Bacterial Genomics

Introduction to Computational & Quantitative Biology

Anne-Catrin Uhlemann, MD, PhD
Department of Medicine / Division of Infectious Diseases

March 22nd 2017
Overview

• Sequencing technologies
• Analysis approaches
  - Bacterial whole genome sequencing
  - Microbiome analyses
• Some practical examples
Recap: Sanger sequencing

- Dideoxynucleotide sequencing
- Chain-termination method
History of bacterial genome sequencing

• First closed genome:
  - 1995
  - *Haemophilus influenzae*

• Since 2008:
  - 5-log decrease in cost

• 2013:
  - 2,264 finished genomes
  - 4,067 draft genomes
Sequencing technologies –
A moving target

Illumina
HiSeq
Miseq

SOLiD
Ion Torrent

454
PaqBio

GnuBio
Helioscope

Nanopore?
Traditional versus Next-Generation Sequencing

**SANGER SEQUENCING:**

- Single sequence read per bp

**NGS:**

- Multiple sequence reads per bp
High-throughput Next generation sequencing by synthesis

Genomic DNA → Cut DNA → Add Linkers

Input library → Flow cell → In Situ PCR → Sequencing → An image of hundreds of extended molecules

©2013 by American Society of Hematology
Read length

Chemistries limit read length, are constantly being improved
- short < 50 consecutive bases
- mid-length 51 - <400
- long > 400 (< 1000)
Depth of coverage

Numbers can be misleading!
## Some comparisons

<table>
<thead>
<tr>
<th></th>
<th>Sequence yield</th>
<th>Run time</th>
<th>Length</th>
<th>The Good</th>
<th>The Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Illumina</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HiSeq</strong></td>
<td>300 Gb</td>
<td>8.5 days</td>
<td>2 x 100 bp</td>
<td>Low error rate</td>
<td>Relatively short reads</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Paired end</td>
<td></td>
</tr>
<tr>
<td><strong>Miseq</strong></td>
<td>8 Gb</td>
<td>27 hrs</td>
<td>2 x 250 bp</td>
<td></td>
<td>Higher indels</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GC-problems</td>
</tr>
<tr>
<td><strong>454</strong></td>
<td>450 – 700 Mb</td>
<td>10 – 23 hrs</td>
<td>Up to 1 kb</td>
<td>Read length speed</td>
<td>Indels</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Difficulty</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>homopolymers</td>
</tr>
<tr>
<td><strong>Ion Torrent</strong></td>
<td>600 Mb – 2Gb</td>
<td>3 – 7 hrs</td>
<td>200 – 400bp</td>
<td>SNP’s</td>
<td>Bias with AT-rich</td>
</tr>
<tr>
<td><strong>PaqBio</strong></td>
<td>230 Mb</td>
<td>2 hrs</td>
<td>4.5 kB</td>
<td>Long reads +DNA</td>
<td>SNP’s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>methylation</td>
<td></td>
</tr>
</tbody>
</table>

Starting concentrations: 10 ng to 700 ng per sample
Bacterial sequencing applications

16S rRNA sequencing

Metagenomics

Single cell sequencing
Single isolate bacterial sequencing

• Comparative sequencing
  - SNPs / indels that determine virulence
  - evolution
  - outbreak investigations
  - compare presence / absence

• De novo assembly
  - only way to determine new gene content
  - not optimal for variant calling
K-mers

- Sequence of K base calls
- K = positive integer
- Only consecutive bases are used
- Reads with high sequence similarity must share K-mers in overlapping regions
- Shared K-mers are easier to find than overlaps
- Fast detection of shared K-mer content reduces computational cost/time
- Disadvantage: lower sensitivity in overlap regions
Problem: repetitive regions and de Bruijn Graph
Vertices are k-mers
Edges are pairwise alignments

Vertices are (k-1)-mers
Edges are k-mers

Hamiltonian cycle
Visit each vertex once (harder to solve)

Eularian cycle
Visit each edge once (easier to solve)
Some terminology

- FASTQ: text based format storing sequence data and quality scores
- FASTA file: sequence in text format
- SAM file: tab-delimited text file that contains sequence alignment data
- BAM file: binary version of a SAM file
Figure 1. An example workflow for high-throughput whole genome sequencing in bacteria.

