Amplicon sequencing using next-gen technology

STAMPS 2015
Hilary G. Morrison
Joe Vineis, Nora Downey, Bette Hecox-Lea, Kim Finnegan
Questions

• Metagenomic vs. amplicon sequencing?
• What is the best marker to target?
• How do I design primers?
• What kind of errors or biases are there?

Examples: amplicon sequencing on two Illumina platforms
Metagenomic vs. marker data

• Can infer community composition based on shotgun data or 16s reads extracted from shotgun data
• MG-RAST pipeline includes this step
• 16s reads are ~0.1% of the data
• Relative abundance may be more reliable (no PCR bias)

• HMP dataset test*:
  – extract 16s reads based on BLAST against reference database
  – use GAST pipeline to assign taxonomy to them
  – determine abundance of resulting taxa
  – compare to v4v5 marker gene results

*Pam Lescault
HMP metagenomic vs. marker gene results

- While there were taxa found in metagenomic data that were not represented in amplicon data, most were artifacts. One exception was *Rothia*, found in relatively high numbers with reads that assembled to a full 16s.

114 taxa represent >99.5% of total abundance
Targets

• Ribosomal RNA (SSU, LSU)
• Internal transcribed spacers (ITS)—fungi
• Functional genes
  – nirS
  – butyrate kinase, butyrate transferase
  – others? depending on size of reference db
Which variable regions to target?

V6 (967F-1046R) 60 bp
V4 (518F-806R) 288 bp
V4V5 (518F-1064R) 550 bp
V9 (1380/1389-1513) for eukaryotes

http://vamps.mbl.edu/resources/primers.php
ITS (fungi)

- **ITS1** ~350-400 bp
- **ITS2** ~250-300 bp
- **ITS1-ITS2** ~550-800 bp

Diagram showing the locations of primers ITS1F, ITS1R/ITS2F, and ITS2R on the 18S, 5.8S, and 28S regions.
Target-specific primer design

• Start with known primers
• Check matches against SILVA or other rRNA database
  – Phyla missed?
  – Proportion missed?
• Add variants or degenerate sites to expand coverage
Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies

Anna Klindworth\textsuperscript{1,2}, Elmar Pruesse\textsuperscript{1,2}, Timmy Schweer\textsuperscript{1}, Jörg Peplies\textsuperscript{3}, Christian Quast\textsuperscript{1}, Matthias Horn\textsuperscript{4} and Frank Oliver Glöckner\textsuperscript{1,2,*}
Database sources

SILVA
Welcome to the SILVA rRNA database project
A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data, free for academic use.
SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (Bacteria, Archaea and Eukarya).

3,194,778 Total
739,633 SSURef

GREENGENES
16S rRNA gene database and workbench compatible with ARB
greengenes.lbl.gov

UNITE
A molecular database for the identification of fungi
Number of UNITE barcoding sequences: 7802 ITS sequences of 2120 species from 442 genera.
Number of fungal ITS sequences in database (UNITE + INSD): 342448
Example: bacterial-archaeal universal primer v. 1

GGACTAC[ACG][CG]GGGTATCTAAT
Example: bacterial-archaeal universal primer v.2

GGACTAC[ACG][CG]GGGTATCTAAT → GGACTAC[ACT][ACG]GGGT[AT]TCTAAT

- **Bacteria**
- **Archaea**
- **Eukarya**
V6 primers

967F is pool of 4 oligos with 2-16X degeneracy
- $CNACGCGAAGAACCTTANC$
- $CAACGCGMARAACCTTACC$
- $ATACGCGARGAACCTTACC$
- $CTAACCGANGAACCTYACC$

1046R: 4 oligos or 1 oligo with 16X degeneracy
- $CGACAGCCATGCANCACCT$
- $CGACACCCCATGCANCACCT$
- $CGACGGCCATGCANCACCT$
- $CGACGACCATGCANCACCT$
- $CGACRRCCATGCANNCACCT$
### V6

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
<th>Target</th>
<th>RefDB exact match</th>
</tr>
</thead>
<tbody>
<tr>
<td>967F1-PP</td>
<td>CNACGCGAAGAACCTTANC</td>
<td>Bacteria</td>
<td>63.3%</td>
</tr>
<tr>
<td>967F-AQ</td>
<td>CTAACCGANGAACCTYACC</td>
<td>Deferribacteres, Aquificales</td>
<td>0.1%</td>
</tr>
<tr>
<td>967F-U12</td>
<td>CAACGCGMARAACCTTACC</td>
<td>Additional Proteobacteria</td>
<td>14.0%</td>
</tr>
<tr>
<td>967F-U3</td>
<td>ATACGCGARGAACCTTACC</td>
<td>Additional Bacteroides; Spirochaetes</td>
<td>12.7%</td>
</tr>
<tr>
<td>1046R</td>
<td>CGACRRCCATGCANACCT</td>
<td>Bacteria</td>
<td>95.4%</td>
</tr>
</tbody>
</table>

