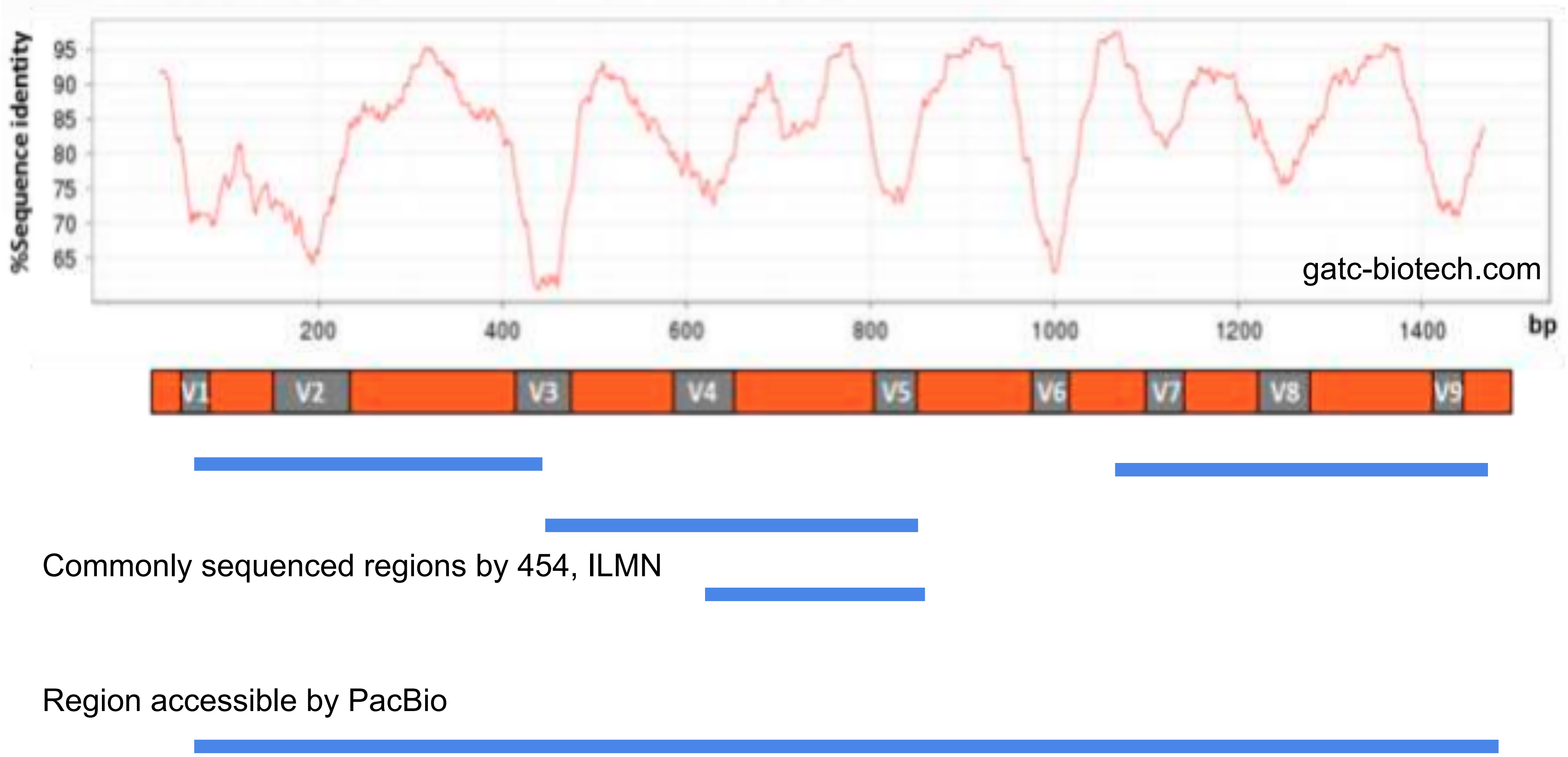
[PBIL] Full length 16S rDNA vs subregion sequencing for microbial community analyses

DS
06 Aug 2015
P2-Tech

Microbiome research and rDNA analysis



gatc-biotech.com

Commonly sequenced regions by 454, ILMN

Region accessible by PacBio

Which is best in terms of taxonomic accuracy, particularly for strain-level ?

