Metagenomic analysis of differential gene abundance across ocean, soil, and oil sands

Justin Chu[1], W. Evan Durno[2], Craig Hammett[3], Kateryna Ievdokymenko[4], Jake Lever[1], and Thuy Nguyen[5]

[1]Genome Sciences Center, [2]Bioinformatics, [3]Forest Sciences & Conservation, [4]Genome Sciences & Technology, [5]BC Center for Excellence in HIV/AIDS

Objectives

- Which genes are differentially abundant between environmental subsets and what do they do?
- How do you normalize gene abundance between samples of variable read length?
- What taxa live in our ocean's waters and how do they compare to terrestrial samples?
- Can we detect increased gene abundance for detoxification in the oil sands samples?

Introduction

Metagenomics is a field that allows us to study microbial flora that are unculturable (99% of all microbes). This is done by filtering an environmental sample for total DNA for direct sequencing utilizing direct shotgun sequencing (Illumina HiSeq was used). The gene abundance can be correlated to microbial abundance allow us to study the biodiversity of an environment was well as which genes may be biologically relevant to that environment..

Data

Total sample set was 92 isolates (48 open ocean, 42 inlet, 6 soil, and 2 oil sands). The open ocean samples were collected from five locations at 5 different depths while the inlet samples were collected at 6 different depths. Samples were taken between 2008-11.

This full data set was used to analyze the riboflavin synthesis cluster of interest for potential abundance in the contaminant-rich oil sand samples. Thus, the detoxification genes (*rib* gene cluster) were tested with greater statistical power than the remaining genes. Thus, F-Tests for these genes were more able to reject their null hypothesis than the remaining genes.

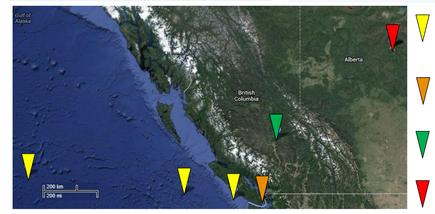
Due to computational burden a subset of 656 304 genes were selected for counting in 38 samples. This subset was chosen due to availability from the Uniprot database with corresponding Gene Ontology identification.

Samples used for *rib* subset

Environment	Total Samples
Open Ocean	42
Inlet	48
Soil	6
Oil sands	2

Samples used for total gene abundance

Location	Site	Total Samples
Open ocean	P4	7
Open ocean	P12	2
Open ocean	P16	10
Open ocean	P20	8
Inlet	SI03	6



Ocean samples: 50km to 2000km from shore along the 49th parallel

Inlet: Saanich Inlet, estuary near Victoria, BC

Soil: Boreal forest near Williams Lake, BC

Oil Sands: Suncor tailing ponds near Fort McMurray, AB

Methods

Alignment

The sequence data from the 38 samples was processed using the MetaPathways pipeline that identifies possible genes from the RefSeq database. RAPSearch2 was then used to align the short-reads against the RefSeq genes protein sequences. These alignments were then used to generate read counts per gene for each sample as well as residue alignment counts.

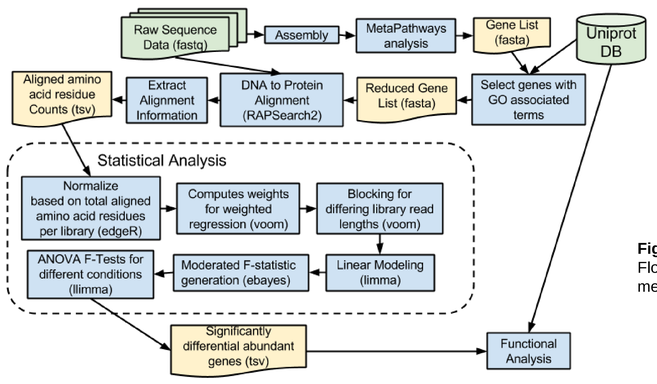


Figure 1. Flowchart of methodology

Simulated Data Set

As several sample libraries varied in read length, there was a concern that alignments would decrease in the samples with shorter reads. A simulated dataset was generated to test methods for different read lengths. 500 chunks of 500,000 reads from the oil-sands samples (150bp reads) were selected. New libraries were generated by truncating these 500 chunks of reads to 50 and 100bp. An additional set of libraries was generated with reads truncated to lengths uniformly distributed between 50bp and 150bp. All four sets of libraries (50bp, 100bp, 150bp and varied 50-150bp) were aligned using RAPSearch2 to a smaller set of 335 genes and read counts were generated.