Keep level of degeneracy low (2-16X)
Are rare, but important groups missed?
Primers with 1-2 mismatches often work—pay attention to Tm
# V4V5

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
<th>Target</th>
<th>RefDB exact match</th>
</tr>
</thead>
<tbody>
<tr>
<td>518F</td>
<td>CCAGCAGCYGCGGTAAN</td>
<td>Bacteria</td>
<td>97.0%</td>
</tr>
<tr>
<td>926R1</td>
<td>CCGTCAATTCNTTTRAGT</td>
<td>Bacteria</td>
<td>96.2%</td>
</tr>
<tr>
<td>926R3</td>
<td>CCGTCAATTTCCTTGAGT</td>
<td>Verrucomicrobia, Delta proteobacteria, Actinobacteria</td>
<td>2.3%</td>
</tr>
<tr>
<td>926R4</td>
<td>CCGTCTATTCCTTTGANT</td>
<td>Epsilon proteobacteria</td>
<td>1.3%</td>
</tr>
</tbody>
</table>

Why not use CCGTCWATTYNTTTRANT?

128 variants and only 40 match a known bacterial sequence!
Fusion primer modifications

- Fusion with platform-specific sequencing primer sequences (vs. ligation)
- Illumina fusion primers include bridge adapter regions
- Phosphorothioate linkages (stability)
- Barcoding: 5-12 nt
- Indexing: 6-12 nt
Illumina primer design
Illumina amplicon strategy

• Fused primer approach, not adapter ligation
  – Primers much longer than 454 versions
  – Include bridge adapter sequences
  – Minimize cost of barcoded primer sets
    • Many labs now have EMP 515F/806R sets
targets V4; >1000 indices available
    • Combination of 12 indices x 8 inline barcodes creates
96-plex set for other targets
Illumina amplicon targets

• How long are the reads?
  – HiSeq/NextSeq: 100/150 PE
  – MiSeq 250 PE and longer

• How good is the quality towards the end?
  – Undetected errors will inflate alpha-diversity
  – Merged reads require overlap
150 nt PE NextSeq500

967F

Read 1

~60 bp; complete overlap

Read 2

V6R: 1046R

Never run only v6 amplicon libraries on PE 150
Longer reads on MiSeq

V4F: 518F

50-100 nt overlap

V5R: 926R
Why still v6?

- Original 454 pyrotag target-lots of experience
- Sequence overlap with v6v4 datasets
- Cheaper to do
- Greater sampling depth on NextSeq
  - 300-400 million clusters per flowcell possible
  - 3M reads at 96-plex
  - 300K reads at 960-plex
Fusion primer design

5’ adapter

“pad/linker”* “515F”

EMP sequencing primer 1

5’ AATGATA CGGCA CCAC CGAG AT CTAC AC- TATGGTAA TT GT- GTGCCAG CMGCCGCGGTAA… rRNA

rRNA… TAATCTWTGGVHCATCAGG- CCGACTGACTGA- TTGCGTGCGATC- TAGACGATA CGGCAGAAGACGAAC

“806R-rc” “pad/linker-rc”* index-rc 3’ adapter-rc

EMP sequencing primer 2
RC is indexing sequencing primer

*designed to have no homology to 16s or Illumina adapters/primers
Design considerations

• First 4 bases of read 1 used for cluster finding (no change)
• Next 8 bases (up through base 12) need to be high-complexity because of phasing/prephasing calculation
• This is also true of read 2
• Thus, need to have randomness in cycles 1-12 of both reads
• Add it by design, by offset start, by combining low and high complexity libraries on run
In-line bar codes vs. indices

• In-line bar codes cut into sequencing reads, but:
• Cluster finding requires random distribution of A,C,G,T at start of read 1
• We designed our primers to provide this diversity
  – 5’ NNNN (1-4)
  – Careful mix of bar codes (5-9)
• Didn’t work
  – Clusters are found, but amplicons are still ‘low complexity’ samples
  – Diversity needed at first 12 bases of both read 1 and read 2
• Had to add high proportion of PhiX or other ‘random’ library
Fusion primer design

967F_PP_N4TCAGC: one of 4 primers with TCAGC barcode; forward read

5’ AATGATACGGCGACCACCAGATCTACAC-­‐TCTTTCCCTACACGACGCTCTTCCGATCT-­‐NNNNNTCAGC-­‐CNACGCGAAGAACCTTANC 3’

1046R_index_1: single degenerate primer with index; reverse read

5’ CAAGCAGAAGACGGCAGACGACGAGAT-­‐CGTGAT-­‐GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-­‐CGACRRCCATGCANCACCT 3’

