Statistical simulations guide construction of a fast and accurate metagenomic annotation pipeline

Stephen Nayfach
iSEEM2 Call
August 6, 2014

Motivation

- No clear consensus on best methods for functionally annotating unassembled metagenomic data
- No end-to-end statistical evaluation of metagenomic annotation pipelines
- Unclear if commonly used annotation pipelines (MG-RAST, IMG/M) produce accurate estimates of functional abundances
- Conclusions made from previous analysis looking only at TPR/PPV may not hold when looking at more meaningful performance metrics

Outline

- **Methods: overview of simulation framework**
- Read translation
- Score/E-value thresholds
- Search algorithm
- Sequencing depth
- MG-RAST benchmark

Methods: Simulation Framework

Mock Communities
- Species abundance distributions modeled from gut microbiome samples
- Genomes annotated with various databases: Sfams, Pfams, KOs, Metacyc
- Expected protein family abundance distributions

Simulated metagenomic libraries
- Libraries of varying length: 1 community; 27 libraries; 25-500 bp; 1% uniform error rate; ~1.3M reads/library
- Libraries from different communities: 10 communities; 101 bp reads; Illumina error model; 1.5M reads/library
- Deep sequenced library: 1 community; 101 bp reads; Illumina error model; 100M reads

Read Translation
- 6FT
- 6FT-split +/- length filtering
- Gene finders (Prodigal, FGS, MGM)

Search
- RAPsearch2, HMMer3, BLAST

Read classification
- Best-hit by read or by ORF
- Score/E-value threshold
- Rank of taxonomic exclusion (species, genus, etc.)

Estimation of relative abundance
- +/- target length normalization
- Count statistic (# of classified reads or # of aligned residues from classified reads)

Performance evaluation
- L1 distance between predicted and expected relative abundance distributions

Methods: Default parameters

Mock Communities

- One community
- Sfams database

Simulated metagenomic libraries

- Libraries of varying length: 1 community; 27 libraries; 25-500 bp; 1% uniform error rate; ~1.3M reads/library

Read Translation

- 6FT-split

Search

- RAPsearch2

Read classification

- Best-hit by read
- Optimal score threshold
- Taxon-level exclusion

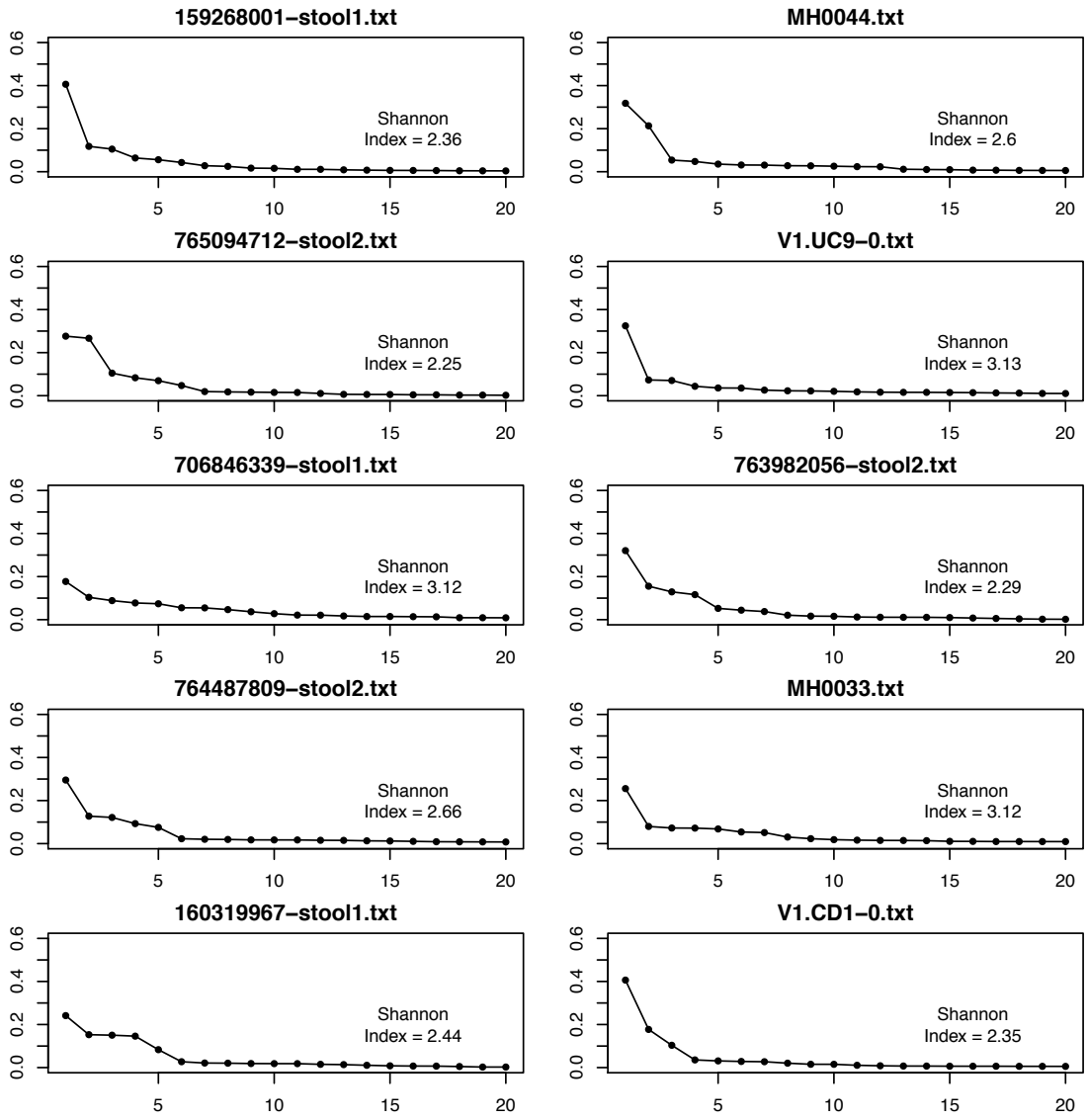
Estimation of relative abundance

- Gene length normalization
- Counting # of aligned residues from classified reads