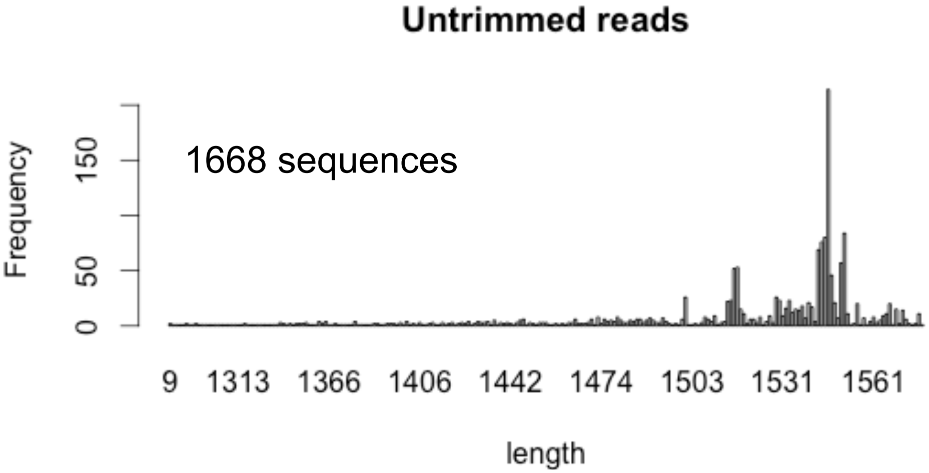
Obtaining high-quality sequences for each genus

1. Grab all matches to those names in GreenGenes (human_assoc_gold_strains_gg16S_aligned.fasta NOT gg_13_5)

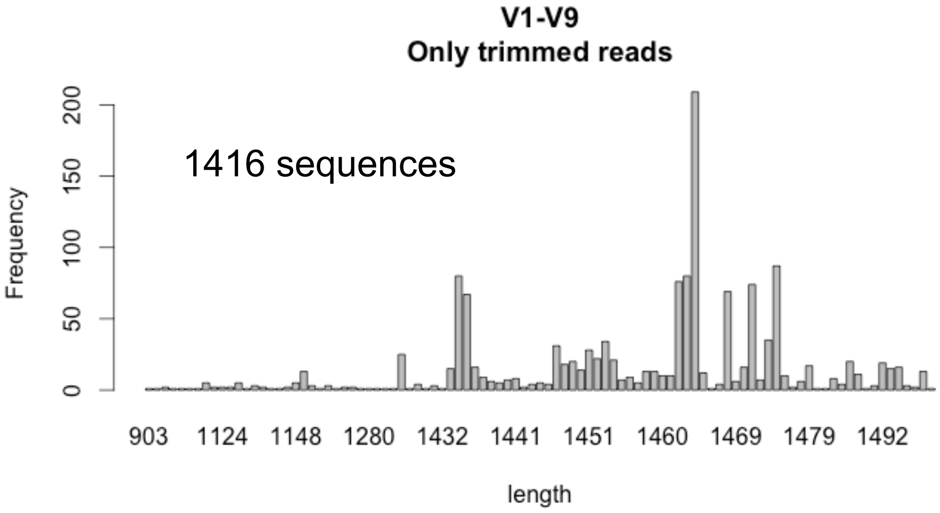
2. Truncate sequences to V1-V9, discard those without matches
   a. FlexBar

Summary of the file of orgs pulled from the database

- Sequential trimming with Flexbar (first V1, then V9)
- Grabbing only sequences that were trimmed in each case (flag with -g and then grep Flexbar_removal in header)