Similar for v4v5 suite
Primer synthesis

• >80 nt long
• Purification needed
  – Desalting only
  – PAGE (used for first set; expensive)
  – No other option because of oligo length
• Pooled 4 967F primers for each barcode (V6)
• Pooled 3 926R primers for each index (V4V5)
• Final set of 8 forward/barcoded and 12 reverse/indexed primers = 96 unique combinations for each domain
• Working stocks @10 uM each
Illumina amplicon generation and sequencing
Amplification process

• DOs:
  – High quality DNA template: e.g. 260/280 ~2
  – Replicate reactions using high-fidelity polymerase
  – Negative controls
  – Pool, clean, size-select
    • Qiagen MinElute (FLX; Illumina v6),
    • Ampure (Titanium; Illumina v4v5);
    • Pippin Prep (Illumina pools)
  – Quantitation
    • e.g. PicoGreen, qPCR
  – Visualization
    • Bioanalyzer, Caliper, gel

• DON’Ts:
  – MDA or nested amplification
Size selection

v4v5 amplicons (Ampure)

v6 amplicons (Pippin)
What can go wrong?

• Contamination or sample mix up
• PCR error (enzyme misincorporation)
• PCR ectopic amplification (non-rRNA)
• Chimera formation
• Low biomass samples—run enough cycles, you’ll get a (meaningless) product
Effect of low-complexity libraries

- Clusters found correctly
- Then read hits highly conserved sequence
- **Quality scores reduced**
- Mitigate with lower cluster density and HC library spike-in
  - over 50% PhiX on early Miseq runs
- Illumina has software fix for problem on Miseq
  - now ~2% PhiX on Miseq runs
  - not corrected on Hiseq
Control gDNA: known community

human microbiome-centric

Product Information Sheet for HM-276D

Genomic DNA from Microbial Mock Community B (Even, High Concentration), v5.1H, for Whole Genome Shotgun Sequencing

Catalog No. HM-276D

For research use only. Not for human use.

Contributor and Manufacturer:
Sarah K. Highlander, Associate Professor, Department of Molecular Virology and Microbiology; Baylor College of Medicine, Houston, Texas

Product Description:
HM-276D contains genomic DNA from 20 bacterial strains containing equimolar (Even) ribosomal RNA operon counts (1,000,000 copies per organism per µL). This mock community can be used for whole genome shotgun sequencing and the recommended amount to use per experiment is 20 µL. The bacterial strains that DNA was extracted from are listed in Table 1.

Material Provided:
Each vial contains approximately 100 µL of the bacterial genomic DNA mixture suspended in Tris-HCl, pH ~ 7.5. The concentration is shown on the Certificate of Analysis. The vial should be centrifuged prior to opening.

Packaging/Storage:

now also available through ATCC
Error rate

- Generated data from HMP mock community gDNA
- Expected sequences recovered
- Aligned unique sequences to reference
- Sequencing errors, not quality scores

- Overall error: 0.52%
How does Illumina tag sequencing compare to pyrotag sequencing?
Analysis of v6 data from real samples

Sandra Mclellan samples: Illumina (HiSeq) overrepresents Actinomyces; misses Arcobacter almost entirely

Samples are: 3 HiSeq v6 amplicon datasets and 2 454 datasets (v6v4 and v6 only)
Do fusion primers bias recovery?

2-step PCR test

– round 1, 16s primers only (20 cycles)
– round 2, full-length fusion primers (5 cycles)
  • 1 ul of 1:50 dilution of round 1 product
PCR results

16s only, 20 cycles
16s only, +5 cycles
16s only, negative
16s only, 20 cycles
16s only, +5 cycles
16s only, negative
**Sequencing results**

<table>
<thead>
<tr>
<th>Sample, conditions</th>
<th>Count read1</th>
<th>Count Arcobacter</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS_WWTP_1_25_11_Bv6v4</td>
<td>35,612</td>
<td>5,525</td>
<td>15.50%</td>
</tr>
<tr>
<td>SS_WWTP_1_25_11_ACGCA</td>
<td>214,950</td>
<td>2,309</td>
<td>1.10%</td>
</tr>
<tr>
<td>SS_WWTP_1_25_11_GACTC</td>
<td>240,520</td>
<td>2,606</td>
<td>1.10%</td>
</tr>
<tr>
<td>SS_WWTP_1_25_11_GTATC</td>
<td>392,570</td>
<td>5,445</td>
<td>1.40%</td>
</tr>
<tr>
<td>SS_WWTP_1_25_11_2step</td>
<td>6,106,654</td>
<td>900,824</td>
<td>14.80%</td>
</tr>
<tr>
<td>SS_WWTP_4_11_11_2step</td>
<td>5,779,659</td>
<td>824,789</td>
<td>14.30%</td>
</tr>
</tbody>
</table>

