SAGEX: The Single-cell Amplified Genome EXtrapolator

W. E. Durno[1,2], Alyse K. Hawley[1], Connor Morgan-Lang[1,2], and Steven J. Hallam[1,2]

[1] Department of Microbiology and Immunology, [2] Graduate Program in Bioinformatics, University of British Columbia, Vancouver, BC, Canada

Overview

Single-cell amplified genome (SAG) sequencing promises to provide a powerful adjunct to shotgun sequencing of environmental DNA (e.g. metagenome) in determining uncultivated microbial community structure and function. Resulting SAG sequences can be unambiguously assigned to cognate taxa providing robust fragment recruitment platforms for determining gene frequency distribution patterns and population structure in natural and human engineered ecosystems. Unfortunately the multiple displacement amplification process used to generate SAG sequences results in substantial gaps in genome coverage. For example, by counting Clusters of Orthologous Groups (COGs) [1], we show that for a collection of 129 SAGs genome completeness ranges between 5-80%. Here, we propose a bioinformatic remedy to incompleteness inspired by [2], the single-cell amplified genome extrapolator (SAGEX) that leverages available SAG sequence information to extrapolate genome coverage using assembled metagenome sequences generating a population genome bin. SAGEX is capable of high Positive Predictive Value (PPV) while maintaining sensitivity making it an excellent predictor of related donor genotypes. SAGEX is implemented in C/C++ and requires no additional libraries beyond the compiler. The pipeline is fast (usually running in under 5 minutes) and can be used to optionally write kmer counts and a kmer PCA readable into R for visualization.

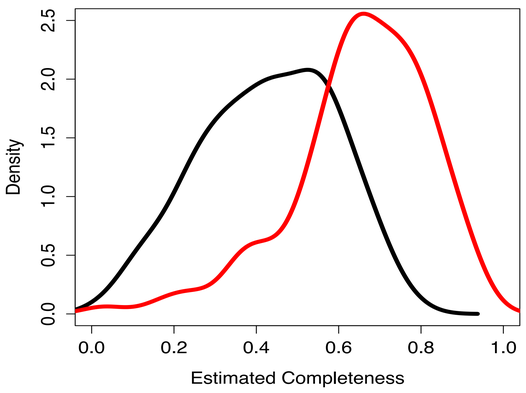


Figure 1. SAGEX confidently extrapolates SAGs.
Prior to running SAGEX on 129 SUP05 SAGs sourced from the Saanich Inlet water column (black) [3], estimated completeness (by counting COGs) prior to extrapolation averages 40% (red). Using high PPV settings, SAGEX increases average completeness to 70%. Relaxing PPV settings can increase completeness of the extrapolated genome.