- Annotation
- Identification of phages, MGEs


http://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1002874
Single cell bacterial sequencing
The example of *S. aureus* USA300
California: Initial documentation of CA-MRSA epidemic / USA300

**CA-MRSA:**
- No hospitalization past 6 months
- Not nursing home
- Culture + < 48hrs
- Not dialysis
- Not homeless
- Skin & Soft tissue infections
- Invasive ~5% cases

Chambers et al.
**S. aureus** clonal typing

- **Multi-locus sequence typing**
  - Sequence types (ST), clonal complexes

- **Pulsed-field gel electrophoresis (PFGE)**
  - Restriction digest with *Sma*I
  - Compare pattern to reference

- **Spa-typing**
  - Polymorphic X (short sequence repeat) region of protein A gene
  - Variable 24-bp repeats upstream of C-terminal cell wall attachment
  - Rapid strain typing/clonal assignment compared to PFGE or MLST

USA300 genome composition

Genomic & biological features:

Core genome
- Increased expression of core virulence genes
  PSM, $\alpha$-toxin

Mobile genetic elements
- Small SCCmec IV
- $\varphi 2$ / PVL toxin
- SaPi5
- ACME I
detoxifies host antimicrobials
- derived from USA500?

Diep et al. The Lancet 2005
USA300 sample selection (n = 387)

161 cases

Clinical USA300 n = 131

Serial infections n = 22

USA300 colonization n = 138

Environmental n = 51

Outside HH contacts n = 15

Total, n = 357

161 controls

USA300 colonization n = 14

Environmental n = 14

Outside HH contacts n = 2

Total, n = 30
Whole genome sequencing
387 isolates

Sequencing:
- Mate-paired libraries
  100 bp paired-end
- Illumina Hi-Seq
- Coverage 100 to 170 fold

Mapping:
- Reference genome FPR3757
- Exclusion unmapped reads, MGEs
- Repeat Scout
- SNP calling

Phylogenetic tree:
- Core genome
- Concatenated SNPs
- RAxML
Phylogenetetic tree

All life is related by common ancestry.

Phylogeny: pattern of historical relationships

Tree: mathematical structure used to model the evolutionary history of a group of organisms
Tree Notation

- **Node**: Human, Mouse
- **Clade**: Mouse
- **Branch Length**: Human
- **Root**: Fly
- **Branch**: Fly
Genome composition of *S. aureus*
Mobile elements do not follow tree like evolution

Chromosomal Genome:
1. Stable core
   - MLST
2. Variable core
   - Surface proteins
     - spa-types
   - Some virulence factors
3. Mobile genetic elements
   - Integrated Pro-phages
   - Pathogenicity islands
   - Transposons and Insertion sequences

Plasmids:
- Resistance genes
SNP calling (dataset n = 375)

- Mapped to reference
- Mapped genome ~90%
- MGEs excluded (need to analyze separately)
- 12,451 SNPs
- Coverage 3-fold per SNP base needed
Comparison of multiple isolates per person