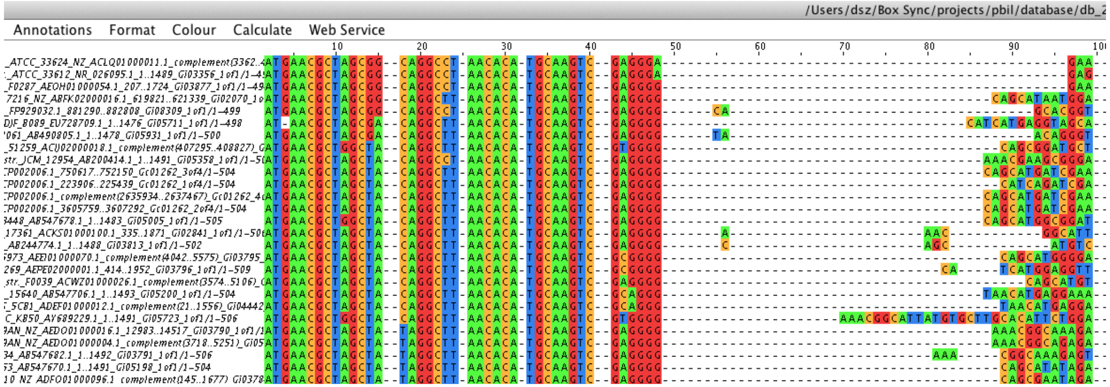
1668 sequences



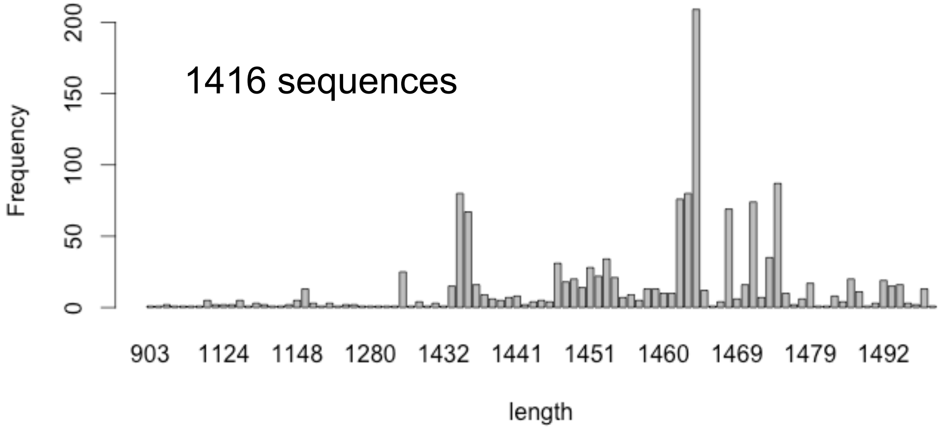
1416 sequences

Trimming the alignment with trimAl

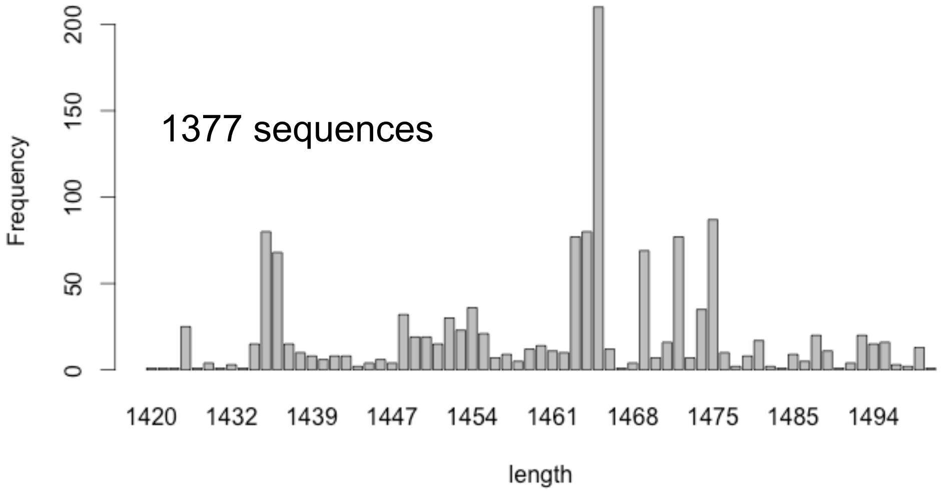
1. Align sequences
2. Visually identify primer site with Jalview
3. Cut at alignment coordinates with trimAl



Flexbar trimmed



trimAl trimmed



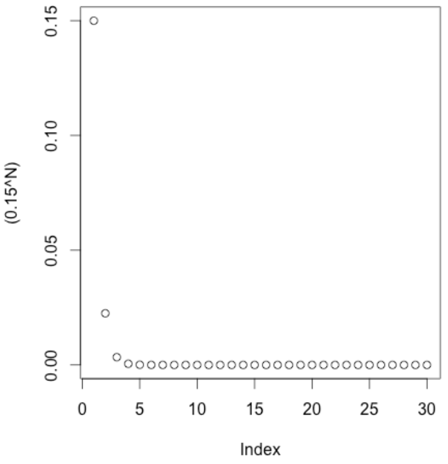
Phylogenetic signal by OTU clustering

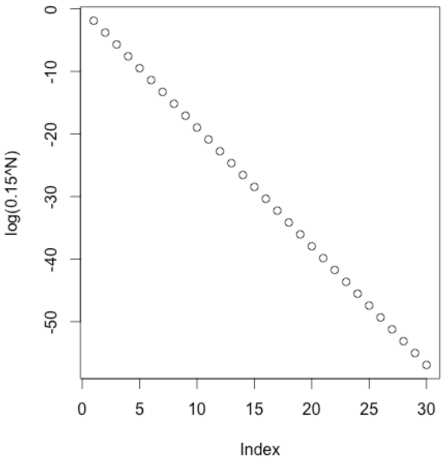
File Name	16S Region	# of unique sequences	# of OTUs @ 99% similarity	# of chimeras @ 99% similarity
db_27F_1492R_final.fasta	V1-V9	806	220	45
db_27F_338R_trimal.fasta	V1-V2	554	262	14
db_27F_534R_trimal.fasta	V1-V3	617	248	14
db_341F_926R_trimal.fasta	V3-V5	452	185	35
db_515F_806R_trimal.fasta	V4	335	178	8