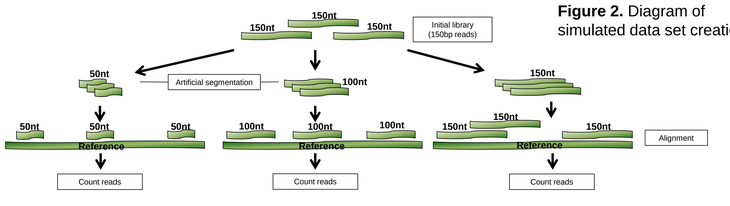


Figure 2. Diagram of simulated data set creation

Results

Differential Gene Abundance

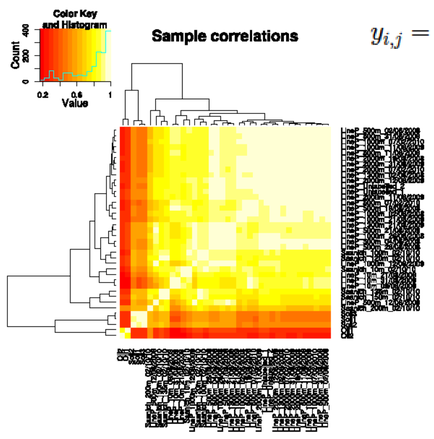


Figure 7. Heatmap of the correlation matrix of our 38 samples' aligned residue counts.

$$y_{i,j} = \mu + \beta_i + (ave.length) \times \tau + \varepsilon_{i,j}$$

Figure 8. Linear model used in analysis. Mu is the oil sands baseline. Beta is the environmental effect. Tau is the blocking variable for read length.

Having decided to utilize blocking to deal with bias for read length because it effectively reduces false positives, we continued to search for bias in our correlation structure with a heat map (Figure 7). Samples tended to correlate within environments more than between, which suggested effects could be found. No other major sources of bias were found.

To test for differential gene abundance between environments, calculated significance with Voom & Limma using the stated model (Figure 8). After F-tests were performed, it was discovered that 54.7% of genes were differentially abundant.

Gene Functions

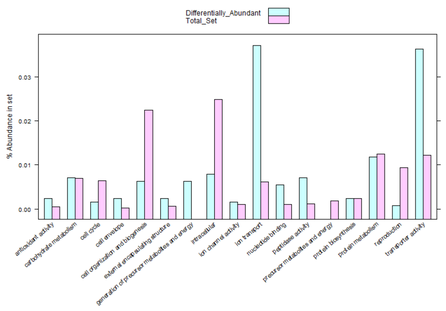


Figure 10. Toptable results for differential gene abundance with respect to gene abundance with a false discovery rate (FDR) of 1e-20 yielded 2409 genes.

GO annotations were calculated as a proportion of the significant hits and compared to the abundance of each GO term from the entire gene set.

Over represented GO ids include candidate de-toxification proteins such as cell envelope, generation of precursor metabolites, and ion transport. While under represented genes in the hits included 'house-keeping' proteins such as cell organization, and intracellular annotations.

GO ID	Description	Counts in Hits	Total Counts
GO:0030313	cell envelope	3 (0.24%)	6 (0.02%)
GO:0006811	ion transport	47 (3.71%)	202 (0.61%)
GO:0006091	generation of precursor metabolites and energy	8 (0.63%)	61 (0.18%)
GO:0000166	nucleotide binding	7 (0.55%)	32 (0.10%)
GO:0016209	antioxidant activity	3 (0.24%)	16 (0.05%)
GO:0030312	external encapsulating structure	3 (0.24%)	20 (0.06%)

Taxonomies

Functional protein annotations can be assigned to a particular taxonomic group based on sequence similarity. We used a platform MEGAN4 to facilitate the analysis of taxonomic content of metagenomic data. Taxonomic signal of many hits are collapsed to their lowest common ancestor (LCA) on the NCBI Taxonomy phylogenetic 'tree-of-life.'

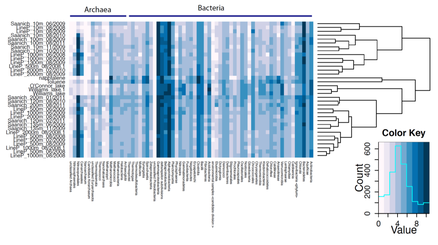


Figure 9. Heatmap of the variance of the taxon ID sequences aligned for bacteria and archaea between samples.

