Size matters when using local similarity searches to estimate microbial average genome size and relative abundance in metagenomes

Florent Angly, Rob Edwards, Robert Schmieder, Rebecca Vega-Thurber, Dana Willner, Forest Rohwer
Size matters when using local similarity searches to estimate microbial average genome size and relative abundance in metagenomes.

**Viral ecology**

- **High diversity**
  - **Mimivirus**
    - 1,700,000 kb
    - 750 nm
  - **Hepadnavirus**
    - 3.2 kb
    - 18 nm

- **Global impact**

- **Ubiquity**

- **Metagenomic**

- **Species composition**

**Average genome size**
Size matters when using local similarity searches to estimate microbial average genome size and relative abundance in metagenomes

**Methods**

**Overview** Calculate average genome size by finding local similarities (with BLASTn) between metagenomic random shotgun sequences and a database of complete genome sequences.

**Problems**

1) Local similarities allow annotation of a portion of a sequence, not its entirety

2) A metagenomic sequence can have equally good hits to several database sequences

3) The random nature of DNA libraries means that larger genomes are overrepresented

**Solutions**

Filter hits by alignment relative size:

Keep all hits for each query sequence and give them a statistically meaningful weight:

\[
 w = \frac{1}{E}
\]

Normalize relative abundance by the genome length \( t \):

\[
 r = \frac{w}{t}
\]
Size matters when using local similarity searches to estimate microbial average genome size and relative abundance in metagenomes

**Benchmark**

- GAAS on >4000 artificial metagenomes
- Simulated viral pyrosequenced metagenomes with 1% error rate and species following a power-law rank-abundance distribution
- 90% of the species simulated to be unknown
- Cumulative error calculated as:

\[
R = \frac{\sum r_i^2}{n}
\]

The GAAS approach yields better results (lower error)
Size matters when using local similarity searches to estimate microbial average genome size and relative abundance in metagenomes

In practice

Average viral size in various biomes

Classical approach

GAAS approach

Viral composition of the Sargasso Sea

Biome

- OCEAN
- SOIL
- HOST-ASSOCIATED
Size matters when using local similarity searches to estimate microbial average genome size and relative abundance in metagenomes

Conclusion

- GAAS is a new free software bioinformatic tool
- GAAS implements a novel methodology to calculate average genome size in random shotgun libraries
- GAAS calculates a more accurate community composition than other tools by estimating species relative abundance
- GAAS has the potential for making important discoveries and changing our perception of viral community composition
- GAAS is applicable for a wide variety of complete genomic sequences:
Size matters when using local similarity searches to estimate microbial average genome size and relative abundance in metagenomes.
Size matters when using local similarity searches to estimate microbial average genome size and relative abundance in metagenomes
Size matters when using local similarity searches to estimate microbial average genome size and relative abundance in metagenomes