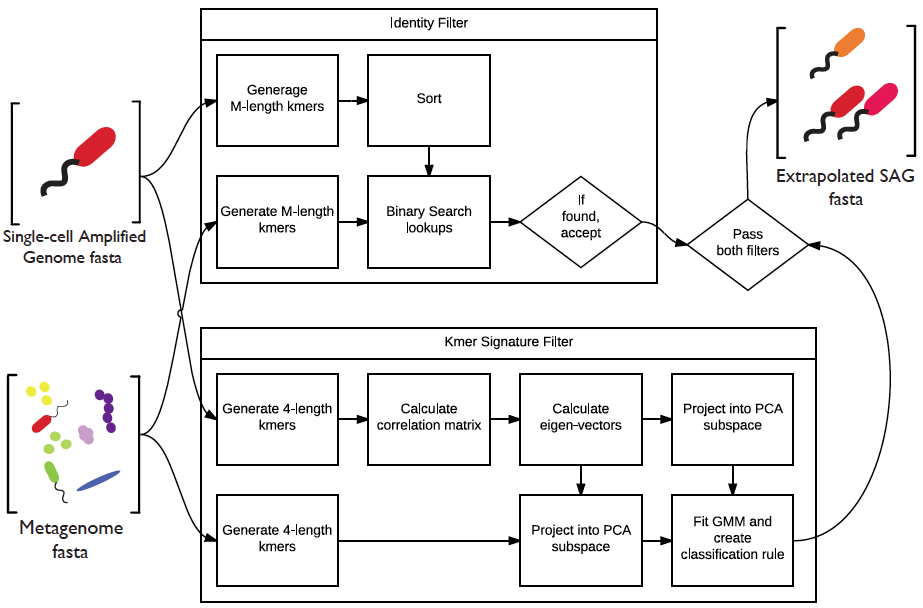


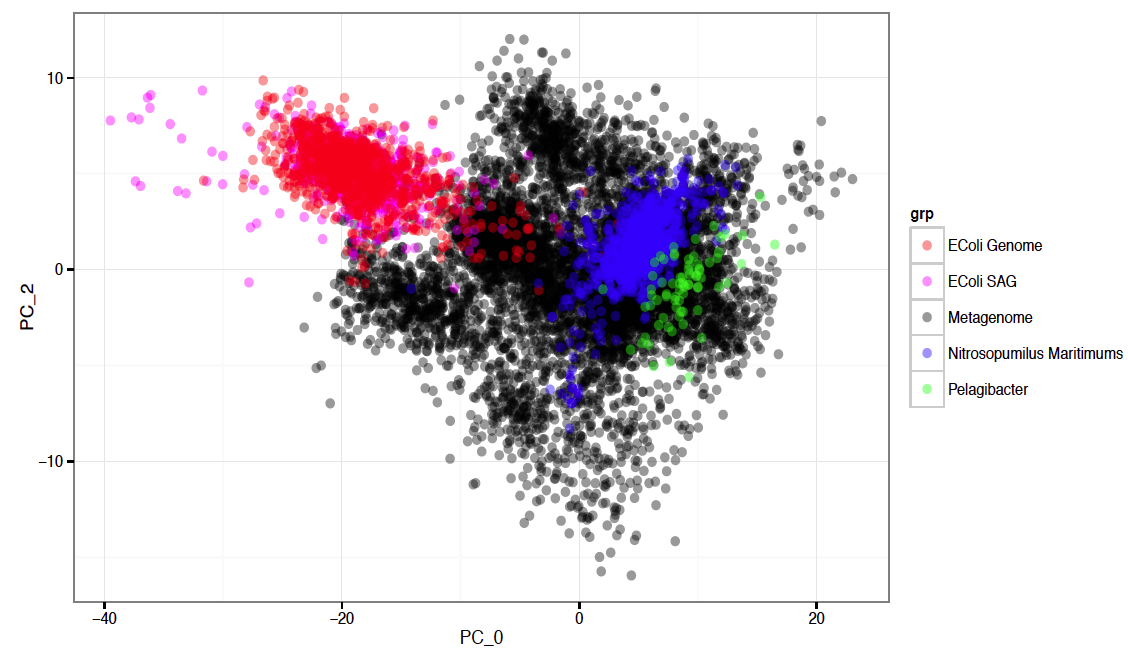
**Figure 1. SAGEX confidently extrapolates SAGs.**

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.



