SAGEX: The Single-cell Amplified Genome EXtrapolator (ROUGH, do not distribute)

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 Genomes (SAGs) are the sequenced genomes of invdividual cells and reperesent an important source of information in understanding microbial communities. Unfortunately the amplification process required to bring DNA up to sequenceable levels often results in excluding massive portions of genomes. By counting Clusters of Orthologous Groups (COGs), we estimate that many SAGs are missing the majority of their genomes (figure 1). We propose a bioinformatic remedy to this issue of genome incompletess. SAGEX (Single-cell Amplified Genome EXtrapolator) uses contigs from an assembled metagenome to extrapolate the SAG. Error analyses (figure 4) suggest that the extrapolated genome is an excellent predictor of phylogenetically nearest genomes.

SAGEX is capable of high Positive Predictive Value (PPV) while maintaining sensitivity. This ensures that, if used correctly, SAGEX will allow the user to confidently study populations through SAGs without sufferring the burden of imcomplete SAGs due to amplification bias.

SAGEX is implemented in C/C++ and requires no additional libraries beyond the compiler. The pipeline is fast (usually running in under 5 minutes). It may optionally write kmer counts and a kmer PCA (figure 2) readable into R for visualization.

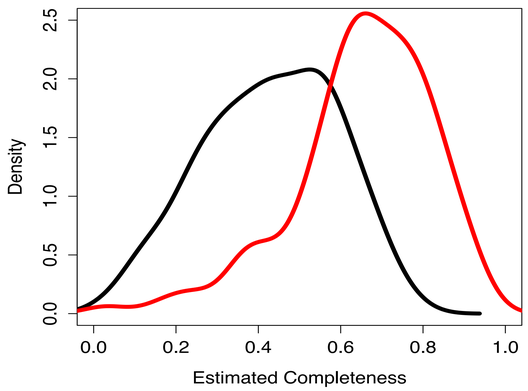


Figure 1. SAGEX confidently extrapolates SAGs.
Prior to running SAGEX on 129 SUP05 SAGs (black), estimated completness (by counting COGs) is lower than after extrapolating (red). On high PPV settings, SAGEX can bring in an average of 23% more COGs. Relaxing PPV will allow more greater completeness to the extrapolated genome.

How it works

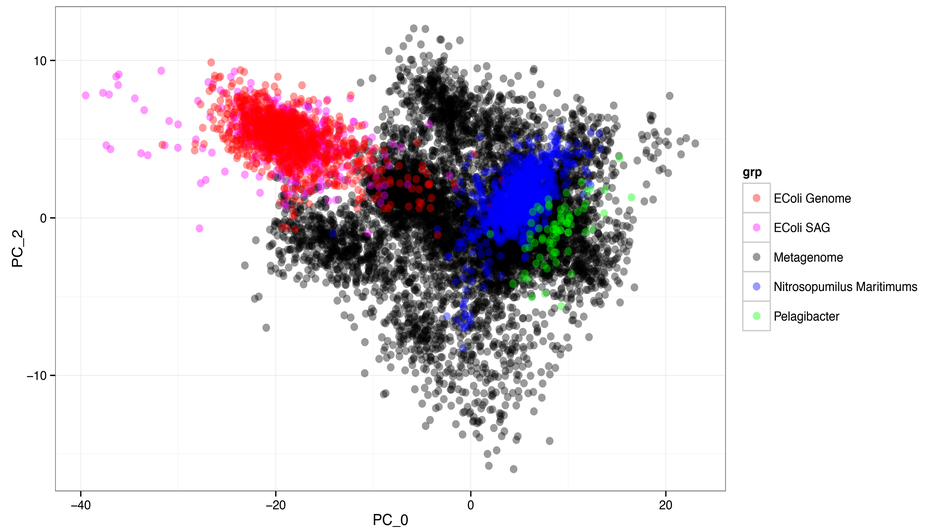


Figure 2. Genome's kmer signatures.
SAGEX requires contigs to pass two tests before calling them hits. First is an identity filter. Second is a kmer signature filter. Various genome kmer signatures are pictured above demonstrating the intuition behind the second test. Notice how the EColi's genome and 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.

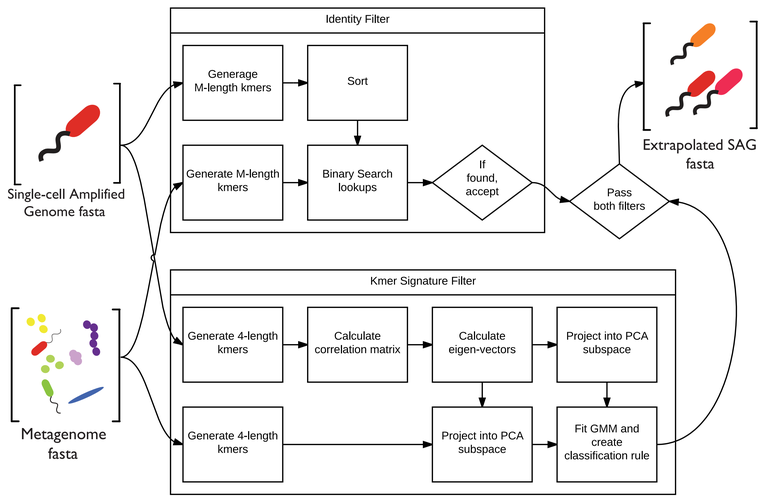


Figure 3. The SAGEX pipeline.
SAGEX accepts two input FASTAs, an assembeld SAG and assembled metagenome, and provides a single output FASTA, a subset of the metagenome conigs recruited by the SAG. The two tests which every hit must pass are expanded here. The identity filter test requires every hit to 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

