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

• 2020:
  - too many to count?
Sequencing technologies – A moving target

- Illumina
- NoaSeq
- NextSeq
- HiSeq
- Miseq

- SOLiD
  - Ion Torrent

- 454
  - PaqBio

- GnuBio
  - Helioscope

- Nanopore?
Traditional versus Next-Generation Sequencing

SANGER SEQUENCING:

NGS:

1 SEQUENCE READ PER BP

MULTIPLE SEQUENCE READS PER BP
High-throughput Next generation sequencing by synthesis
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 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>Illumina Miseq</strong></td>
<td>8 Gb</td>
<td>27 hrs</td>
<td>2 x 250 bp</td>
<td>Higher indels</td>
<td>GC problems</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Read length Speed</td>
<td></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>SNP’s</td>
<td>Indels</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Speed</td>
<td>Difficulty 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 methylation</td>
<td>SNP’s</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 always optimal for variant calling
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

- Annotation
- Identification of phages, MGEs
**K-mers**

- Sequence of K base calls (DNA that is k long)
  - ATGC = 4-mer, ATGCTG = 6-mer
  - all of a sequence's subsequences of length
- Only consecutive bases are used
- Reads with high sequence similarity must share K-mers in overlapping regions
- Shared K-mers are easier to find then overlaps
- Fast detection of shared K-mer content reduces computational cost / time
- Disadvantage: lower sensitivity in overlap regions
De Bruijn graph assembly

• AACC
  AACC
  ACCG
  CCGG
  CGGT
  GGTT
  GTTA

• GGTT
  GTTA
  TTAT
  TATA
  ATAC

AACC GGT TAT A

Spades: uses k=31 ->127
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
USA300 genome composition

Genomic & biological features:

Core genome
- Increased expression of core virulence genes
  PSM, α-toxin

Mobile genetic elements
- Small SCCmec IV
- ϕ2 / PVL toxin
- SaPi5
- ACME I
  - detoxifies host antimicrobials
- derived from USA500?

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

- Clinical USA300 (n = 131)
  - Serial infections (n = 22)
  - USA300 colonization (n = 138)
    - Environmental (n = 51)
      - Outside HH contacts (n = 15)
      - Total, n = 357
- 161 cases

- USA300 colonization (n = 14)
  - Environmental (n = 14)
    - Outside HH contacts (n = 2)
    - Total, n = 30
- 161 controls

Total, n = 357
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
Phylogenetic 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**
- **Branch Length**
- **Root**
- **Branch**
- **Fly**
- **Clade**
- **Human**
- **Mouse**
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/tRNA</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>9274 1#15</th>
<th>7748 4#95</th>
<th>9266 1#16</th>
<th>7750 7#31</th>
<th>7748 4#96</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2079_1</td>
<td>2079_2</td>
<td>2079_2</td>
<td>2079_2</td>
<td>2079_3</td>
<td>2079_3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>108951</td>
<td>Intergenic</td>
<td>-</td>
<td>-</td>
<td>G</td>
<td>T</td>
<td>6</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>N</td>
<td>T</td>
<td>T</td>
<td>T</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>6</td>
<td>T</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>6</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>N</td>
<td>T</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>T</td>
<td>N</td>
<td>T</td>
<td>T</td>
<td>T</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>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>110608</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipopr</td>
<td>N</td>
<td>C</td>
<td>T</td>
<td>5</td>
<td>N</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>110610</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipopr</td>
<td>S</td>
<td>A</td>
<td>T</td>
<td>5</td>
<td>N</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>110613</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipopr</td>
<td>S</td>
<td>A</td>
<td>T</td>
<td>5</td>
<td>N</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>110709</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipopr</td>
<td>S</td>
<td>C</td>
<td>T</td>
<td>3</td>
<td>N</td>
<td>T</td>
<td>N</td>
<td>T</td>
<td>N</td>
<td>T</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>2031077</td>
<td>Intergenic</td>
<td>-</td>
<td>-</td>
<td>T</td>
<td>G</td>
<td>72</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>G</td>
<td>G</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>2</td>
<td>N</td>
<td>N</td>
<td>C</td>
<td>N</td>
<td>C</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>110614</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipopr</td>
<td>N</td>
<td>A</td>
<td>G</td>
<td>5</td>
<td>N</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>N</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>110615</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipopr</td>
<td>S</td>
<td>A</td>
<td>G</td>
<td>5</td>
<td>N</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>N</td>
<td>G</td>
</tr>
<tr>
<td>110620</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipopr</td>
<td>A</td>
<td>G</td>
<td>5</td>
<td>N</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>N</td>
<td>G</td>
</tr>
<tr>
<td>110634</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipopr</td>
<td>A</td>
<td>C</td>
<td>3</td>
<td>N</td>
<td>C</td>
<td>N</td>
<td>C</td>
<td>N</td>
<td>C</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>110718</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipopr</td>
<td>T</td>
<td>A</td>
<td>3</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>110628</td>
<td>CDS</td>
<td>SAUSA300_0100</td>
<td>staphylococcal tandem lipopr</td>
<td>T</td>
<td>A</td>
<td>3</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</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>3</td>
<td>N</td>
<td>A</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>N</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>N</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>N</td>
</tr>
<tr>
<td>1630757</td>
<td>Intergenic</td>
<td>-</td>
<td>-</td>
<td>T</td>
<td>A</td>
<td>215</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>A</td>
</tr>
<tr>
<td>107426</td>
<td>Intergenic</td>
<td>-</td>
<td>-</td>
<td>G</td>
<td>A</td>
<td>21</td>
<td>A</td>
<td>A</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>960514</td>
<td>Intergenic</td>
<td>-</td>
<td>-</td>
<td>C</td>
<td>A</td>
<td>13</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>A</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>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</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
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