Performance evaluation

- L1 distance between predicted and expected relative abundance distributions

Methods: Species abundance distributions of 10 mock communities



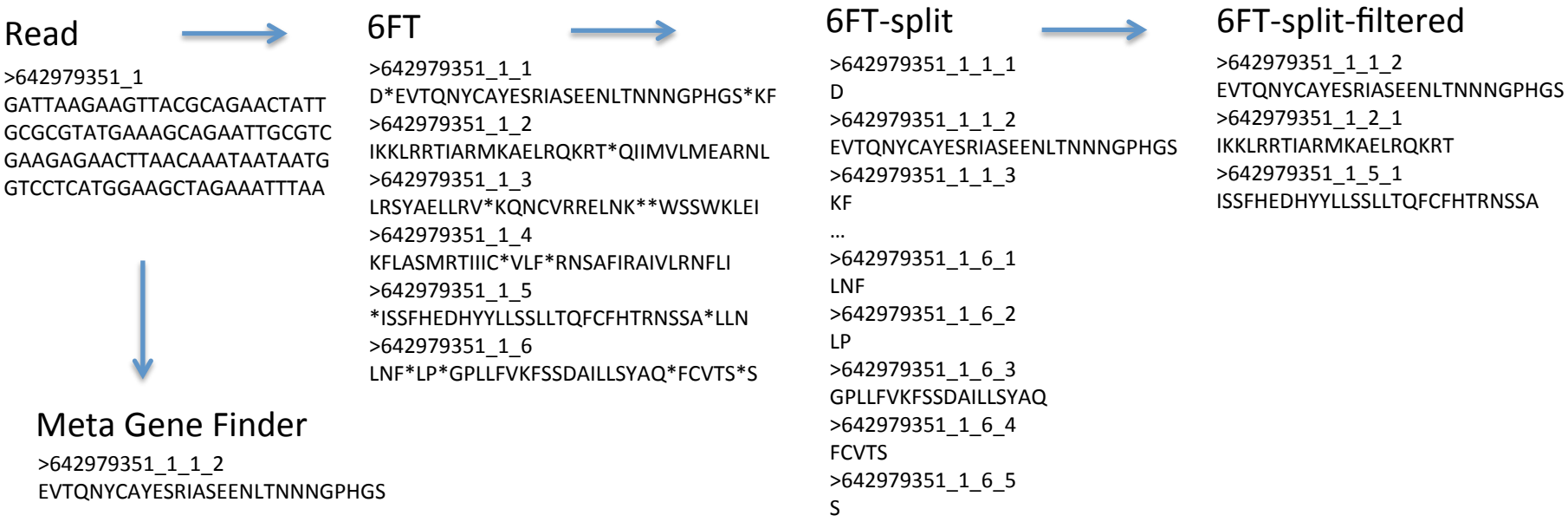
Example output from simulation pipeline

community	read_length	orf_type	minorf	fam_name	db_name	thresh	taxlevel	classlevel	countstat	norm_gene_len	L1	L2	R2
1	100	orf_split	0	sfams	rapsearch	20	species	read	aln	F	0.32	0.03	0.75
1	100	orf_split	0	sfams	rapsearch	21	species	read	aln	F	0.32	0.03	0.75
1	100	orf_split	0	sfams	rapsearch	22	species	read	aln	F	0.32	0.03	0.75
1	100	orf_split	0	sfams	rapsearch	23	species	read	aln	F	0.32	0.03	0.75
1	100	orf_split	0	sfams	rapsearch	24	species	read	aln	F	0.32	0.03	0.75
1	100	orf_split	0	sfams	rapsearch	25	species	read	aln	F	0.32	0.03	0.75
1	100	orf_split	0	sfams	rapsearch	26	species	read	aln	F	0.32	0.03	0.75
1	100	orf_split	0	sfams	rapsearch	27	species	read	aln	F	0.31	0.03	0.75
1	100	orf_split	0	sfams	rapsearch	28	species	read	aln	F	0.30	0.03	0.75
1	100	orf_split	0	sfams	rapsearch	29	species	read	aln	F	0.29	0.03	0.74
1	100	orf_split	0	sfams	rapsearch	30	species	read	aln	F	0.29	0.03	0.74
1	100	orf_split	0	sfams	rapsearch	31	species	read	aln	F	0.29	0.03	0.74
1	100	orf_split	0	sfams	rapsearch	32	species	read	aln	F	0.29	0.03	0.74
1	100	orf_split	0	sfams	rapsearch	33	species	read	aln	F	0.29	0.03	0.74
1	100	orf_split	0	sfams	rapsearch	34	species	read	aln	F	0.29	0.03	0.74

Outline

- Methods: overview of simulation framework
- **Read translation**
- Score/E-value thresholds
- Search algorithm
- Sequencing depth
- MG-RAST benchmark

