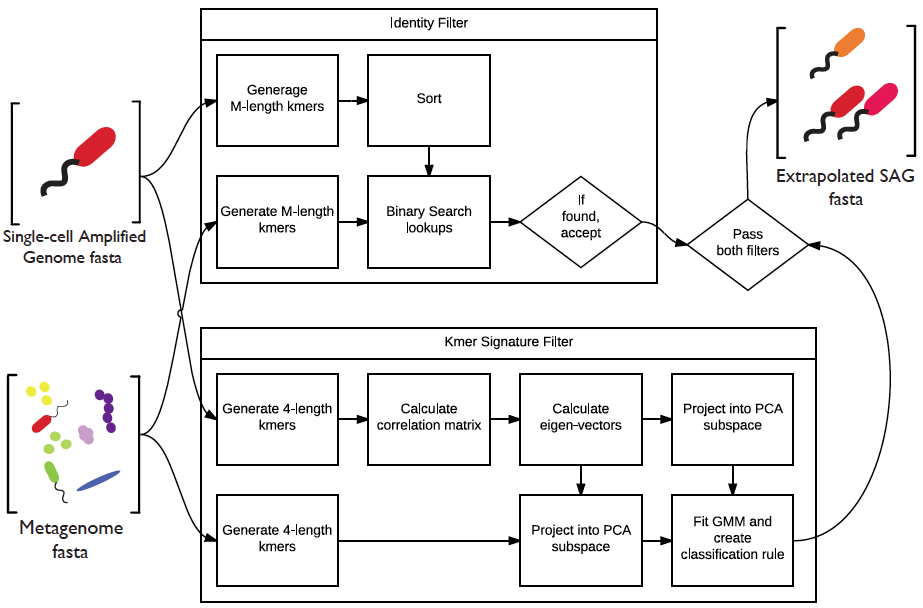


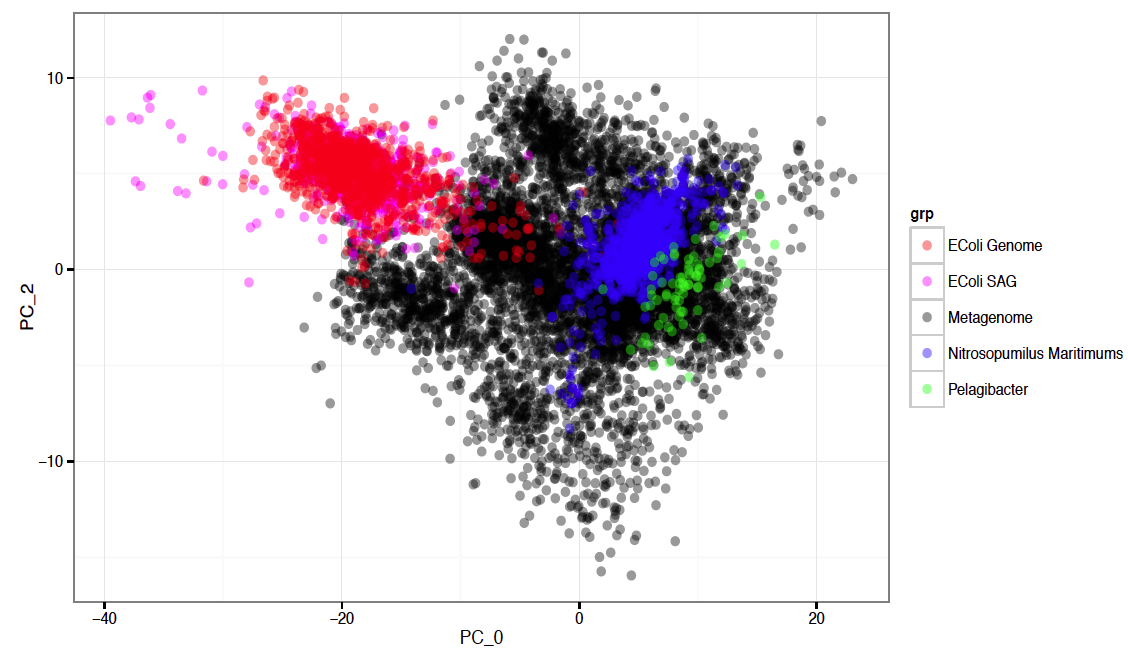
**Figure 1. SAGEX confidently extrapolates SAGs.**

