Karp: Accurate and fast taxonomic classification using pseudoalignment

Mark Reppell, John Novembre

Department of Human Genetics

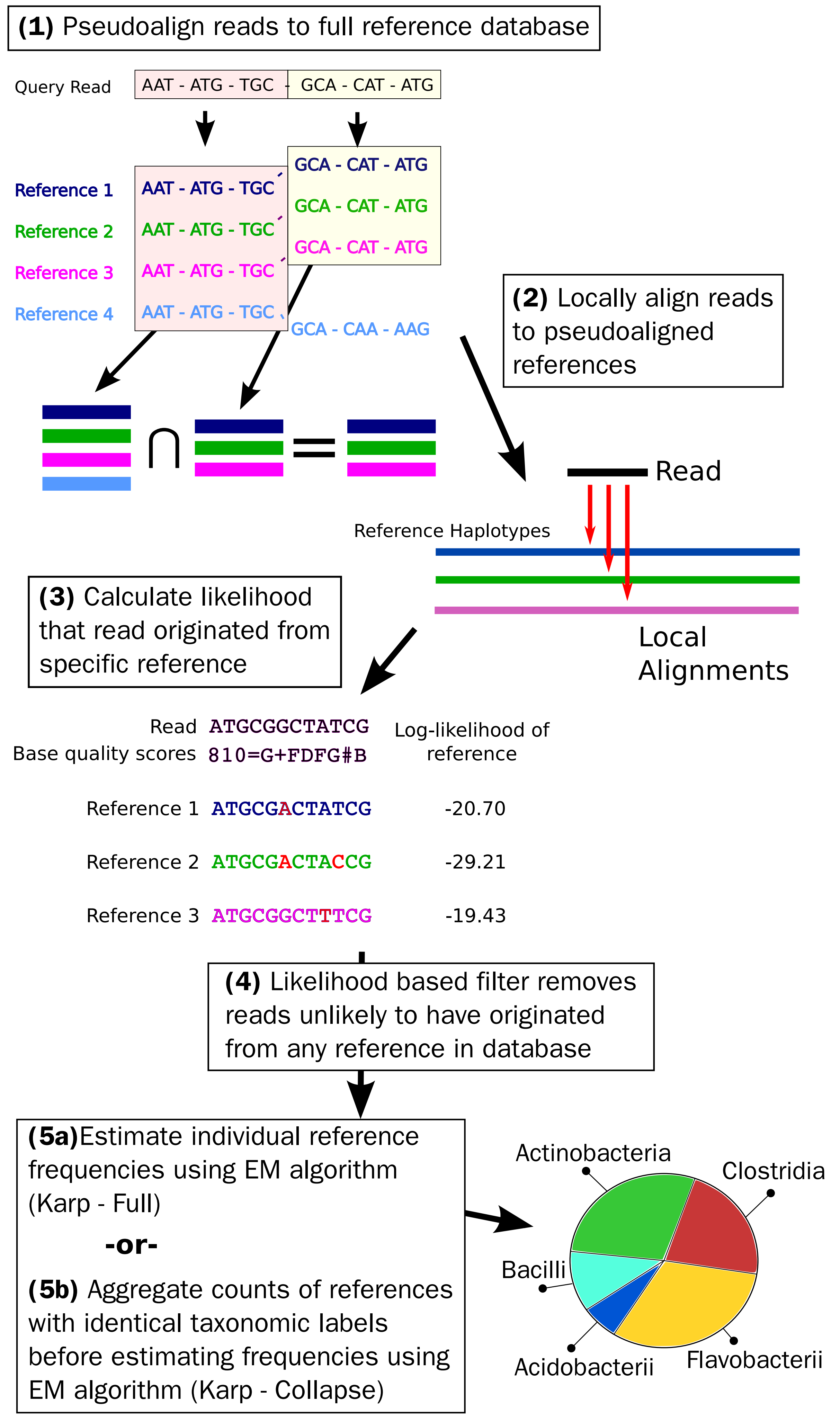
THE UNIVERSITY OF CHICAGO

Software available at https://github.com/mreppell/Karp

Introduction and Background

- Pooled DNA samples arise in many contexts, i.e. microbiome or artificial selection experiments
- Accurately classifying the frequencies of contributors is critical for finding factors associated with pool composition
- Karp is a novel method for classifying the relative frequencies of organisms in a pooled DNA sample that combines pseudoalignment with likelihood-based estimation
- Karp builds on Kessner et al (2013) which introduced a likelihood-based inference procedure for estimating sample composition, and Bray et al's (2016) pseudoalignment, an efficient alignment-free classification method