Read translation

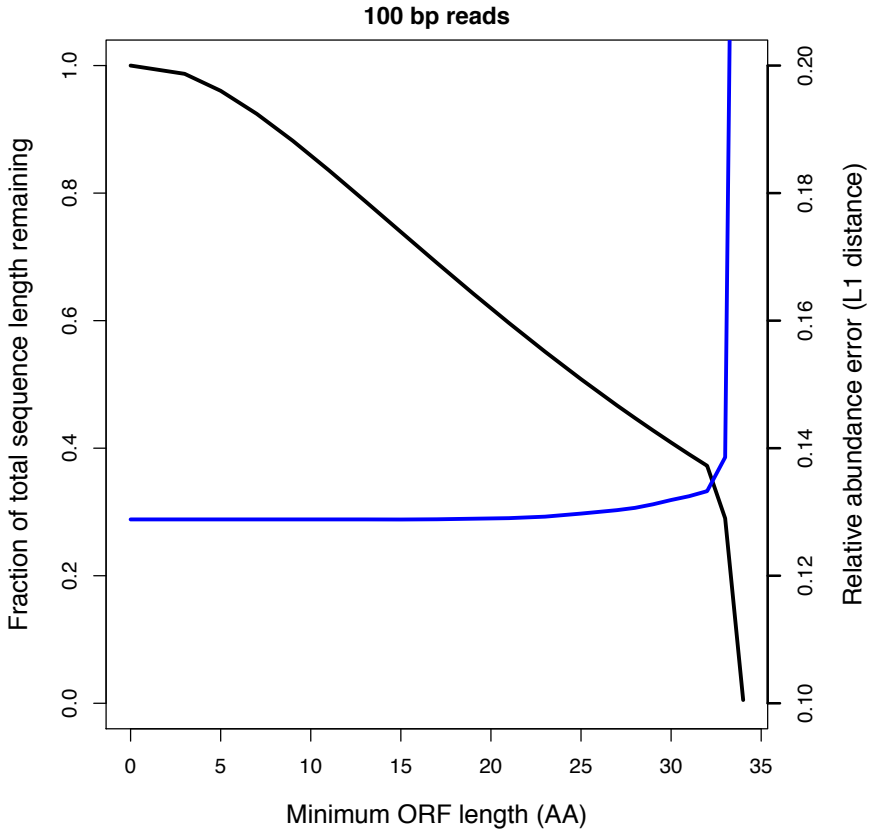


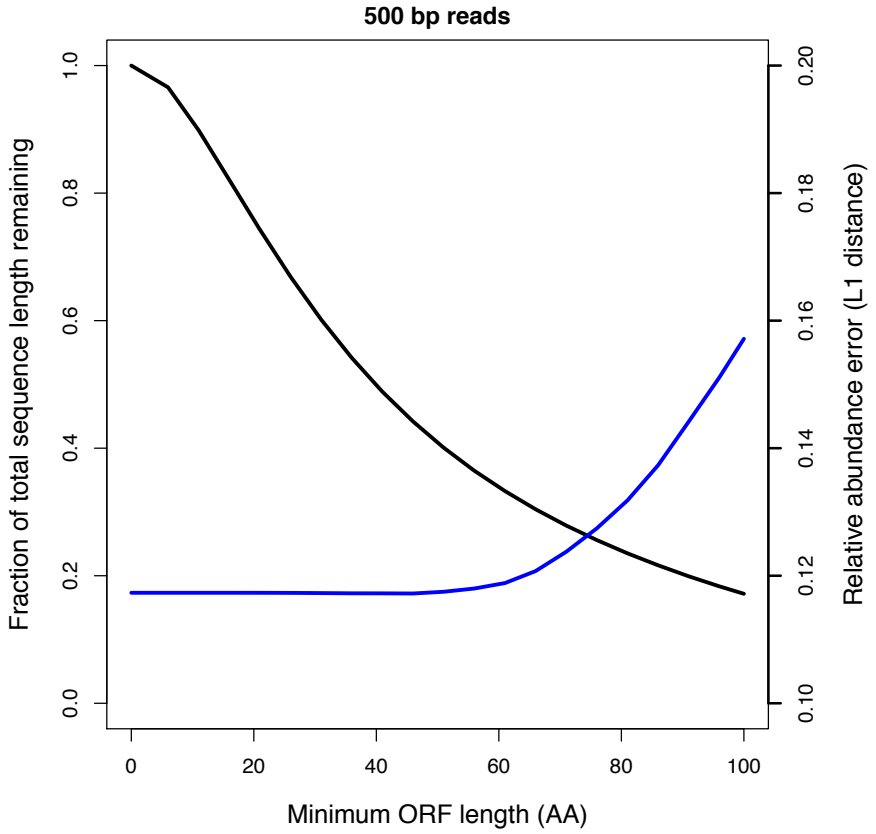
ORF ‘splitting’ does not reduce downstream

	L1 Error for Protein Family Abundance Estimates	
Read length	6FT	6FT-split
50	0.166	0.167
60	0.153	0.154
70	0.144	0.144
80	0.137	0.138
90	0.132	0.132
100	0.129	0.129
150	0.118	0.118
250	0.112	0.112
500	0.114	0.117

	% Increase in Error for Protein Family Abundance Estimates Relative to 6FT	
Read length	6FT	6FT-split
50	0.0	0.3
60	0.0	0.5
70	0.0	0.4
80	0.0	0.2
90	0.0	-0.2
100	0.0	0.1
150	0.0	0.1
250	0.0	0.0
500	0.0	2.7

Filtering short ORFs reduces data volume without impacting relative abundance estimates





ORF prediction dramatically reduces data volume at a small cost to accuracy

L1 Error for Protein Family Abundance Estimates				
Read length	6FT	FGS	MGM	Prodigal
50	0.166	n/a	n/a	n/a
60	0.153	n/a	0.311	0.342
70	0.144	0.259	0.272	0.166
80	0.137	0.153	0.245	0.155
90	0.132	0.146	0.225	0.148
100	0.129	0.141	0.209	0.142
150	0.118	0.128	0.162	0.128
250	0.112	0.119	0.131	0.117
500	0.114	0.124	0.121	0.117

