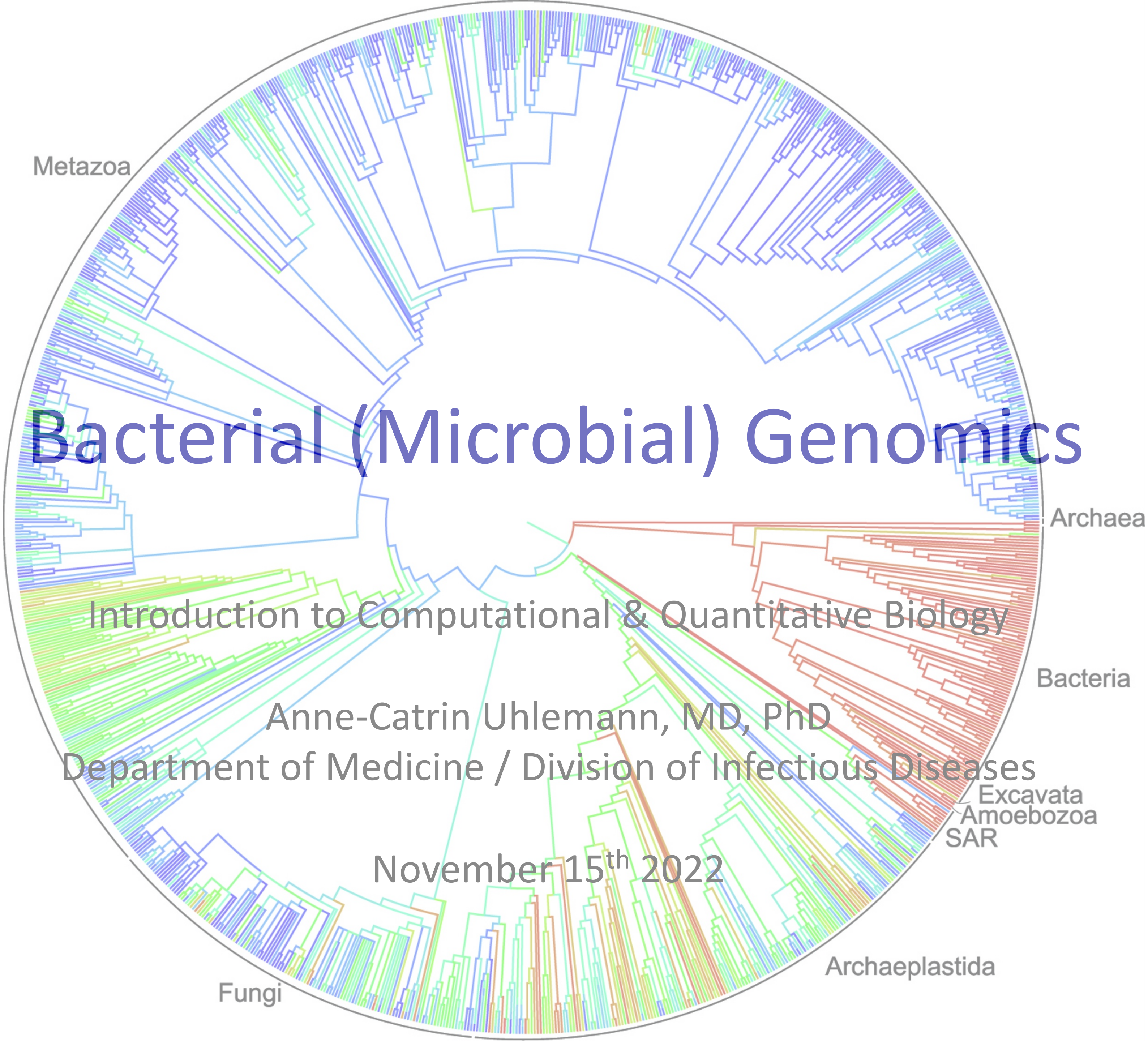
Bacterial (Microbial) Genomics

Introduction to Computational & Quantitative Biology

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

November 15th 2022



Overview

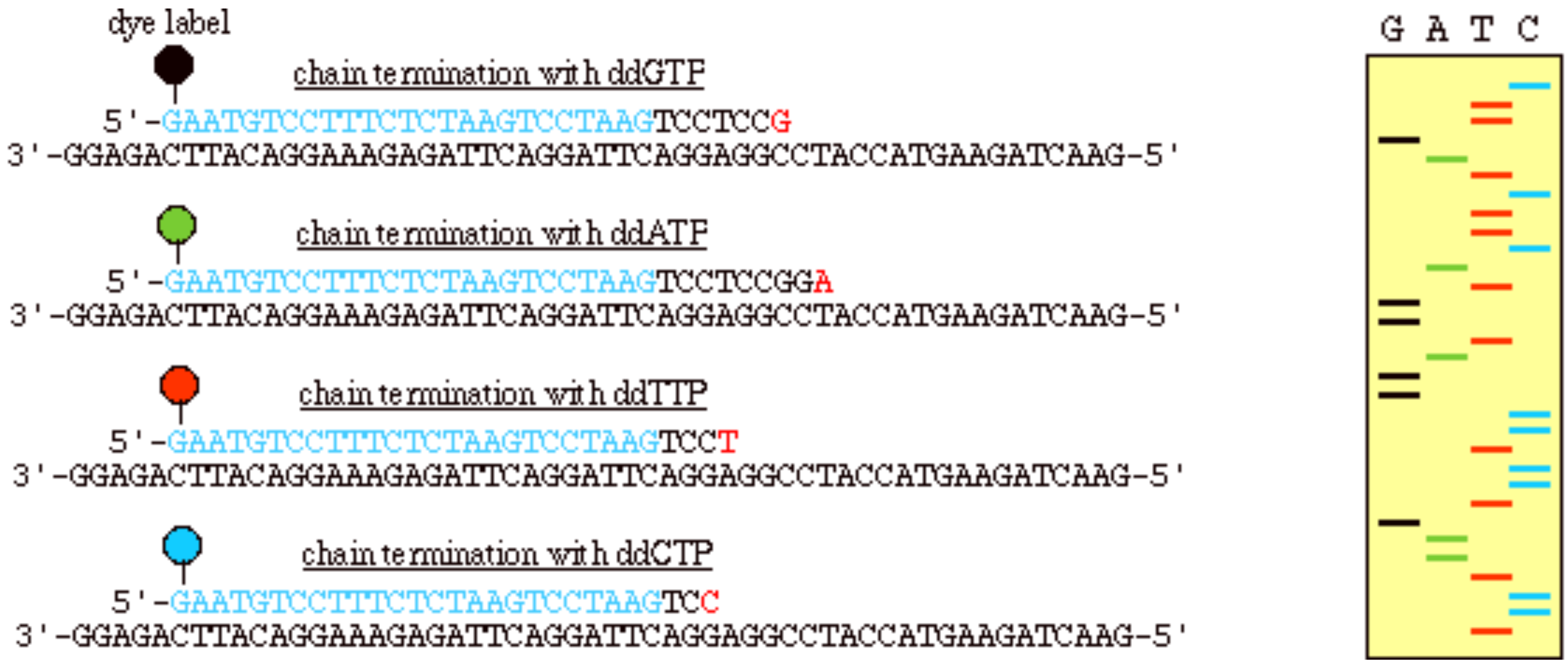
- Sequencing technologies
- Analysis approaches

  - Bacterial whole genome sequencing

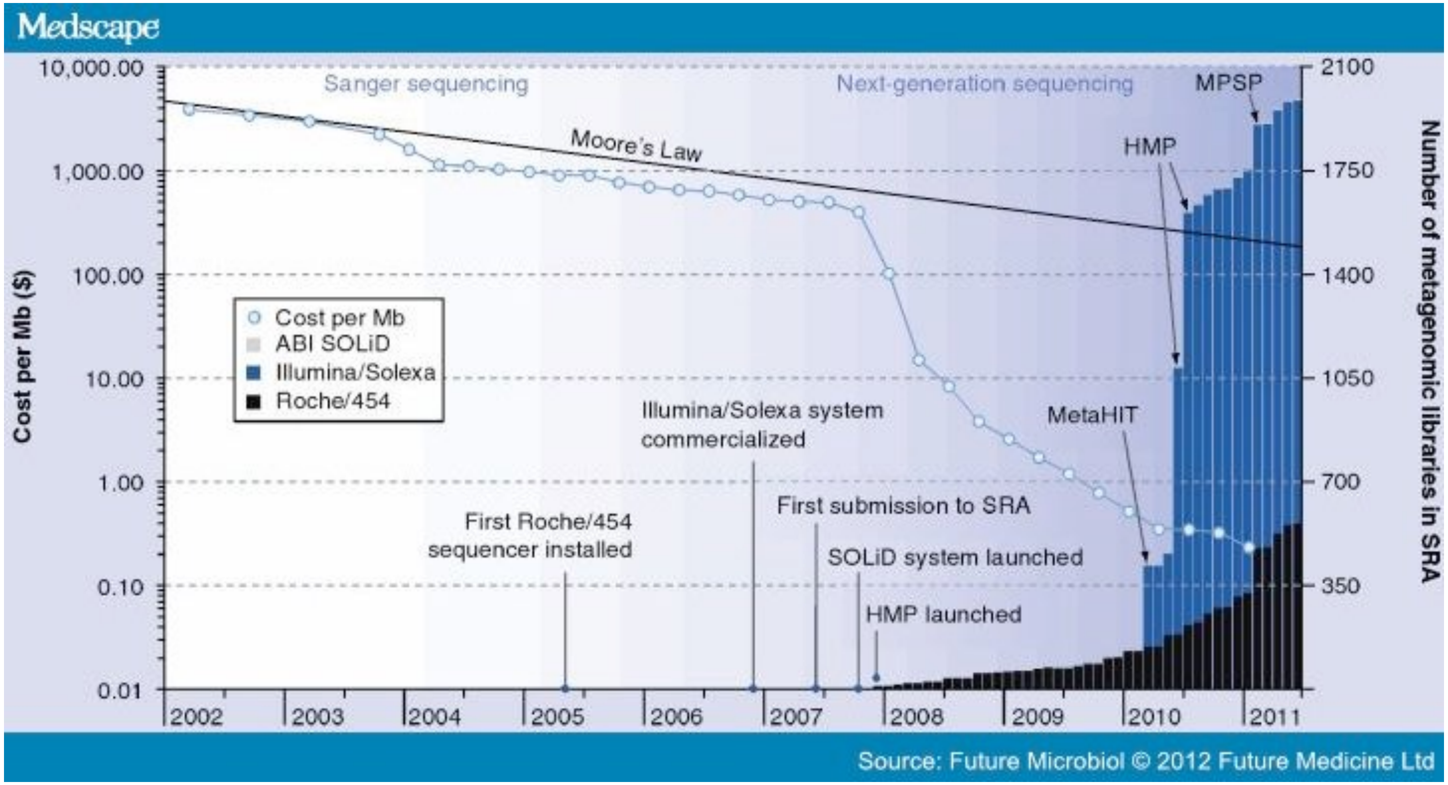
  - Microbiome analyses
- Some practical examples
- Maybe a few words on COVID-19 at the end

Recap: Sanger sequencing

- Dideoxynucleotide sequencing
- Chain-termination method

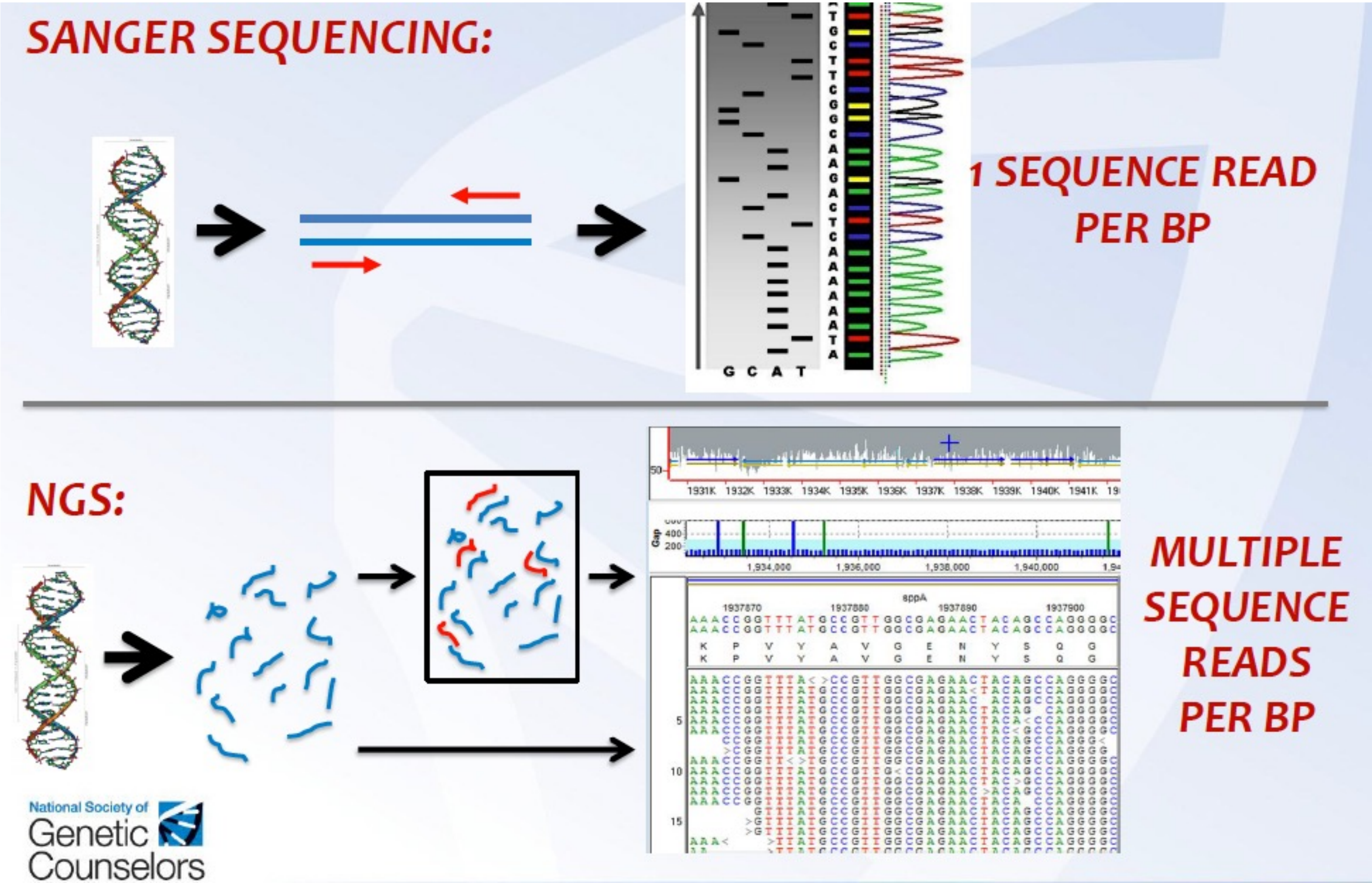

dye label
chain termination with ddGTP
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
chain termination with ddATP
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGA
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
chain termination with ddTTP
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCT
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
chain termination with ddCTP
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCC
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
G A T C

Microbial genomics – byproducts of the race for the human genome



Source: Future Microbiol © 2012 Future Medicine Ltd

Traditional versus Next-Generation Sequencing



SANGER SEQUENCING:

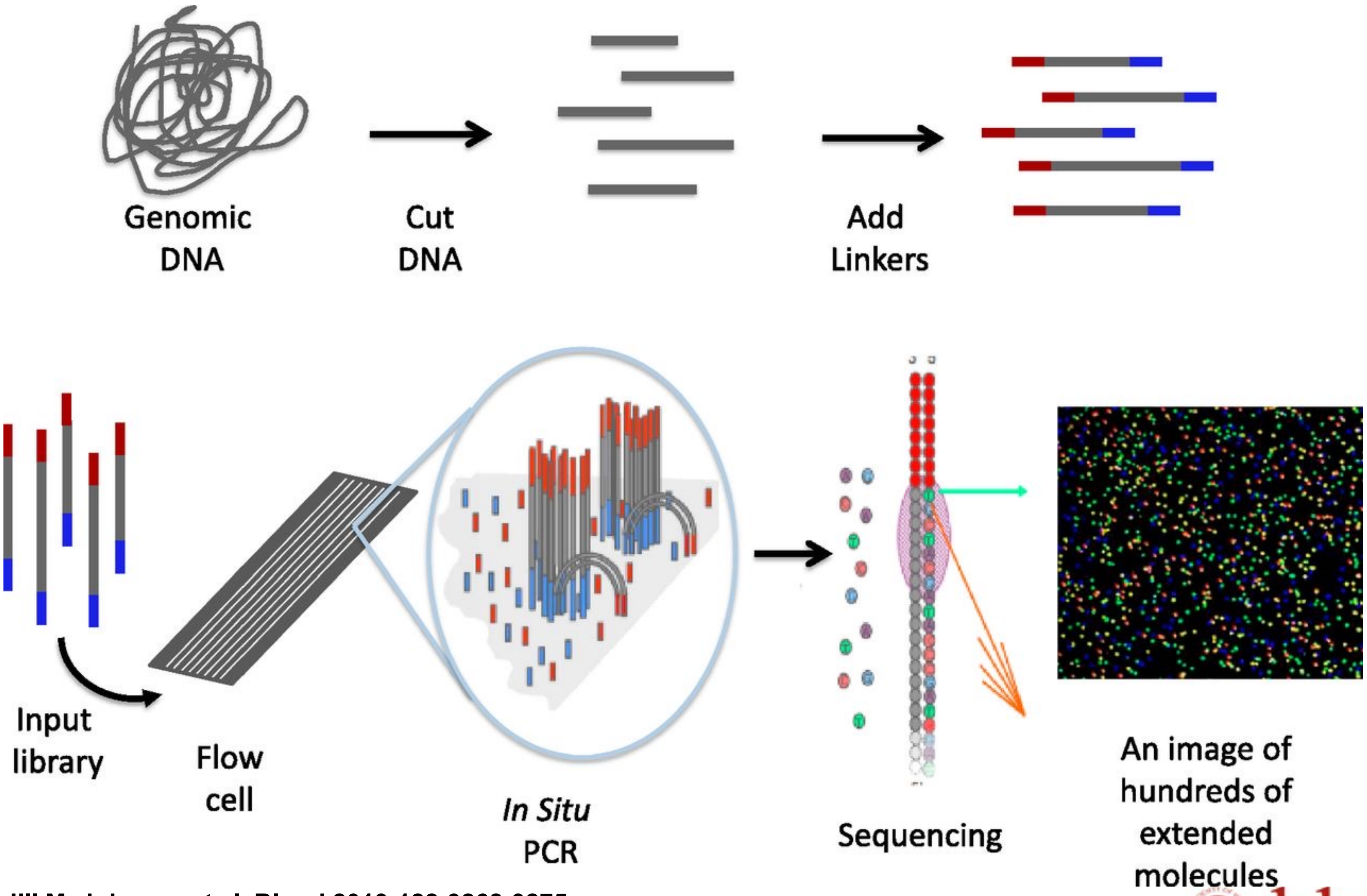
1 SEQUENCE READ PER BP

NGS:

MULTIPLE SEQUENCE READS PER BP

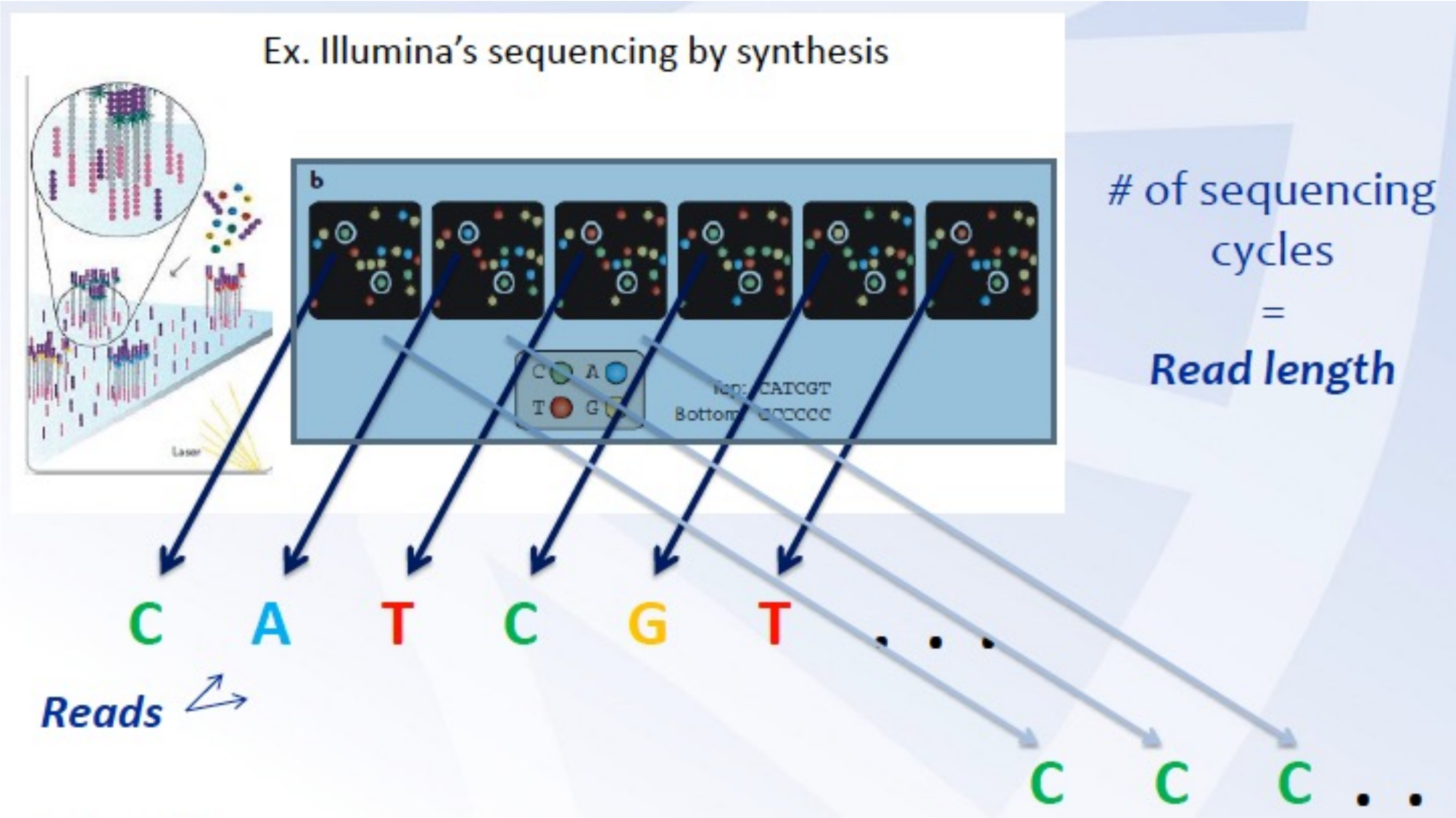
National Society of Genetic Counselors

High-throughput Next generation sequencing by synthesis



Jill M. Johnsen et al. Blood 2013;122:3268-3275

Read length



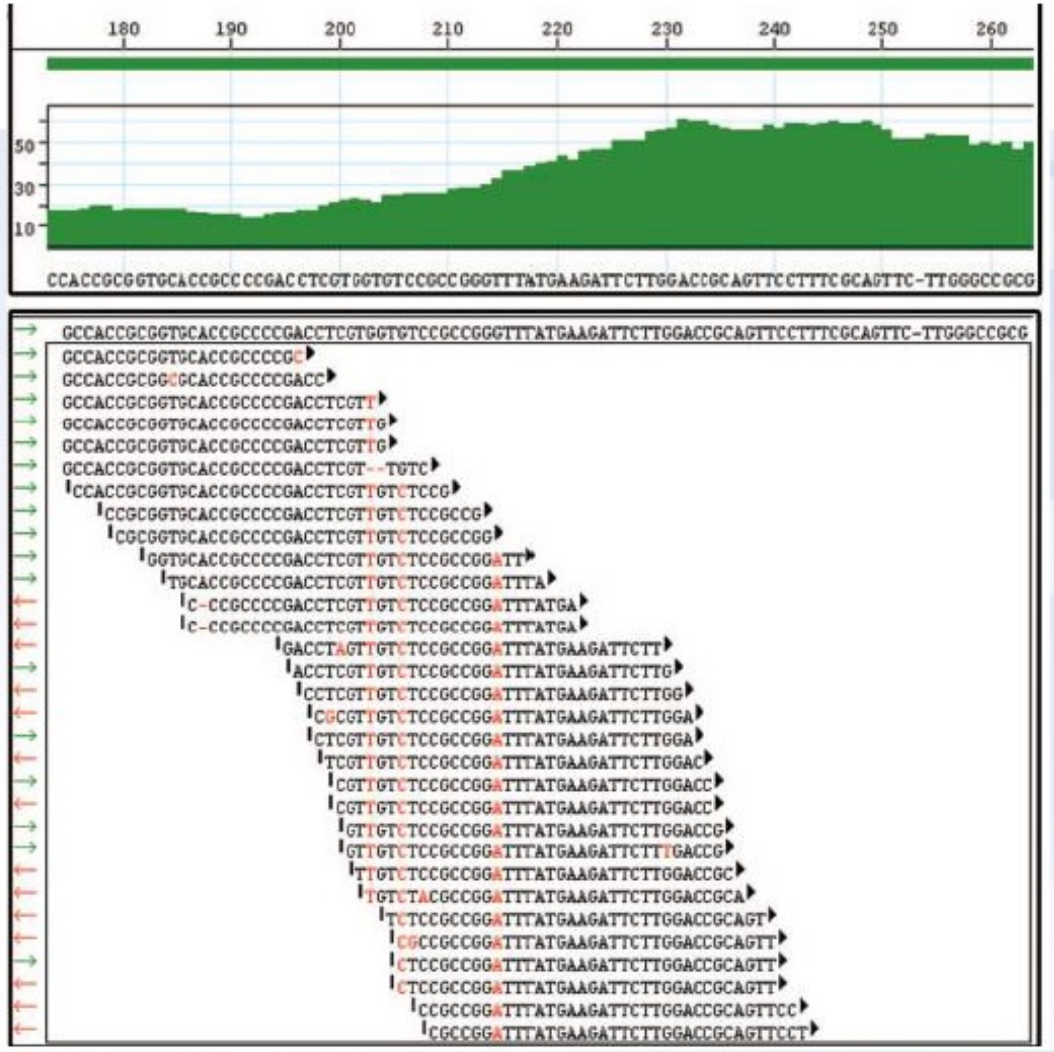
Ex. Illumina's sequencing by synthesis

# of sequencing cycles = *Read length*

Chemistries limit read length, are constantly being improved

- short	< 50 consecutive bases
- mid-length	51 - <400
- long	> 400 (< 1000)

Depth of coverage



```
CCACCGCGGTGCACCGCCCCGACCTCGTGGTGTCCGCCGGGTTTATGAAGATTCTTGGACCGCAGTTCCTTTCGCAGTTC-TTGGGCCGCG

→ GCCACCGCGGTGCACCGCCCCGACCTCGTGGTGTCCGCCGGGTTTATGAAGATTCTTGGACCGCAGTTCCTTTCGCAGTTC-TTGGGCCGCG
→ GCCACCGCGGTGCACCGCCCCGC
→ GCCACCGCGGCGCACCGCCCCGACC
→ GCCACCGCGGTGCACCGCCCCGACCTCGTT
→ GCCACCGCGGTGCACCGCCCCGACCTCGTTG
→ GCCACCGCGGTGCACCGCCCCGACCTCGTTG
→ GCCACCGCGGTGCACCGCCCCGACCTCGT--TGTC
→  CCACCGCGGTGCACCGCCCCGACCTCGTTGTCTCCG
→     CCGCGGTGCACCGCCCCGACCTCGTTGTCTCCGCCG
→      CGCGGTGCACCGCCCCGACCTCGTTGTCTCCGCCGG
→         GGTGCACCGCCCCGACCTCGTTGTCTCCGCCGGATT
→           TGCACCGCCCCGACCTCGTTGTCTCCGCCGGATTTA
←             C-CCGCCCCGACCTCGTTGTCTCCGCCGGATTTATGA
←             C-CCGCCCCGACCTCGTTGTCTCCGCCGGATTTATGA
←                      GACCTAGTTGTCTCCGCCGGATTTATGAAGATTCTT
→                       ACCTCGTTGTCTCCGCCGGATTTATGAAGATTCTTG
←                        CCTCGTTGTCTCCGCCGGATTTATGAAGATTCTTGG
←                         CGCGTTGTCTCCGCCGGATTTATGAAGATTCTTGGA
→                         CTCGTTGTCTCCGCCGGATTTATGAAGATTCTTGGA
←                          TCGTTGTCTCCGCCGGATTTATGAAGATTCTTGGAC
→                           CGTTGTCTCCGCCGGATTTATGAAGATTCTTGGACC
←                           CGTTGTCTCCGCCGGATTTATGAAGATTCTTGGACC
→                            GTTGTCTCCGCCGGATTTATGAAGATTCTTGGACCG
→                            GTTGTCTCCGCCGGATTTATGAAGATTCTTTGACCG
←                             TTGTCTCCGCCGGATTTATGAAGATTCTTGGACCGC
←                              TGTCTACGCCGGATTTATGAAGATTCTTGGACCGCA
←                                TCTCCGCCGGATTTATGAAGATTCTTGGACCGCAGT
←                                 CGCCGCCGGATTTATGAAGATTCTTGGACCGCAGTT
→                                 CTCCGCCGGATTTATGAAGATTCTTGGACCGCAGTT
←                                 CTCCGCCGGATTTATGAAGATTCTTGGACCGCAGTT
←                                   CCGCCGGATTTATGAAGATTCTTGGACCGCAGTTCC
←                                    CGCCGGATTTATGAAGATTCTTGGACCGCAGTTCCT
```

Numbers can be misleading!

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

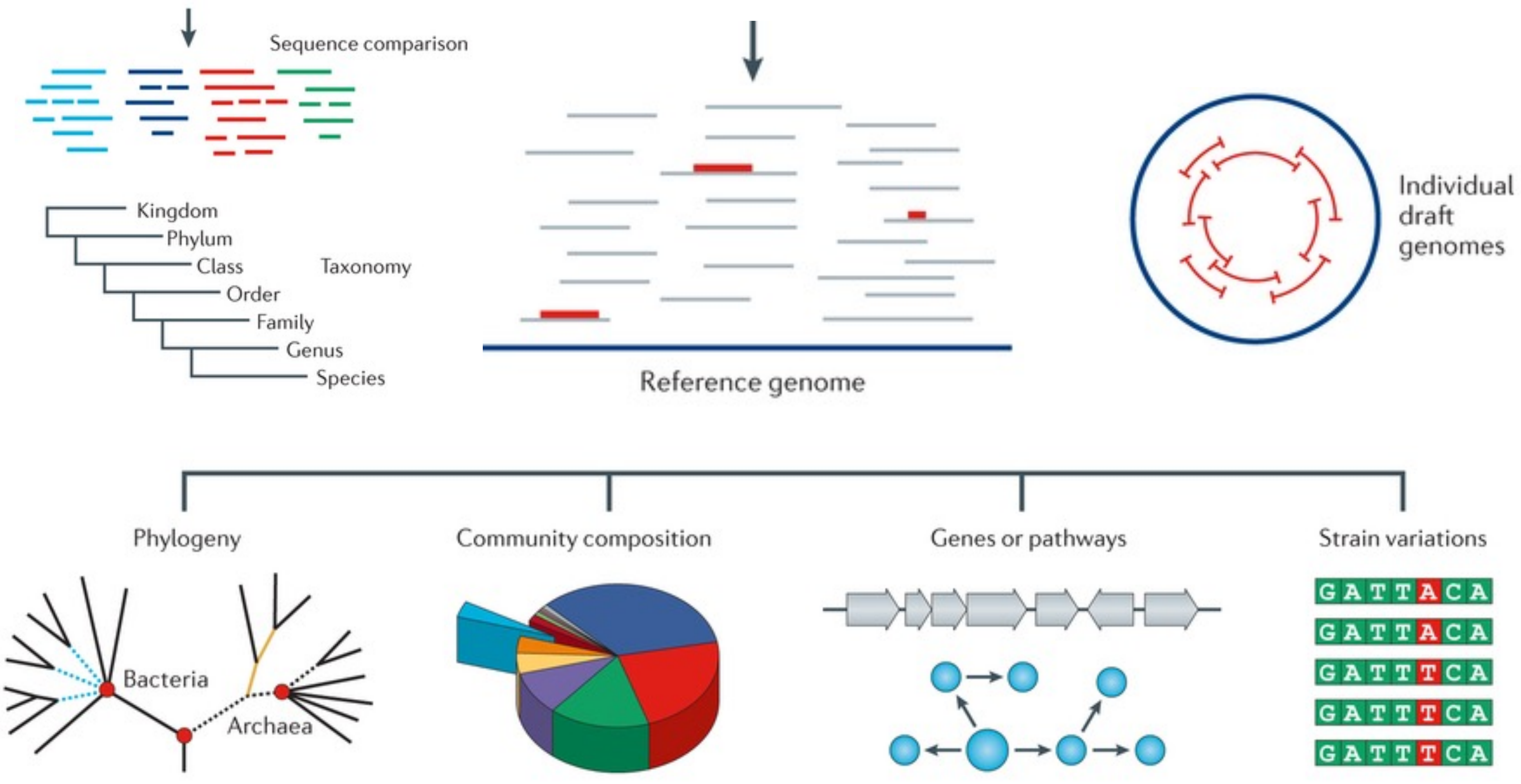
Bacterial genomics overview

Bacterial sequencing applications

16S rRNA sequencing

Metagenomics

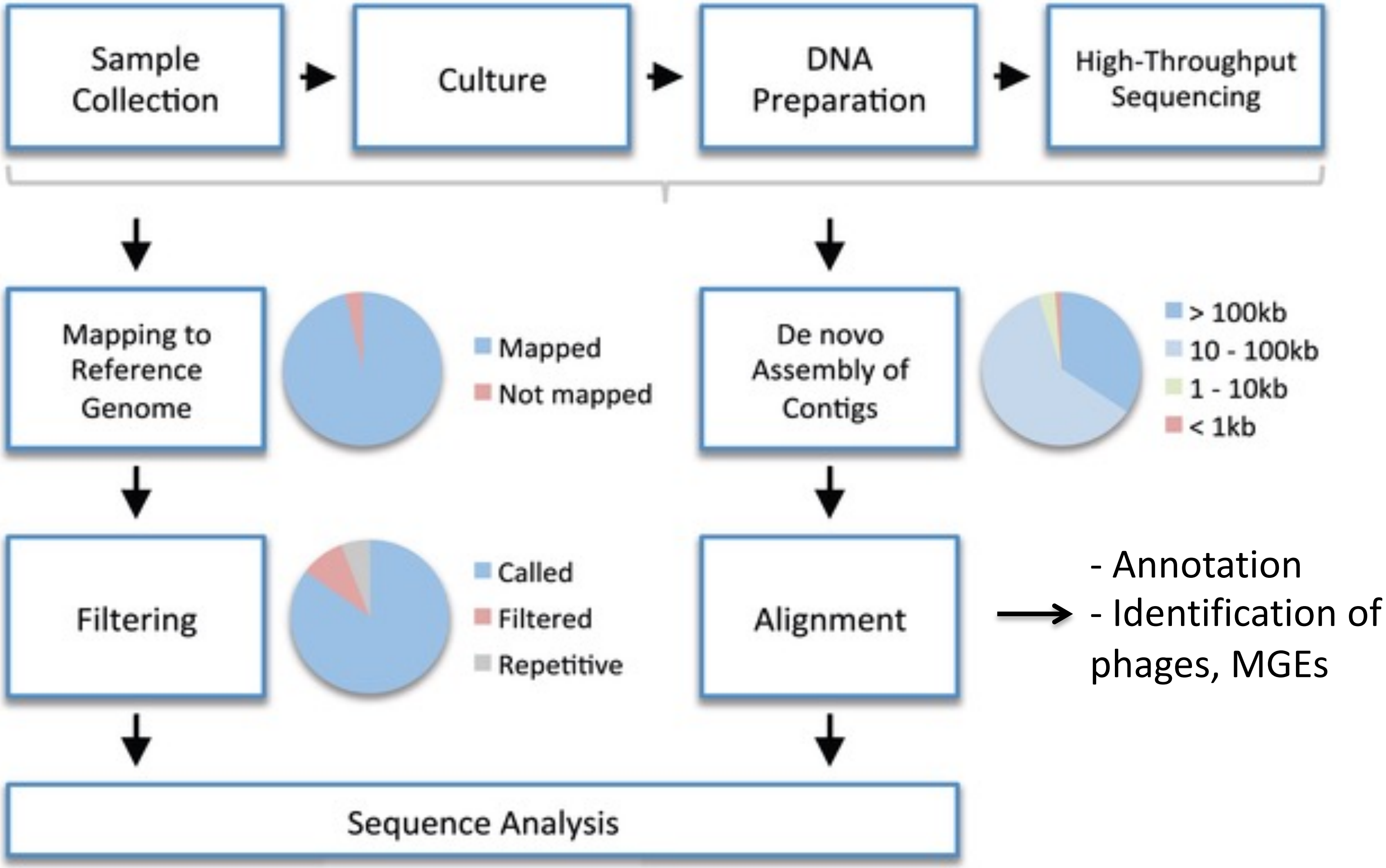
Single cell sequencing



Single isolate bacterial sequencing

- Comparative sequencing
  - SNPs / indels that determine virulence
  - evolution
  - outbreak investigations
  - compare presence / absence mobile elements
- *De novo* assembly
  - only way to determine new gene content
  - not always optimal for variant calling

Figure 1. An example workflow for high-throughput whole genome sequencing in bacteria.



Sample Collection → Culture → DNA Preparation → High-Throughput Sequencing

Mapping to Reference Genome

🟦 Mapped
🟥 Not mapped

Filtering

🟦 Called
🟥 Filtered
⬜ Repetitive

De novo Assembly of Contigs

🟦 > 100kb
⬜ 10 - 100kb
🟩 1 - 10kb
🟥 < 1kb

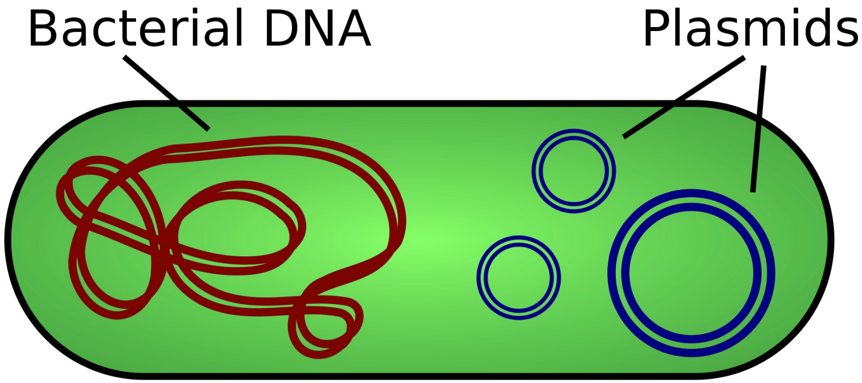
Alignment

- Annotation
- Identification of phages, MGEs

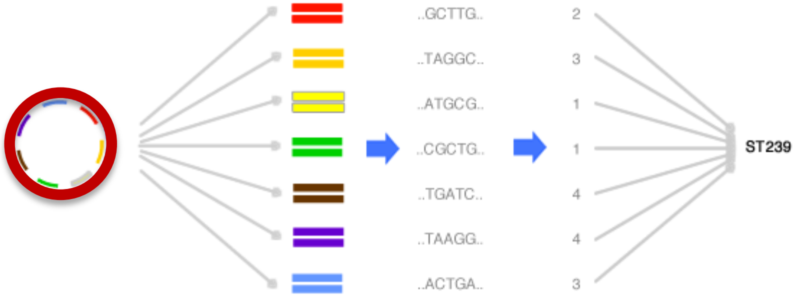
Sequence Analysis

Wilson DJ (2012) Insights from Genomics into Bacterial Pathogen Populations. PLOS Pathogens 8(9): e1002874. doi:10.1371/journal.ppat.1002874
http://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1002874

Typing schemes of bacterial DNA

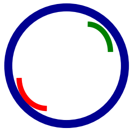


Multi-locus sequence typing (housekeeping genes)



