc

**Group name column**

- Column: 2
- Column: 3
- Column: 4
- Column: 5
- Column: 6
- Column: 7
- Column: 8
- Column: 9
- Column: 10
- Column: 11
- Column: 12
- Column: 13
- Column: 14
- Column: 15
- Column: 16
- Column: 17
- Column: 18
- Column: 19
- Column: 20

Select each column that contains counts

**Input has a header line**

- Yes
- No

**X-axis label**

- Sample Size

**Y-axis label**

- Group

**Label beta_diversity curves by rownames of X**

- Yes
- No

**Diversity index to compute**

- $d = \frac{(b+c)^2}{2(a+b+c)}$

Order sites by increasing number of species
Doing It

- QC and Preparation
- Classify Reads
- Normalize
- alpha diversity
- beta diversity
- **Downstream Analysis and Visualization**
• Additional information about each sample
• Parallel **tabular** files (e.g. LEfSe: cat on top of input)
• BIOM format (not handled by many tools)

Use converters between tabular files and **BIOM**, various ways of pulling in and out metadata values
Combine Metadata with Relative Abundances

**Concatenate datasets**

**Concatenate Dataset**

- **Dataset**
  - 1: Dataset
  - **Select**
    - 508: Column frequencies on data 500

**Remove beginning of a file**

**Remove first**

- **Remove first 1 lines from**
  - 517: Concatenate datasets on data 508 and data 516

This tool removes a specified number of lines from the beginning of a dataset.
LEfSe - LDA Effect Size

Biological hypothesis
- Differential analysis
- comparative analysis
- biomarker discovery
- known biological structure
Two (or more) conditions

High throughput experiments
- 16S
- WGS
- mRNA

Quantitative estimations
- Taxon abundances
- Functional abundances (COGs, KEGG...)
- Gene expression

LEfSe

Visualization of differential features ranked by effect size

Representation of relevant features on taxonomic or phylogenetic trees

Prior Knowledge
- Known relation between feat.
- dependency between feat.
- biological feature groupings

Plot of features with statistically significant differences between conditions

https://bytebucket.org/biobakery/galaxy_lefse/wiki/lefse_ove.png
Format Data for LEfSe

Upload a tabular file of relative abundances and class labels (possibly also subclass and subjects labels) for LEfSe – See samples below - Please use Galaxy Get-Data/Upload-File, Use File-Type = tabular

Select whether the vectors (features and meta-data information) are listed in rows or columns

- Rows

Select which row to use as class
- #1: time

Select which row to use as subclass
- no subclass

Select which row to use as subject
- #2: #ID

Per-sample normalization of the sum of the values to 1M (recommended when very low values are present)
- Yes

Execute

What it does

LDA Effect Size (LEfSe) (Segata et al. 2010) is an algorithm for high-dimensional biomarker discovery and explanation that identifies genomic features (genes, pathways, or taxa) characterizing the differences between two or more biological conditions (or classes, see figure below). It emphasizes both statistical significance and biological relevance, allowing researchers to identify differentially abundant features that are also consistent with biologically meaningful categories (subclasses). LEfSe first robustly identifies features that are statistically different among biological classes. It then performs additional tests to assess whether these differences are consistent with respect to expected biological behavior.

Specifically, we first use the non-parametric factorial Kruskal–Wallis (KW) sum-rank test to detect features with significant differential abundance with respect to the class of interest; biological significance is subsequently investigated using a set of pairwise tests among subclasses using the (unpaired)
Visualization: Phinch

https://github.com/PitchInteractiveInc/Phinch

https://github.com/blankenberg/Phinch

http://www.bx.psu.edu/~dan/Phinch/
Overview

- Brief Introduction to Metagenomics
- Doing It
- Challenges, Solutions, and Future Work
Decreasing load times for list of saved histories

- Operations that can be done in the Database should be done in the Database.