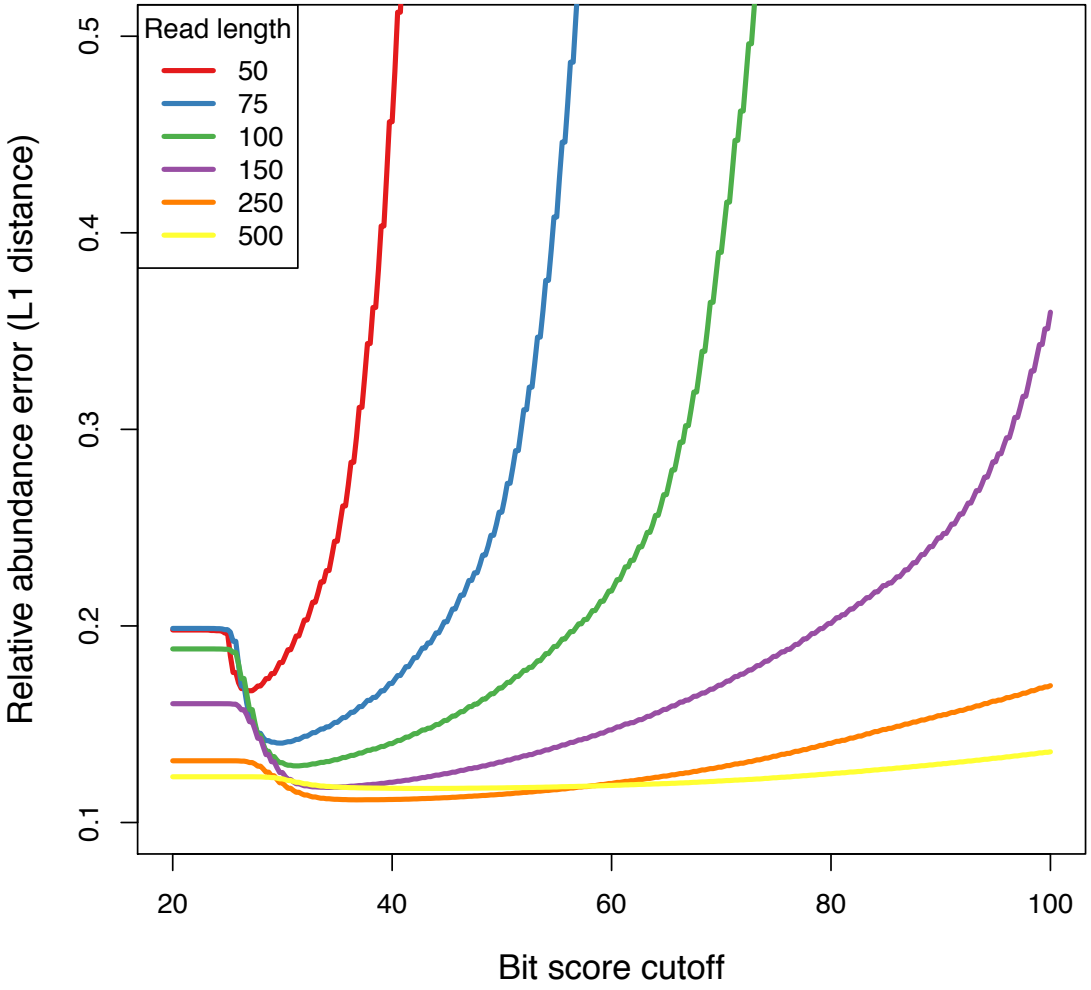
% Increase in Error for Protein Family Abundance Estimates Relative to 6FT				
Read length	6FT	FGS	MGM	Prodigal
50	0.0	n/a	n/a	n/a
60	0.0	n/a	103.204	123.325
70	0.0	80.1	89.3	15.4
80	0.0	11.4	78.5	12.9
90	0.0	10.4	70.2	12.2
100	0.0	9.3	62.2	10.6
150	0.0	8.9	37.2	8.4
250	0.0	6.9	17.8	4.7
500	0.0	8.2	5.9	2.8

	% reduction of sequence volume relative to 6-frame translation		
Read length	FGS	MGM	Prodigal
100	84.6	92.7	85.8
150	84.5	89.6	85.9
250	84.5	88.0	86.2
500	84.5	87.3	86.5

Outline

- Methods: overview of simulation framework
- Read translation
- **Score/E-value thresholds**
- Search algorithm
- Sequencing depth
- MG-RAST benchmark

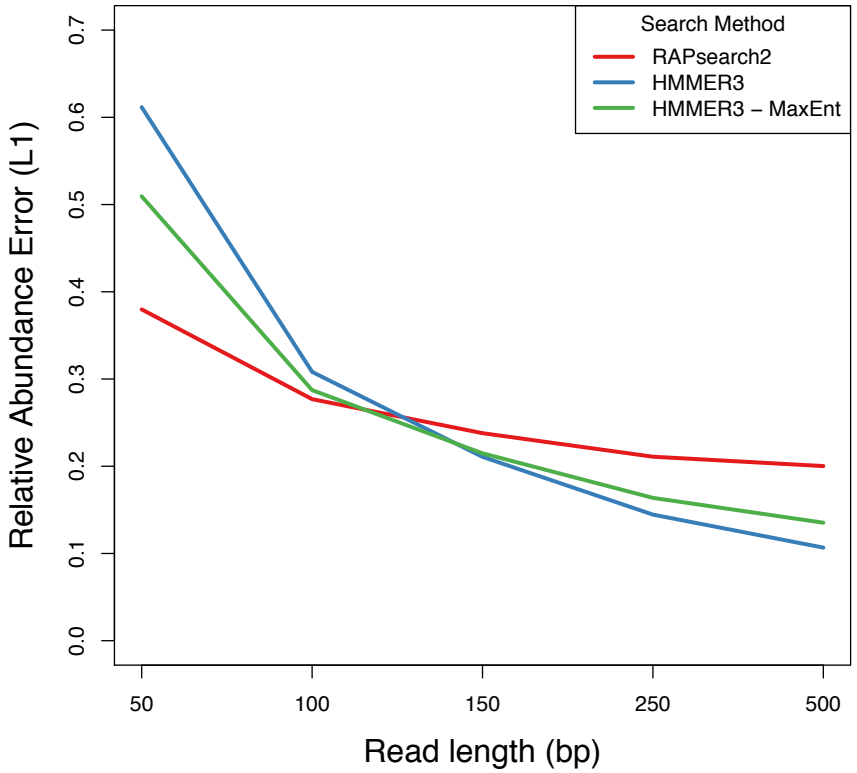
Precise bit score thresholds are necessary for accurate estimates of protein family abundance for short-read libraries



Outline

- Methods: overview of simulation framework
- Read translation
- Score/E-value thresholds
- **Search algorithm**
- Sequencing depth
- MG-RAST benchmark