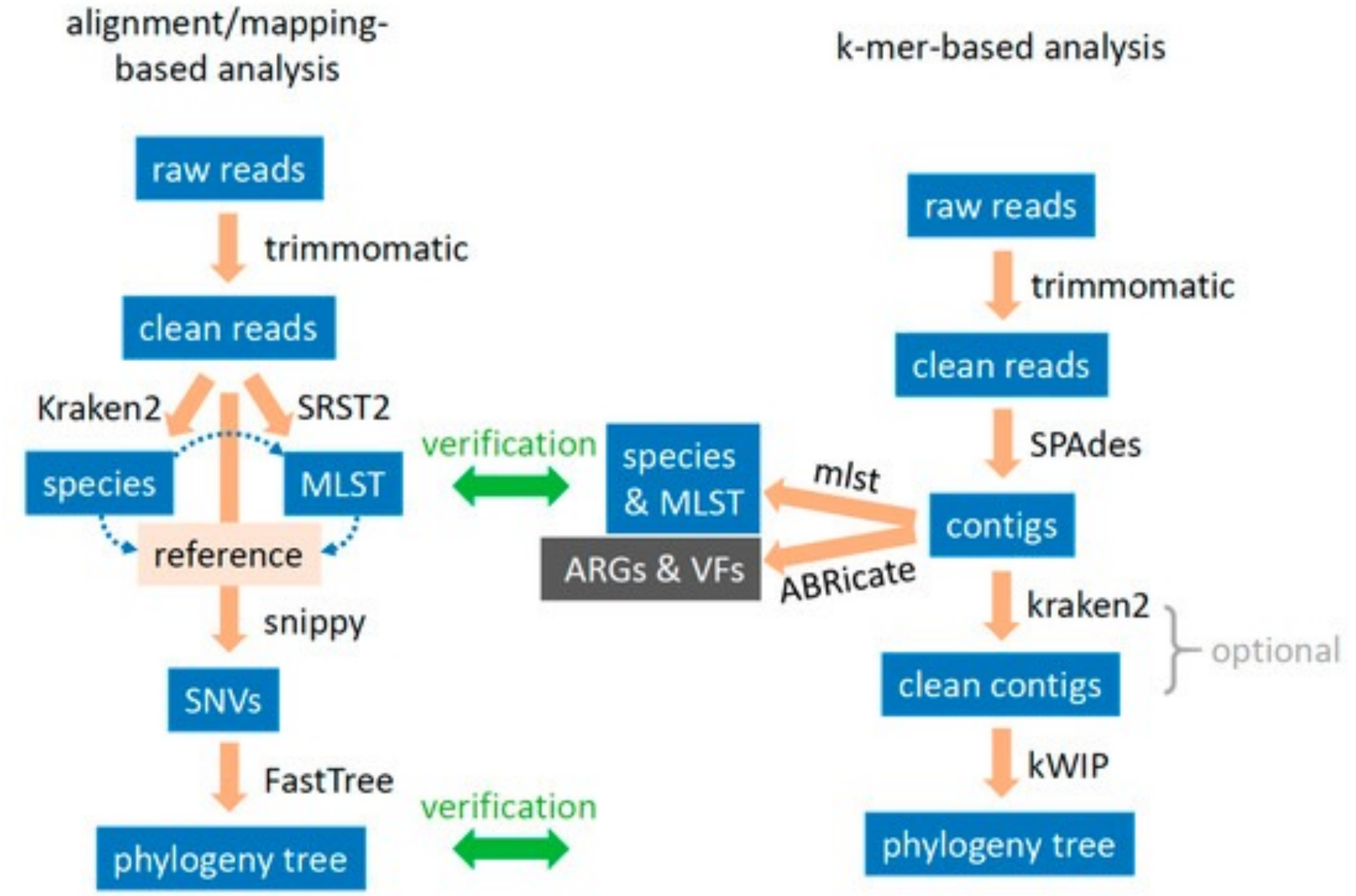
Plasmid PCR typing

- *rep* gene
- $\mathit{bla}_{\mathrm{KPC}}$ gene mutations



MLST reflects clonal type quite well

Workflows for comparative analyses



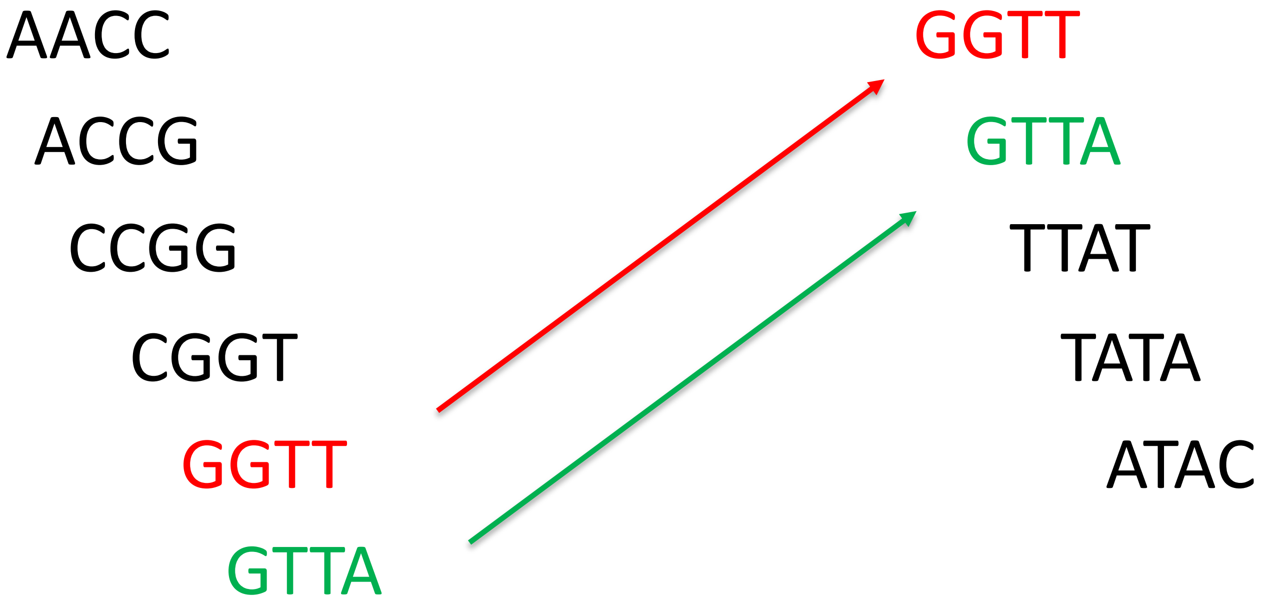
Huang et al. Microorganisms 2019

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

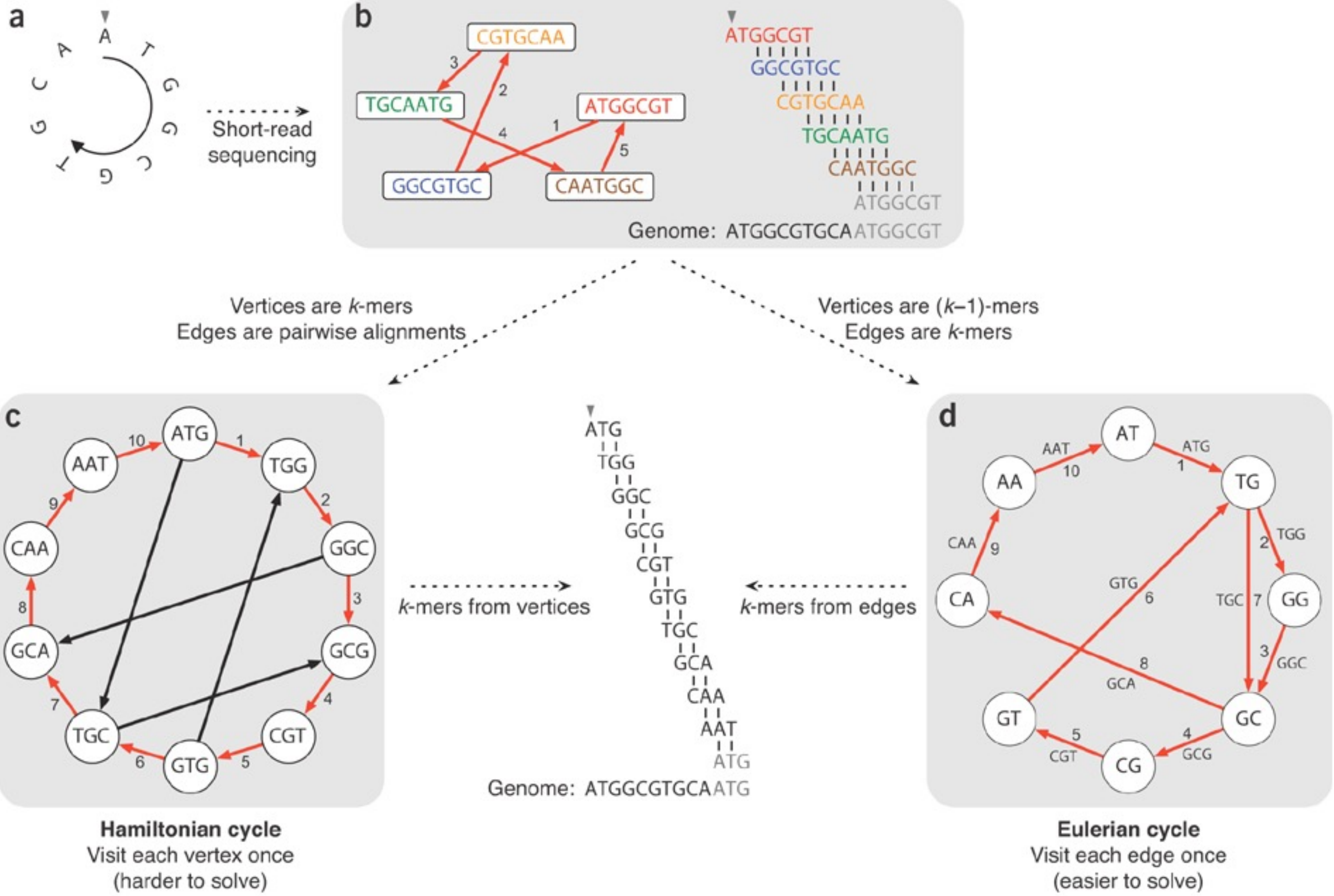
- AACCGGTTA
- GGTTATAC


AACC
ACCG
CCGG
CGGT
GGTT
GTTA

GGTT
GTTA
TTAT
TATA
ATAC

AACCGGTTATAC

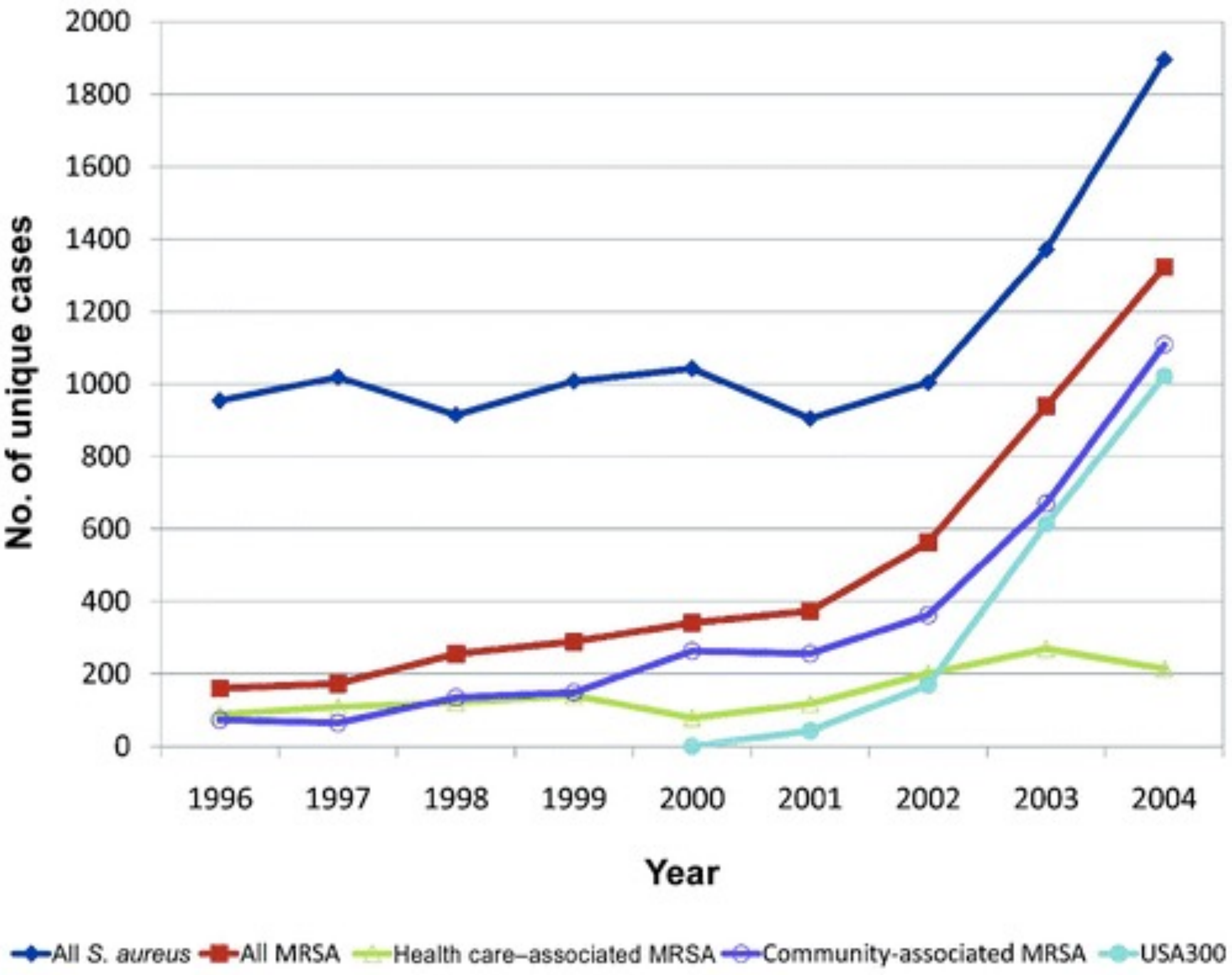
Spades: uses k=31 ->127



Compeau et al. Nat Biotechnol 2011

Bacterial sequencing application: Evolution of *S. aureus* USA300

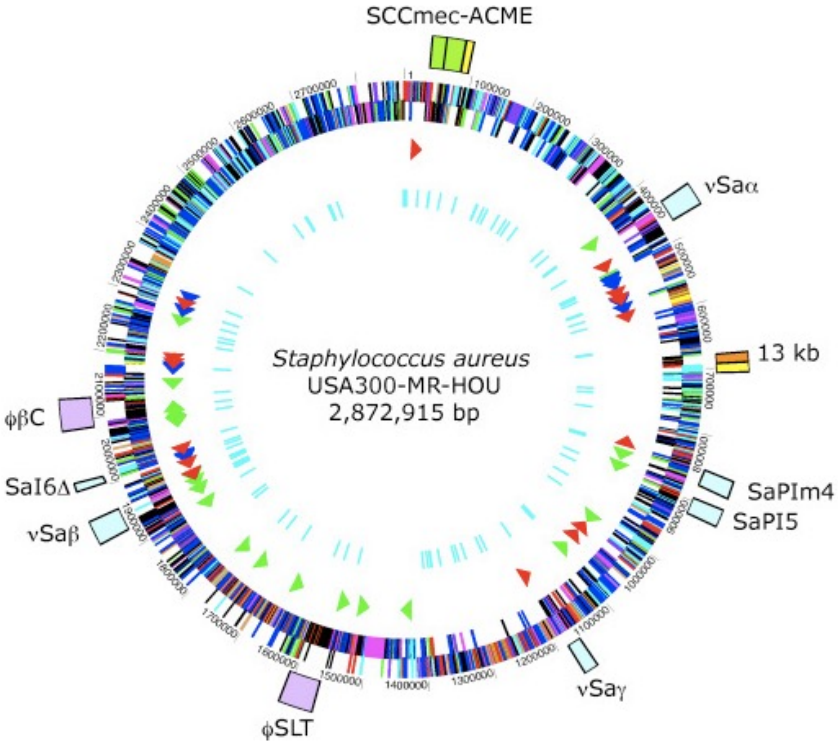
California: Initial documentation of CA-MRSA epidemic / USA300



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

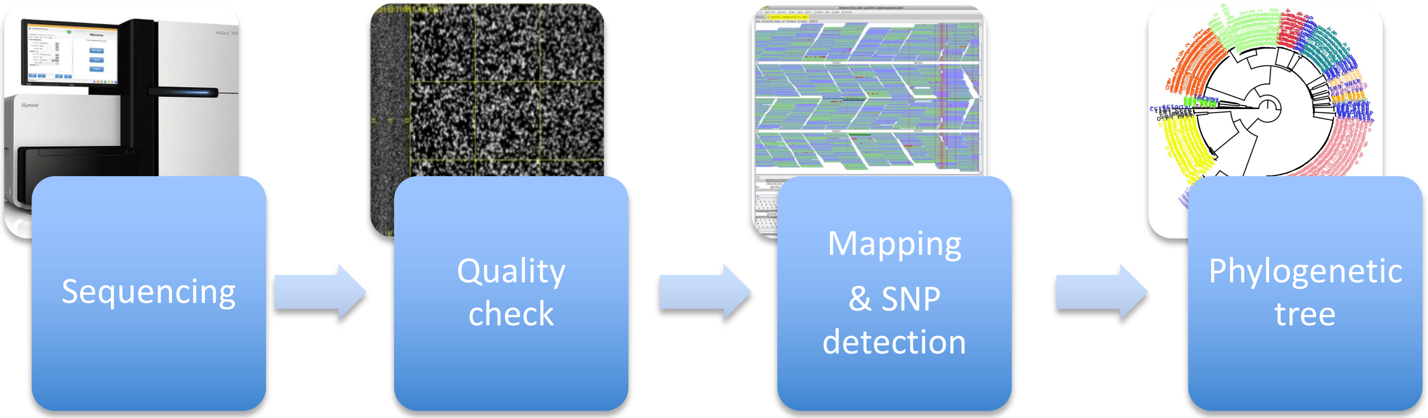
Mobile genetic elements

- Small SCC*mec* IV
- φ2 / PVL toxin
- SaPi5
- ACME I

  detoxifies host anti-microbials
- derived from USA500?

Diep et al. The Lancet 2005

Whole genome sequencing
387 isolates



Sequencing

Quality check

Mapping & SNP detection

Phylogenetic tree

Sequencing:

- Large dataset of infectious and household isolates
- Mate-paired libraries
  100 bp paired-end
- Illumina Hi-Seq
- Coverage 100 to 170 fold

Mapping:

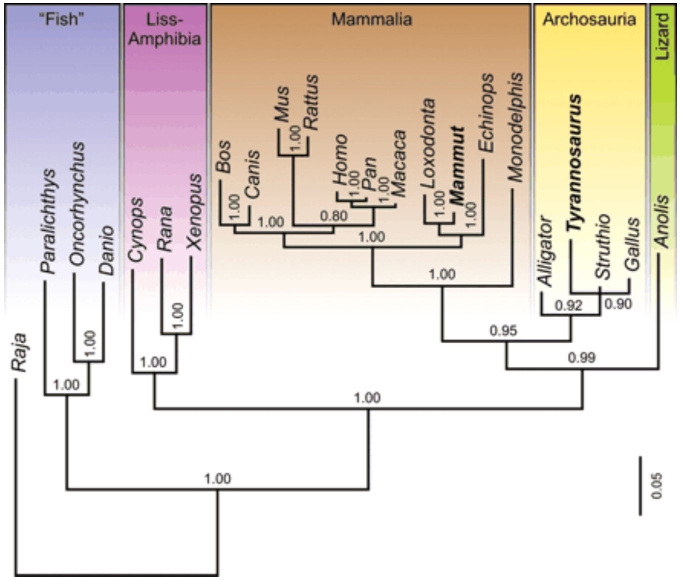
- Reference genome FPR3757
- Exclusion unmapped reads, MGEs
- Repeat Scout
- SNP calling

Phylogenetic tree

- Core genome
- Concatenated SNPs
- RAxML

Phylogenetic tree

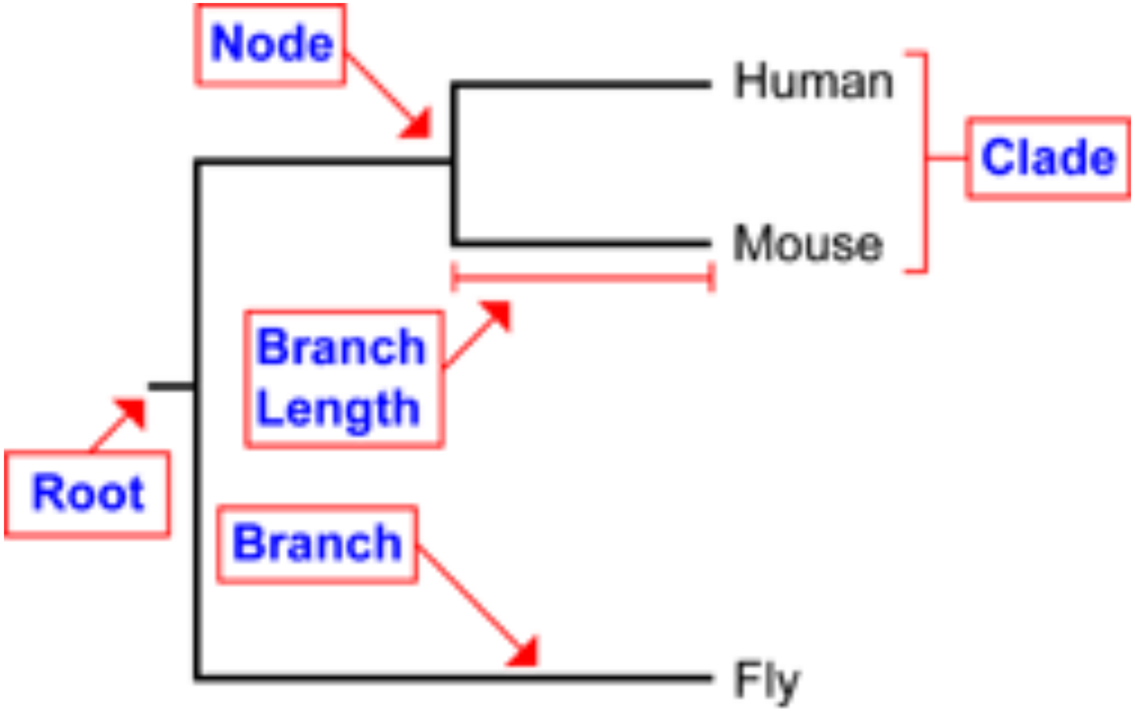
All life is related by common ancestry.