The density of estimated completeness of 129 SUP05 SAGs (black) is compared with the density of estimated completeness of their extrapolated SAGs (red). On high-PPV settings, SAGEX achieved an average of 20% increased completion. SAGEX is capable of greater completion of low-PPV settings. SAGEX recruited from a Saanich Inlet metagenome. Completion was estimated by counting COGs.



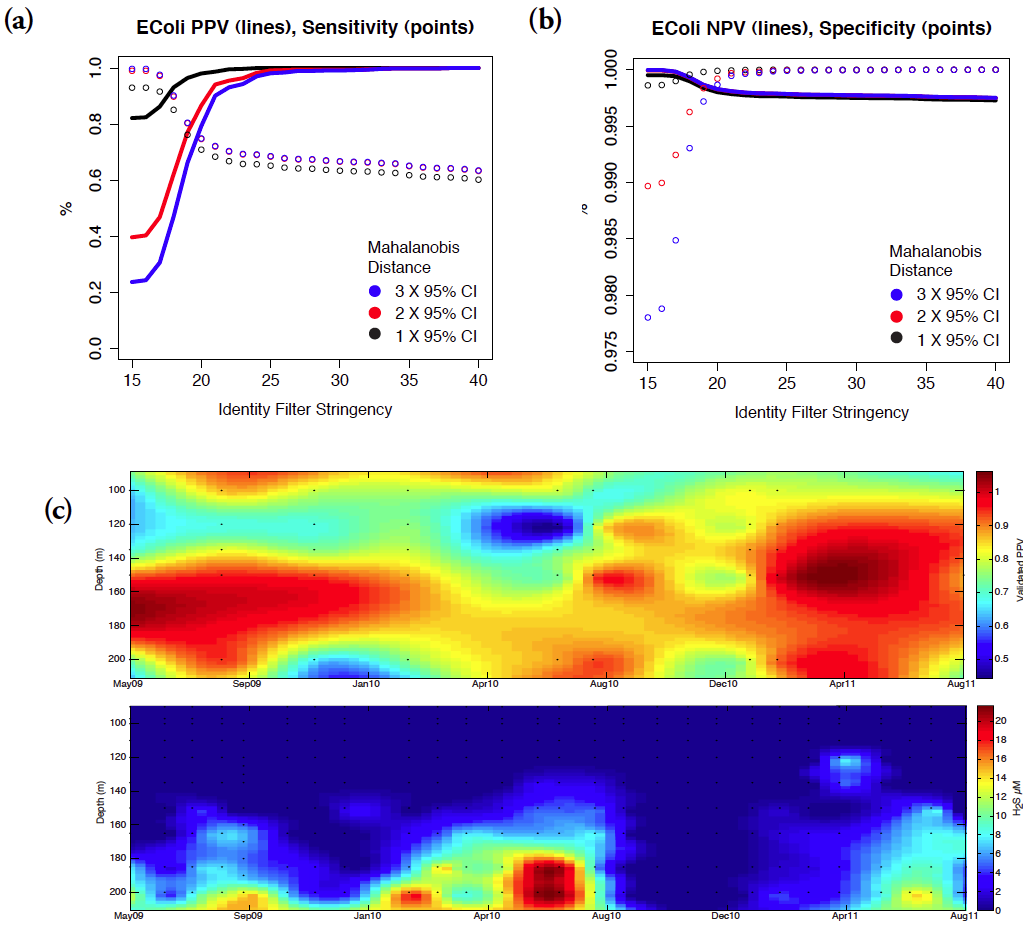
**Figure 2. The SAGEX pipeline.**

The pipeline has two major components: the identity filter and the kmer signature filter. The identity filter requires a contig from the metagenome to share at least an M-length subsequence with any SAG contig before it may be accepted as a hit. The contig must also pass the kmer signature filter by falling within the general region of the SAG’s tetranucleotide after PCA dimension reduction.