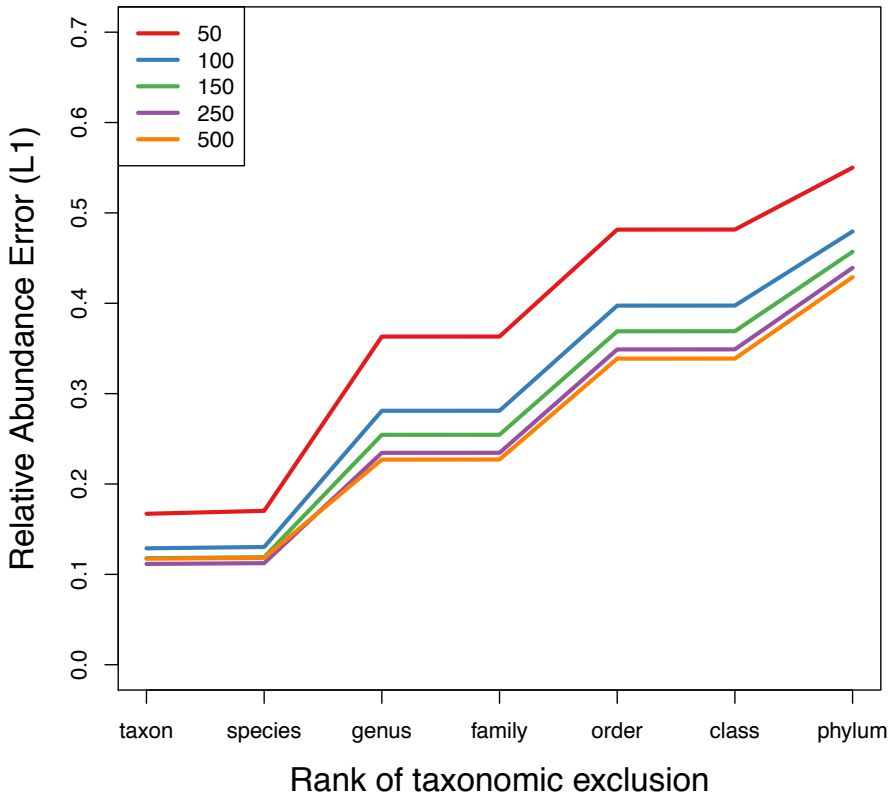
BLAST (i.e. pairwise search) better for short reads
HMMER (i.e. profile search) better for long reads

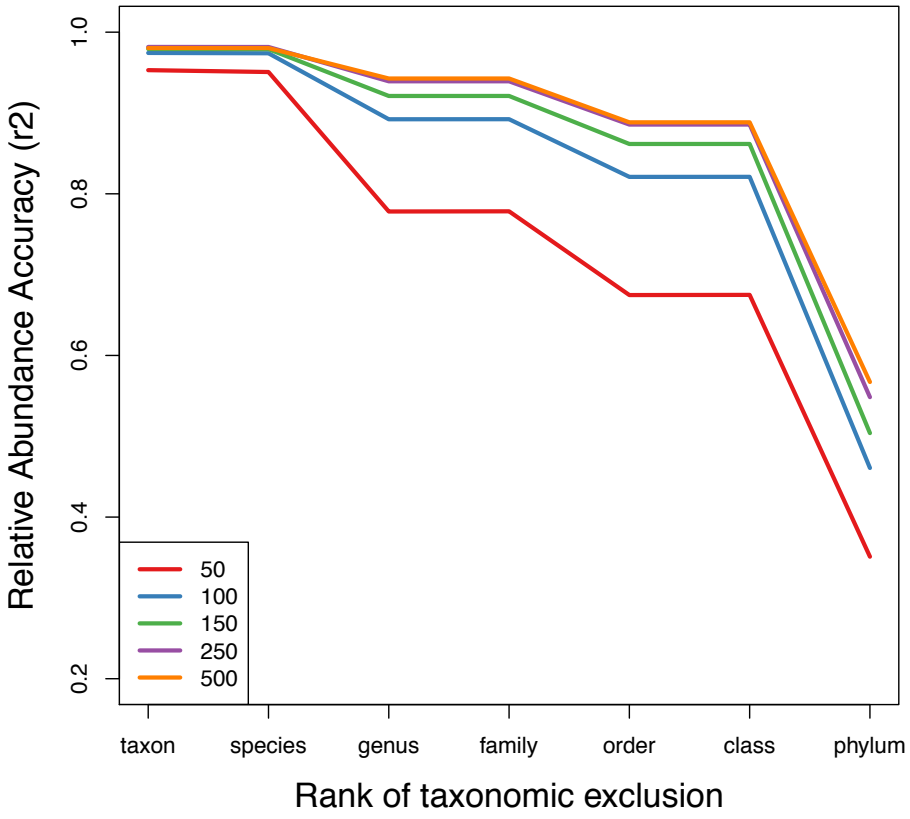


	Relative abundance error (L1 distance)		
	blast	rapsearch	hmmsearch
50	**0.37**	0.38	0.61
100	**0.25**	0.28	0.31
150	**0.20**	0.24	0.21
250	0.16	0.21	**0.14**
500	n/a	0.20	**0.11**

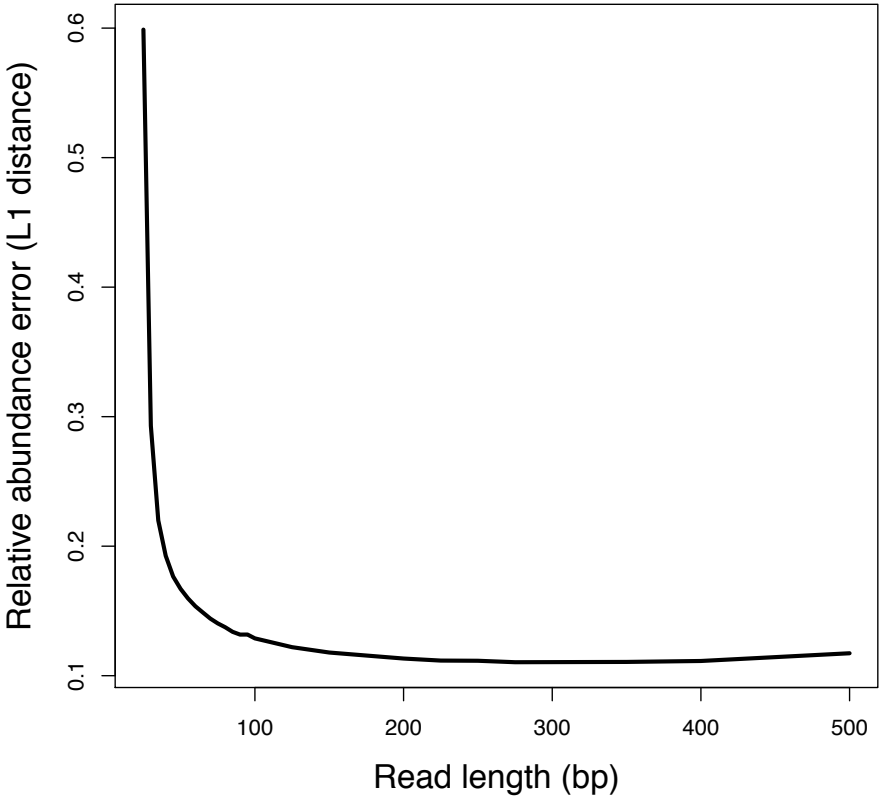
- Results shown are for **Pfam**
- For Sfams, HMMER was much worse than RAPsearch2 across all read lengths

