**Abstract**

**Introduction**

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% (Figure 1). Here, we propose a bioinformatics remedy to incompleteness inspired by [2], the singlecell amplified genome extrapolator (SAGEX) that leverages available SAG sequence information to extrapolate genome coverage using assembled metagenome sequences generating a Population Genome Bin (PGB). 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.

SAGEX allows each SAG to be more valuable, because fewer SAGs will be needed to understand a PGB. We demonstrate this point by utilizing SAGEX on organism XXX. We also explore the possibility of using SAGEX to understand organism abundance by counting recruited Reads Per Kilobase per Million mapped reads (RPKM) by leveraging our large collection of SUP05 SAGs.

*REMARKS:*

The hooks for selling SAGEX are two-fold:

1. Users get more out of their SAGs via greater completion (extrapolation). Fewer SAGs may be required to perform an experiment or their sequencing can represent more taxonomic variety.
2. Combining with RPKM demonstrates where relatives of the SAG have been across several metagenomes.

**Methodology**

*REMARKS:*

Currently we are combining methodology and discussion until a target journal is decided. The general sections of this section follow.

1. A description of how SAGEX works (figure 2)

2a. Cover general usage

2b. Identity filter

2c. Kmer signature filter (figure 3)

2d. Optional byproducts: kemrs, R-plottable PCA

1. Error Analysis (figure 4)

2a. Ecoli error analysis covers ideal circumstances

2b. Sell the excellent PPV-sensitivity trade-off

2c. The 129 SUP05 SAGs are used to validate PPV, a powerful addition

1. Application

3a. Describe completion after extrapolation (figure 1)

1. Discussion

4a. Describe caveats of usage

4b. Explore completion of MDM???

SAGEX is a classifier. The pipeline (Figure 2), in optimized default behavior, has a simple interface. It accepts an assembled SAG fasta and an assembled metagenome fasta as input, and outputs contigs from the metagenome classified to be part of the extrapolated SAG. The SAG is training data for the classifier. The metagenome is the data to be classified.

The classifier works by leveraging two intuitions suggesting a contig may belong to the SAG’s PGB, (1) they should share an identical subsequence, and (2) they have a similar kmer signature (Figure 3). A contig must pass a test for each to be accepted into the extrapolated SAG. Parameters have been optimized to allow for 90% PPV and approximately 60% sensitivity. SAGEX is implemented in C/C++ and requires no additional libraries beyond a compiler allowing the software to remain easy to compile.

The classification rule

The identity filter with stringency M requires each candidate metagenome contig to share an exact M-length subsequence with at least one SAG contig. It works by hashing all M-length SAG contigs as integers (M-mers) into a sorted array for binary search by the metagenome’s contig’s M-mers. Since M is large (i.e. 20), binary search is used instead of constant time look-ups because unique storage of all possible M-mers is space-inefficient, and logarithmic time look-ups are sufficiently fast for the task.

The kmer signature filter with distance D and clusters K requires candidate contigs to have a kmer signature similar to the SAG. In precise language, the metagenome and SAG contigs have their tetranucleotide (4-mers) frequencies calculated and are thus represented as 256-length integer lists which are then converted to proportions to control for contig length variation. The correlation matrix of the metagenome 4-mers is calculated, of which eigenvectors are generated, and both the SAG and metagenome 4-mers are dimensionally reduced via Pricipal Component Analysis (PCA) to three dimensions is the basis generated by the metagenome. A Guassian Mixture Model (GMM) with K clusters is fit to the SAG’s 4-mers in the PCA-subspace. If K = 1, the null distribution of the GMM is known and the 95% confidence region is calculable as the Mahalanobis distance follows a Chi-Squared distribution. If K > 1, the null distribution is estimated via simulation, and the 95% confidence region is defined as the minimum Mahalanobis distance to a Gaussian centroid. Any metagenome 4-mer which falls within D times the Mahalanobis distance cut-off passes the filter.

It should be noted that SAGEX has some features of convenience due to the variety of computations required to produce classifications. SAGEX can be used to calculate 4-mers and write them to disk. It can also write the PCA to disk in an R-readable format ready for easy visualization.

Error analysis

The two rules, inspired by heuristics of genomic science (???), are very effective when tuned. An error analysis (figure 4a,b) suggests that SAGEX is capable of above 95% PPV with at least 60% sensitivity. The error analysis is validated with a large collection of 129 SUP05 SAGs.

The error analysis was done by using an EColi MG1655 SAG as training data, its genome as true positives, and a Saanich Inlet metagenome as false positives. The lab-cultivated strain is ideal for validating the tool because it allows for minimal variation in the kmer signature. The analysis demonstrates the favorable exchange between PPV and sensitivity but also shows SAGEX to be a highly specific classifier in that its Negative Predictive Value (NPV) and specificity are always very high (above 97.5%).

Because EColi MG1655 was never expected to exist in Saanich Inlet, the initial error analysis is perhaps trivially favourful. To validate the claims of high PPV, we estimated PPV via a more realistic means using SAGEX on SUP05 SAGs. Our validated PPV estimate is calculated by deciding hits as valid by aligning hit contig Open Reading Frames (ORFs) to our database of 129 SUP05 SAG ORFs with BLAST. Validated PPV estimates were then calculated for all 129 SUP05 SAGs across a 48 metagenomes from the Saanich Inlet (Figure 4c). The metagenomes were sampled across depth and time, containing the region where SUP05 is expected to reside. It was seen that average validated PPV estimates were highest (about 90%) in SUP05’s region. This tells us that SAGEX maintains high PPV when the metagenome contains an abundance of members from the SAG’s family.

An ideal application

SAGEX can be used to get maximum value per SAG. If used correctly, fewer SAGs will be needed to understand a PGB. Thus sequencing can be strategically distributed across different taxa, allowing a better understanding of the whole community. Alternatively, it may simply reduce sequencing costs by allowing fewer SAGs to sequenced. To demonstrate this value, we’ve taken a SAG with only XXX representatives in our SAG collection and run it through SAGEX with an appropriate metagenome. By analyzing the extrapolated genome, we can see …

RPKM recruitment

As an alternative application of SAGEX, we’ve used it to attempt to measure abundance of a SAG’s PGB in alternative metagenomes. To do this we’ve calculated the recruited RPKM per metagenome. A plot of the average recruited RPKM (Figure 5) across our SUP05 SAGs via our Saanich Inlet metagenomes demonstrates how SAGEX suggest the presence of SUP05 in the correct locations.

**Conclusion**

SAGEX is a classifier trains on a Single-cell Amplified Genome (SAG) and recruits from a metagenome to help the user better understand the Population Genome Bin (PGB). SAGEX adds value to SAG-driven analyses by complementing the incomplete SAG with extrapolated content from the metagenome. This empowers the user to taxonomically diversify sequencing of SAGs or simply reduce sequencing costs by sequencing fewer SAGs. SAGEX is a powerful classifier capable of high PPV (90% or higher) while maintaining sensitivity (60% or higher). The error analysis has been validated with a set of 129 SUP05 SAGs. SAGEX’s ideal application has been exemplified through an analysis of XXX revealing XXX. It has also been demonstrated that SAGEX is capable of indicating existence of a PGB lineage in a set of metagenomes.

**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).