Overview of Karp



(1) Pseudoalign reads to full reference database

(2) Locally align reads to pseudoaligned references

(3) Calculate likelihood that read originated from specific reference

Read ATGCGGCTATCG
Base quality scores 810=G+FDFG#B
Log-likelihood of reference
Reference 1 ATGCGACTATCG -20.70
Reference 2 ATGCGACTACCG -29.21
Reference 3 ATGCGGCTTTCG -19.43

(4) Likelihood based filter removes reads unlikely to have originated from any reference in database

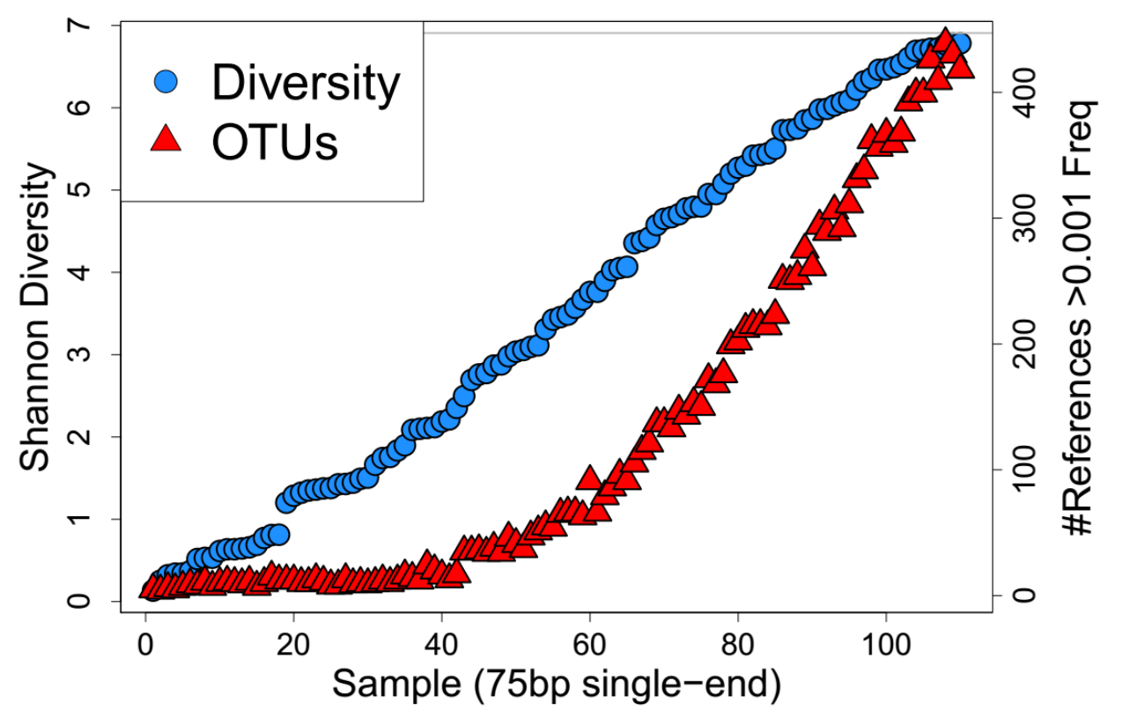
(5a) Estimate individual reference frequencies using EM algorithm (Karp - Full)

-or-

(5b) Aggregate counts of references with identical taxonomic labels before estimating frequencies using EM algorithm (Karp - Collapse)

