Bioinformatics D: Amplicon Sequencing

DAVID L. TABB, PH.D.

Overview

- Designing PCR primers
- Measuring microbiomes and metagenomes
- Characterizing diversity
- Quantifying with species composition

Dos and Don’ts of primer selection

- Do employ primers melting at similar temps.
- Do design target-*specific* primers.
- Do design *efficient* primers (near 2x)
- Do not inhibit *Taq* DNA polymerase.
- Do not allow substantial homology among primer sequences.

CW. Dieffenbach et al. *Genome Res*. (1993) 3: S30-S37

Interaction Challenges of PCR primer design

A-C. Primer-Primer Interactions

D. Hairpin structures

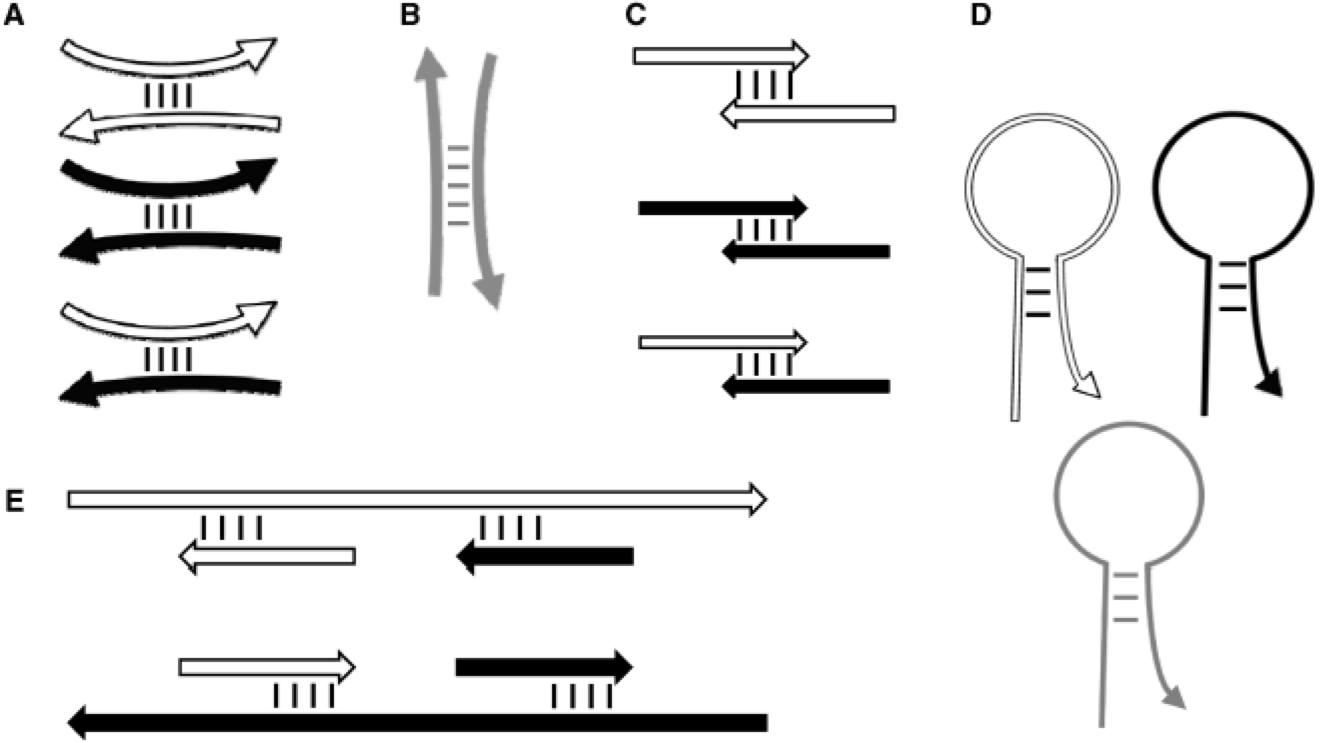
E. Primer-Template Interactions

White: Forward primer

Black: Reverse primer

Both internal and end interactions are possible

Challenge rises as number of primers increases.