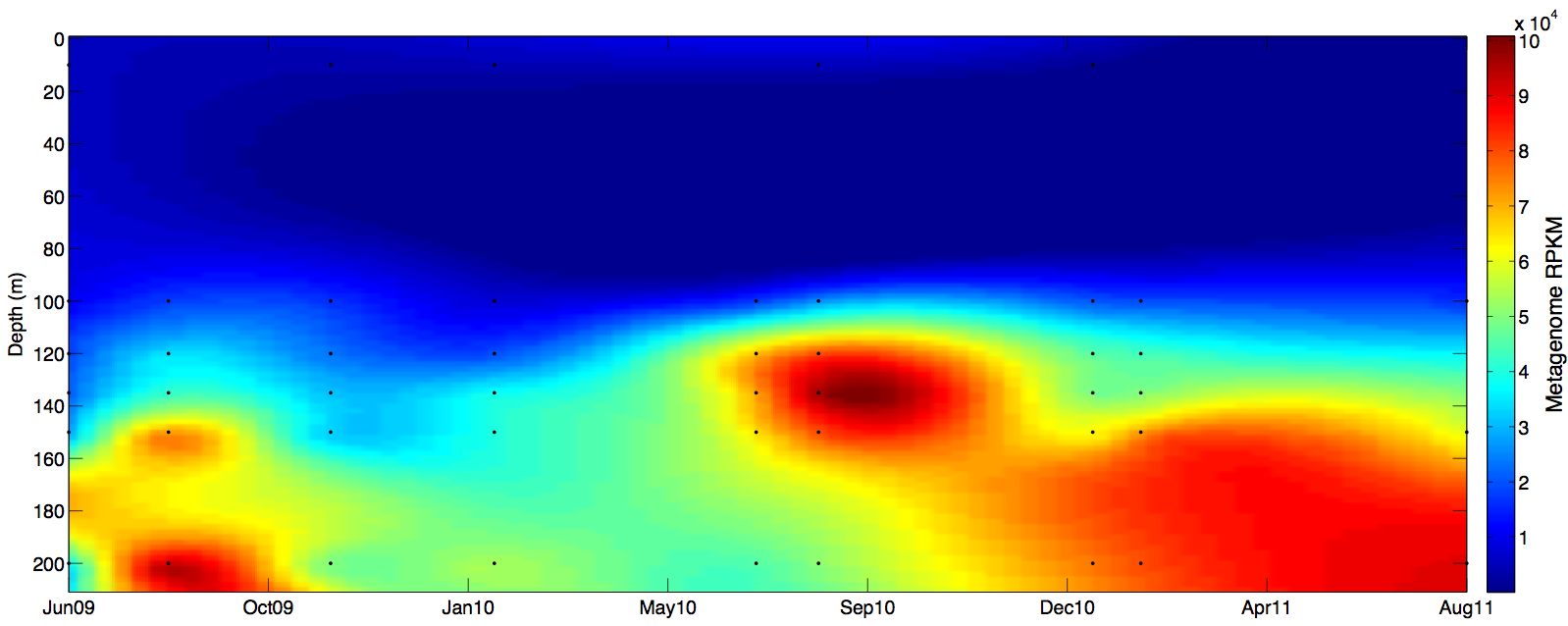
**Figure 3. Genome’s kmer signatures.**

The kmer signature filter is based on the assumption that the SAG and its genome fall within a similar region after their tetranucleotides are transformed via PCA. This idea is verified by showing the overlapping kmers of an EColi MG1655 SAG and genome. The model also assumes the kmers roughly follow the shape of a Gaussian Mixture Model (GMM). Several genomes’ kmer signatures verify this intuition.



**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.**

The 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.