16S rRNA Simulations

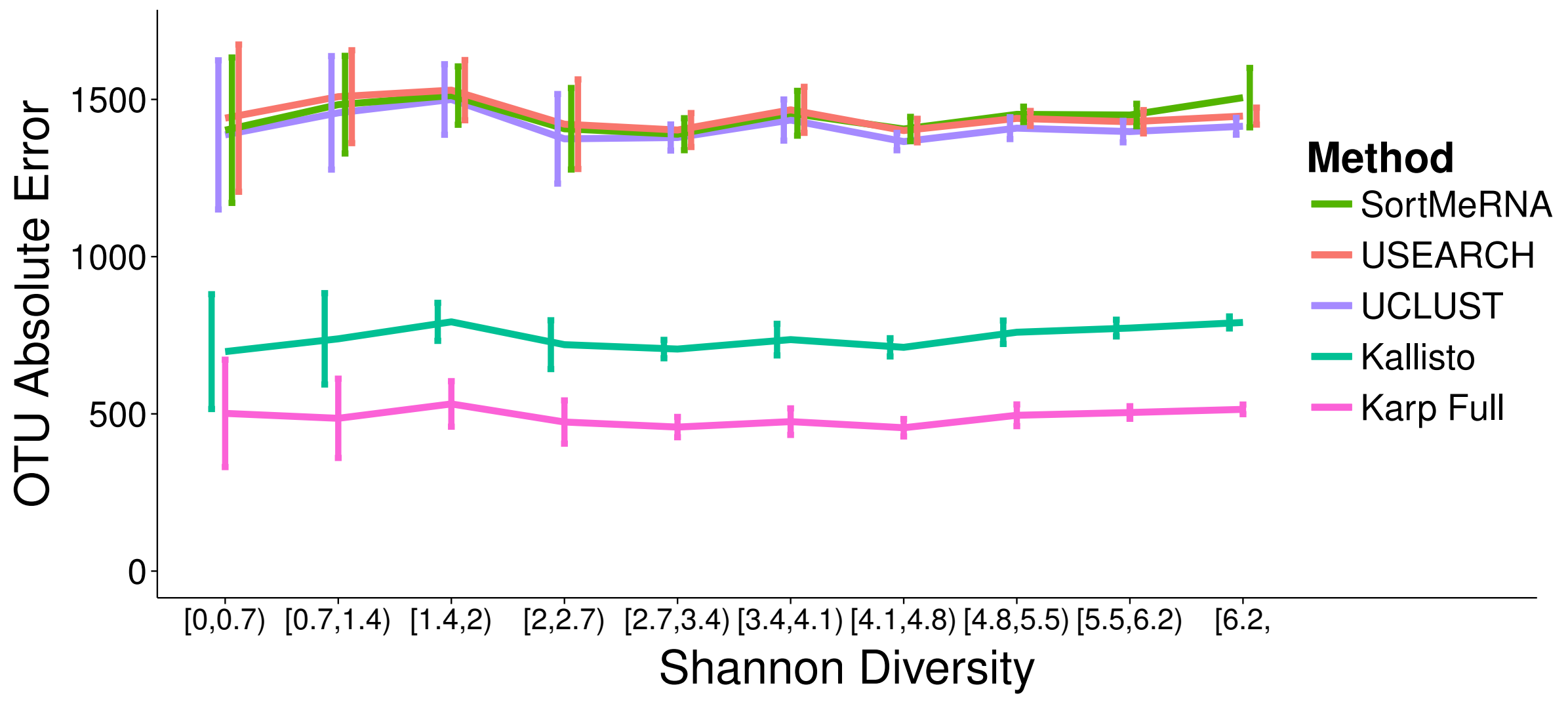
- We simulate samples comprised of 1,000,000 total reads drawn from 1,000 different GreenGenes v13.8 references
- Each read is a partial copy of a reference haplotype with errors introduced according to base quality scores
- 3 scenerios with different empirical error distributions: 75bp single-end samples, 151bp paired-end samples, and 301bp paired-end samples
- Frequency of references in each sample chosen to vary across range of possible Shannon diversities (entropies)