Untergasser et al. *Nucl. Acids Res*. (2012) 40:e115

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

Two key software implementations

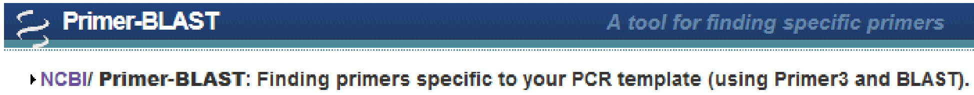
OLIGO VERSION 7

- Also useful for siRNA and restriction analysis
- http://www.oligo.net/
- W. Rychlik and RE Rhoads. *Nucl. Acids Res*. (1989) 17: 8543-8551.

Oligo® .net
Primer Analysis Software

PRIMER3WEB VERSION 4.0.0

- Engine behind NCBI Primer-BLAST
- http://bioinfo.ebc.ee/mprimer3/
- S. Rosen and H. Skaletsky. (2000) ISBN 978-1-59259-192-3

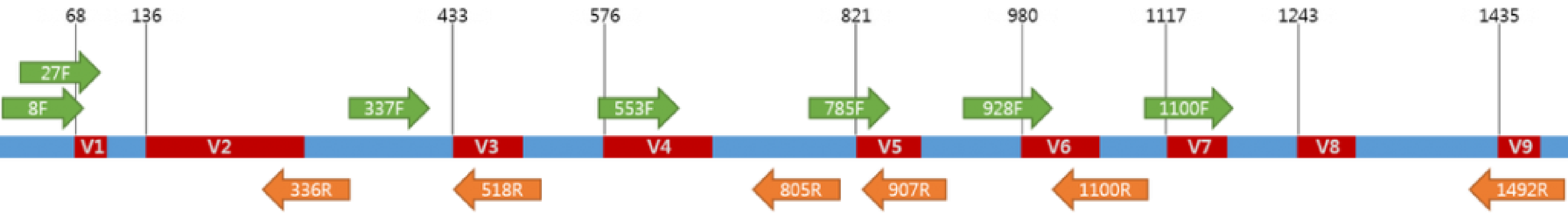

Primer-BLAST A tool for finding specific primers

NCBI/ Primer-BLAST: Finding primers specific to your PCR template (using Primer3 and BLAST).