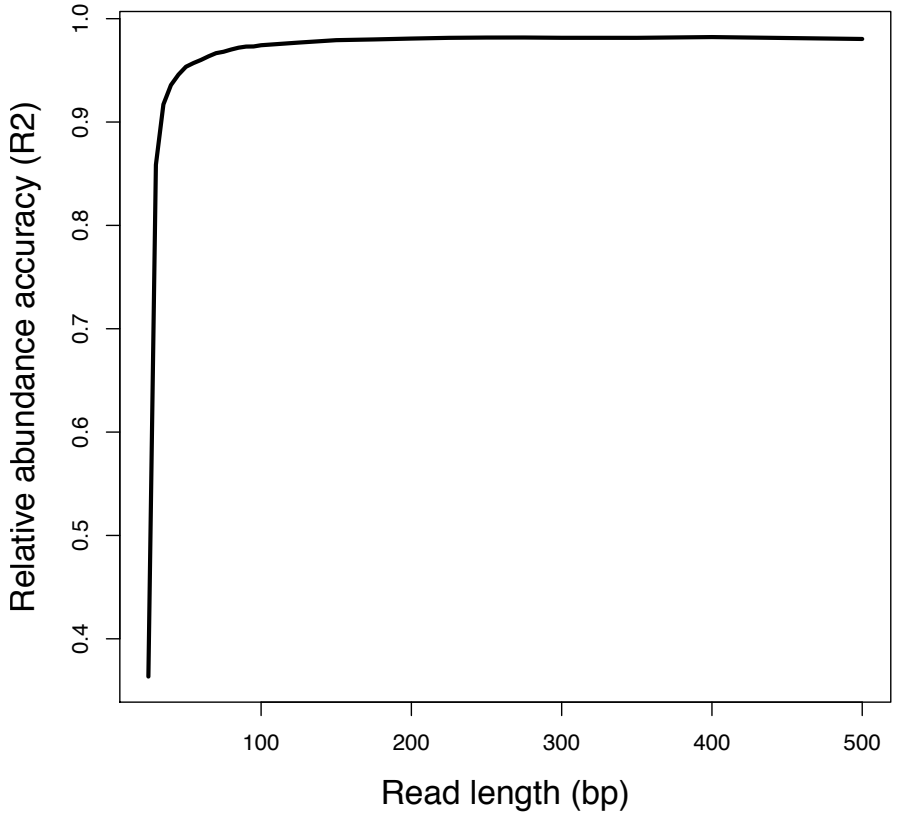
Effect of ‘taxonomic exclusion’