<table>
<thead>
<tr>
<th>Position_in</th>
<th>CDS/rRNA/tf</th>
<th>CDS_name</th>
<th>product</th>
<th>Synonymous</th>
<th>Ref_base</th>
<th>SNP_base</th>
<th>Total</th>
<th>7748_4#94</th>
<th>9266_1#14</th>
<th>9266_1#15</th>
<th>7748_4#95</th>
<th>9266_1#16</th>
<th>7790_7#31</th>
<th>7748_4#96</th>
</tr>
</thead>
<tbody>
<tr>
<td>108951</td>
<td>Intergenic</td>
<td>-</td>
<td>-</td>
<td>G</td>
<td>T</td>
<td>6</td>
<td></td>
<td>2079_1</td>
<td>2079_1</td>
<td>2079_1</td>
<td>2079_1</td>
<td>2079_1</td>
<td>2079_1</td>
<td>2079_1</td>
</tr>
<tr>
<td>349157</td>
<td>CDS</td>
<td>SAUSA300_0297</td>
<td>putative lipoprotein</td>
<td>S</td>
<td>C</td>
<td>T</td>
<td></td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>N</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>349159</td>
<td>CDS</td>
<td>SAUSA300_0297</td>
<td>putative lipoprotein</td>
<td>N</td>
<td>C</td>
<td>T</td>
<td></td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>N</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>1074221</td>
<td>Intergenic</td>
<td>-</td>
<td>-</td>
<td>G</td>
<td>T</td>
<td>21</td>
<td></td>
<td>T</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>2217199</td>
<td>Intergenic</td>
<td>-</td>
<td>-</td>
<td>C</td>
<td>T</td>
<td>1</td>
<td></td>
<td>T</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>110608</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipoprotein</td>
<td>N</td>
<td>C</td>
<td>T</td>
<td></td>
<td>N</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>110610</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipoprotein</td>
<td>S</td>
<td>A</td>
<td>T</td>
<td></td>
<td>N</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>110613</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipoprotein</td>
<td>S</td>
<td>A</td>
<td>T</td>
<td></td>
<td>N</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>N</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>110709</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipoprotein</td>
<td>S</td>
<td>C</td>
<td>T</td>
<td></td>
<td>N</td>
<td>T</td>
<td>N</td>
<td>T</td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
<tr>
<td>2003077</td>
<td>Intergenic</td>
<td>-</td>
<td>-</td>
<td>T</td>
<td>G</td>
<td>72</td>
<td></td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>349259</td>
<td>CDS</td>
<td>SAUSA300_0297</td>
<td>putative lipoprotein</td>
<td>S</td>
<td>T</td>
<td>C</td>
<td></td>
<td>2</td>
<td>N</td>
<td>N</td>
<td>C</td>
<td>N</td>
<td>C</td>
<td>N</td>
</tr>
<tr>
<td>110614</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipoprotein</td>
<td>N</td>
<td>A</td>
<td>G</td>
<td></td>
<td>5</td>
<td>N</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>N</td>
</tr>
<tr>
<td>110616</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipoprotein</td>
<td>N</td>
<td>A</td>
<td>G</td>
<td></td>
<td>5</td>
<td>N</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>N</td>
</tr>
<tr>
<td>110620</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipoprotein</td>
<td>N</td>
<td>A</td>
<td>G</td>
<td></td>
<td>5</td>
<td>N</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>N</td>
</tr>
<tr>
<td>110634</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipoprotein</td>
<td>N</td>
<td>A</td>
<td>C</td>
<td></td>
<td>3</td>
<td>N</td>
<td>C</td>
<td>C</td>
<td>N</td>
<td>C</td>
<td>N</td>
</tr>
<tr>
<td>110718</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipoprotein</td>
<td>N</td>
<td>T</td>
<td>A</td>
<td></td>
<td>3</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>A</td>
<td>N</td>
</tr>
<tr>
<td>110628</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipoprotein</td>
<td>N</td>
<td>T</td>
<td>A</td>
<td></td>
<td>3</td>
<td>N</td>
<td>A</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>N</td>
</tr>
<tr>
<td>349265</td>
<td>CDS</td>
<td>SAUSA300_0297</td>
<td>putative lipoprotein</td>
<td>S</td>
<td>G</td>
<td>A</td>
<td></td>
<td>3</td>
<td>N</td>
<td>A</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>N</td>
</tr>
<tr>
<td>2179424</td>
<td>Intergenic</td>
<td>-</td>
<td>-</td>
<td>G</td>
<td>A</td>
<td>40</td>
<td></td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>1630757</td>
<td>Intergenic</td>
<td>-</td>
<td>-</td>
<td>T</td>
<td>A</td>
<td>215</td>
<td></td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>A</td>
</tr>
<tr>
<td>1074250</td>
<td>Intergenic</td>
<td>-</td>
<td>-</td>
<td>G</td>
<td>A</td>
<td>21</td>
<td></td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>N</td>
<td>A</td>
</tr>
<tr>
<td>960514</td>
<td>Intergenic</td>
<td>-</td>
<td>-</td>
<td>C</td>
<td>A</td>
<td>13</td>
<td></td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>1433531</td>
<td>CDS</td>
<td>SAUSA300_1302</td>
<td>ATPase family protein</td>
<td>2</td>
<td>G</td>
<td>A</td>
<td></td>
<td>1</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>