https://www.ncbi.nlm.nih.gov/tools/primer-blast/

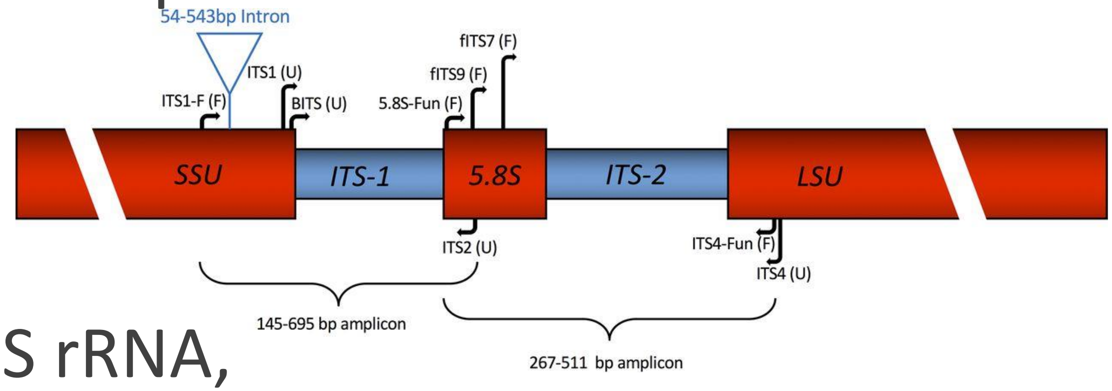
Characterizing microbiomes

Different taxa, different target



help.ezbiocloud.net/16s-rrna-and-16s-rrna-gene/

- Bacteria: 16S rRNA and cpn60



- Fungi: 18S rRNA, 28S rRNA, and Internal Transcribed Spacer

R Sinha et al. *Nature Biotech*. (2017) 35: 1077-1086

DL Taylor et al. *Appl. and Enviro. Microbio*. (2016) 82: 7217-7226

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

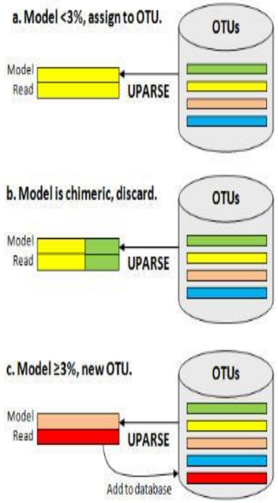
Key definition for amplicon sequencing

- Operational Taxonomic Unit: a cluster of 97% similar sequences, represented by a single consensus sequence.
- Natural sequence variation within species and sequencing errors may yield variation. We must allow for sequence variety.
- Each OTU is a different species; *we may only recognize its phylum, class, or order*.

Nguyen et al. *Nat. Biofilms and Microbiomes* (2016) 2: 16004.

Software for NGS -> OTUs

- MOTHUR (2009): scales to NGS, histograms distinct sequences, clusters OTUs at different distance thresholds, reports diversity
- QIIME (2010): emphasizes modular pipe-line, visualizes findings and metrics
- UPARSE (2013): excludes read singletons and filters chimeras during clustering



Schloss et al. *Appl. and Enviro. Microbio*. (2009) 75: 7537-7541.
Caporaso et al. *Nat. Methods* (2010) 7: 335-336.
R.C. Edgar. *Nat. Methods* (2013) 10: 996-998.

Reference taxonomies for 16S

- Ribosomal Database Project (1997) has grown to 3.3M 16S and 126K 28S rRNAs.
- Greengenes (2006) drew attention to removal of chimeric sequences from DB.
- SILVA (2007) grew from ARB toolkit, dividing into small and large subunit sequences.

rdp

GG GREENGENES
The 16S rRNA Gene Database and Tools

silva
high quality ribosomal RNA databases

BL. Maidik et al. *Nucl. Acids Res*. (1997) 25: 109-110.

TZ. DeSantis et al. *Appl. and Enviro. Microbio*. (2006) 72: 5069-5072.

E. Pruesse et al. *Nucl. Acids Res*. (2007) 35: 7188-7196.

Open-ended technologies

- Metagenomics: sequence random inserts from all DNA in a community of microbes.
- Metatranscriptomics: sequence random cDNA from all mRNA in a community.

“What are these bacteria capable of and what are they doing?”

JM Di Bella et al. *J. Microbiol. Methods*. (2013) 95: 401-414

Diversity and Quantitation

Estimating diversity

“Compositional differentiation and similarity of groups is often analysed by partitioning a regional or ‘gamma’ diversity measure into *within*- and *between*-group components, ‘alpha’ and ‘beta’.”

- α: diversity within samples of a single cohort
- β: diversity among cohorts of samples
- γ: diversity of the population

Jost et al. *Diversity and Distns* (2010) 16:65-76.

Human Microbiome Project. Nature (2012) 486: 207-214.

Did I sequence enough reads?

- Rarefy: "take a random subset of a given size of the original sample"
- "Rarefaction curves can be used to understand the depth of sampling of a community compared with its total diversity."