Analysis of Riboflavin Synthesis genes

To study microbial adaptation to toxic environments we looked at the riboflavin cluster (*Rib*) to analyze abundance in the toxic environment of oil sands tailing ponds. We allocated increased computational resources to these genes for all 98 samples; thus resulting in higher statistical power. Riboflavin is a precursor of flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD), which work as co-factors for a wide variety of enzymes.

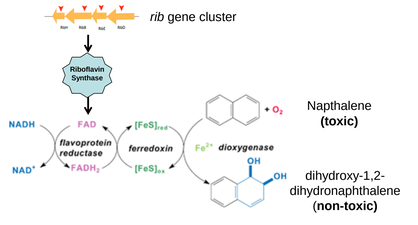


Figure 11. The *rib* gene cluster produces riboflavin synthase which helps de-toxify toxic substances such as napthalene

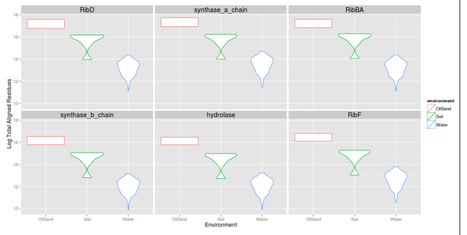


Figure 12. Scatterplot of differential abundance for the six genes in the *rib* cluster among all 98 samples

Sources of Bias

Blocking by Read Length

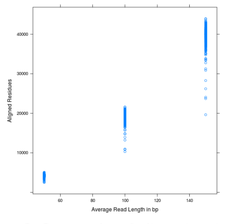


Figure 3. Scatterplot of aligned residues for 3 artificial datasets of varying read lengths

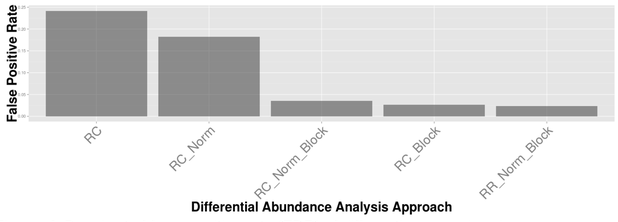


Figure 4. Bar chart of false positive rate (FDR) of differential gene abundance analyses with read counts (RC), aligned residues (RR), normalization (Norm), and blocking for read length (Block)

Using our simulated dataset, we evaluated bias due to read length by comparing a set of samples at one read length to the same set of samples at different read lengths. As the set of libraries are generated from the same original data, with only the read length varying, few significant genes should be identified. Various approaches were assessed on their ability to account for read length differences. Figure 4 shows that representing gene abundance by aligned residues, normalizing for library size, and blocking for read lengths obtains the smallest false positive rate.

Normalizing with Aligned Residues

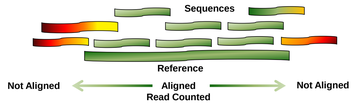


Figure 5. Partial read alignments can make it difficult to determine true gene abundance due to ambiguity of partial alignments.

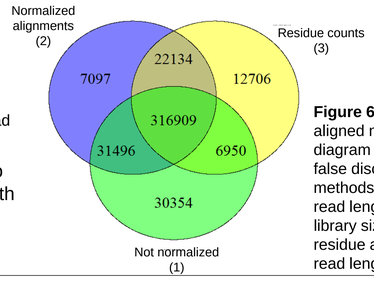


Figure 6. Comparison of methods using our aligned metagenomic sample data. Venn diagram of differentially abundant genes with a false discovery rate (FDR) of 1e-5 between three methods of: (1) alignment counts blocked for read length, (2) alignment counts normalized to library size and blocked for read length, and (3) residue abundance normalized and blocked for read length.

Conclusions

- Significant differential gene abundance between samples was detected. In total 358 699 of 656 304 genes were differentially abundant between environments with reference to the oil sands with a FDR of 1e-5.
- We compared normalization against blocking for read length and discovered that blocking produced fewer false positives.
- Our significant differentially abundant genes were enriched for genes that could be of interest in detoxification, including cell envelope, antioxidant activity, and ion transport.
- We were able to detect significant differential gene abundance riboflavin synthesis genes in the oil sand samples.
- We were able to illustrate different taxonomic abundances across environments.