Effect of read length

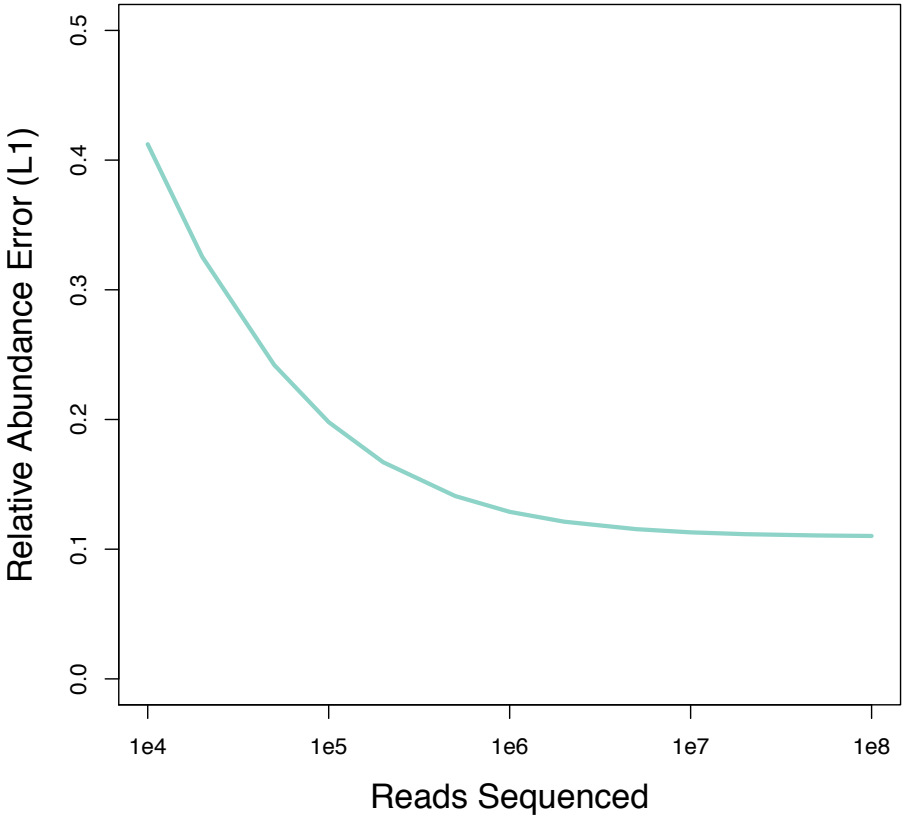


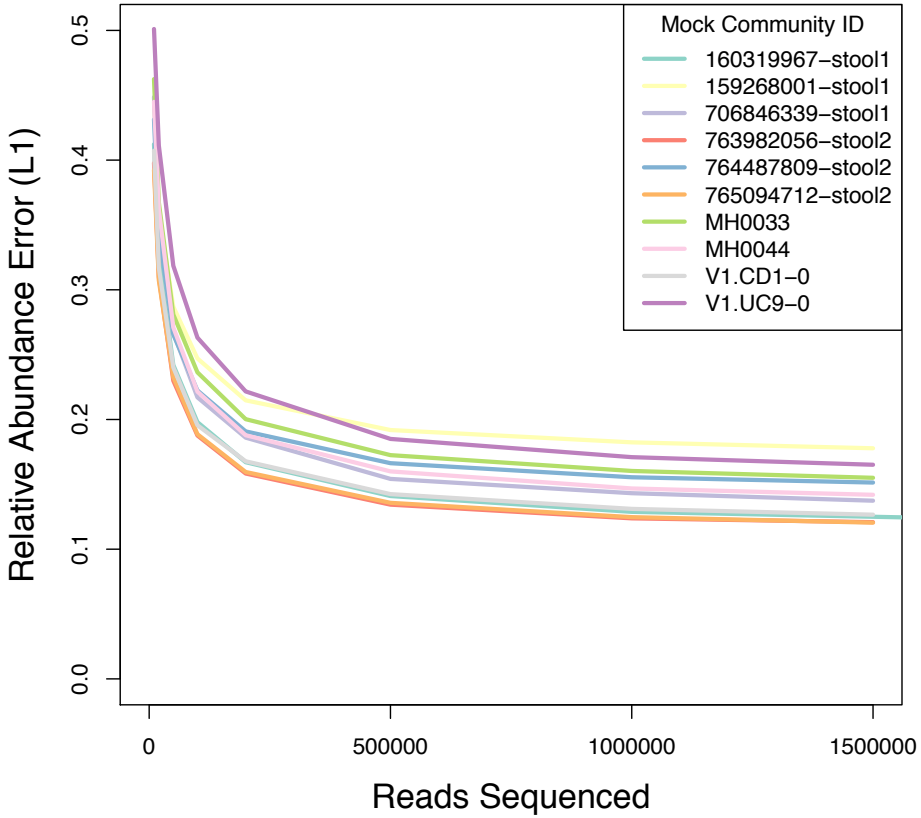


Outline

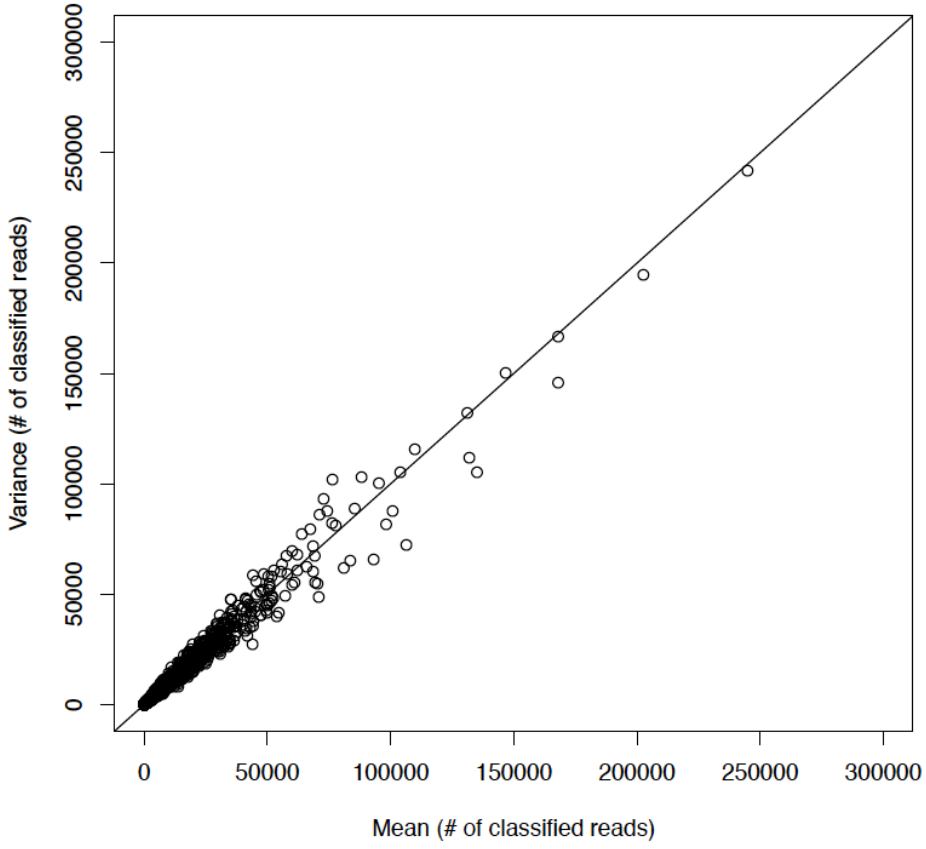
- Methods: overview of simulation framework
- Read translation
- Score/E-value thresholds
- Search algorithm
- **Sequencing depth**
- MG-RAST benchmark

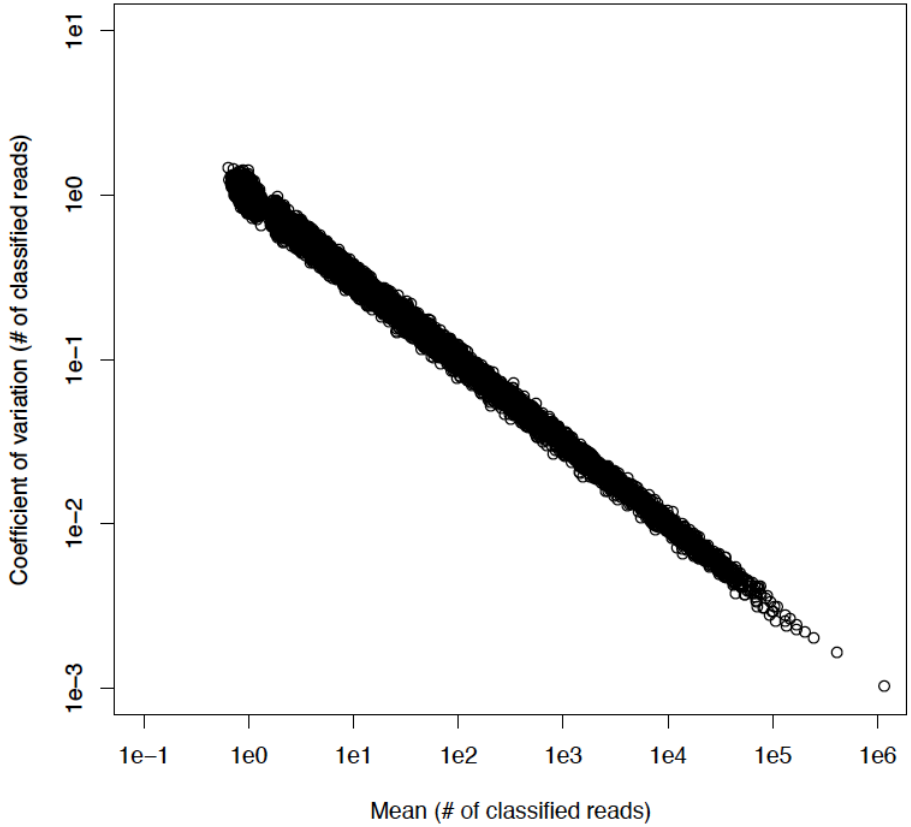
Effect of sequencing depth on accurate estimates of within-sample functional abundance