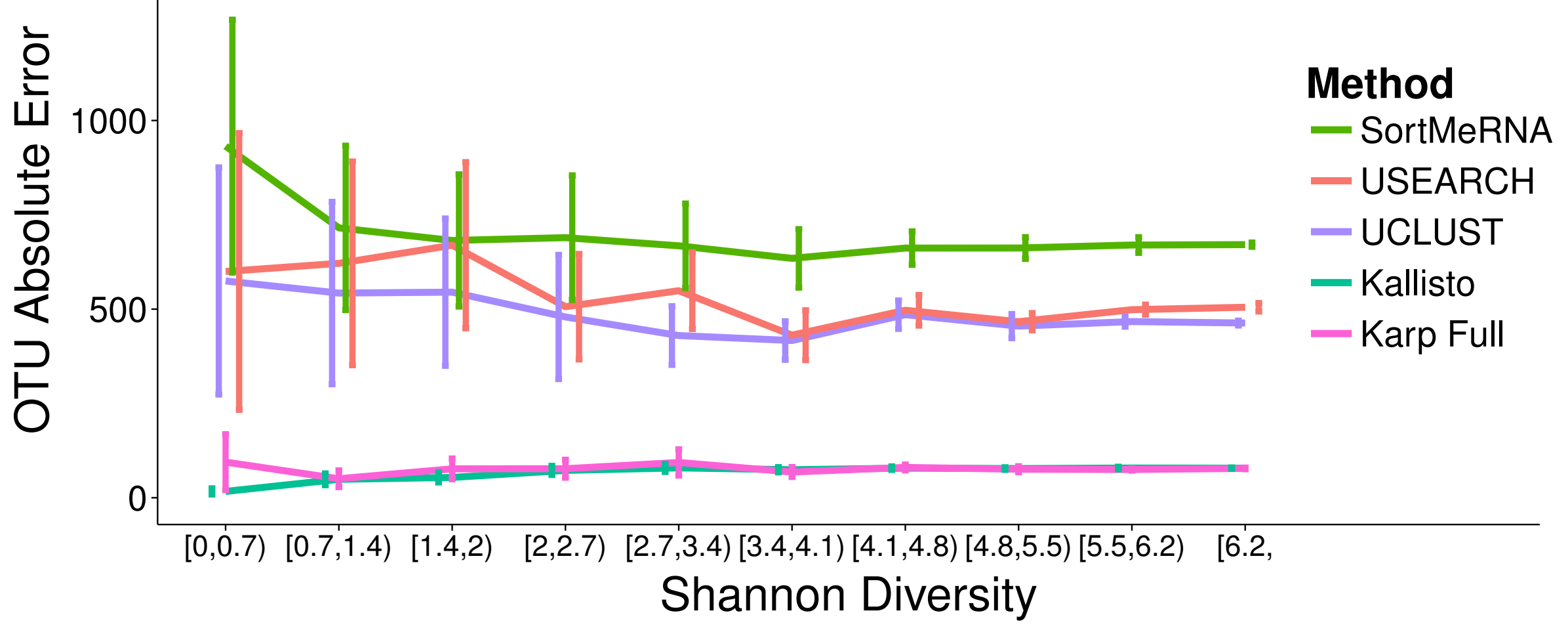
- Error metric = scaled summed absolute difference between true and estimated counts for each reference