Concern: high number of SNPs in isolates samples at the same time from same person based on “N” (i.e. inability to call sequence) in “non-mutant”
Suspicious: clustering in one gene/region
Distribution of N’s across SNPs

NumberOfNs

- Intergenic between Secretary antigen
- SWA, C1AP domain protein
- SAUSA300_2482 & hydroxymethyl-glutaryl-CoA reductase SAUSA300_2483
- Intergenic between acetyl-CoA carboxylase, biotin carboxyl
- Carrier protein SAUSA300_1476 & IS30 transposase SAUSA300_1477
- Intergenic between sdiD and sdiE
- deoC

0 2000 4000 6000 8000 10000 12000
Possible explanations for N’s

• Repetitive sequences
• Does not overtly match deletions/insertions
• Difficulty of mapping repeats (read length only 100bp – disadvantage of Illumina)
• Duplications or recombination?
• Identified new ORF in “intergenic region” transposase type (published in other genomes)
How were N’s addressed in other studies?

- No mention!
- Between the lines:

  “Unmapped reads and sequences that were not present in all genomes were not considered as part of the core genome, and therefore SNPs from these regions were not included in the analysis... as were SNPs falling in high-density SNP regions, which could have arisen by recombination. The core genome was curated manually to ensure a high-quality data set...”
Distances

• Once we compute the distances, how do we find a good tree?
• There are several methods.
Trees are like mobiles

• The same tree can be represented in different ways, by permuting the branches.
Different trees

• Topology (no lengths):
  – Cladogram: relative common ancestry without specifying lengths.

• Topology + lengths:
  – Additive trees: incorporate the length of the branch representing the amount of evolutionary change.
Methods of constructing trees

• Alignment of homologous sequences
  -> concatenated SNPs

• Tree methods:
  1. Distance methods.
     - Minimal Evolution.
     - Least Squares.
     - UPGMA.
     - Neighbor-Joining.
  2. Parsimony.
  3. Likelihood. PHYLIP (Felsenstein).
  4. Bayesian methods.
Maximum likelihood estimation

• Accommodates time structure of temporally-spaced sequences
• Tips have isolation date; internal nodes are unknown -> arbitrary starting times (order on tree)
• Substitution rate used to scale times into units of expected number of substitutions per site
• Likelihood of the model; standard multi-dimensional optimization -> maximum likelihood
• “single rate dated tips”
Maximum likelihood (II)

• Allows hypothesis testing and model comparison via likelihood ratio test
• Test if one hypothesis provides better fit (nested hypotheses)
• Use for step-wise change in substitution rate
• Multiple dated tips models
Bootstrapping

• How much do we trust a tree that we have constructed?

• A simple method for parsimony, distance or ML is bootstrapping.
  – Select some random positions, with repetition.
  – Construct another tree with the bootstrapped data.
  – Repeat many times.
  – Check the consistency of the results.