Phylogeny: pattern of historical relationships

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

Tree Notation



Genome composition of *S. aureus*
Mobile elements do not follow tree like evolution

Chromosomal Genome:

1. Stable core

   - MLST

2. Variable core

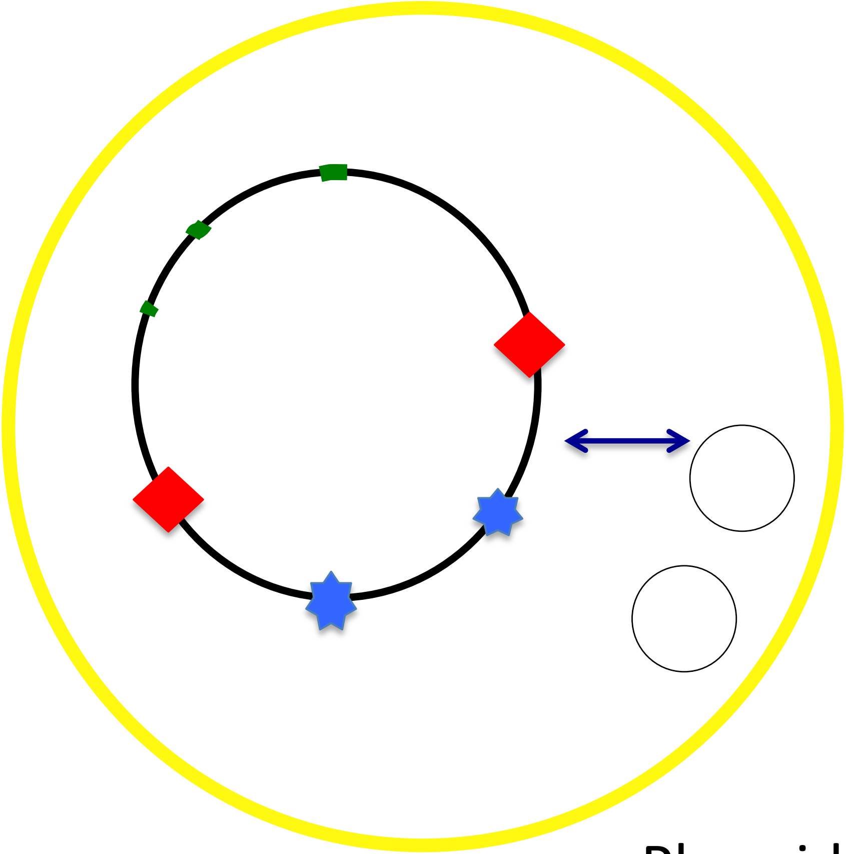
- Surface proteins

  spa-types

- Some virulence factors

3. Mobile genetic elements

- Integrated Pro-phages
- Pathogenicity islands
- Transposons and Insertion sequences



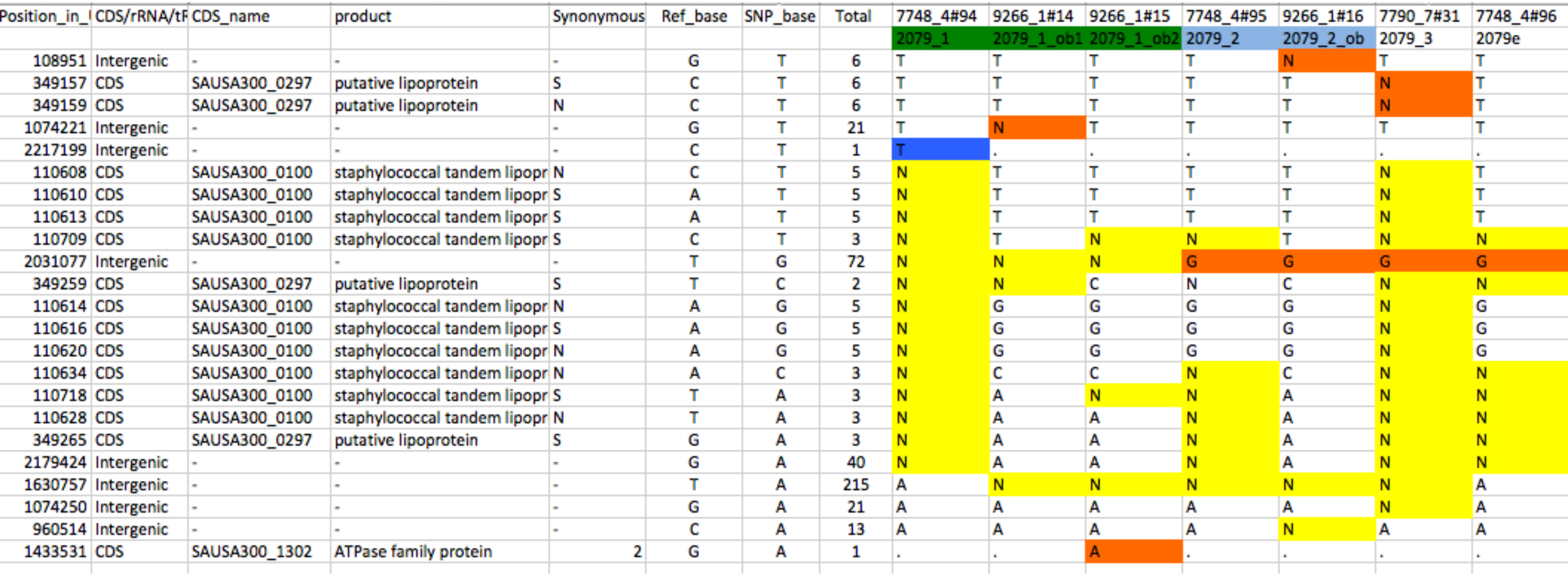
Plasmids:

- 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

Validation of SNP calls: Comparison of multiple isolates per person

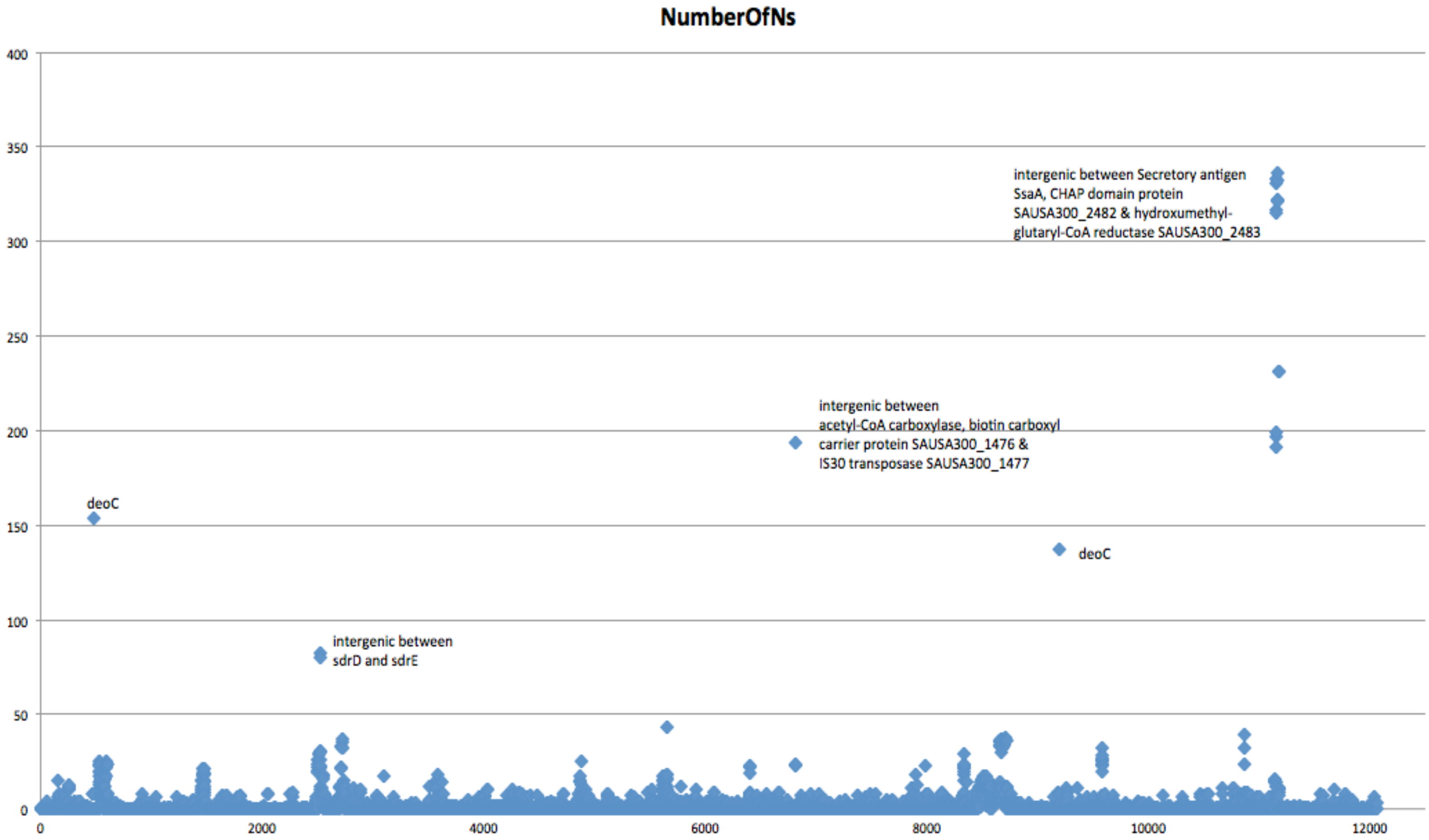


Position_in_	CDS/rRNA/tR	CDS_name	product	Synonymous	Ref_base	SNP_base	Total	7748_4#94	9266_1#14	9266_1#15	7748_4#95	9266_1#16	7790_7#31	7748_4#96
								2079_1	2079_1_ob1	2079_1_ob2	2079_2	2079_2_ob	2079_3	2079e
108951	Intergenic	-	-	-	G	T	6	T	T	T	T	N	T	T
349157	CDS	SAUSA300_0297	putative lipoprotein	S	C	T	6	T	T	T	T	T	N	T
349159	CDS	SAUSA300_0297	putative lipoprotein	N	C	T	6	T	T	T	T	T	N	T
1074221	Intergenic	-	-	-	G	T	21	T	N	T	T	T	T	T
2217199	Intergenic	-	-	-	C	T	1	T	.	.	.	.	.	.
110608	CDS	SAUSA300_0100	staphylococcal tandem lipopr	N	C	T	5	N	T	T	T	T	N	T
110610	CDS	SAUSA300_0100	staphylococcal tandem lipopr	S	A	T	5	N	T	T	T	T	N	T
110613	CDS	SAUSA300_0100	staphylococcal tandem lipopr	S	A	T	5	N	T	T	T	T	N	T
110709	CDS	SAUSA300_0100	staphylococcal tandem lipopr	S	C	T	3	N	T	N	N	T	N	N
2031077	Intergenic	-	-	-	T	G	72	N	N	N	G	G	G	G
349259	CDS	SAUSA300_0297	putative lipoprotein	S	T	C	2	N	N	C	N	C	N	N
110614	CDS	SAUSA300_0100	staphylococcal tandem lipopr	N	A	G	5	N	G	G	G	G	N	G
110616	CDS	SAUSA300_0100	staphylococcal tandem lipopr	S	A	G	5	N	G	G	G	G	N	G
110620	CDS	SAUSA300_0100	staphylococcal tandem lipopr	N	A	G	5	N	G	G	G	G	N	G
110634	CDS	SAUSA300_0100	staphylococcal tandem lipopr	N	A	C	3	N	C	C	N	C	N	N
110718	CDS	SAUSA300_0100	staphylococcal tandem lipopr	S	T	A	3	N	A	N	N	A	N	N
110628	CDS	SAUSA300_0100	staphylococcal tandem lipopr	N	T	A	3	N	A	A	N	A	N	N
349265	CDS	SAUSA300_0297	putative lipoprotein	S	G	A	3	N	A	A	N	A	N	N
2179424	Intergenic	-	-	-	G	A	40	N	A	A	N	A	N	N
1630757	Intergenic	-	-	-	T	A	215	A	N	N	N	N	N	A
1074250	Intergenic	-	-	-	G	A	21	A	A	A	A	A	N	A
960514	Intergenic	-	-	-	C	A	13	A	A	A	A	N	A	A
1433531	CDS	SAUSA300_1302	ATPase family protein	2	G	A	1	.	.	A	.	.	.	.

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”

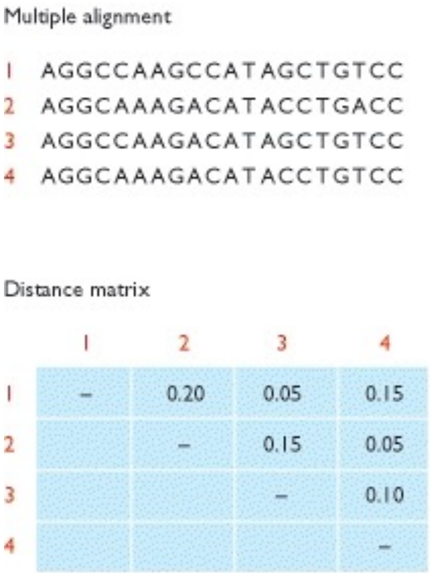
  transposase type (published in other genomes)

How are N’s addressed in the literature?

- Usually 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…”

SNP matrices, distances and trees



Multiple alignment

```
1  AGGCCAAGCCATAGCTGTCC
2  AGGCAAAGACATACCTGACC
3  AGGCCAAGACATAGCTGTCC
4  AGGCAAAGACATACCTGTCC
```

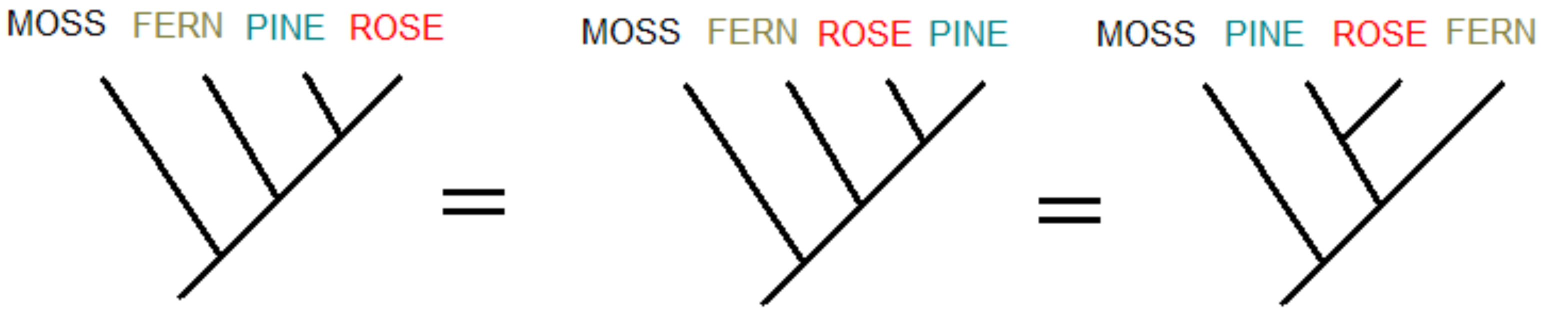
Distance matrix

	1	2	3	4
1	–	0.20	0.05	0.15
2		–	0.15	0.05
3			–	0.10
4				–

- 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.

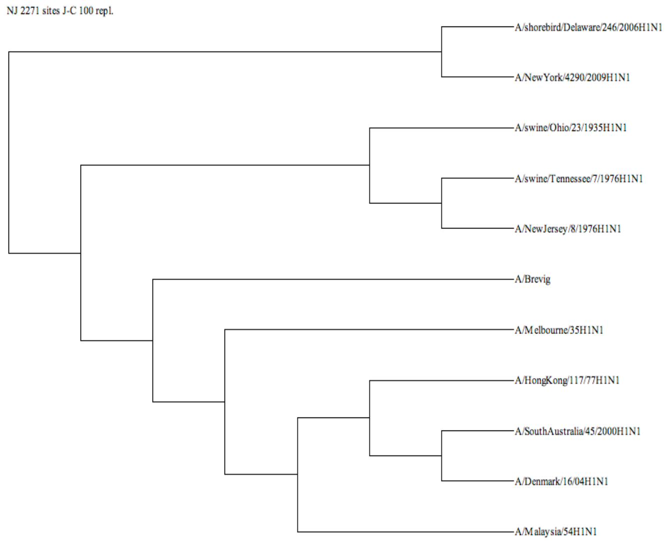


Different trees

- Alignment of homologous sequences

  -> concatenated SNPs

- 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

1. Distance methods
   - Minimal Evolution
   - Least Squares
   - UPGMA
   - Neighbor-Joining
2. Parsimony
3. Likelihood. PHYLIP (Felsenstein)
4. Bayesian methods

Maximum likelihood estimation

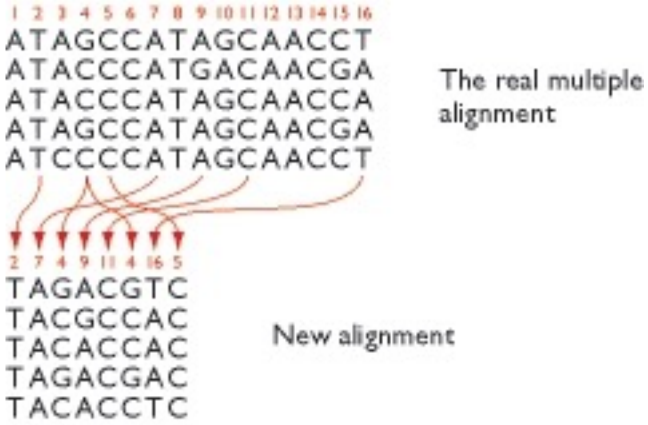
- Principle: Choose the tree which makes the data most probable
- Each position evolves independently
- Accommodates time structure of temporally-spaced sequences
- Tips have isolation date; internal nodes are unknown -> arbitrary starting times (order on tree)
- Substitution rate used to scale times 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

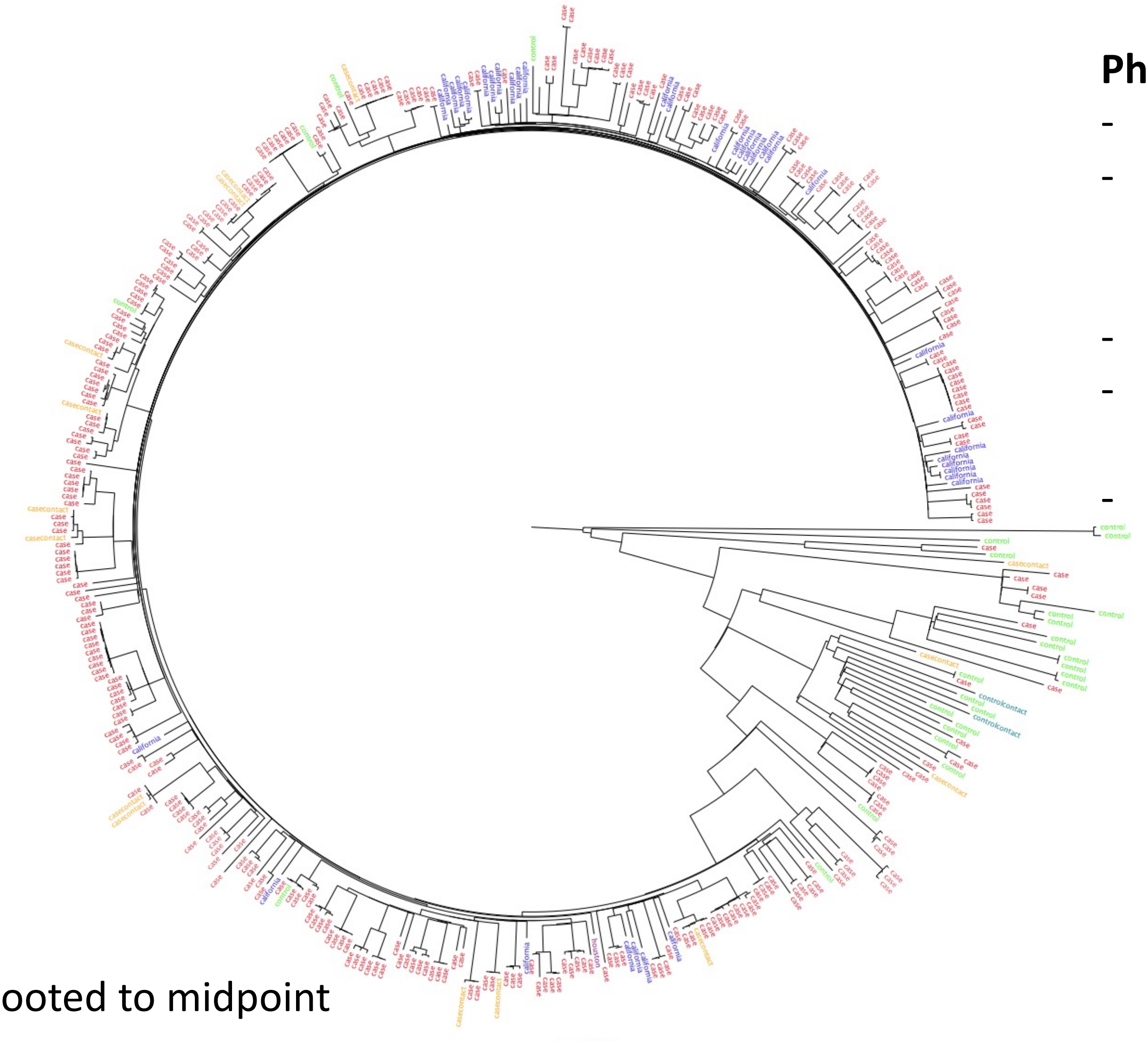
- How much do we trust a tree that we have constructed?
- A simple method for parsimony, distance or ML is bootstrapping.
  - Select some random positions, with repetition.
  - Construct another tree with the bootstrapped data.
  - Repeat many times.
  - Check the consistency of the results.
- In Bayesian methods, one can estimate the confidence by looking at posterior probabilities.



```
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
ATAGCCATAGCAACCT
ATACCCATGACAACGA
ATACCCATAGCAACCA
ATAGCCATAGCAACGA
ATCCCCATAGCAACCT

2 7 4 9 11 4 16 5
TAGACGTC
TACGCCAC
TACACCAC
TAGACGAC
TACACCTC
```