- Principle: Choose the tree which makes the data most probable
- Each position evolves independently
- 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
Maximum likelihood (II)

• Allows hypothesis testing and model comparison via likelihood ratio test
• Test if one hypothesis provides better fit (nested hypotheses)
• Problem: can be time / computationally intensive
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

Cases
Case contact
Control
California isolates

Rooted to midpoint
Mobile genome analysis - PVL in USA300 core lineage differs from other ST8

Distribution of pairwise distances

MSSA, USA500

220 SNPs
Mapping of mobile genetic elements matches core phylogeny

ACME & SaPI5
- Present
- Absent

SCCmec
- IVa
- IVb
- IVc
- IVg
- MSSA

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

~ 53 SNPs

A

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)
  \[ E[d_{root,i}] = m(t_i - t_{root}) = mt_i - mt_{root} \]
• 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

Correlation coefficient 0.4853
$R^2$: 0.2355

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

Uhlemann 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

- 1927: PP = 1
- 1993: PP = 1
- 1995: PP = 0.92

USA300 subclade

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

The First Revolution
Whole-genome shotgun

- Strand to be sequenced
- Shotgun
- Sequencing
- Assembly
- Genomic sequence
- 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

- 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)
Short versus long-range sequencing to build better genome scaffolds
MinION
portable real-time sequences

- Oxford Nanopore: nanopore and motor protein
- Real-time data visualization
<table>
<thead>
<tr>
<th>Temperature</th>
<th>Voltage</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>MinION: 36.99°C</td>
<td>-180.00 mV</td>
<td>Delay: 0 ms</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>unclassified</th>
<th>Out of range 1</th>
<th>Zero</th>
<th>Out of range 2</th>
<th>Single pore</th>
<th>Possible multiple</th>
<th>Unavailable pore</th>
<th>Threshold triggered</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>47</td>
<td>7</td>
<td></td>
<td>371</td>
<td>15</td>
<td>16</td>
<td>0</td>
</tr>
</tbody>
</table>

| 56 | Saturated |

[Graph depicting pore distribution]
Ceftazidime-Avi E-tests

**Phenotype 1**
- Meropenem: 128
- CAZ-AVI: 3

**Phenotype 2**
- Meropenem: 2
- CAZ-AVI: >256

**Phenotype 3**
- Meropenem: >128
- CAZ-AVI: 12
Increased KPC read depth in phenotype 3 isolates
Integrated plasmid blaOXA-1, CTX-M-1
ompK35
ompK36
Rcs gene locus (phosphorelay, membrane integrity, virulence)

IS5376-KPC-2 \((Tn4401e)\)
tus-merRPCA-tnsB
xer-C
Integrated plasmid blaOXA-1, CTX-M-1
ompK35
ompK36
Rcs gene locus

Phenotype 2
NR5632
p1
p2
D179Y
IS5376-KPC-2 D179Y
tus-merRPCA-tnsB
xer-C
Transposition of KPC-gene into P2
Novel transposon organization

Novel stop codon in *ompK35*
IS1 insertion at -48 of *ompK36*
Shortening of *rcs*-encoding locus -> adaption of virulence / host?
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, much improved with long-read sequencing
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)

- Escherichia
- Bacteroides
- Veillonella
- Planomicrobium
- Serratia
- Rhodococcus
- Streptococcus
- Pantoea
- Enterobacter
- Staphylococcus

Sample Group/Timepoint

**Color Key**

0.1 - 0.3 Value
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

Lozupone & Knight 2005 AEM 71:8228
Principal Coordinate Analysis

- Visualization of beta diversity matrix
- Transform distance matrix into new set of orthogonal axes
- 2D or 3D
QIIME / QIITA

- 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

- PICRUSSt 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 amplified and sequenced to generate reads. These reads are then grouped into operational taxonomic units (OTUs) based on similarity. The resulting OTUs are classified using databases such as GreenGenes, myRDP, and Silva. The community composition is then analyzed to determine which organisms are present, and their relative abundance in the community is assessed. The OTU phylogeny is also used to infer the evolutionary relationships among the taxa. Community function is determined by comparing the sequences to reference genomes using tools like KEGG and SEED, and the relative abundance of gene pathways in the community is evaluated.
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
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