• In Bayesian methods, one can estimate the confidence by looking at posterior probabilities.
Phylogeny of ST8 and the emergence of USA300

Phylogenetic tree:
- 433 isolates
- Additional isolates
  2005 / 2006 study
  California SSTIs
- 12,212 SNPs in core genome
- Maximum likelihood tree with 1000 bootstraps
- Homoplasy index 0.007

Rooted to midpoint

Cases
Case contact
Control
California isolates
PVL in USA300 core lineage differs from other ST8

Distribution of pairwise distances

MSSA, USA500

220 SNPs
Very limited variation in ACME suggests clonal dissemination.
Mapping of mobile genetic elements matches core phylogeny

A

ACME & SaPI5
- present
- absent

SCCmec
- IVa
- IVb
- IVc
- IVg
- absent

PVL-prophage (Sa2int)
- subtype 1 (USA500)
- subtype 2
- subtype 3
- subtype 4
- subtype 5

~ 53 SNPs

B

USA300 FPR3757* (SCCmecIVa SaPI5+ ACME+)

SCCmec

ACME

SCCmec

ACME remnant (70 bp)

7748_4#59* (SCCmecIVa SaPI5+ ACME-)
What determines “strain similarity”

- Substitution rate
  - Root-to-tip analysis
  - Bayesian reconstruction (subset of isolates)
- Pairwise SNP distance
Root-to-tip linear regression

• First estimates rooted phylogeny
  - matrix pairwise genetic distance using empiric model of substitution
  - matrix used for neighbor-joining tree

• Second linear regression between time of sampling of each tip and genetic distance (sum of reconstructed branch length)

\[ E[d_{root,i} \mid t_i - t_{root}] = mt_i - mt_{root} \]

• Root of tree picked to maximize \( R^2 \) value of regression
• Advantage: fast visualization
• Not the final model!
Root-to-tip analysis to estimate date of ancestry

- Substitution rate/site/year: $1.56 \times 10^{-6}$
- Time most common recent ancestor: ~1995

Correlation coefficient 0.4853
$R^2$: 0.2355

Uhleman et al. PNAS 2014
USA300 substitution rate comparable to other MRSA clones

Corresponds to ~4 SNPs per year
Bayesian interference of evolutionary rate

• Phylogeny as how to assign probability to different trees given that we observed some sequences.
  – We can think that we do not know the right history but a few histories can be compatible.
  – P(T|D): probability of a tree given the data.
    • that is the inverse of likelihood: P(D|T).

• Uses Markov chain Monte Carlo

• To estimate substitution rates includes:
  - tree topology
  - times of ancestral nodes
  - substitution are (?)
  - substitution parameters (transition/transversion)
Phylogeographic reconstruction

Support for root in Fort Washington neighborhood (site of CUMC)

Uhlemann et al. PNAS 2014
Have unique genomic USA300 subpopulations emerged?

Stop codons:

- **wbrA**
  - 29 isolates, 13 households, 20 months apart
  - Tryptophan-repressor binding protein, NADP(H)-quinone-oxidoreductase, oxidative stress response?

- **ebh**
  - 10 isolates, 6 households
  - ECM binding
Expansion of Fluoroquinolone-resistant clone (*gyrA / grlA* SNPs)

CDC survey
- Decrease in FQ-susceptibility: 63% to 45% from 2004 – 2008
- National prescription data overlap with FQ-R prevalence
- Additional 15 non-synonymous SNPs associated with gyrA/grlA
Time scaled evolution of USA300

USA300 subclade

ST8  SaPi5  ACME & PVL
53 non-synon SNPs
20 years of bacterial genome sequencing

The First Revolution
Whole-genome shotgun

- Sanger shotgun sequencing
  - Sequencing by synthesis
  - Amplified templates generated in vivo
  - Requires onerous colony picking and plasmid preparation

  For example, ABI capillary sequencer (ABI)

The Second Revolution
High-throughput sequencing

- Illumina sequencing
  - Sequencing by synthesis
  - Amplified templates generated in vitro
  - High accuracy but short read lengths

  For example, MiSeq (Illumina)

- 454 sequencing
  - Sequencing by synthesis
  - Amplified templates generated in vitro
  - High accuracy outside homopolymers but short read lengths