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

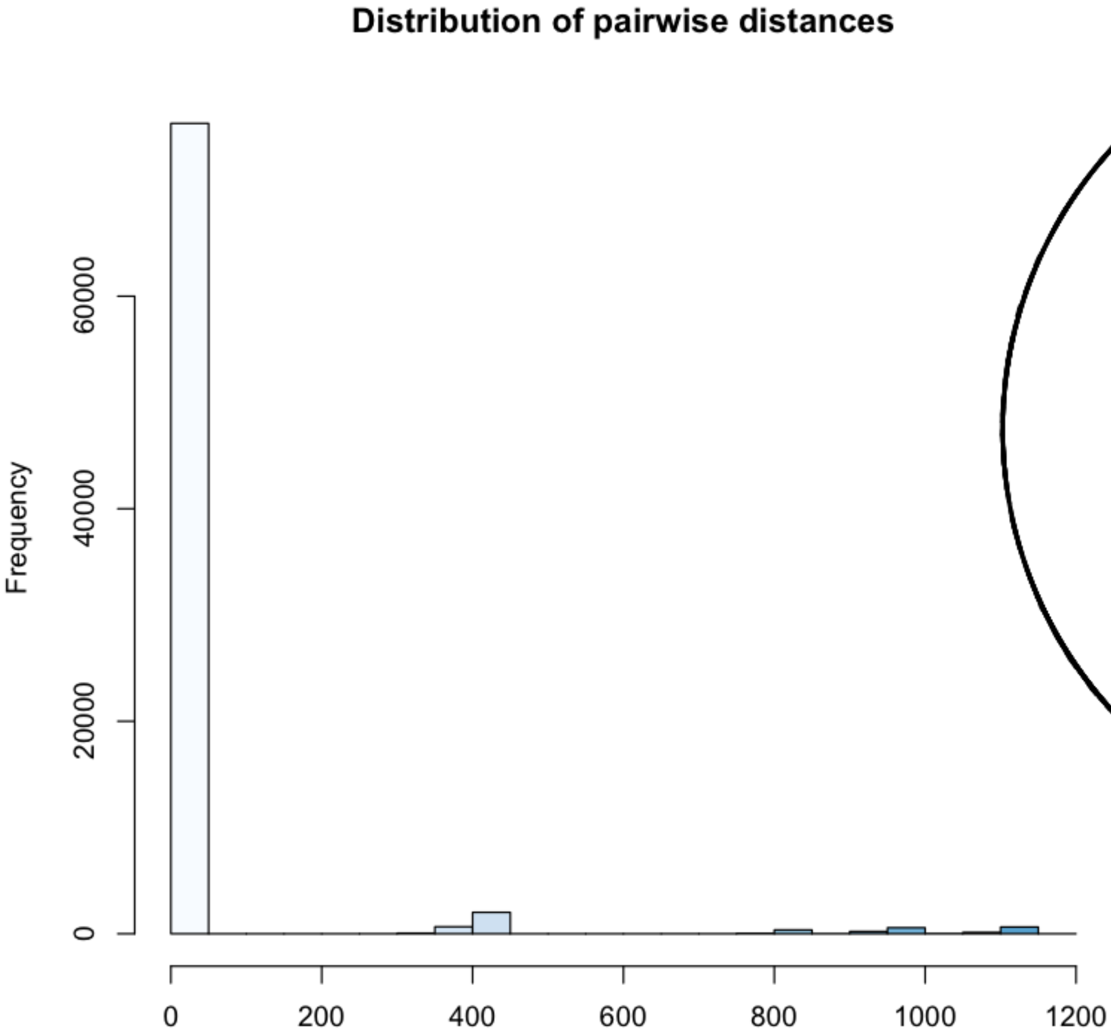
Case contact

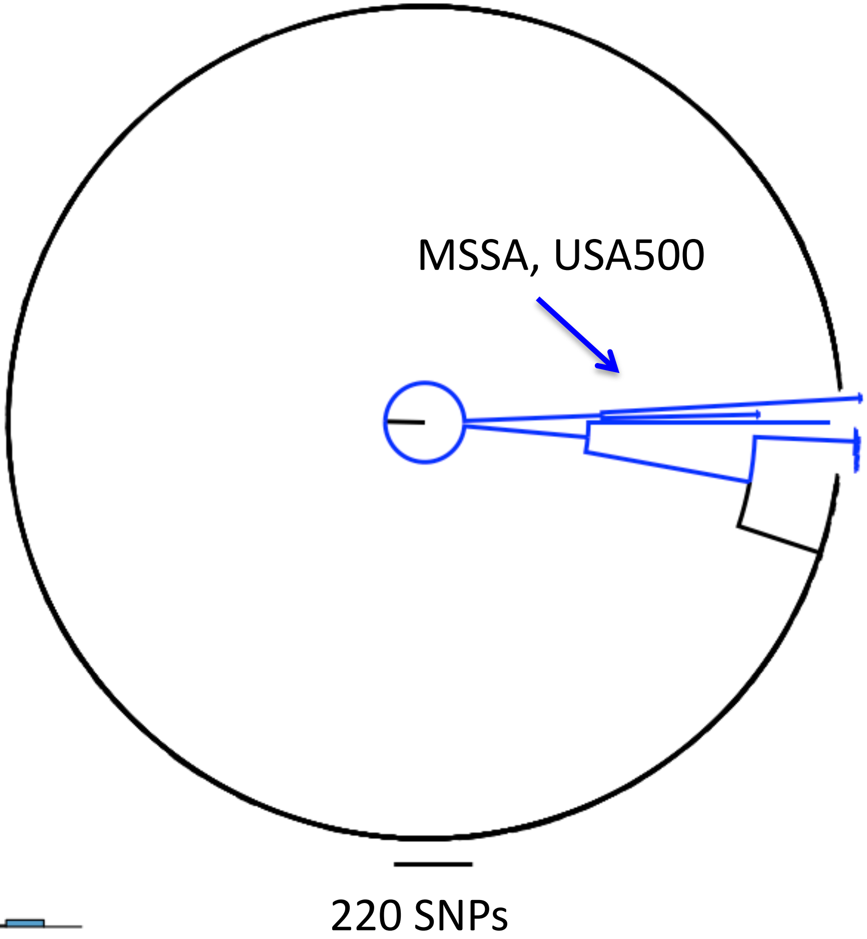
Control

California isolates

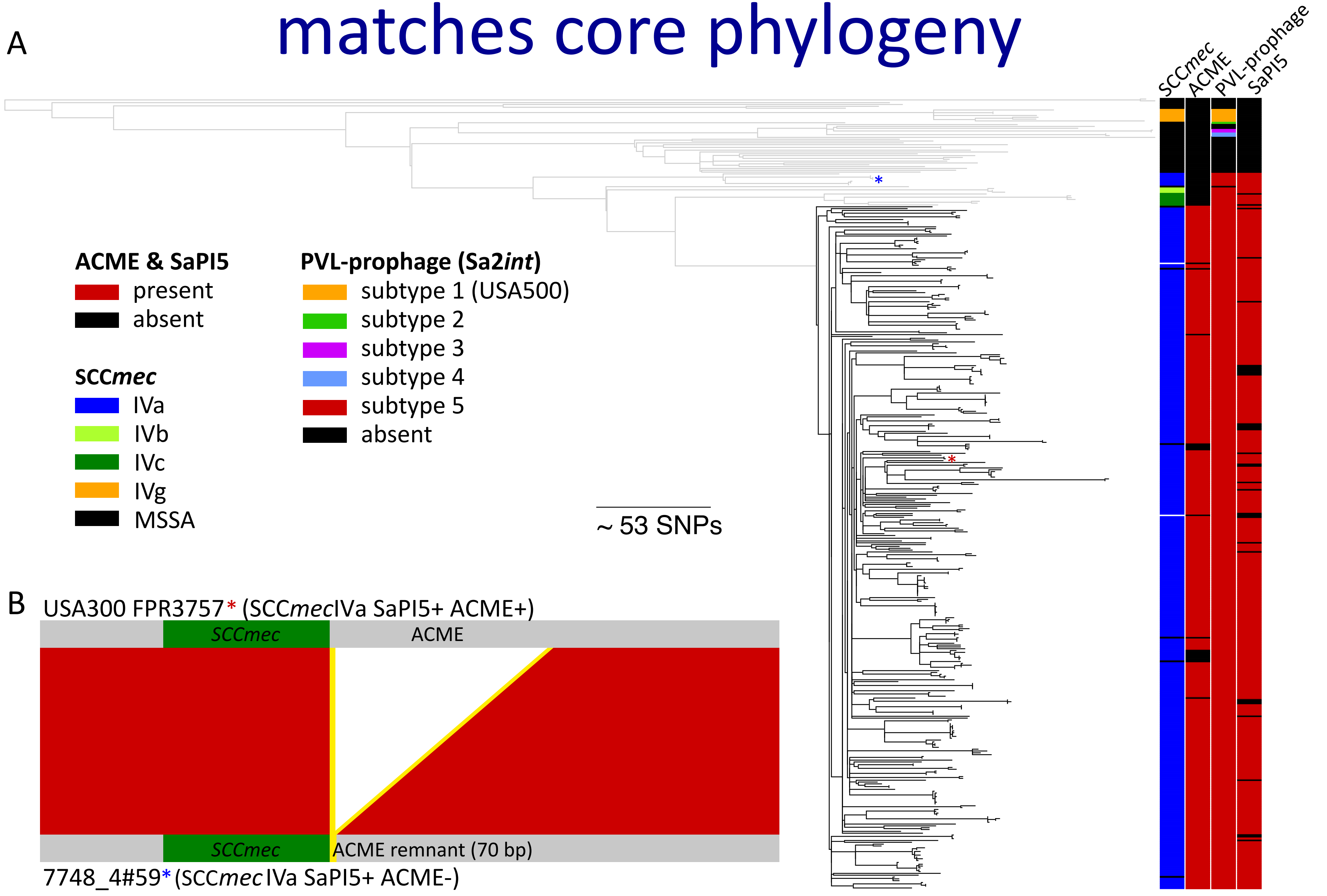
Rooted to midpoint

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





Mapping of mobile genetic elements matches core phylogeny



ACME & SaPI5

- 🟥 present
- ⬛ absent

SCC*mec*

- 🟦 IVa
- 🟩 IVb
- 🟩 IVc
- 🟧 IVg
- ⬛ MSSA

PVL-prophage (Sa2*int*)

- 🟧 subtype 1 (USA500)
- 🟩 subtype 2
- 🟪 subtype 3
- 🟦 subtype 4
- 🟥 subtype 5
- ⬛ absent

USA300 FPR3757* (SCC*mec*IVa SaPI5+ ACME+)

7748_4#59* (SCC*mec* IVa 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

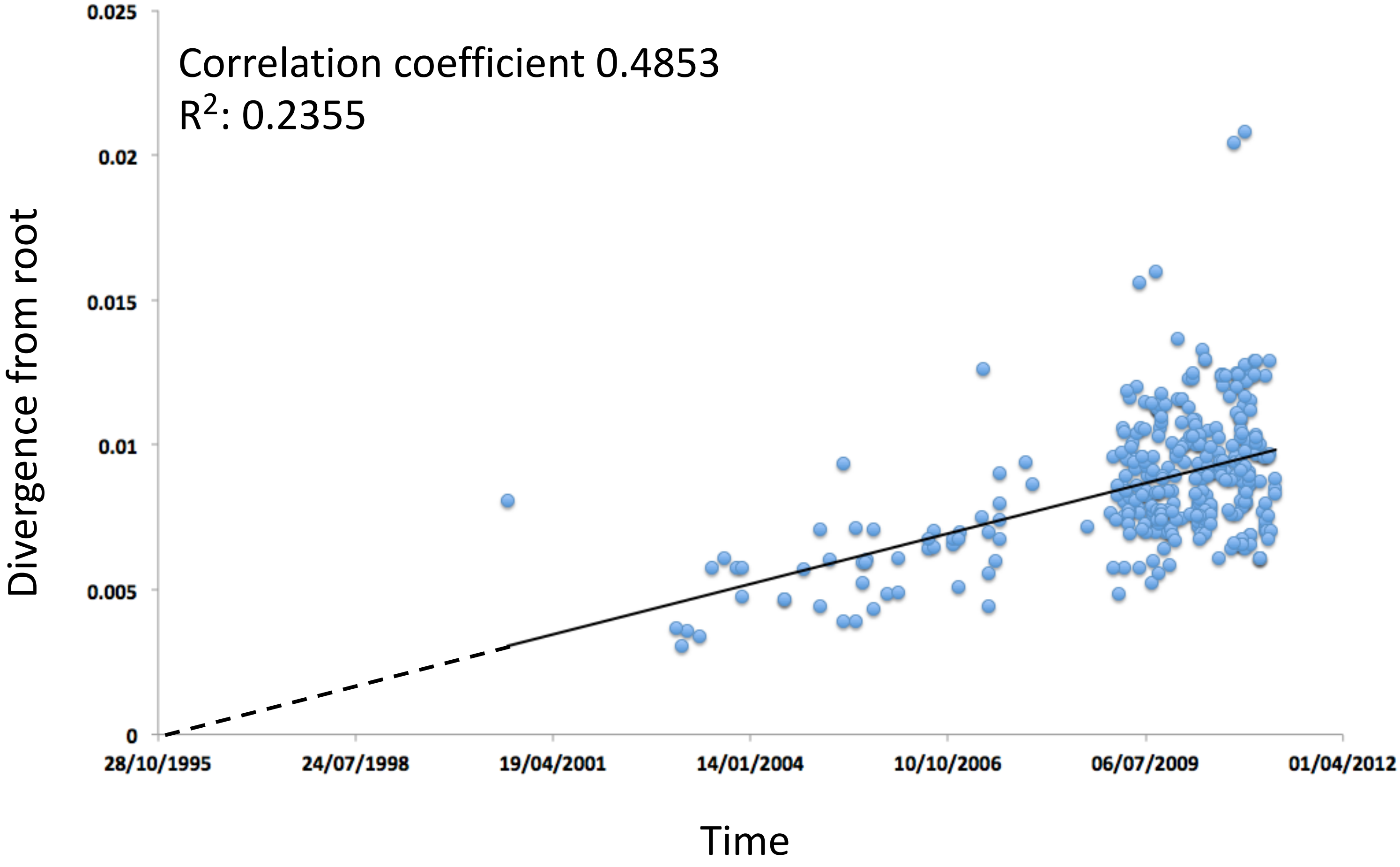
  - matrix used for neighbor-joining tree

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

  $$\mathrm{E}[d_{root,i}] = \mathrm{m}(t_i - t_{root}) = \mathrm{m}t_i - \mathrm{m}t_{root}$$

- 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

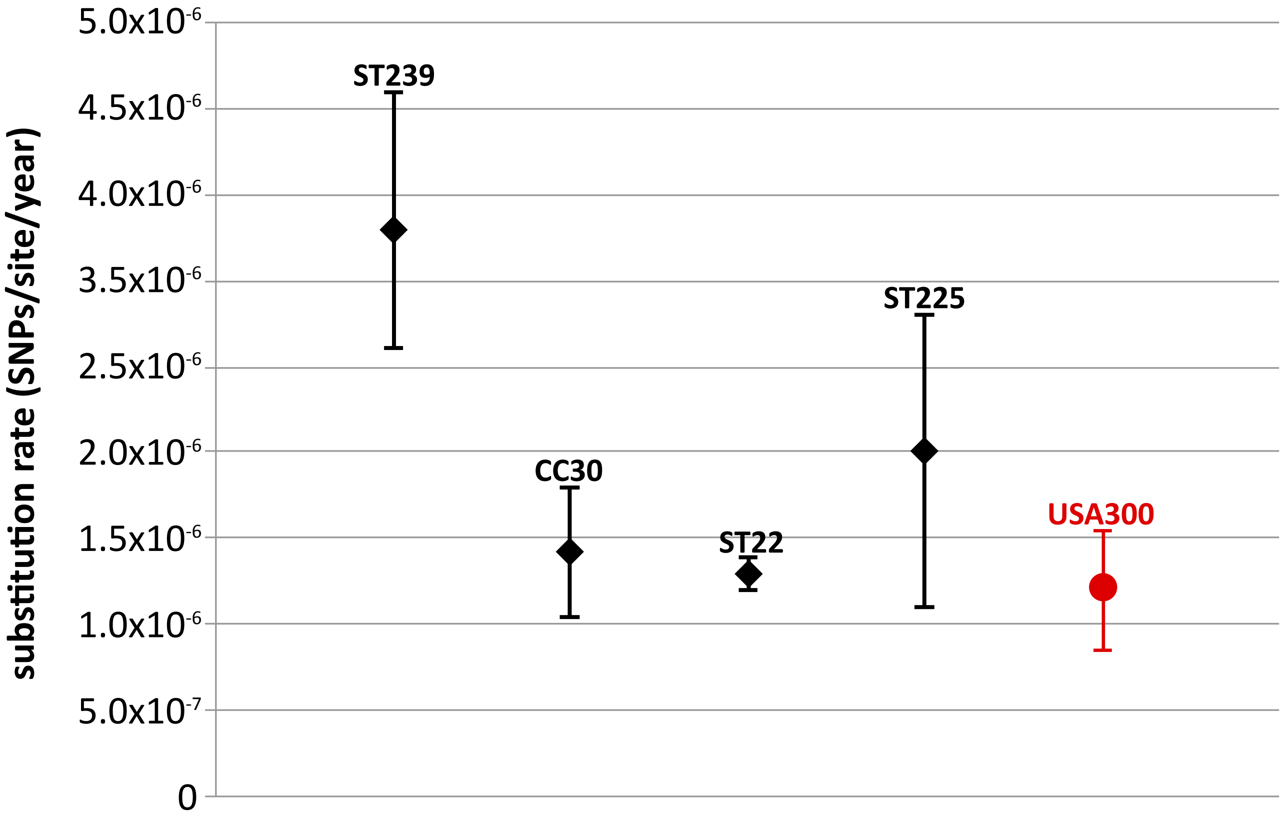


Substitution rate/site/year: 1.56×10^{-6}

Time most common recent ancestor: ~1995

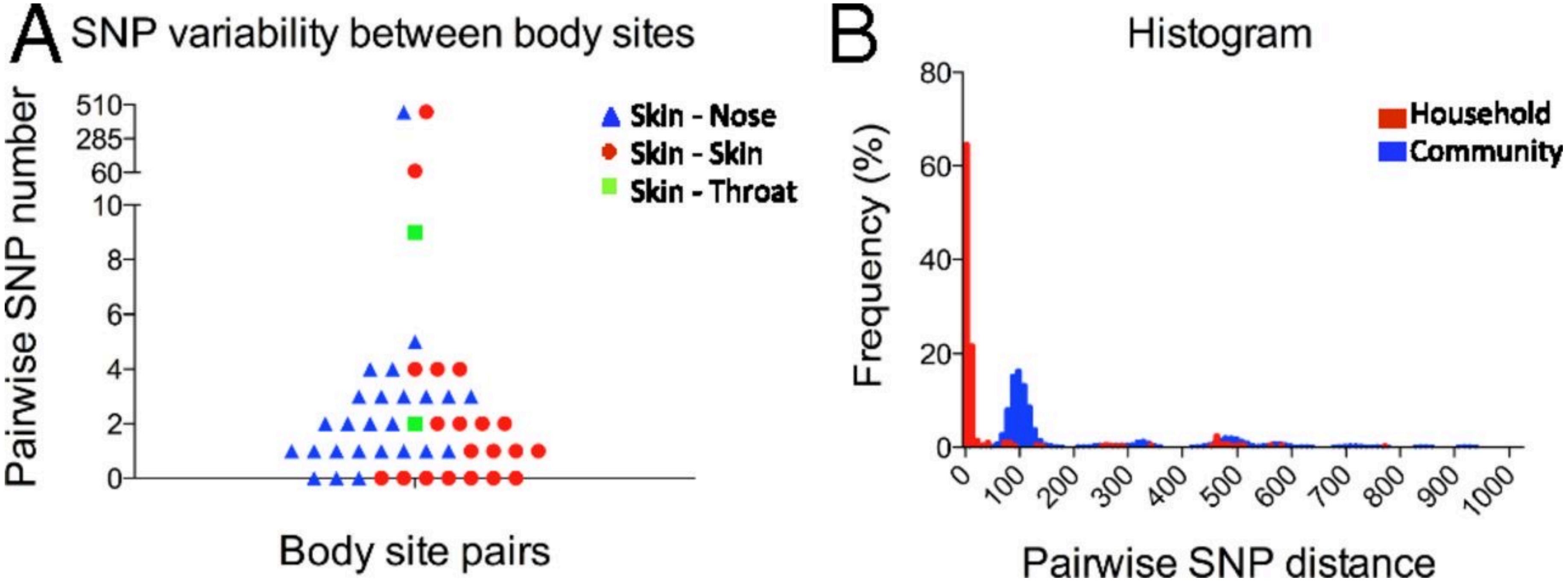
Uhlemann et al. PNAS 2014

USA300 substitution rate comparable to other MRSA clones

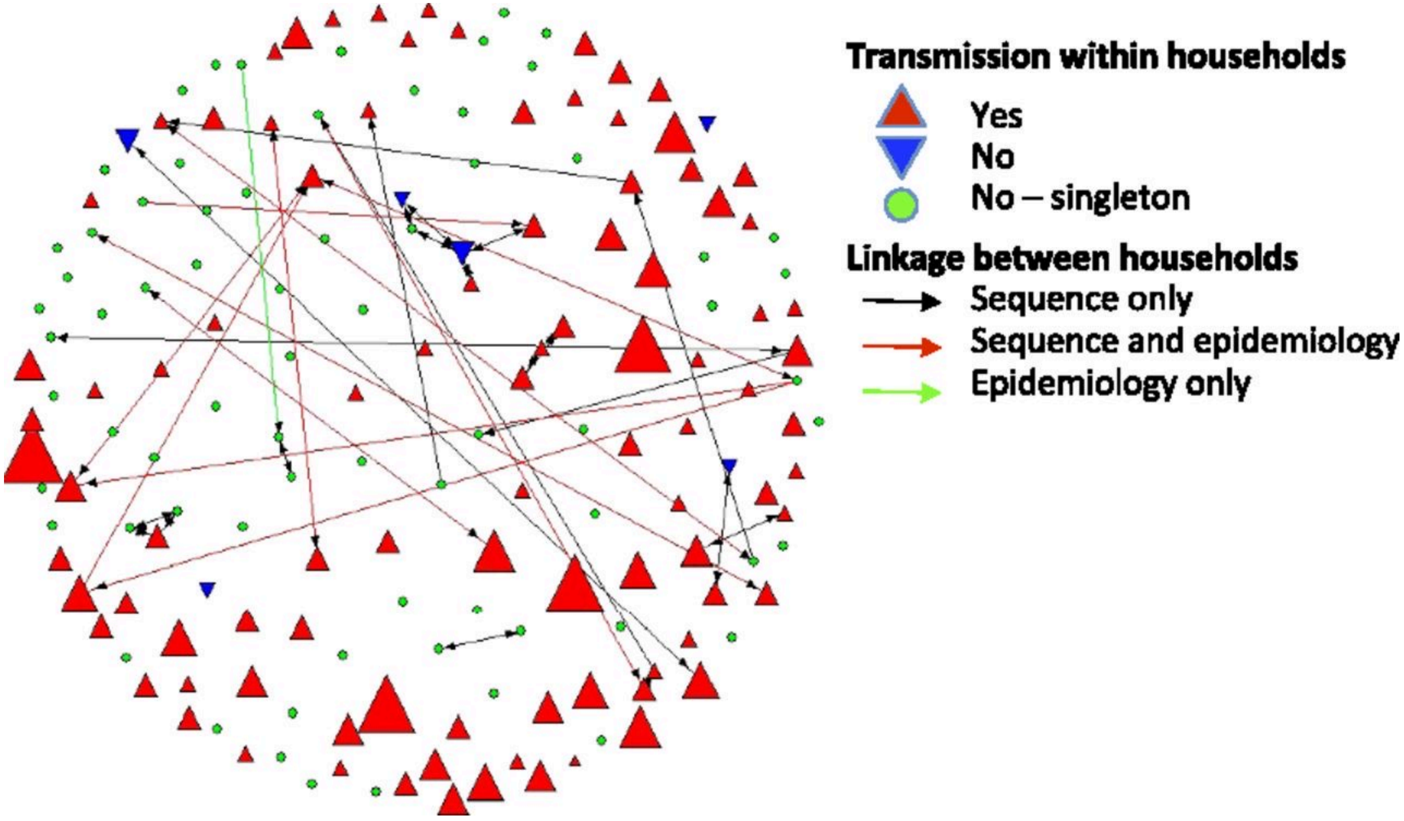


Corresponds to ~ 4 SNPs per year

SNPs for ruling in / out transmission



Applying SNPs to estimate transmission between community households


Transmission within households
Yes
No
No – singleton
Linkage between households
Sequence only
Sequence and epidemiology
Epidemiology only