Use database query rather than dataset iteration to get counts in saved histories grid. In my testing, speedup is 4-10X.
Multiple History View

- Fast view across many histories
- Drag and drop to copy Datasets
Dataset Collection Operations

Collection Operations (Limited) #2434

Merged

martenson merged 10 commits into galaxyproject:dev from jmchilton:fewer_collection_opts 15 days ago

Conversation 13
Commits 10
Files changed 31

jmchilton commented 28 days ago

Overview

This PR introduces Tool-derived framework-level plumbing for dealing collections at the model level instead of at the file level, allowing operations that generate new HDAs and collections without duplicating Dataset objects. Together operations vastly expand the expressiveness of Galaxy workflows.

The Collection Operations:

- Zip (two datasets -> paired collection). Like all these tools, it can be mapped over - so it can easily be used to take two dataset lists and build a list of pairs for instance.
- Unzip (paired collection -> two datasets).
- Filter failed datasets (list -> list). Given a list it produces another list without any failed datasets. (Most commonly requested of these operations.)
- Flatten collection (* -> list). Produces a flat list from any collection, joining identifiers on user supplied character.
Naming of Datasets

Concern
• Default Galaxy names lose Sample names (Tool xyz on Dataset 1,2,3,4)

Solution
• Dataset Collections have an Element Identifier that is maintained throughout jobs
• Tools that make use of \${dataset.name} should now use \${dataset.element_identifier}
Future Work: Scaling Up

- Viewing Large Histories
- Pagination
- Infinite Scrolling
- Dataset Bundling

- Uploading Large Collections
- Better handling of selecting many files
- Importing multiple files from an archive
  - Provide a manifest file that instructs Galaxy to build a Collection
Additional Options Available: Qiime

WIP: Add Qiime wrappers #431

bgruening wants to merge 76 commits into master from qiime

- Conversation 39
- Commits 76
- Files changed 72

bgruening commented on Nov 30, 2015 • edited

This is our WIP PR during the metagenomics hackathon

edit Saskia:
..and the GCC2016 hackathon.

Resources:
Qiime script documentation: http://qiime.org/scripts/index.html
Qiime github repo: https://github.com/biocore/qiime
Qiime test data in github repo (folder qiime_test_data)
**FROGS**: Find Rapidly OTUs with Galaxy\[5\] Solution

Frédéric ESCUDIE*, Lucas AUER*, Maria BERNARD, Laurent CAUQUIL, Katia VIDAL, Sarah MAMAN, Mahendra MARIADASSOU, Guillermina HERNANDEZ-RAQUET, Géraldine PASCAL

*These authors have contributed equally to the present work.

Contact: geraldine.pascal@toulouse.inra.fr

- **User-friendly** + **Accuracy** + **Speed** + **Scalability** = **FROGS**

---

**Core workflow**

- **Pre-process**
  - To merge paired-end reads: Flash trimming: Cutadapt
  - SWARM \[1\]

- **Clustering**
  - Efficient chimera removal: VSEARCH \[2\]

- **Chimera**
  - A wide choice of filters.
  - We advise to filter OTU abundances at 0.005% \[7\].

- **Filter**
  - Double taxonomic affiliation RDP classifier\[10\] until species.
  - Blast \[4\] with multi species.

---

**FROGS = complete analyze**

Data acquisition

<table>
<thead>
<tr>
<th>ID</th>
<th>Taxonomy</th>
<th>Sample A</th>
<th>Sample B</th>
</tr>
</thead>
<tbody>
<tr>
<td>OTU_1</td>
<td>Bacteria;chlorobi</td>
<td>12 895</td>
<td>54</td>
</tr>
<tr>
<td>OTU_2</td>
<td>Bacteria;firmicutes</td>
<td>0</td>
<td>6 812</td>
</tr>
</tbody>
</table>

---

**Great graphics outputs**
Take Home

- You can do **metagenomics** with Galaxy
- Many different modules to swap in and out
- You can do **large-scale multiple sample analysis** with Galaxy
- 500 Samples are no problem
- 5,000+ Samples through API
- **Client-side** is under active development
Acknowledgements

- Everyone Here
- Tool Developers
- Galaxy Committers and Contributors
- Galaxy Community

- Funding: grant number HG005542 from the National Human Genome Research Institute, National Institutes of Health as well as grants HG005133, HG004909 and HG006620 and NSF grant DBI 0543285. Additional funding is provided by Huck Institutes for the Life Sciences at Penn State and, in part, under a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.
Watch out for Drop Gators!