  For example, 454 GS FLX+ (Roche)

The Third Revolution
Single-molecule sequencing

- Pac Bio SMRT sequencing
  - Sequencing by synthesis
  - Single-molecule templates
  - Low accuracy but long read lengths

  For example, PacBio RS (Pacific Biosciences)

- Oxford Nanopore sequencing
  - Nanopore sequencing
  - Single-molecule templates
  - Low accuracy but long read lengths

  For example, MinION (Oxford Nanopore)
Summary bacterial WGS

- Short-read sequencing
- Comparative analysis
  - variant detection of known traits
  - repetitive sequences still challenging
  - achieves ~90% information on genome
  - plasmids even more difficult
- De novo assemblies for novel gene content
Microbiome analyses
Given that fact, it shouldn't surprise us that microbes occupy most of us as well. Seventy to ninety percent of all cells in the human body are bacterial, representing perhaps 10,000 different species. Genetically we get even less real estate: 99 percent of the unique genes in our bodies are bacterial. This population of over 100 trillion microorganisms makes up our microbiome: a collection of microbial communities that has evolved along with homo sapiens to help orchestrate basic life processes, beginning the moment we're born.

So it really shouldn't surprise us that the microbiome plays a major role in health, especially immunity and metabolism. Nor that disrupting this ancient equilibrium could have serious consequences. But that was the last thing on anyone's mind when antibiotics, which destroy bacteria or slow their growth, came on the scene in the 1940s. "We'd just won World War II, we had dropped the bomb, we were invincible, and now we had these incredible drugs," says Blaser. Formerly lethal diseases could be prevented or cured. Surgery grew far safer. Side effects appeared to be few and mild.

Many bacterial species previously not recognized because unculturable with current methods.
Microbiota? Or Microbiome?

Microbiota
16S rRNA
Taxonomic identification

Metagenome
Genes and genomes of microbiota

Microbiome
Genes, genomes, products, host proteins

*Nat. Rev. Urol.* doi:10.1038/nrurol.2014.361
16s rRNA sequencing

- 16S rRNA gene present in all bacterial species
- Highly conserved and variable sequences
- Variable = “molecular fingerprint”
- Amplification with degenerate primers targeting conserved regions
- Large public database for comparisons
Taxonomy assignment

• Challenges:
  multiple matches
  no match (new OUT)

• Some species may share
  >97% similarity, no resolution
  at species level
Output taxa distribution

- Bar chart
- Heat map

Mean Relative Abundance of Genera By Sample Group (Week 1-4)

Sample Group/Timepoint

Color Key

- Escherichia
- Bacteroides
- Veillonella
- Planomicrobiurn
- Serratia
- Rhodococcus
- Streptococcus
- Pantoea
- Enterobacter
- Staphylococcus
Alpha diversity

- Diversity within a sample
  - taxon based
  - phylogeny based
- Richness – number of species present
  - Chao-index
- Evenness – abundance of different species
  - Shannon index
Beta diversity

• Comparisons of samples to each other
• How different are types present?
• Measure of distance / dissimilarity between sample pair
• UniFrac (weighted, unweighted)
UniFraq example
Principal Coordinate Analysis

- Visualization of beta diversity matrix
- Transform distance matrix into new set of orthogonal axes
- 2D or 3D
• Open source bioinformatics platform
  - data analysis from raw reads to figures
• Qiita: online data repository / data analysis platform
Additional thoughts on 16S rRNA

• Uniqueness of variable region determines taxonomic resolution
  - V3 or V3-V4 or V1-V2
  - length of variable region
• Optimal resolution depends on sample composition
• Extraction methods (Gram-positive versus Gram-negative!) may play an important role in full recovery of species
Predictive functional profiling of microbial communities by 16S rRNA genes

- PICRUSt software
- Validated using HMP data
Simulation cannot predict clonal variants within species and genetic content in mobile genetic elements

Presence of arcD (arginine metabolism) results in decreased virulence, immune evasion