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:

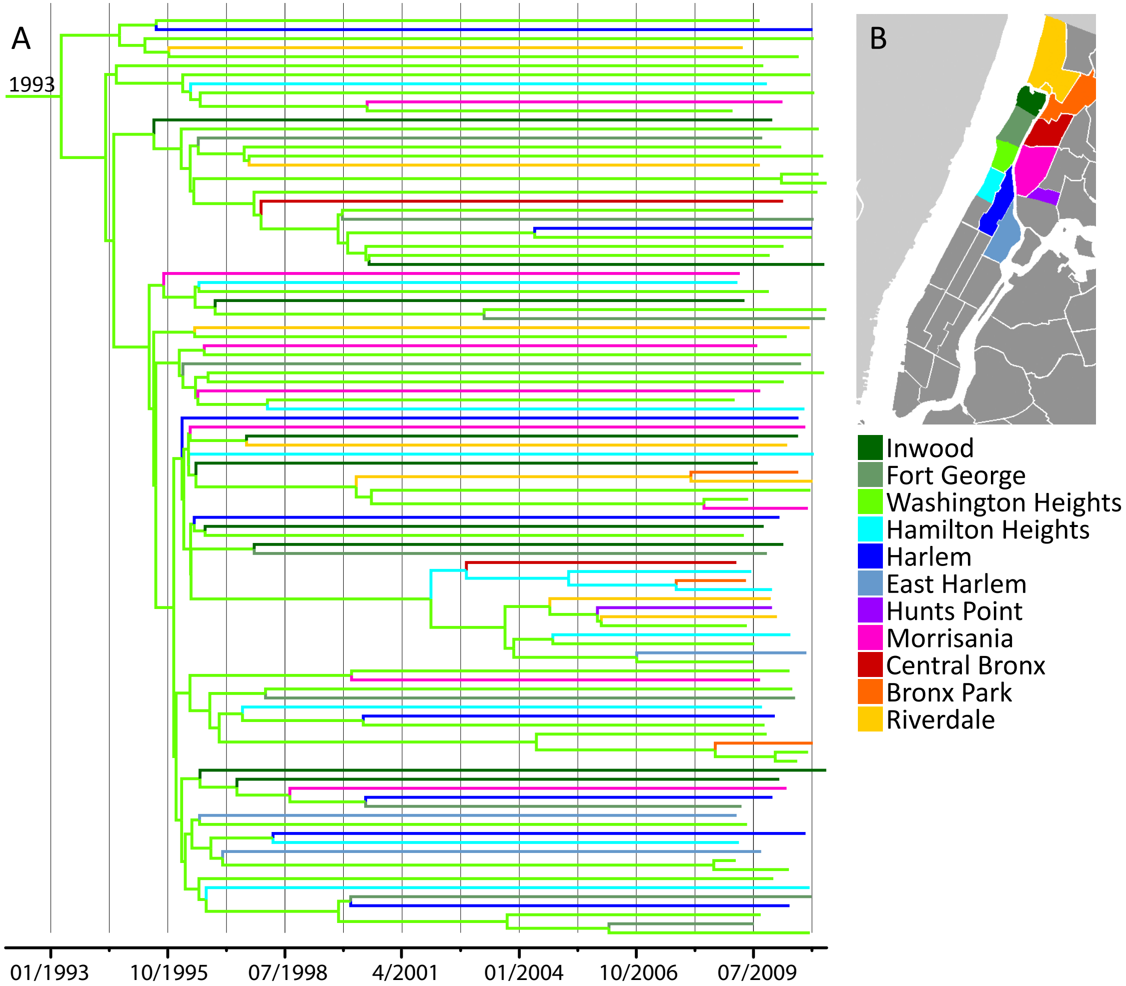
  - tree topology

  - times of ancestral nodes

  - substitution are (?)

  - substitution parameters (transition/transversion)

Phylogeographic reconstruction

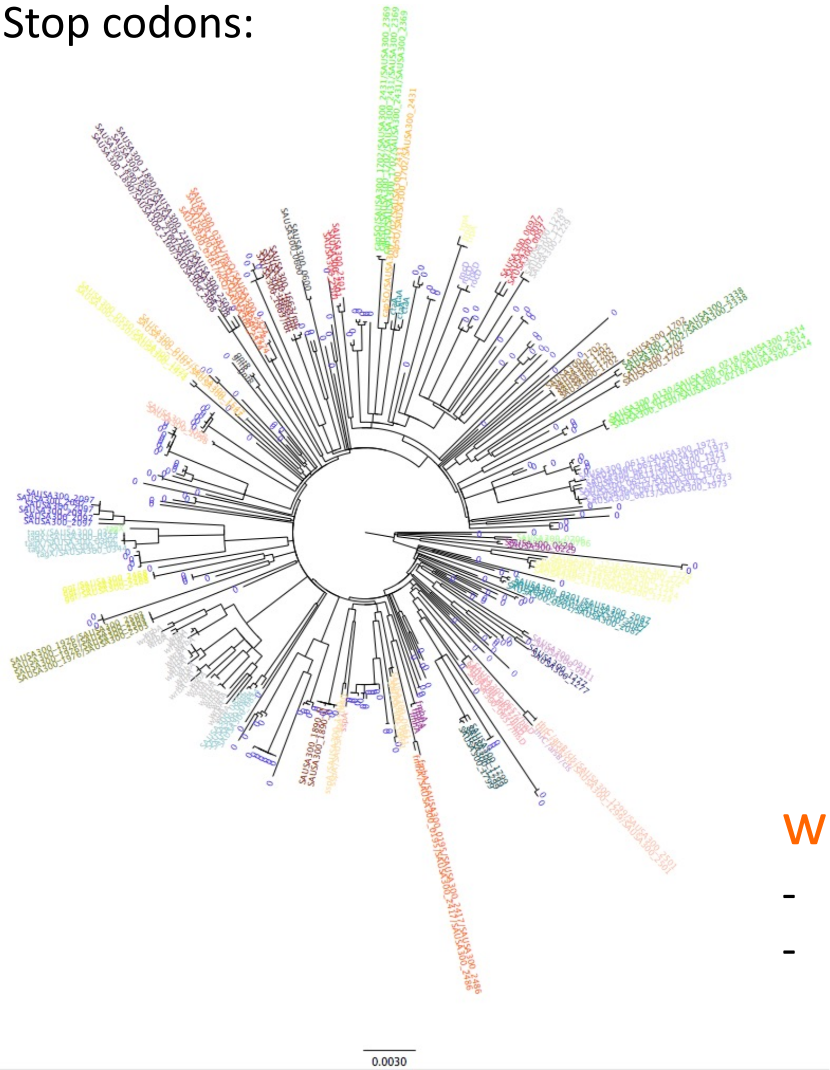


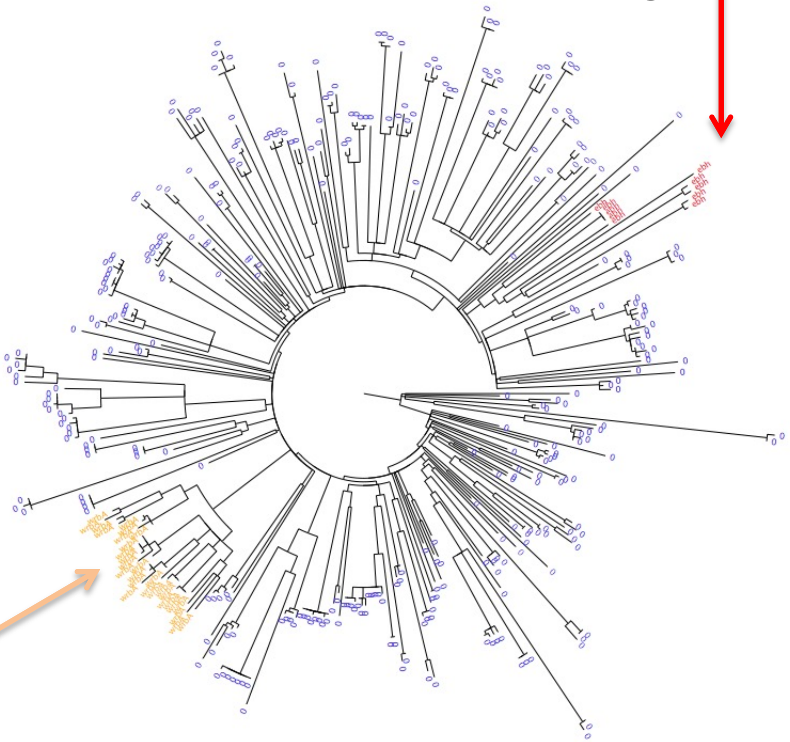
Support for root in Fort Washington neighborhood (site of CUMC)

Uhlemann et al. PNAS 2014

Have unique genomic USA300 subpopulations emerged?

Stop codons:





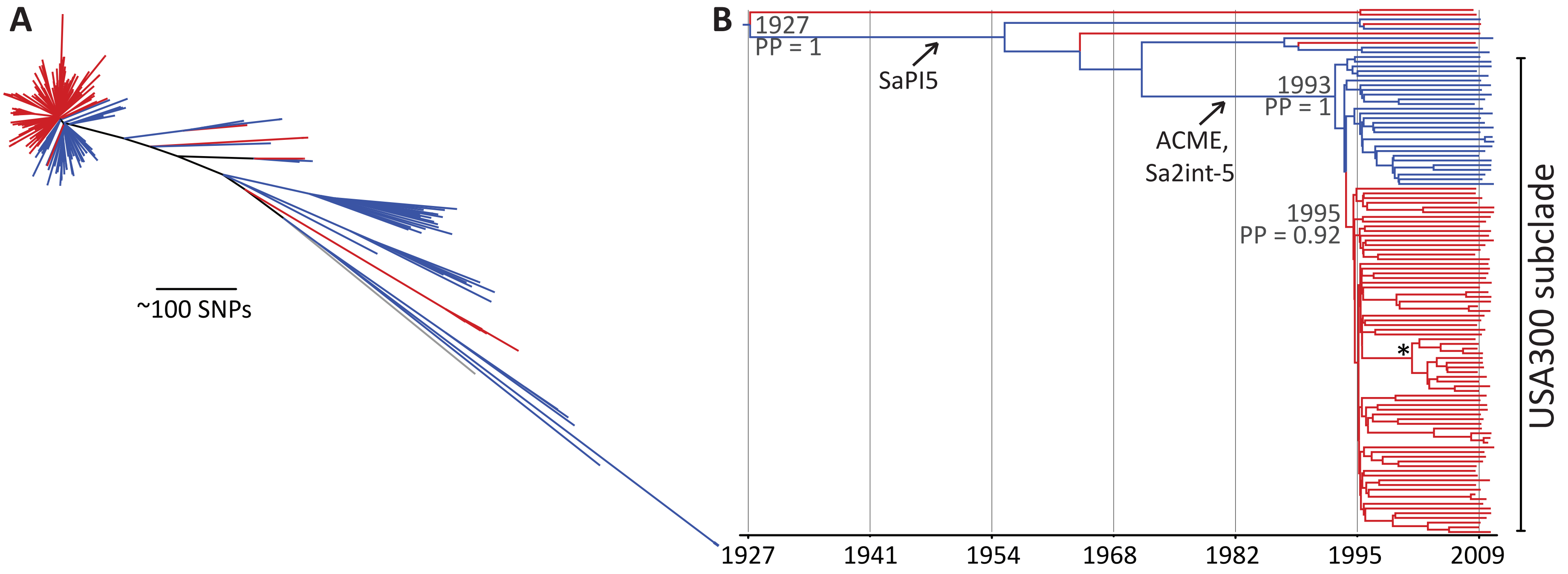
ebh

- 10 isolates, 6 households
- ECM binding

wbrA

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

Expansion of Fluoroquinolone-resistant clone (*gyrA / grlA* SNPs)



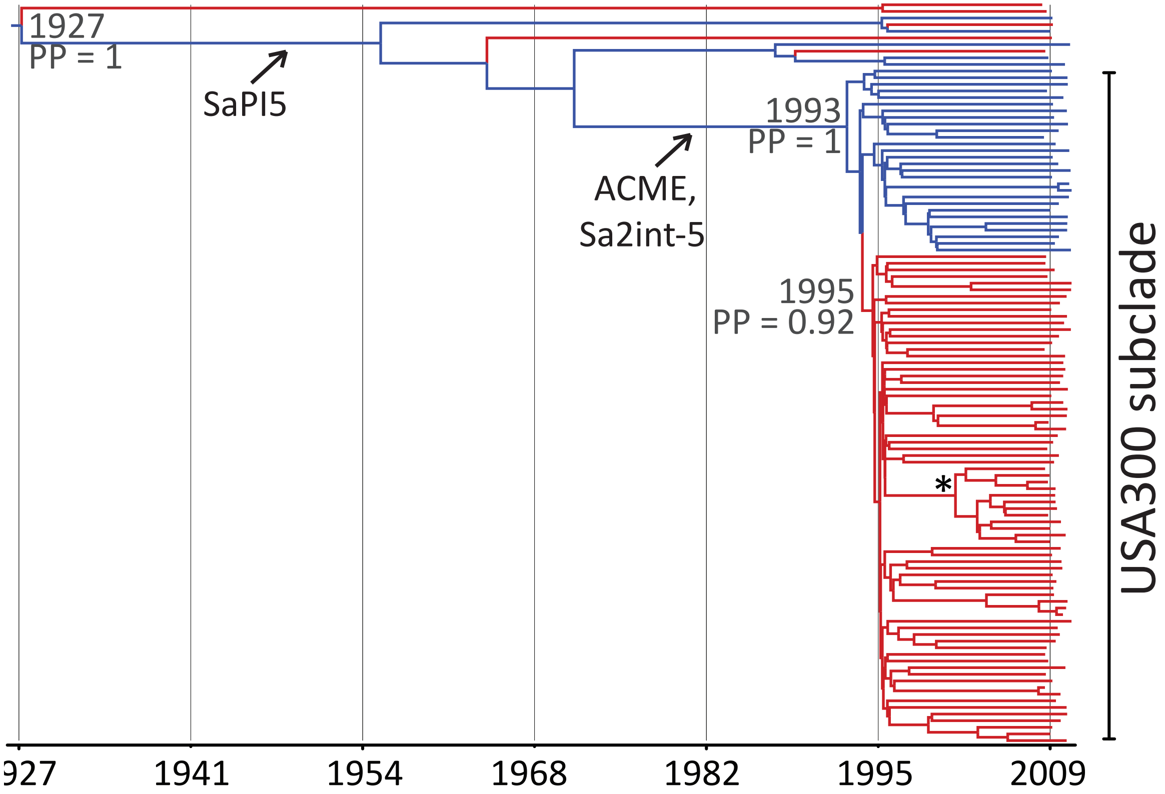
CDC survey

- 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



ST8

SaPi5

ACME & PVL

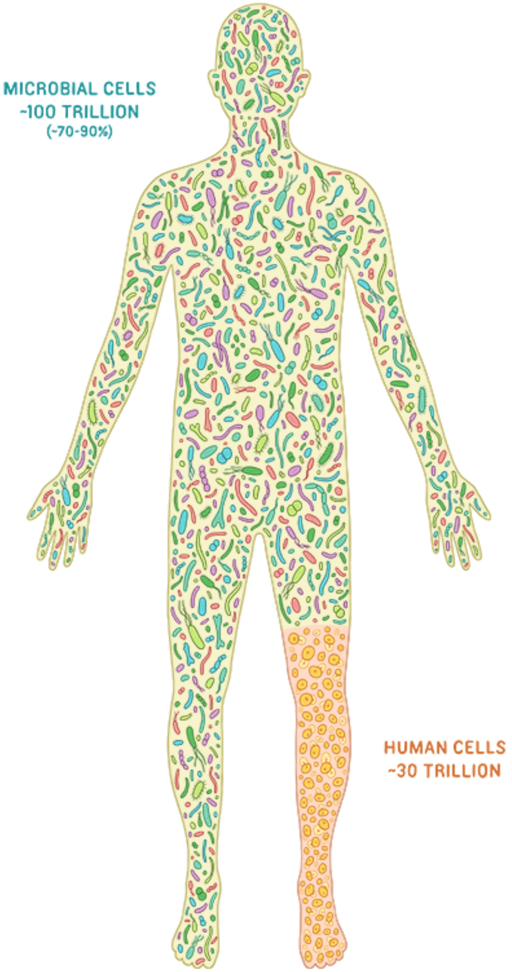
53 non-synon SNPs

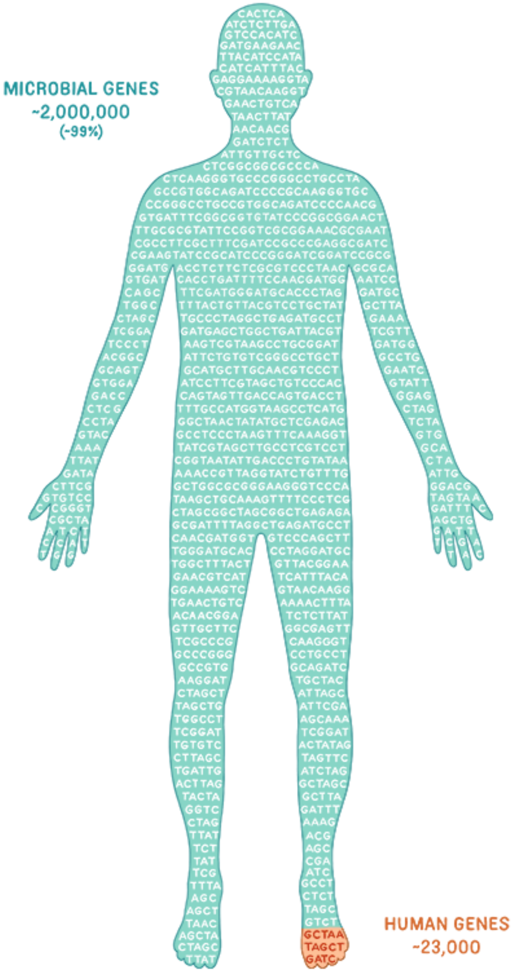
Summary – part 1

- Comparative bacterial short read sequencing informs
  - outbreak info
  - geographic spread / number of introductions
  - evolutionary history
    - acquisition of MGEs
    - acquisition of drug resistance
- Enabled by different analytical approaches from same dataset
- Beware of limitations in sequence included in analyses (~90% of genome for most bacterial species)

Microbiome analyses

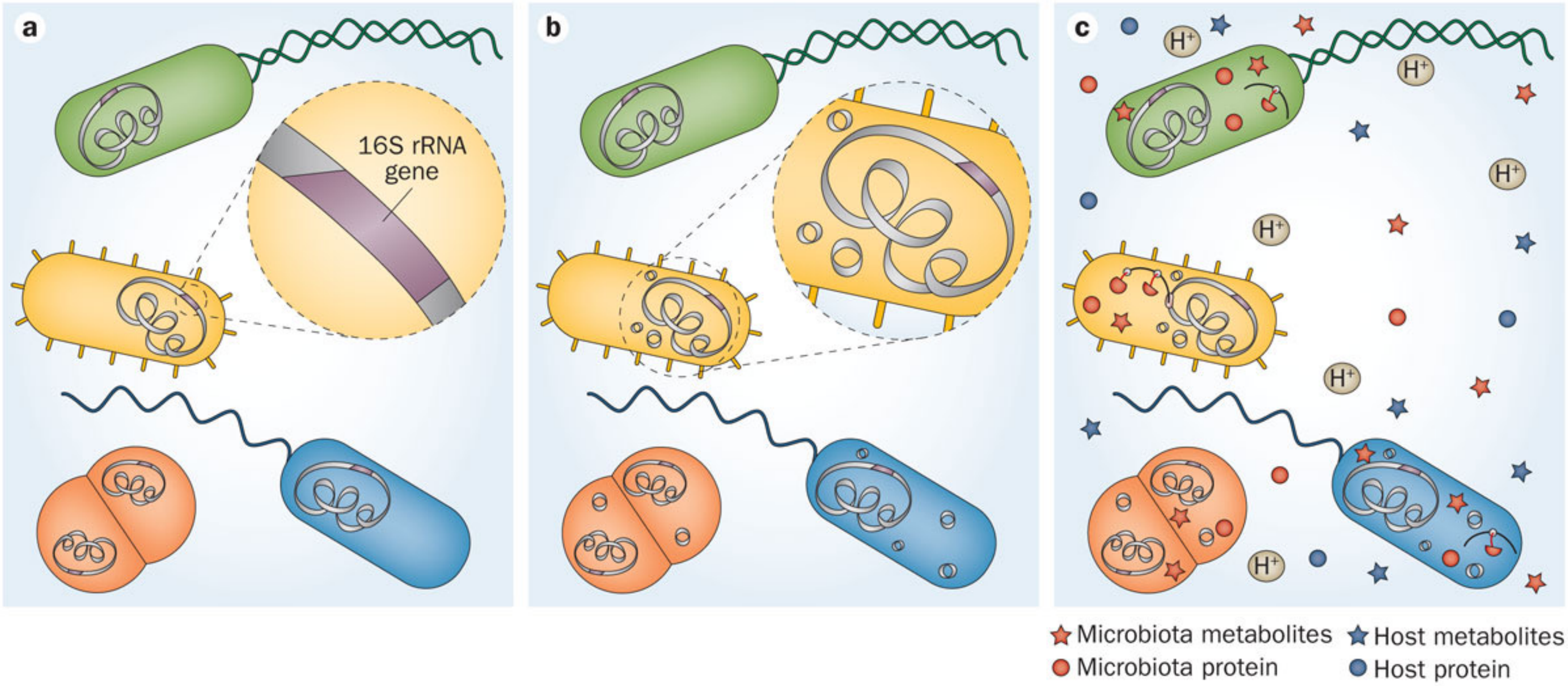
~2,000,000 bacterial genes





Many bacterial species previously not recognized because unculturable with current methods.