Simulation Results

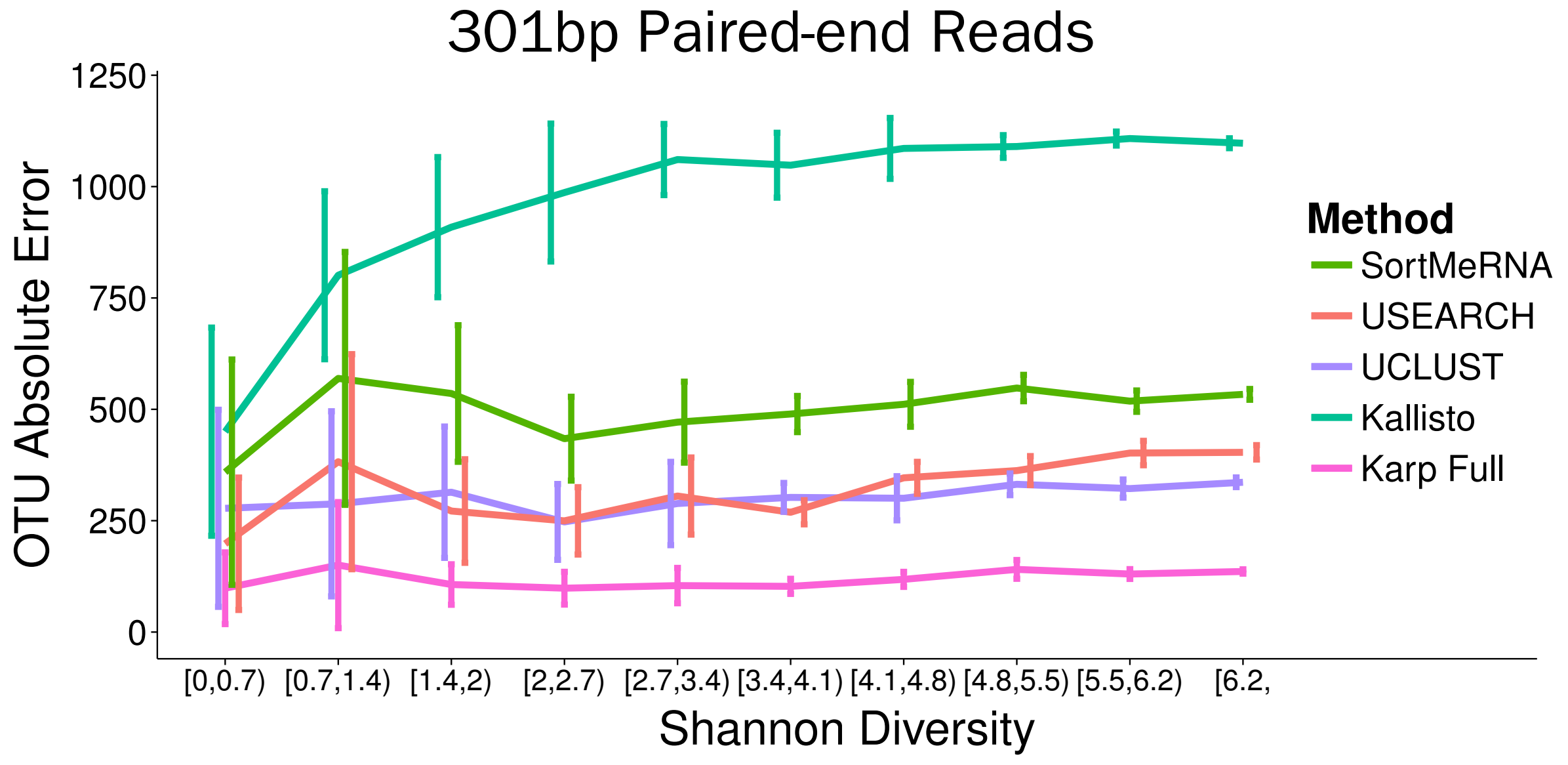
75bp Single-end Reads



151bp Paired-end Reads



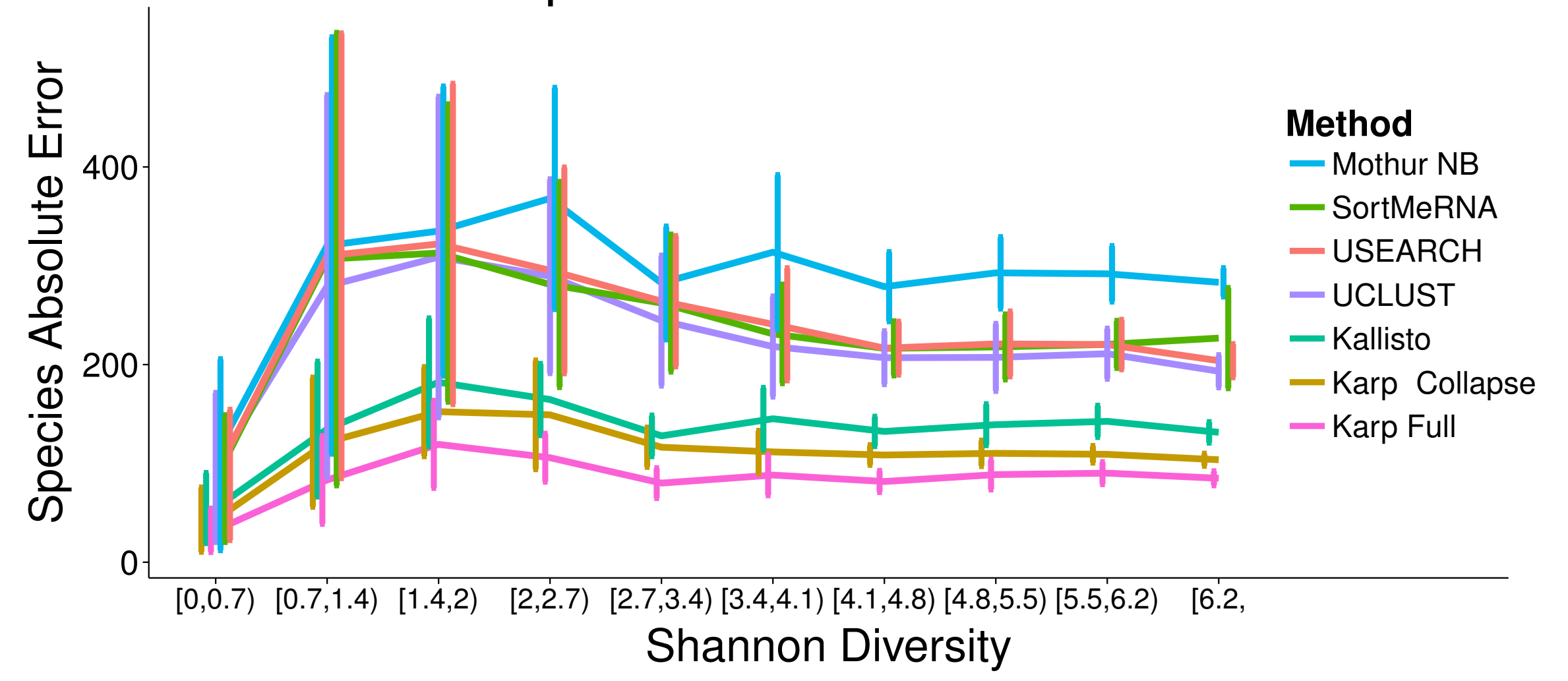
301bp Paired-end Reads



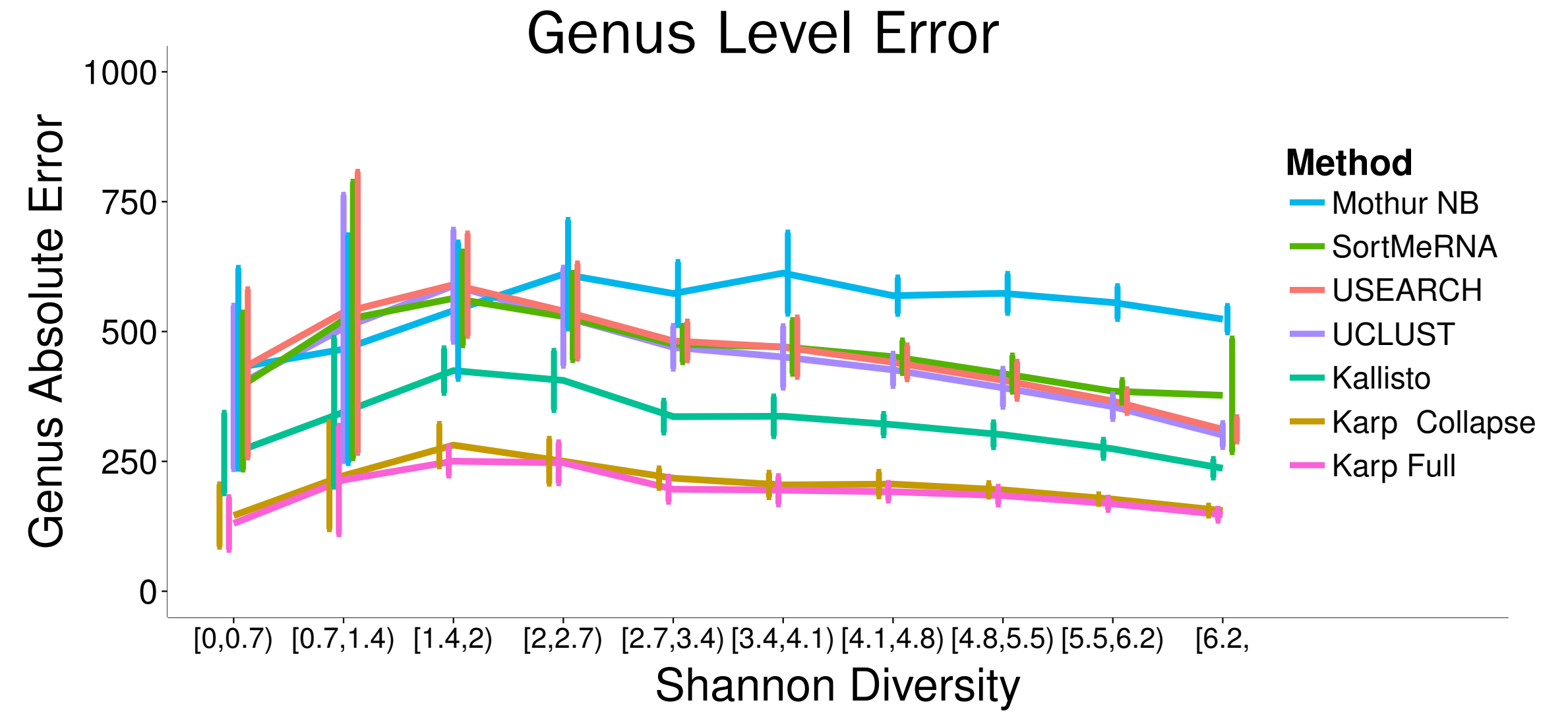
Higher Order Taxonomy

- Often researchers are interested in sample composition at the level of species, genus, or other higher order taxonomic classification, rather than individual reference seqeuences.
- Below we compare the accuracy of competing methods at higher orders in 75bp single-end read samples

Species Level Error



Genus Level Error



Technical Tricks

- Karp is >100x faster than Harp (Kessner 2013) while maintaining inference accuracy. To achieve this Karp uses:
  - Pseudoaligning to avoid >99.9% of alignments
  - Highly parallelized functions
  - Sparse data structure for encoding likelihoods
  - SquareEM for faster EM convergence (Varadhan 2004)
- For taxonomic level estimation rather than individual references, Karp Collapse mode further improves speed

Computational Performance

- Time and memory to classify 1,000,000 75bp single-end reads against full GreenGenes v13.8 database using 12 cores:

Method	Time (Minutes)	Maximum Memory (GB)
Karp Full	179.5	10
Karp Collapse	37.5	10
Kallisto	5.1	10
QIIME UCLUST	84.3	4
QIIME USEARCH	9.5	4
QIIME SortMeRNA	35.8	4
Mothur Naive Bayes**	502.4	16

**to keep memory <16GB limited to 4 cores

We would like to thank the NIH, University of Chicago Resource Computing Center, and the Ober and Gilbert Labs for their support of this project