Expected error rate of PacBio

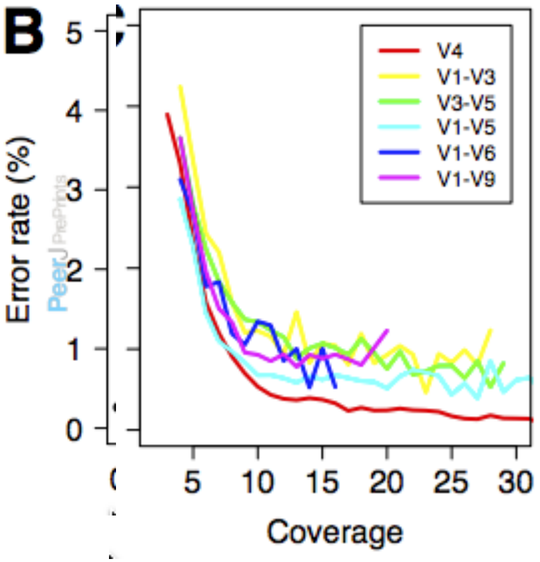
15% raw error rate (Eid et al Science 2009)

- E = $(0.15)^N$
  - where N = # of passes

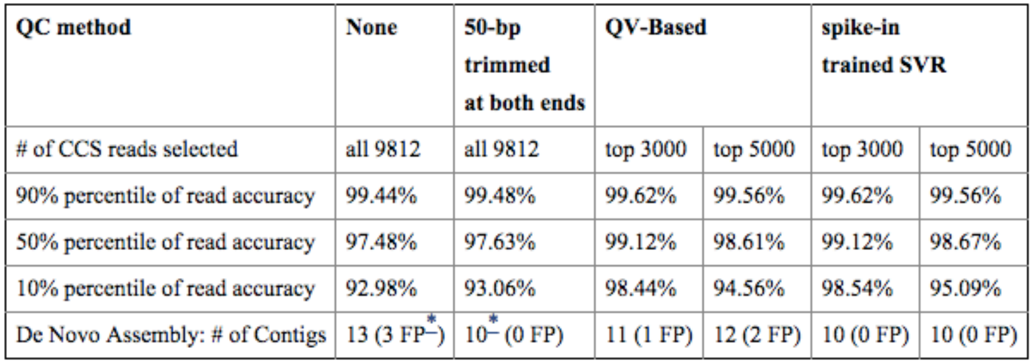




1% ccs error rate (V1-V9) (Schloss et al PeerJ 2015)



2.5% median ccs error rate (plasmid) (Jiao et al J Data Mining Genomics Proteomics 2013)

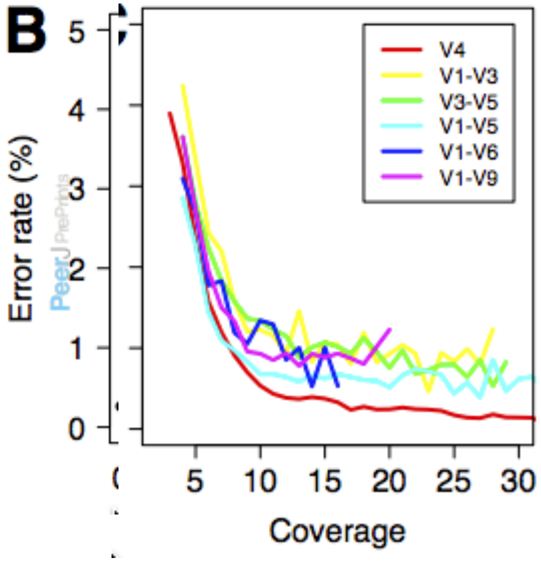


QC method	None	50-bp trimmed at both ends	QV-Based		spike-in trained SVR	
# of CCS reads selected	all 9812	all 9812	top 3000	top 5000	top 3000	top 5000
90% percentile of read accuracy	99.44%	99.48%	99.62%	99.56%	99.62%	99.56%
50% percentile of read accuracy	97.48%	97.63%	99.12%	98.61%	99.12%	98.67%
10% percentile of read accuracy	92.98%	93.06%	98.44%	94.56%	98.54%	95.09%
De Novo Assembly: # of Contigs	13 (3 FP*)	10* (0 FP)	11 (1 FP)	12 (2 FP)	10 (0 FP)	10 (0 FP)

What error rate is necessary to identify strains?