Microbiota? Or Microbiome?



Microbiota
16S rRNA
Taxonomic identification

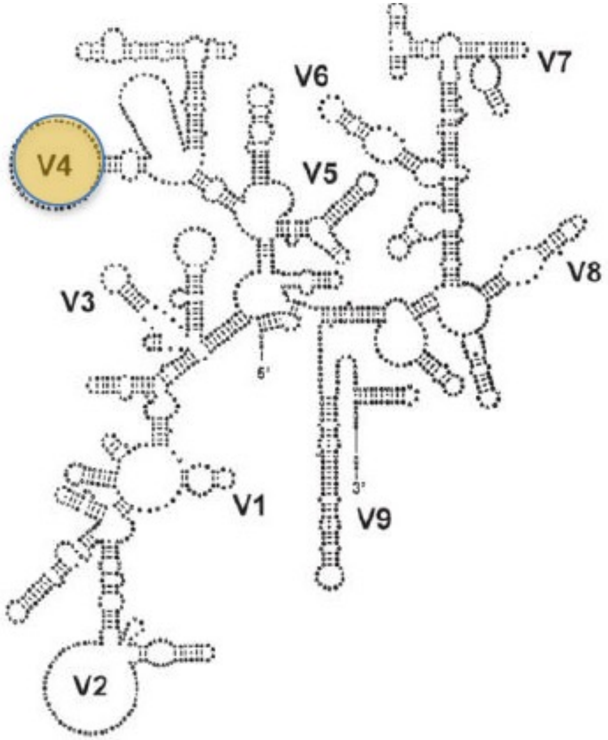
Metagenome
Genes and genomes
of microbiota

Microbiome
Genes, genomes,
products, host
proteins

Whiteside, S. A. *et al.* (2015) The microbiome of the urinary tract—a role beyond infection
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



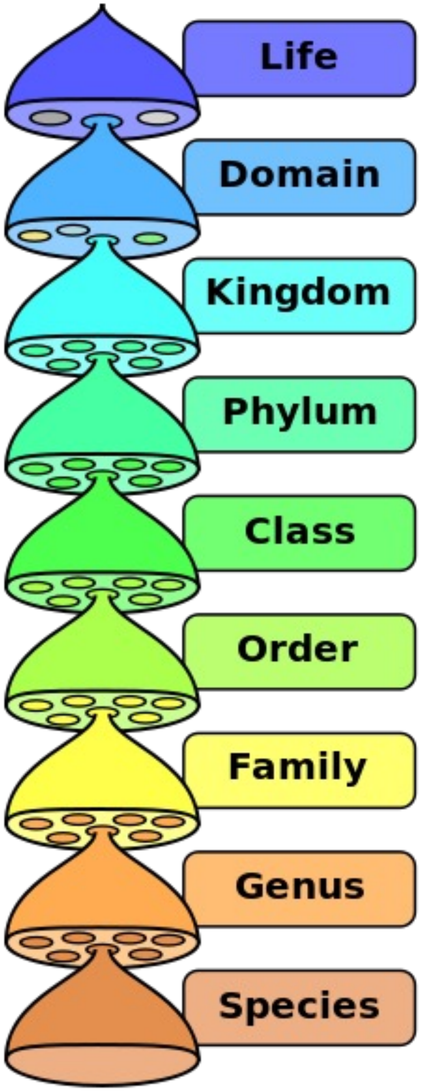
Taxonomy assignment

- Challenges:

  multiple matches

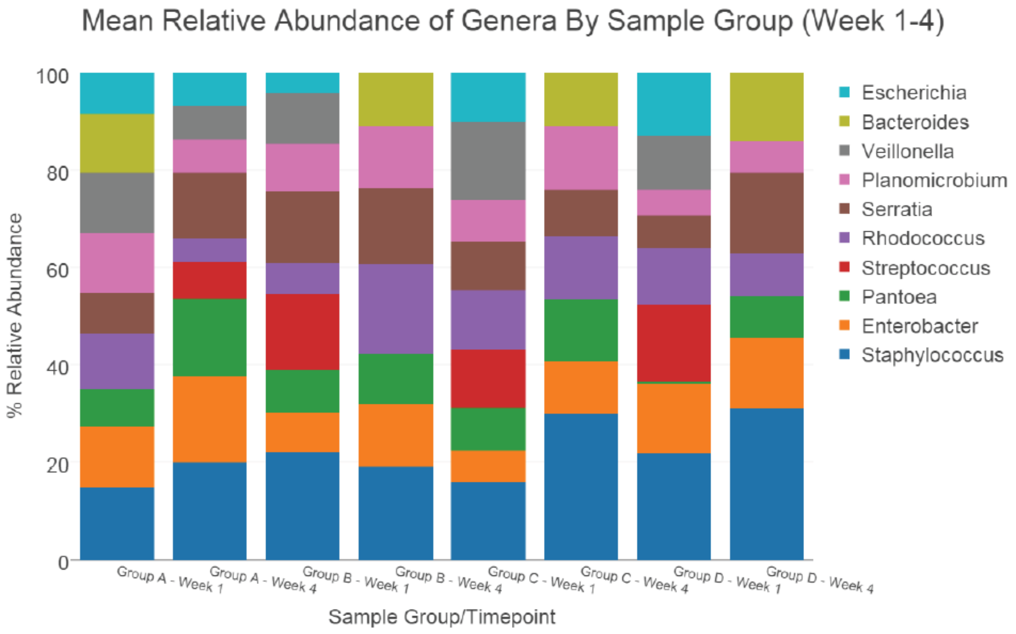
  no match (new OUT)
- Some species may share

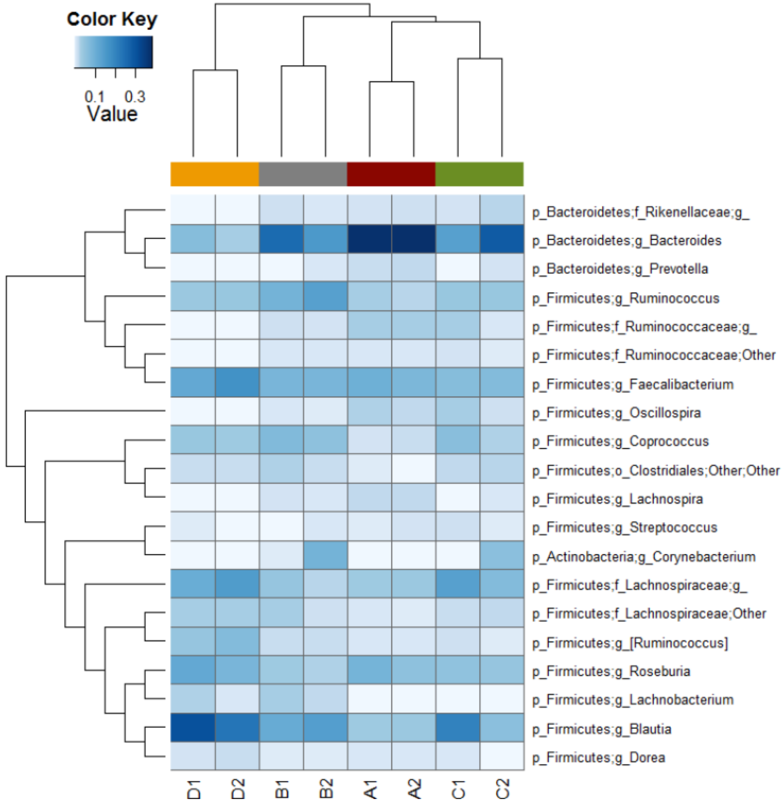
  >97% similarity, no resolution at species level



Output taxa distribution

- Bar chart
- Heat map





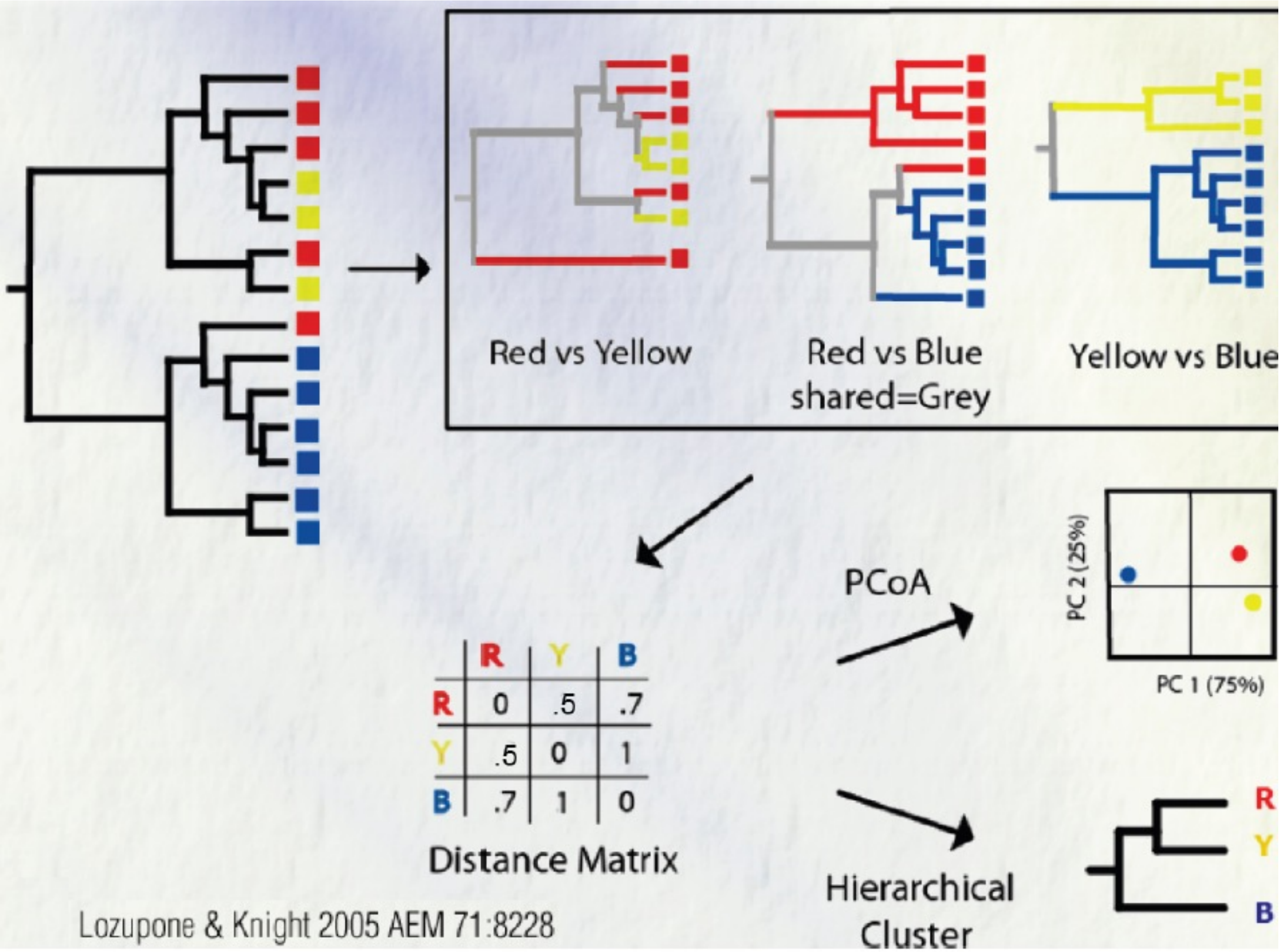
Alpha diversity

- 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

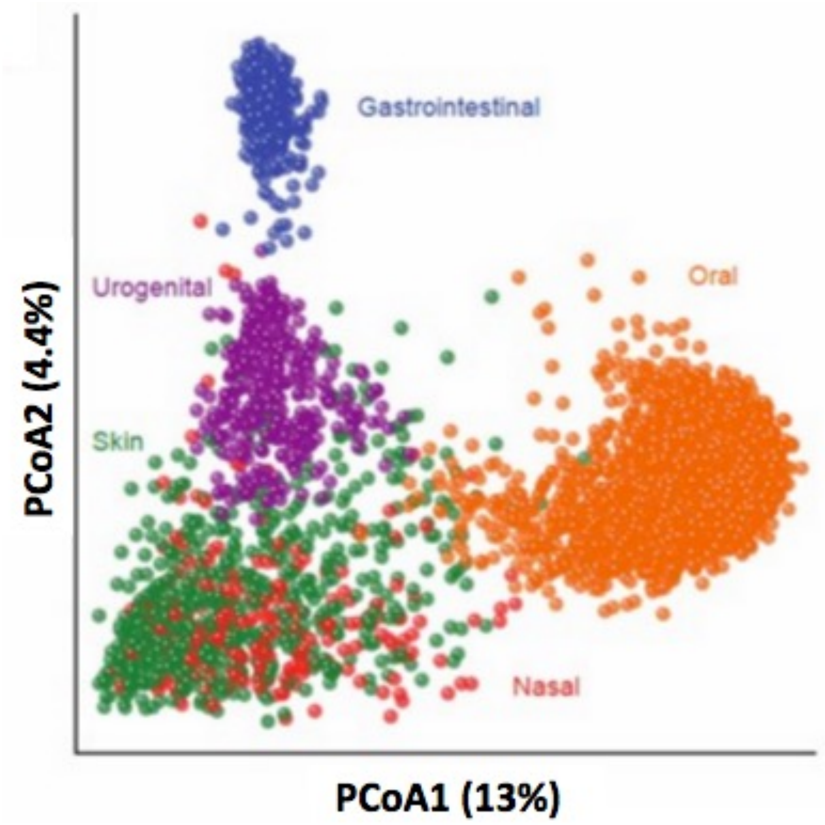


	R	Y	B
R	0	.5	.7
Y	.5	0	1
B	.7	1	0

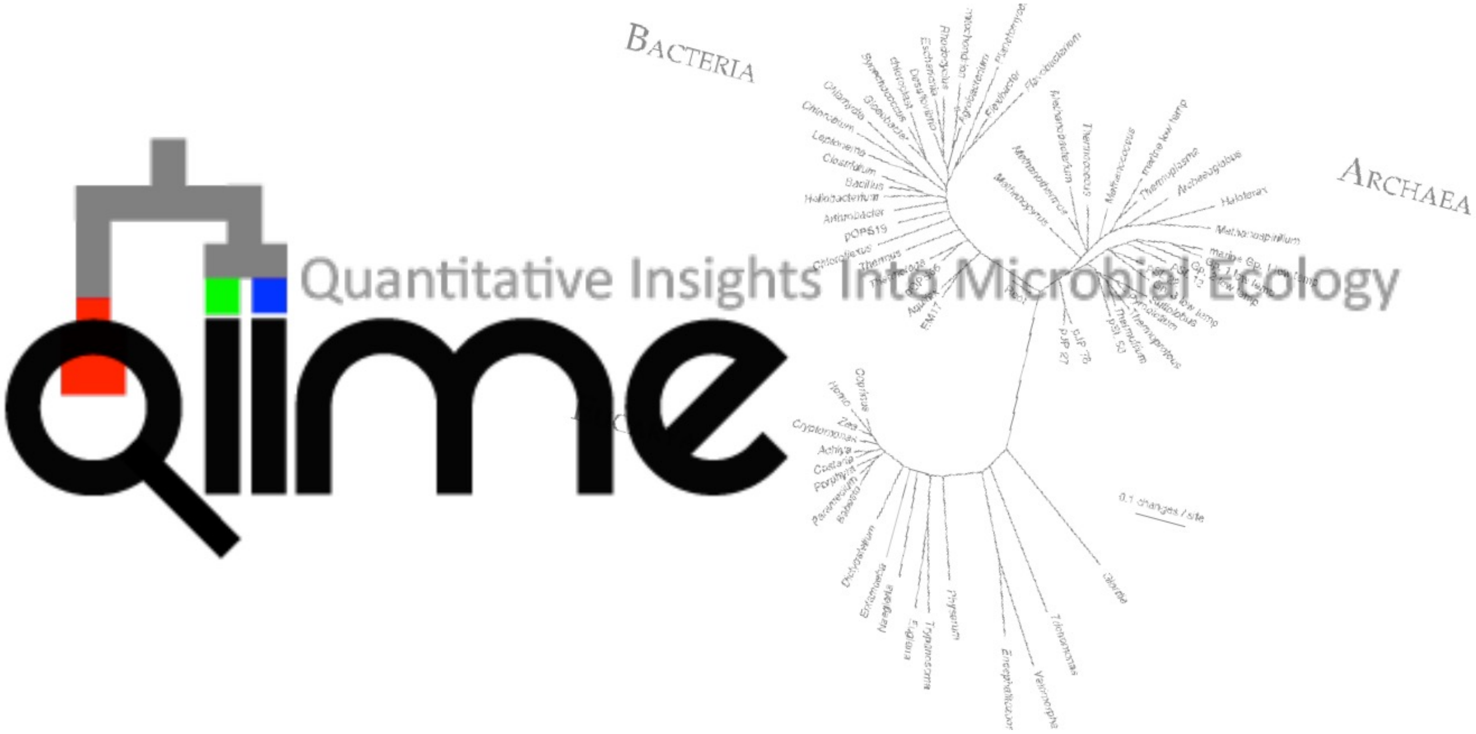
Lozupone & Knight 2005 AEM 71:8228

Principal Coordinate Analysis

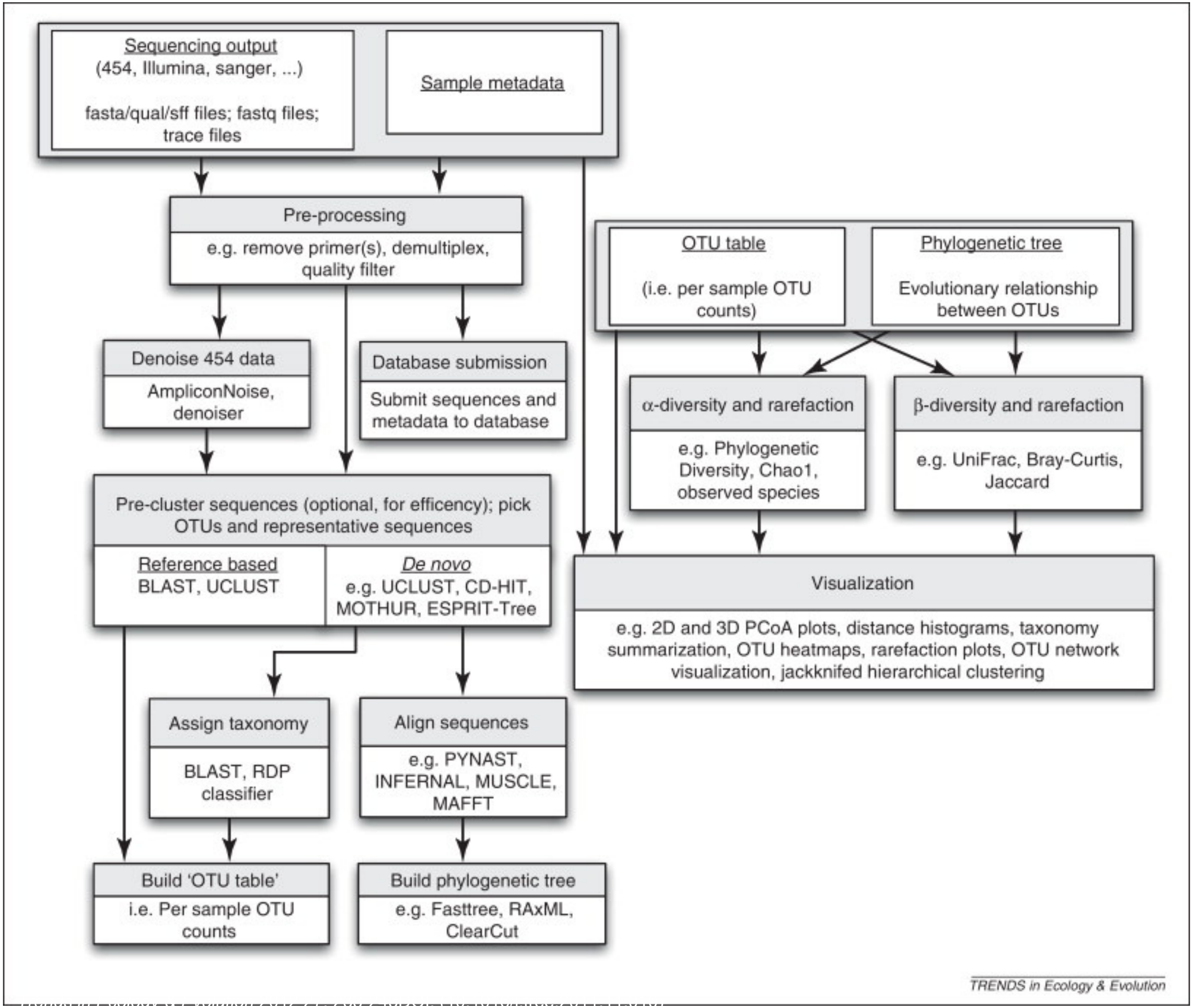
- Visualization of beta diversity matrix
- Transform distance matrix into new set of orthogonal axes
- 2D or 3D