Counted matches to reference in raw data, read 1, each dataset

>gi|343201492 *Arcobacter cibarius*
TGGACTTGACATAGTAAGAACTTTCTAGAGATAGATTGGTGCTGCTTGCAGAAACTTATATATAC
Variables tested when optimizing PCR

- Primer concentration
  - 0.2-0.4 uM
- Magnesium concentration
  - 1.5 mM – 3.67 mM
- Annealing temperature
  - *Tm of primers suggest anywhere from 52°C-60°C*
- Units of enzyme
- Keep primer composition and cycle number constant
## Current protocols

<table>
<thead>
<tr>
<th></th>
<th>bac v6</th>
<th>arc v6</th>
<th>bac v4v5</th>
<th>arc v4v5</th>
</tr>
</thead>
<tbody>
<tr>
<td>HiFi buffer</td>
<td>1X</td>
<td>1X</td>
<td>1X</td>
<td>1X</td>
</tr>
<tr>
<td>MgSO4</td>
<td>2 mM</td>
<td>2 mM</td>
<td>3 mM</td>
<td>3 mM</td>
</tr>
<tr>
<td>dNTPs</td>
<td>0.2 mM</td>
<td>0.2 mM</td>
<td>0.2 mM</td>
<td>0.2 mM</td>
</tr>
<tr>
<td>combined primers</td>
<td>0.2 uM</td>
<td>0.3 uM</td>
<td>0.3 uM</td>
<td>0.3 uM</td>
</tr>
<tr>
<td>Platinum HiFi</td>
<td>4 units</td>
<td>10 units</td>
<td>10 units</td>
<td>10 units</td>
</tr>
<tr>
<td>Template</td>
<td>5-20 ng</td>
<td>5-20 ng</td>
<td>5-20 ng</td>
<td>5-20 ng</td>
</tr>
<tr>
<td>Volume</td>
<td>100 ul</td>
<td>100 ul</td>
<td>100 ul</td>
<td>100 ul</td>
</tr>
<tr>
<td>PCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 min</td>
<td>94°</td>
<td>94°</td>
<td>94°</td>
<td>94°</td>
</tr>
<tr>
<td>30 sec</td>
<td>94°</td>
<td>94°</td>
<td>94°</td>
<td>94°</td>
</tr>
<tr>
<td>45 sec</td>
<td>60°</td>
<td>60°</td>
<td>57°</td>
<td>57°</td>
</tr>
<tr>
<td>1 min</td>
<td>72°</td>
<td>72°</td>
<td>72°</td>
<td>72°</td>
</tr>
<tr>
<td>2 min</td>
<td>72°</td>
<td>72°</td>
<td>72°</td>
<td>72°</td>
</tr>
<tr>
<td>Cycles, step 1</td>
<td>25</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Pooling, clean up</td>
<td>Qiagen MinElute plate</td>
<td>Qiagen MinElute plate</td>
<td>Ampure, 1:1</td>
<td>Ampure, 1:1</td>
</tr>
<tr>
<td>Step 2</td>
<td>4 ul step 1 eluate</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cycles, step 2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Quantitation</td>
<td>Picogreen</td>
<td>Picogreen</td>
<td>Picogreen</td>
<td>Picogreen</td>
</tr>
<tr>
<td>Size selection</td>
<td>Pippin Prep</td>
<td>Pippin Prep</td>
<td>Ampure</td>
<td>Ampure</td>
</tr>
</tbody>
</table>
Illumina amplicon read QC

- Filter out reads based on:
  - Chastity filter (or equivalent; Illumina base caller)
  - Ns
- Then attempt to merge forward and reverse reads
- Keep only those that match perfectly over the region of overlap (v6) or have fewer than 3 mismatches (v4v5)
- Create consensus using basecall with higher Qscore
- Trim away adapters, primers
- Product goes into usual GAST pipeline, but as unique sequences plus frequency info, not all reads
Final comments

- Amplicon sequencing is moving target
  - Technology is changing every year, but amplicon sequencing is still not well supported by next-gen companies
  - Sample sources expanding (animal microbiomes, extreme environments, **low biomass** targets, marine/freshwater/sediments/rock surfaces)
  - There is no perfect set of primers or perfect marker region
  - Processing s/w – changing approaches to filtering, trimming

- So—
  - Think about your experimental design—how much data do you need? Depth of sequencing or more signal? Will the data answer your question?
  - Look at your raw data—What is the accuracy (vs. quality) of the data? Did you get expected taxa from a control template?
    - If not, revise your primers and protocols
    - Never hesitate to discard LQ reads
  - Publish in a timely fashion!
I'm so mad...I just started a new project and the sequencing platform and analysis tools are already obsolete.

Don't feel bad. The reviewers won't ding you because you're behind the curve.

Yeah, we will.

Isn't it fun how technology moves faster than publication?