Ahn et al. JCI Insight. In Press
Bioinformatic methods for functional metagenomics

Studies that aim to define the composition and function of uncultured microbial communities are often referred to collectively as “metagenomic,” although this refers more specifically to particular sequencing-based assays. First, community DNA is extracted from a sample, typically uncultured, containing multiple microbial members. The bacterial taxa present in the sample are then amplified and sequenced to produce short DNA sequence reads (OTUs). These reads are then compared to existing databases, such as GreenGenes, myRDP, and Silva, to identify and classify them. The relative abundances of the identified taxa are then used to construct a phylogenetic tree, which provides insights into the diversity and relationships within the sample.

Variant sequences and SNPs are also identified to understand the genetic diversity within the community. The sequences are then compared to reference genomes to determine the community function, i.e., what the community can do. This is achieved by using databases such as KEGG and SEED to identify and classify the functions associated with the sequences. The relative abundance of these functions is then determined to provide a comprehensive understanding of the community's capability.

Morgan & Huttenhower, PLOS Comp. Biol., 2012
Metagenomics: challenges for high-throughput remain

• Much more starting material required
• Higher sequencing cost (~3-4:1)
  - depends on depth of coverage
• Large amount of data generated, results in high demand on computing infrastructure for data processing and storage
• May still not allow assignment of mobile genetic elements and reliable identification of SNPs
"Other" microbes: mycobiota

- Sequencing of 18S ss rDNA ITS region
- Longer reads needed
- Growing databases UNITE
- Interaction with innate and adaptive immune system

Underhill, Iliev Nat Rev Immun 2014
“Other” microbes - Virome

• Most abundant and fastest mutating genetic elements
• Previously difficult to sequence / analyze given high diversity
• Difficult to extract (enrichment from filtrates, lysis of bacterial and human cells)
• Different types!
  - Eukaryotic, Bacterial, Archaeaic viruses
  - Integrated elements in human host DNA
• Trans-kingdom interaction
• Direct interaction with host / immune signaling
• Phages regulate bacterial content
A. VirusSeeker-Virome

Sequence pre-processing

1. Total raw sequences from sample
2. Remove adapter
3. Stitch read1 and read2
4. Quality control
5. CD-Hit to define unique sequences
6. RepeatMasker and quality filter
7. Reference genome filtering

B. VirusSeeker-Discovery

Sequence pre-processing

1. Total raw sequences from sample
2. Remove adapter
3. Stitch read1 and read2
4. Quality control
5. Reference genome filtering
6. Unmapped reads
7. CD-Hit to define unique sequences
8. Assemble unique reads + top 3 similar reads
9. Contigs
10. Second assembly of pooled contigs
11. Singletons, outliers
12. RepeatMasker and quality filter

Sequential BLAST module from VS-Virome

- BLASTn vs. Virus-only nt DB
  - No hit or hit e > e-5
  - Phage hit e ≤ e-5
- Candidate virus sequences
  - Virus hit e ≤ e-3
- Bacteria genome filtering
  - Mapped reads
- MegaBLAST vs. NCBI NT DB
  - Hit e ≤ e-10
  - No hit or hit e > e-10
- BLASTn vs. NCBI NT DB
  - Hit e ≤ e-10
  - No hit or hit e > e-10
- BLASTx vs. NCBI NT DB
  - Hit e ≤ e-3
  - No hit or hit e > e-3
- Classify using NCBI Taxonomy DB
  - Unassigned
  - Euk. Virus
  - Phage
  - Ambiguous
  - Non-viral

VirusSeeker, a computational pipeline for virus discovery and virome composition analysis
Virology, Volume 503, 2017, 21–30
Microbiome summary

- Fingerprint of bacterial communities
- Relatively affordable and fast
- Does not provide information on unique functional features (MGEs...)
- Metagenomics will be more comprehensive but currently still expensive / data intensive, limiting widespread use. Difficulties assigning plasmids to organisms