How it works

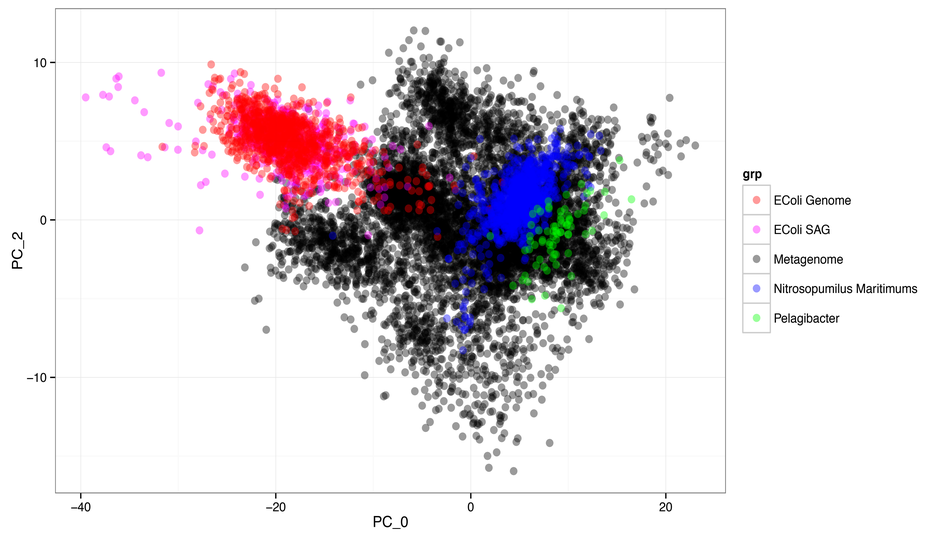


Figure 2. Genome's kmer signatures.
SAGEX requires contigs to pass two tests before calling them hits. First is an identity filter. Second is a kmer signature filter. Various genome kmer signatures are pictured above demonstrating the intuition behind the second test. Notice how the EColi's genome and SAG share very similar kmer signatures. The image is generated by calculating the tetranucleotide frequencies of each contig, converting counts to proportions, transforming all data to the principal component basis of the standardized metagnome, and plotting only two principal components.

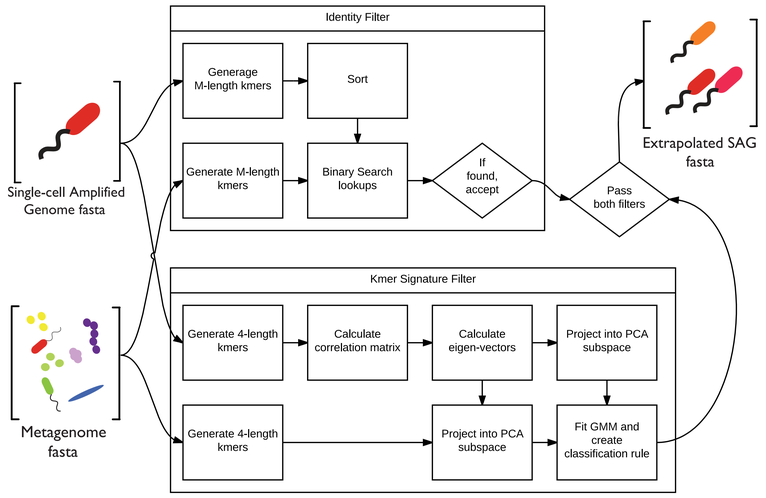


Figure 3. The SAGEX pipeline.
SAGEX accepts two input FASTAs, an assembeld SAG and assembled metagenome, and provides a single output FASTA, a subset of the metagenome conigs recruited by the SAG. The two tests which every hit must pass are expanded here. The identity filter test requires every hit to share at least one M-length substring with any SAG contig. The kmer signature filter requires each every hit to fall within the kmer signature region as estimated by a Gaussian Mixture Model in a kmer-proprotion principal component subspace.