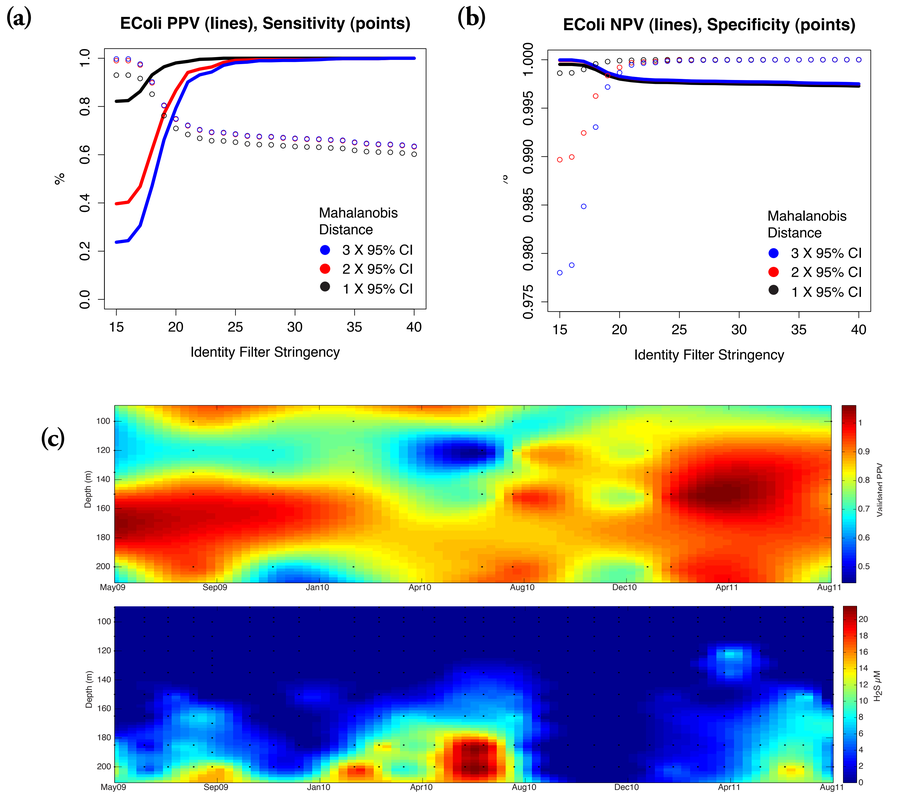
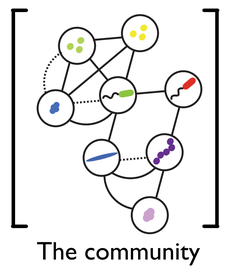


Figure 4. Error analyses of SAGEX.
(a,b) Classification errors were estimated for SAGEX using an EColi MG1655 SAG as training data, a Saanich Metagenome as false hits, and an EColi MG1655 genome as true hits. EColi MG1655 was ideal because it is not expected to occur naturally in the metagenome, thus not confusing true and false hits. (c) SAGEX PPV was validated because (1) EColi is too ideal a case and (2) separate assemblies of the metagnome and EColi genome could have created bias. 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. SAGEX hit ORFs were BLASTed against the collection of other SAGs' ORFs, requiring at least one successful alginment to validate the hit. In (c,top), we have average validated PPV per metagenome. Dots represent sample points in time and depth. Interpolation was used to create the heatmap. In (c,bottom), we have H_2S measurements. SUP05 is known to exist in 4uM H2S quantities, so we percieve that high validated PPV is associated with SUP05 populations. The higher validated PPV in (c,top) top-left is expected to be due to a cousin of SUP05.

Future Directions

Figure 5. Understanding the community.
Microbial ecology is all about relationships. Statistical methods for discovering dependence between microbiota are largely blind to genetic motivation and weak to coincidence. Extrapolated SAGs can confidently provide insights where such methods fall short. SAGEX will be combined with statistical methods to help understand the microbial community as a machine, one relationship at a time.



References

1. THIS PART IS NOT DONE YET.
2. A. L. Barabási, Z. N. Oltvai, Nat Rev Genet 5, 101–113 (2004).
3. Sci2 Team. Science of Science (Sci2) Tool, http://sci2.cns.iu.edu. (2009)
4. C. E. Mason et al., Adv Exp Med and Biol 680, 693 (2010).
5. M. T. Dougherty et al., Com of the ACM 52, 42 (2009).
6. P. D. Karp, S. Paley, and P. Romero. Bioinformatics. 18: S225-32 (2002).
7. Caspi R., et al. Nucleic Acids Res. 38: D473-479 (2009).

Acknowledgments

This work was supported by grants from Genome Canada, the Canadian Institute for Advanced Research (CIFAR), the Tula Foundation, and CIHR/MSFHR Strategic Training Program in Bioinformatics.