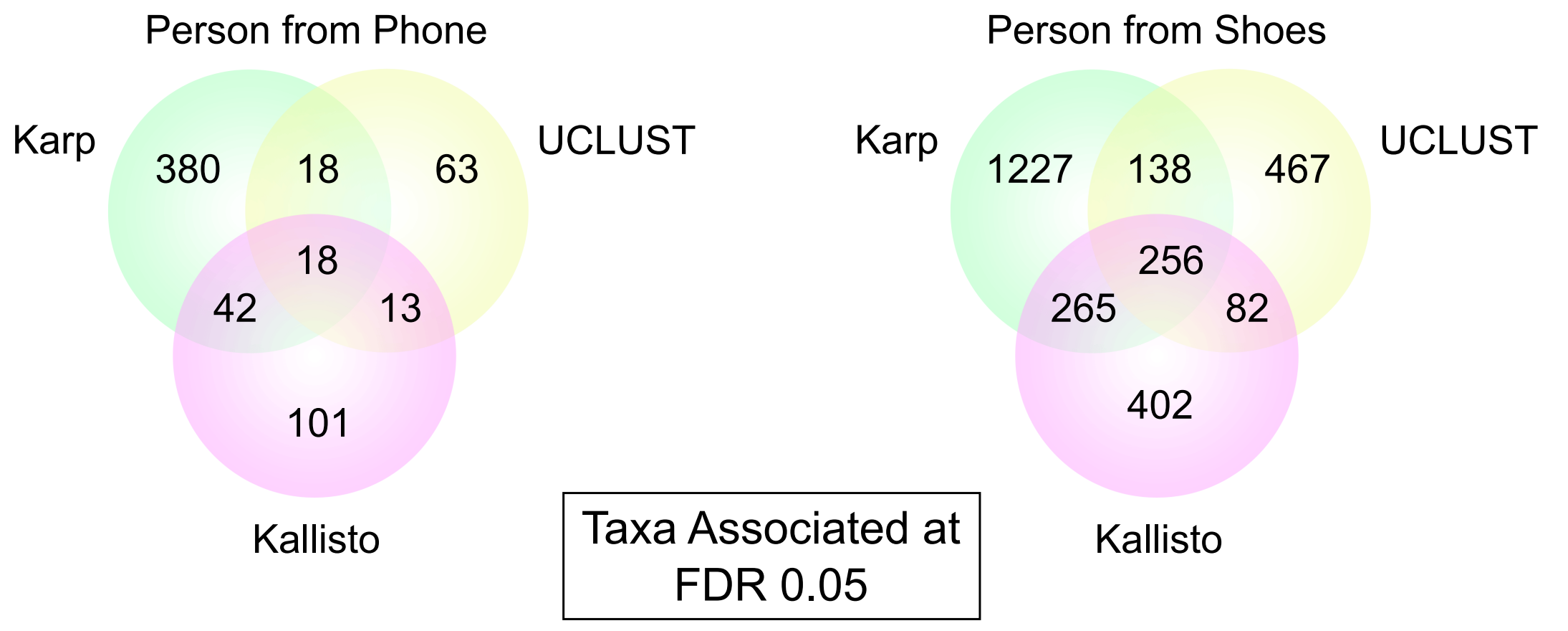
Real 16S Sequence Data

- From Lax et al (2015), samples collected from shoes and phones of two study participants every hour for two 12 hour periods
- V4 region of 16S rRNA gene amplified and sequenced, 151bp single-end reads
- We estimated sample composition with Karp, Kallisto and UCLUST, normalized to equal read depth, then used randomForest method to classify and measured error

Method	OTU Classification Error			
	Person from Phone	Person from Shoe	Phone Side from Person 1	Phone Side from Person 2
Baseline	0.495	0.488	0.481	0.471
Karp	**0.029**	**0.001**	**0.292**	**0.245**
Kallisto	0.036	0.003	0.302	**0.245**
UCLUST	0.04	0.004	0.398	0.278

Method	Genus Classification Error			
	Person from Phone	Person from Shoe	Phone Side from Person 1	Phone Side from Person 2
Baseline	0.495	0.488	0.481	0.471
Karp – Full	0.048	0.005	0.371	0.321
Karp – Collapse	0.046	0.004	**0.363**	**0.287**
Kallisto	0.053	0.006	0.387	0.311
UCLUST	**0.044**	**0.002**	0.392	0.3

- In our classified samples we also tested for differences in the mean abundance of taxa between individuals



Conclusions

- In simulations Karp's estimated compositions are the closest to the true sample compositions across scenerios and taxa level
- In the real 16S data Karp's estimates often result in lower error when clustering samples, suggesting more power to detect important differences between samples
- While very fast, and nearly as accurate as Karp, Kallisto struggles with longer reads. Kallisto uses a stricter definition of pseudoaligning, leading to fewer matches and greater errors with longer reads