Error Analysis

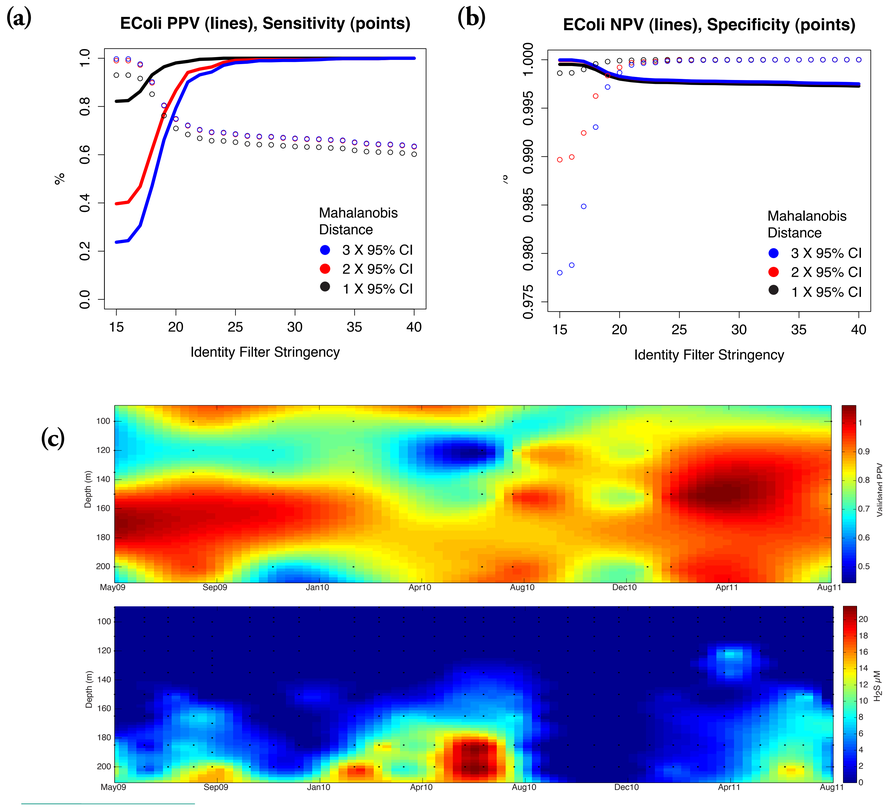
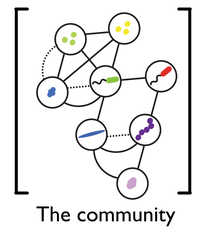


Figure 4. Error analyses of SAGEX.
(a,b) Classification errors were estimated for SAGEX using an E. Coli MG1655 SAG [4] as training data, a Saanich Inlet metagenome [5] as false hits, and an E. Coli MG1655 isolate genome [6] as true hits. E. Coli MG1655 was ideal because it is not expected to occur naturally in the metagenome, thus not confusing true and false hits. **(c)** SAGEX PPV was validated because E. Coli is too ideal a case. We conclude that SAGEX should be used with a SAG and metagenome from the same sample (or similar). To validate SAGEX PPV, the 129 SUP05 SAGs were run against 48 metagenomes at various depths and times. BLAST was used to query SAGEX identified ORFs against a database containing all SUP05 SAG sequences, requiring at least one successful hit for validation. (c,top), The average validated PPV per metagenome. Dots represent sample points in time and depth. Interpolation was used to create the contour plot in Matlab. (c,bottom), H_2S concentrations. SUP05 is a sulfur-oxidizing gamma proteobacterial group known to inhabit anoxic sulfidic waters up to 4µM H_2S, High validated PPV is associated with increased SUP05 abundance and water column H2S concentrations. The higher validated PPV in upper waters appears to be associated with a second group of putative sulfur-oxidizing bacteria, Arctic96BD-19 closely related to SUP05.

Future Directions

Figure 5. Understanding the community.
Microbial ecology is all about relationships. Statistical methods for inferring interactions between uncultivated microbial community members promise to shed new light on distributed metabolic networks in natural and engineered ecosystems. Extrapolated SAGs can provide additional metabolic insights with increased confidence where such methods fall short. Looking forward, combining statistical methods with SAGEX will provide even more powerful "optics" for exploring the microcosmos one relationship at a time.



References

1. C. Rinke et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature 449, 7459 (2013).
2. J. A. Dodworth et al. Single-cell and metagnomic analyses indicate a fermentative and saccharolytic lifestlye for members of the OP9 lineage. Nature Communications 4, 1854 (2012).
3. S. Roux et al. Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell and meta-genomics. eLife 2014;10.7554/eLife.03125.
4. J. Gole et al. Massively parallel polymerase cloning and genome sequencing of single cells using nanoluter microwells. Nature Biotechnology 31, 1126-1132 (2013).
5. E. Zaikova et al. Microbial community dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia. Environmental Microbiology 12, 1 (2010).
6. F. R. Blattner et al. The Complete Genome of the Escherichia coli K-12. Science 277, 533 (1997).

Acknowledgments

This work has also recieved support from the Hakai Beach Institute, and the Monell Foundations.