- 1% error rate for ~1500 nt insufficient for single nucleotide resolution
  - SNPs may be non-randomly distributed

1% ccs error rate (V1-V9) even with 20X coverage (Schloss et al PeerJ 2015)



Potential error rate analyses

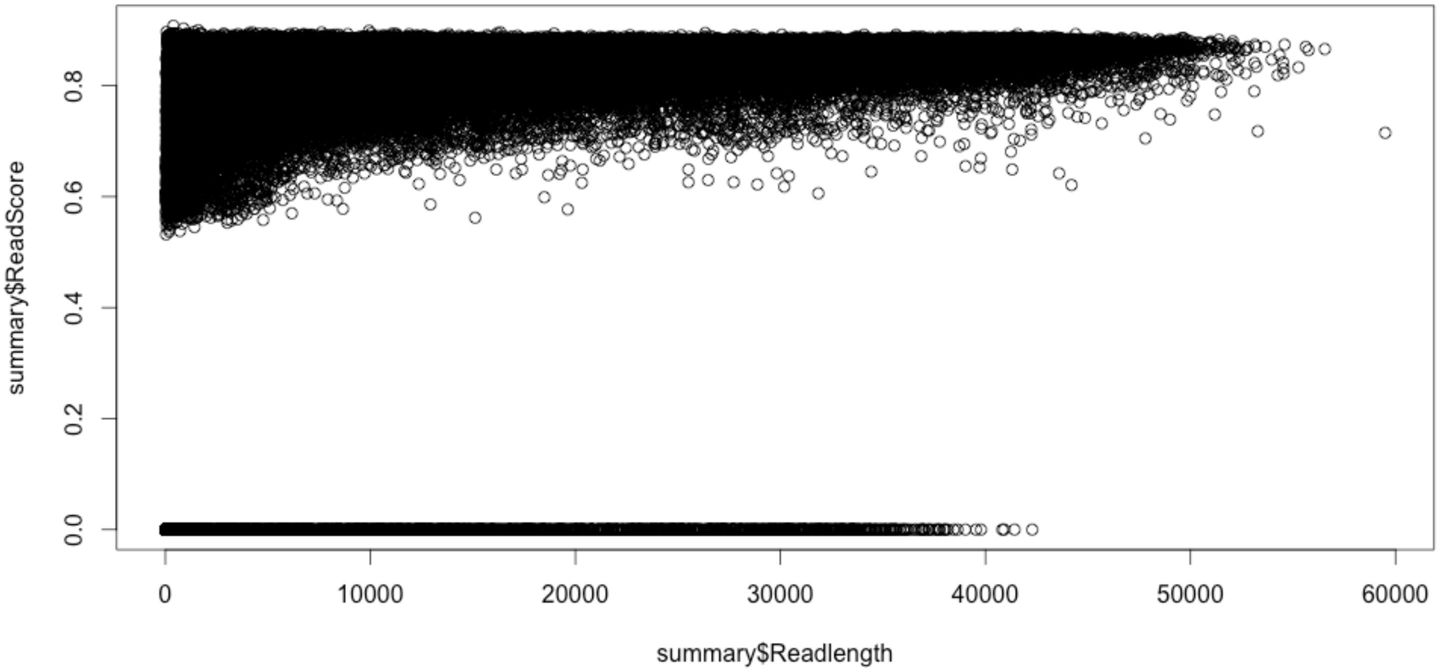
1) Aligning the ccs reads to a database of the mock community and then plotting the # of mismatches against the # of passes (has been done, fig 1C here)

2) Setting the ccs QV threshold in smrt portal to different values (e.g. 99.9%, 99.5%, 99%, etc) and plotting the # of mismatches per sequence against the QV threshold

3) Getting fastqs of the reads before they've been assembled into ccs, breaking each read into the individual passes and then for an individual read measure the error rate with 1 pass, 2 passes, etc.

4) Plotting the fraction of mismatches along the length of the full length 16S to identify the coordinates where there are SNP's (likely because of different copies of the 16S in the genome). Do this analysis with different QV thresholds to see where background error gets rate higher than the SNP signal.

Average QV of the read vs length


summary$ReadScore
summary$Readlength