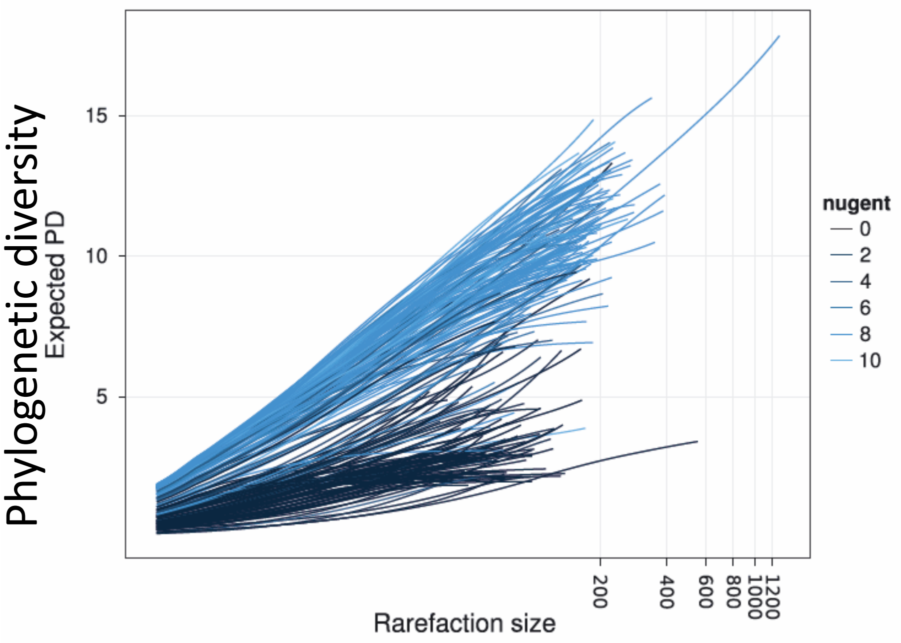


Fig. 5. Rarefaction curve of samples from Srinivasan *et al.* (2012). The Nugent score is a diagnostic score for bacterial vaginosis, with 0 being 'normal' and 10 being classified as BV.

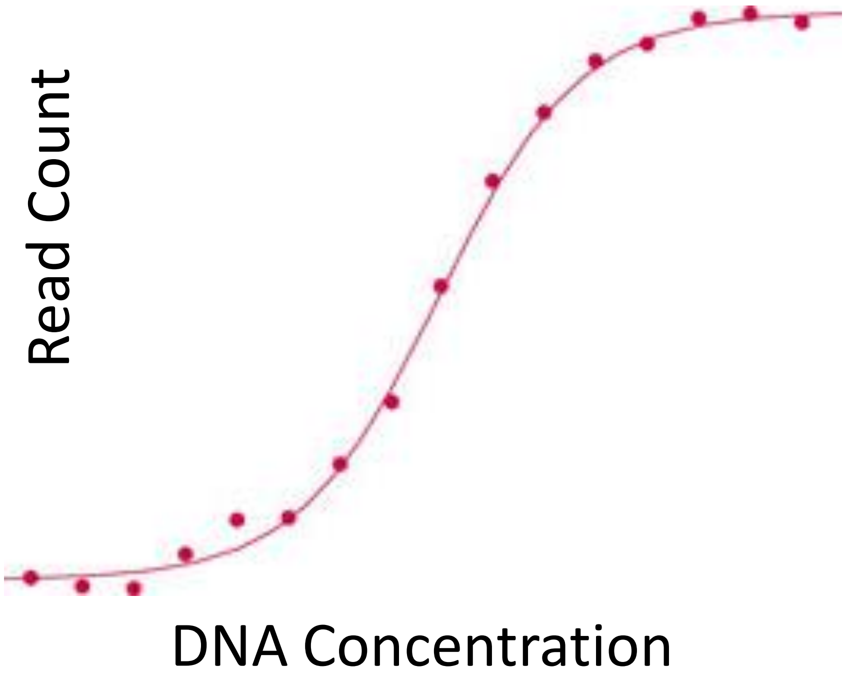
Nipperess and Matsen. *Methods Ecol. Evol.* (2013) 4: 566-572.

Case studies in amplicon quantitation

- What fraction of the *M.tb* microbes in sputum are resistant to TB drugs?
- In this tumor, is the fraction of key genes containing mutations changing over time?
- Do species change in dominance within this bacterial community as a function of pH?

Calibration curves

- Zero reads do not imply total absence of target sequence (Level of Detection).
- Read count does not rise linearly with too few or too many sequence copies (Level of Quantitation).



We call this a “sigmoid” curve.

Why else might read counts mislead us?

- Different target sequences have different primer efficiencies.
- Sequencing errors may cause us to believe sequence variants exist that do not.
- Stochastic noise may cause us to believe cohorts are different when they are not.

Takeaway messages

Bioinformatics tools support key activities in amplicon sequencing:

- Selecting PCR primers
- Clustering reads to determine distinct set of sequences
- Annotating sequence clusters with taxonomy information
- Determining the completeness of sequencing
- Quantifying particular sequence variants