QIIME / QIITA



- Open-source bioinformatics platform
  - data analysis from raw reads to figures
- Qiita: online data repository / data analysis platform



Sequencing output
(454, Illumina, sanger, ...)

fasta/qual/sff files; fastq files; trace files

Sample metadata

Pre-processing

e.g. remove primer(s), demultiplex, quality filter

Denoise 454 data

AmpliconNoise, denoiser

Database submission

Submit sequences and metadata to database

Pre-cluster sequences (optional, for efficency); pick OTUs and representative sequences

Reference based
BLAST, UCLUST

De novo
e.g. UCLUST, CD-HIT, MOTHUR, ESPRIT-Tree

Assign taxonomy

BLAST, RDP classifier

Align sequences

e.g. PYNAST, INFERNAL, MUSCLE, MAFFT

Build ‘OTU table’

i.e. Per sample OTU counts

Build phylogenetic tree

e.g. Fasttree, RAxML, ClearCut

OTU table

(i.e. per sample OTU counts)

Phylogenetic tree

Evolutionary relationship between OTUs

α-diversity and rarefaction

e.g. Phylogenetic Diversity, Chao1, observed species

β-diversity and rarefaction

e.g. UniFrac, Bray-Curtis, Jaccard

Visualization

e.g. 2D and 3D PCoA plots, distance histograms, taxonomy summarization, OTU heatmaps, rarefaction plots, OTU network visualization, jackknifed hierarchical clustering

TRENDS in Ecology & Evolution

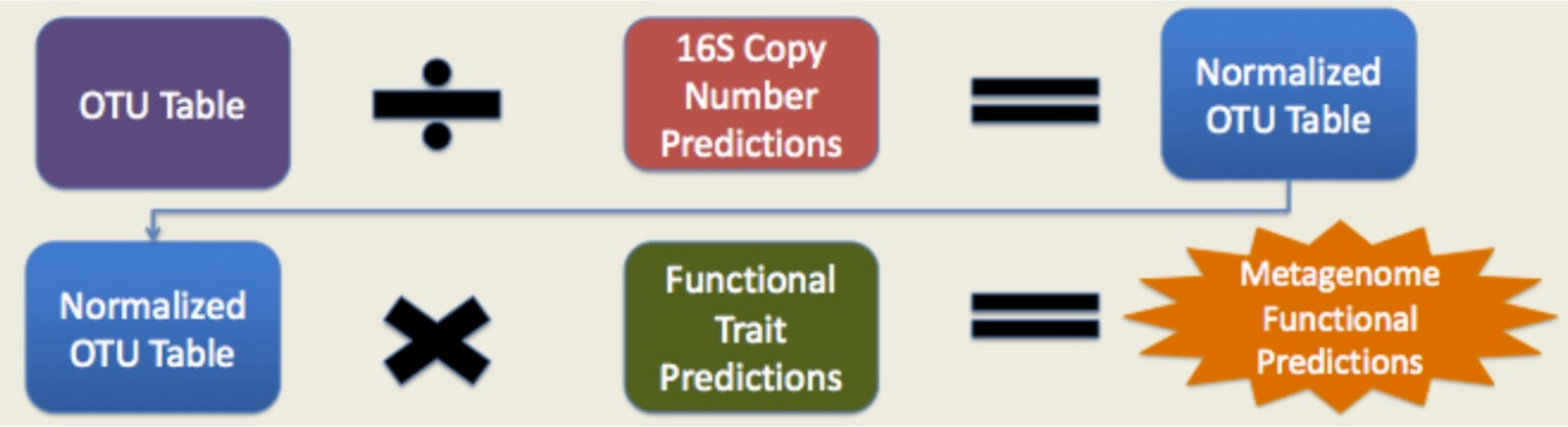
Terms and Conditions

Additional thoughts on 16S rRNA

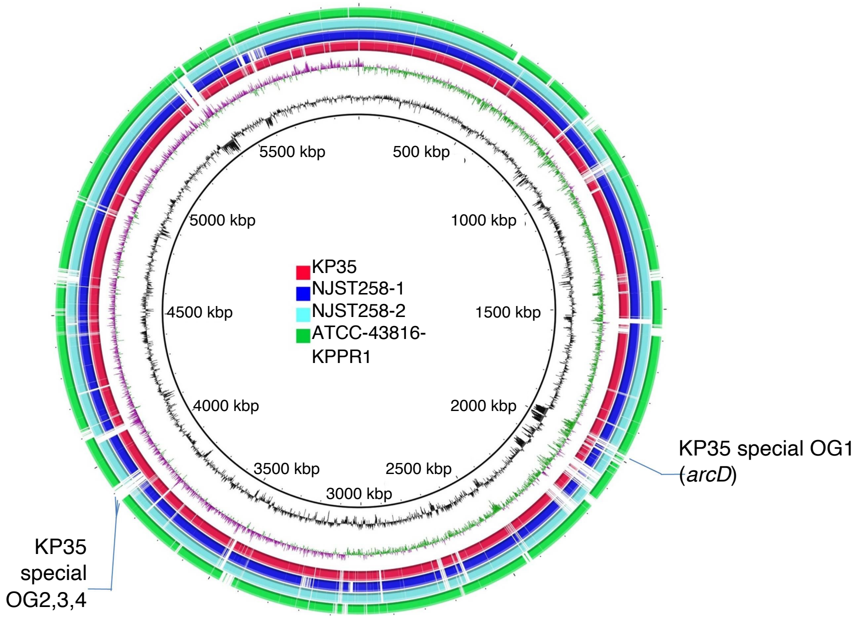
- Uniqueness of variable region determines taxonomic resolution
  - V3 or V3-V4 or V1-V2
  - length of variable region
- Optimal resolution depends on sample composition
- Extraction methods (Gram-positive versus Gram-negative!) may play an important role in full recovery of species

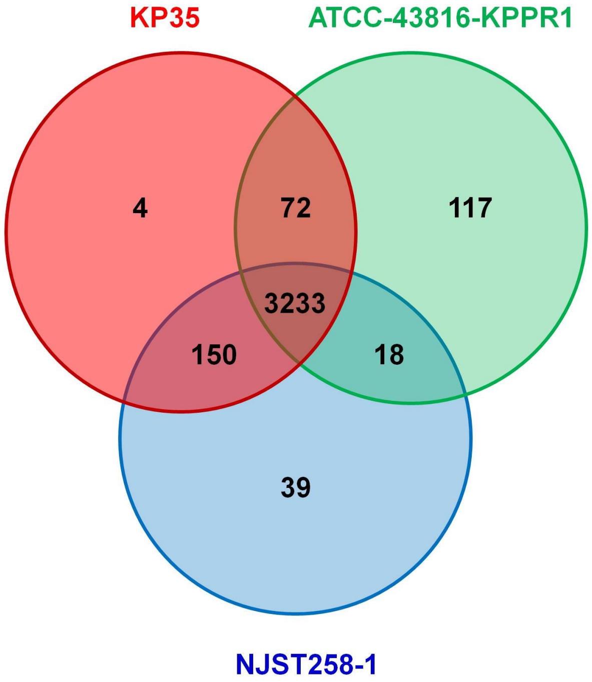
Predictive functional profiling of microbial communities by 16S rRNA genes

- PICRUSt software
- Validated using HMP data


OTU Table
÷
16S Copy Number Predictions
=
Normalized OTU Table
Normalized OTU Table
×
Functional Trait Predictions
=
Metagenome Functional Predictions

Simulation cannot predict clonal variants within species and genetic content in mobile genetic elements

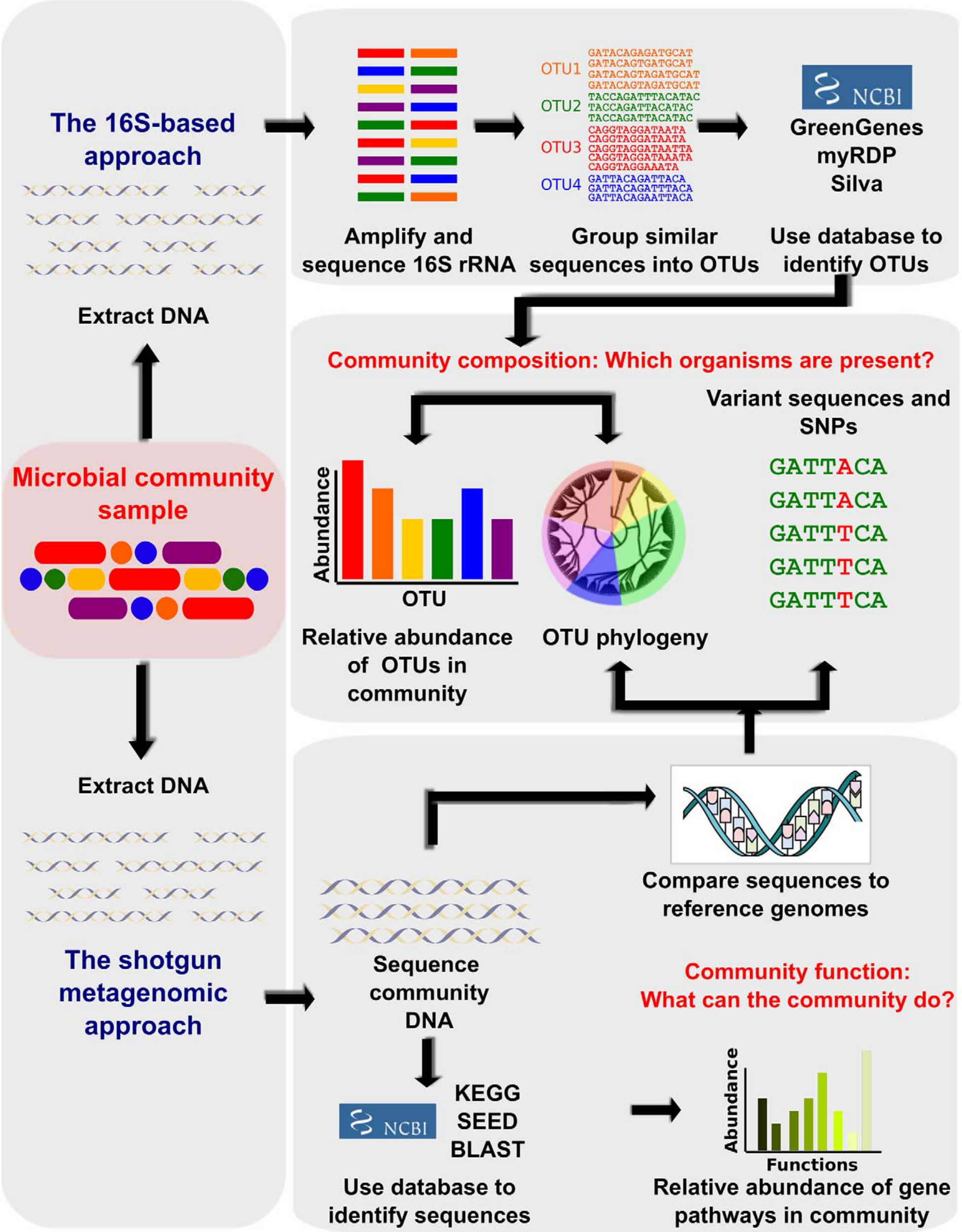




Presence of arcD (arginine metabolism) results in decreased virulence, immune evasion

Ahn et al. JCI Insight.

Bioinformatic methods for functional metagenomics



The 16S-based approach

Extract DNA

Amplify and sequence 16S rRNA

Group similar sequences into OTUs

Use database to identify OTUs

GreenGenes
myRDP
Silva

Community composition: Which organisms are present?

Relative abundance of OTUs in community

OTU phylogeny

Variant sequences and SNPs

Microbial community sample

Extract DNA

The shotgun metagenomic approach

Sequence community DNA

Compare sequences to reference genomes

Community function: What can the community do?

KEGG
SEED
BLAST

Use database to identify sequences

Relative abundance of gene pathways in community

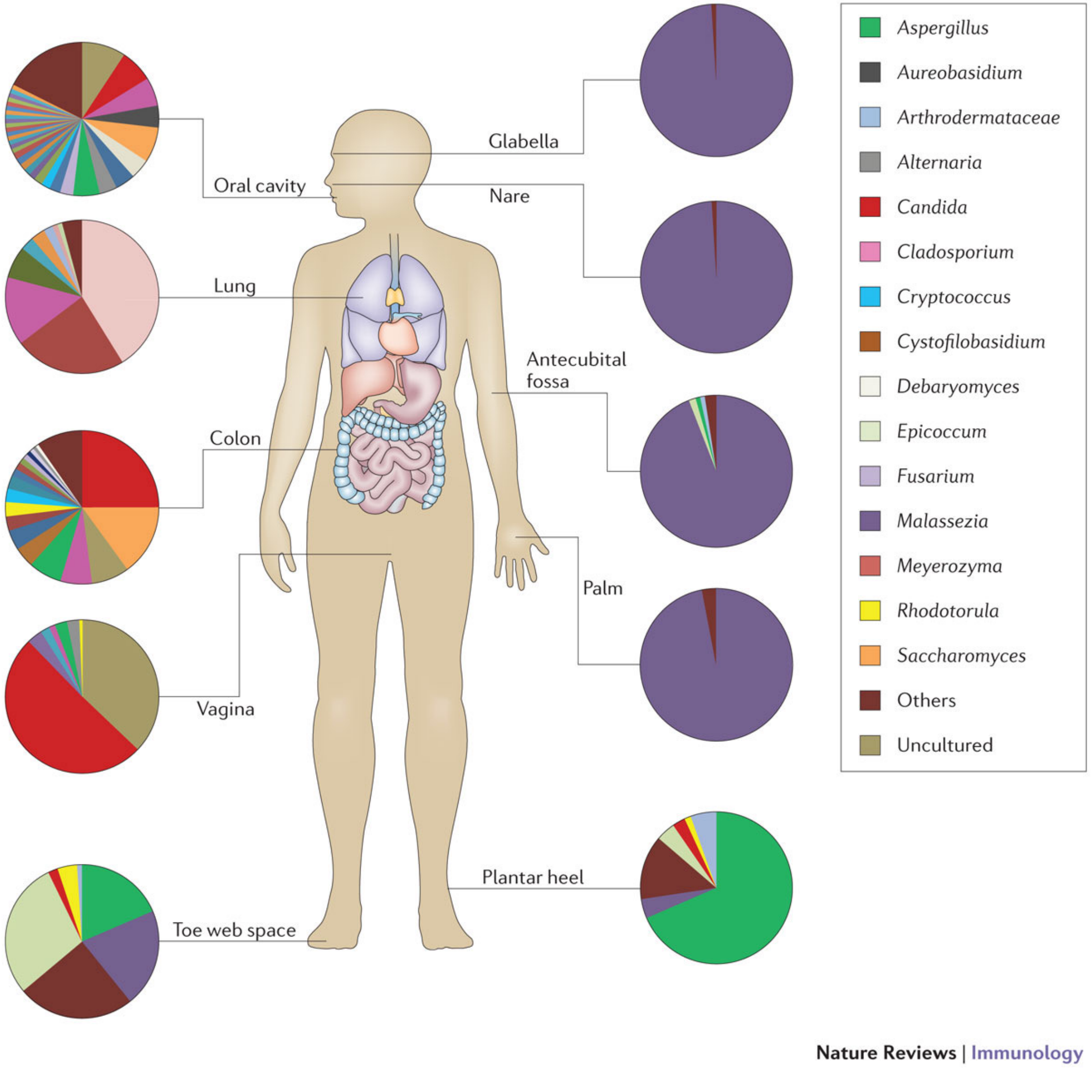
Morgan & Huttenhower,
PLOS Comp. Biol.,
2012

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



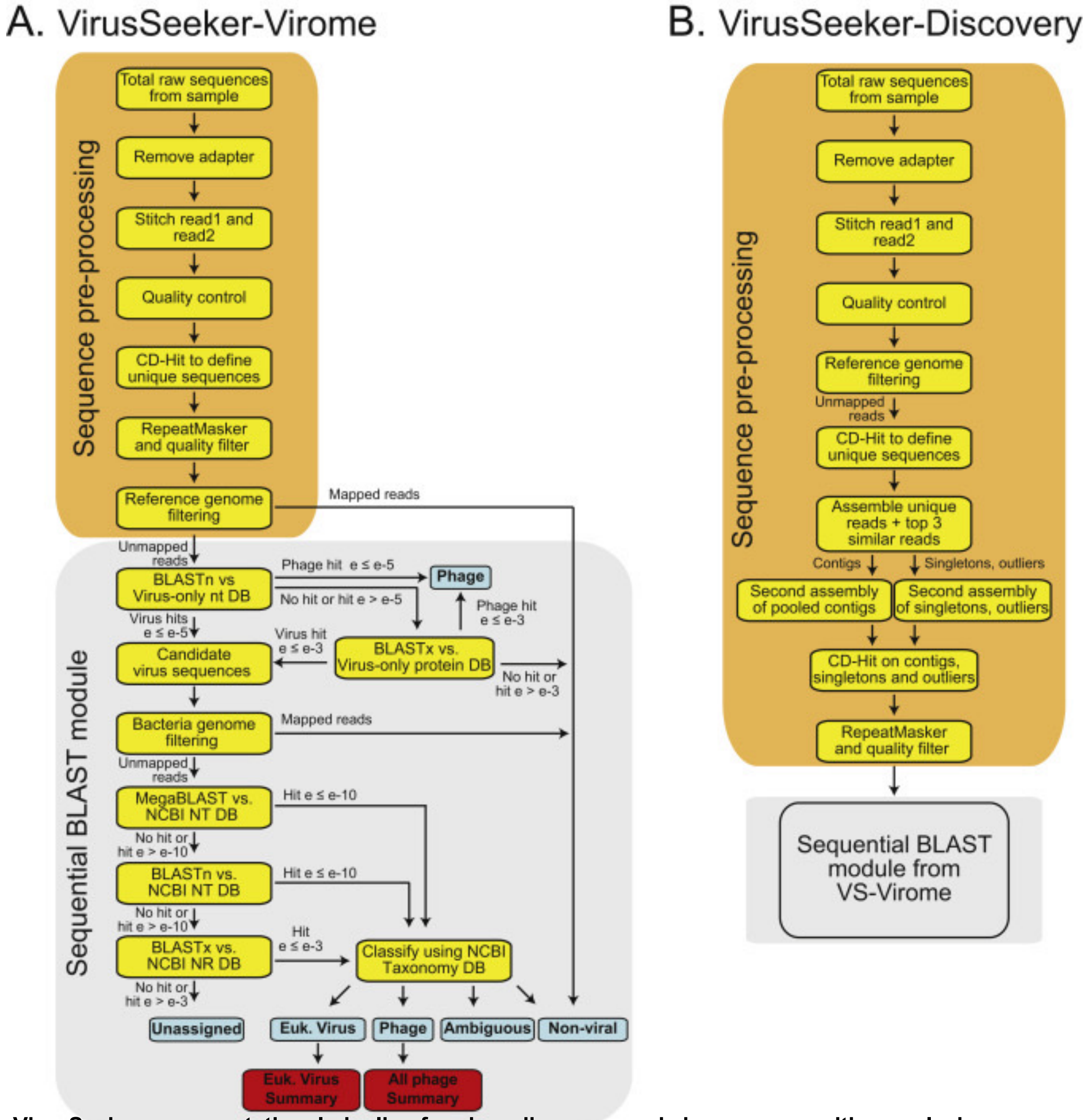
- Sequencing of
  - 18S ss rDNA
  - ITS region
- Longer reads needed
- Growing databases
  - UNITE
- Interaction with innate and adaptive immune system

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

H. Virgin Cell 2015



A. VirusSeeker-Virome

Sequence pre-processing

Total raw sequences from sample

Remove adapter

Stitch read1 and read2

Quality control

CD-Hit to define unique sequences

RepeatMasker and quality filter

Reference genome filtering

Mapped reads

Unmapped reads

Sequential BLAST module

BLASTn vs Virus-only nt DB

Phage hit e ≤ e-5

Phage

No hit or hit e > e-5

Phage hit e ≤ e-3

Virus hits e ≤ e-5

Virus hit e ≤ e-3

Candidate virus sequences

BLASTx vs. Virus-only protein DB

No hit or hit e > e-3

Bacteria genome filtering

Mapped reads

Unmapped 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 NR DB

Hit e ≤ e-3

Classify using NCBI Taxonomy DB

No hit or hit e > e-3

Unassigned

Euk. Virus

Phage

Ambiguous

Non-viral

Euk. Virus Summary

All phage Summary

B. VirusSeeker-Discovery

Sequence pre-processing

Total raw sequences from sample

Remove adapter

Stitch read1 and read2

Quality control

Reference genome filtering

Unmapped reads

CD-Hit to define unique sequences

Assemble unique reads + top 3 similar reads

Contigs

Singletons, outliers

Second assembly of pooled contigs

Second assembly of singletons, outliers

CD-Hit on contigs, singletons and outliers

RepeatMasker and quality filter

Sequential BLAST module from VS-Virome

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