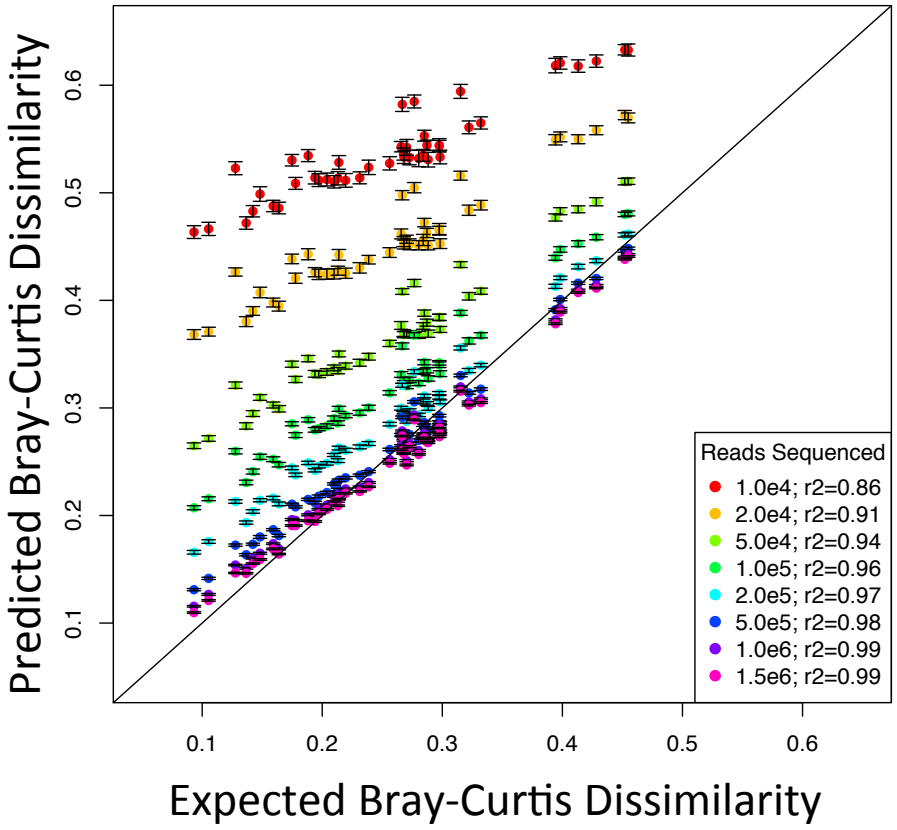


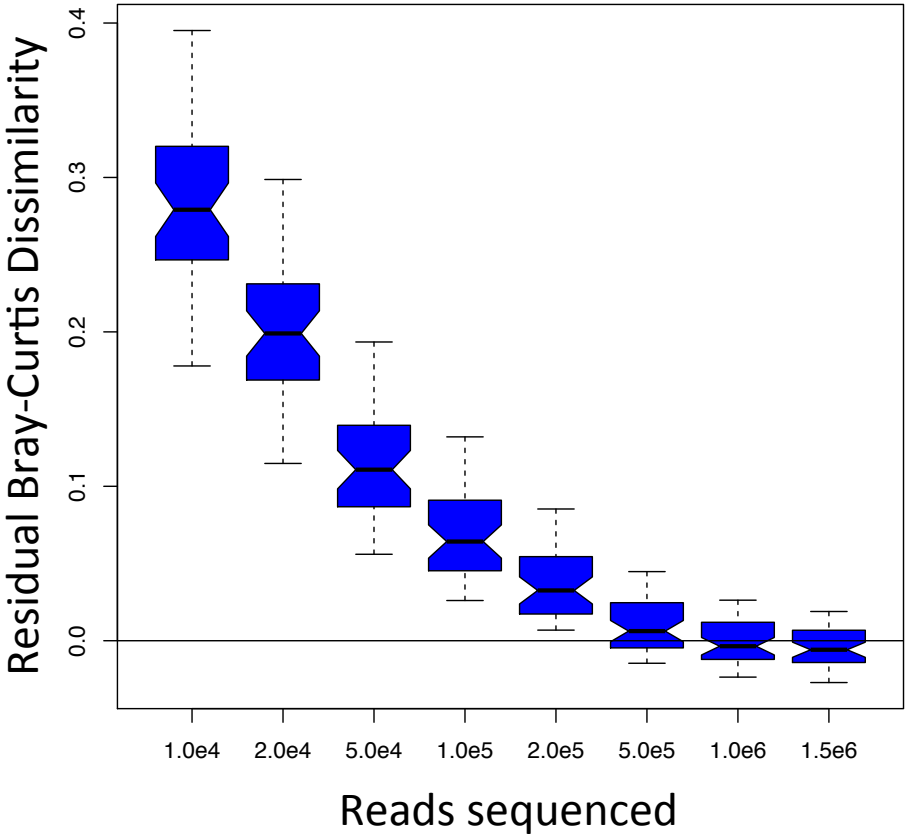
Effect of sequencing depth on individual protein families





Effect of sequencing depth on accurate estimates of between-sample functional abundance





Outline

- Methods: overview of simulation framework
- Read translation
- Score/E-value thresholds
- Search algorithm
- Sequencing depth
- **MG-RAST benchmark**

MG-RAST benchmark: experiment design

Input data

- Simulated metagenomes from 10 mock communities
  - 101 bp reads with Illumina error
  - 1.5 million reads/library

Search database

- For shotmap
  - Protein sequences from M5NR which correspond to KEGG Orthology
    - N=1,912,249
- For MG-RAST
  - All protein sequences from M5NR database
    - N=43,098,145

Truth

- Relative abundance of homologs of KOs in ~700 microbial genomes

MG-RAST benchmark: preliminary results