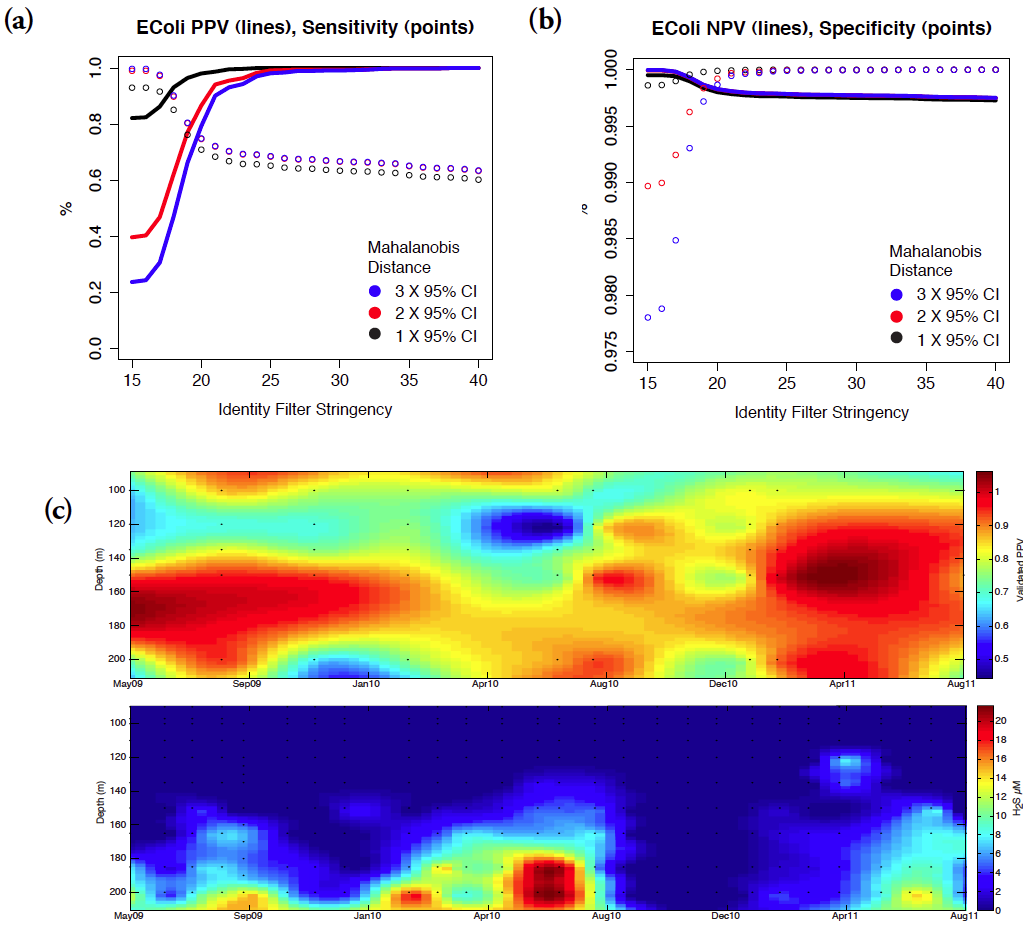
**Figure 2. The SAGEX pipeline.**

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.



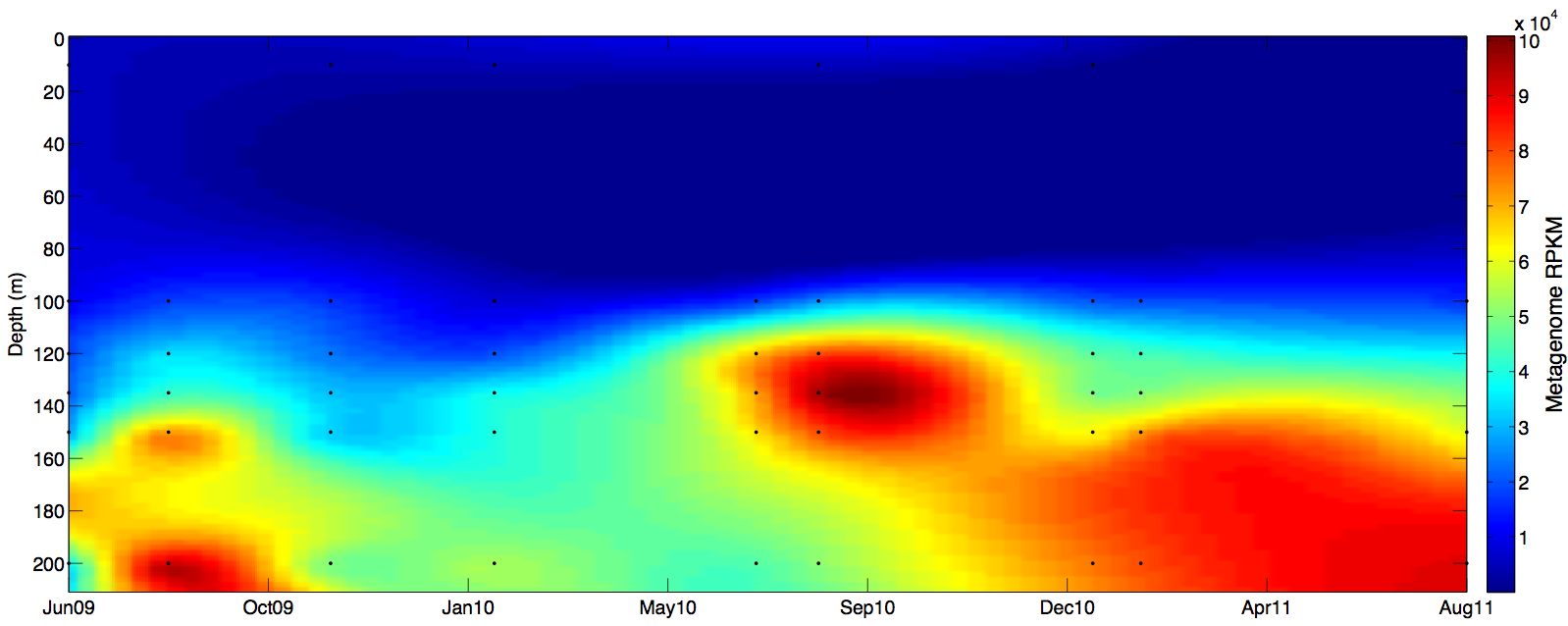
**Figure 2. Genome’s kmer signatures.**

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 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), H2S concentrations. SUP05 is a sulfur-oxidizing gamma proteobacterial group known to inhabit anoxic sulfidic waters up to 4μM H2S, 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.



**Figure 5. SUP05 RPKM recruitment.**

This interpolated surface of average SUP05 RPKM recruitment from metagenomes at varying depths and times demonstrates that SAGEX may be used to infer abundances of relatives to the SAG used to train SAGEX.