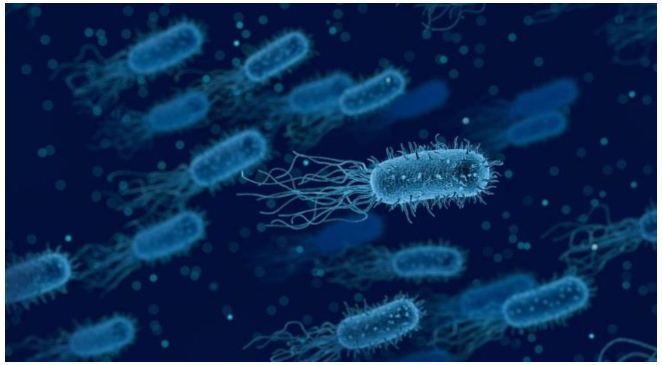
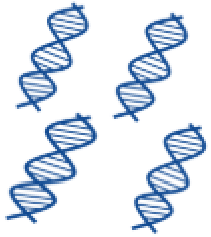
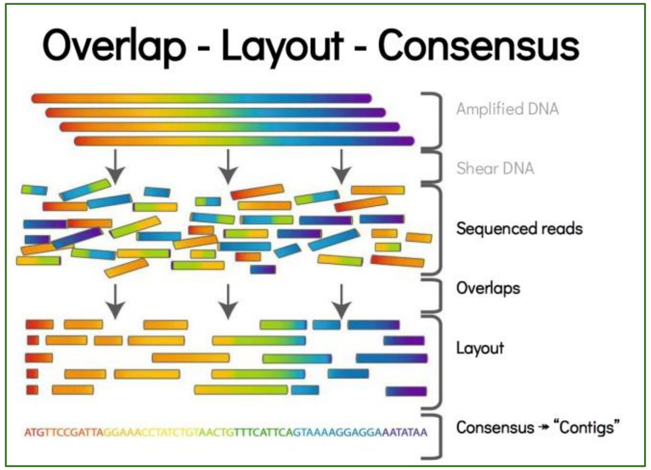
Alphabet Soup of WGS Analysis: MLST and SNP

Heather Blankenship, PhD Candidate Michigan State University,
Bioinformatics Intern Michigan Department of Health and Human Services

WGS Pipeline



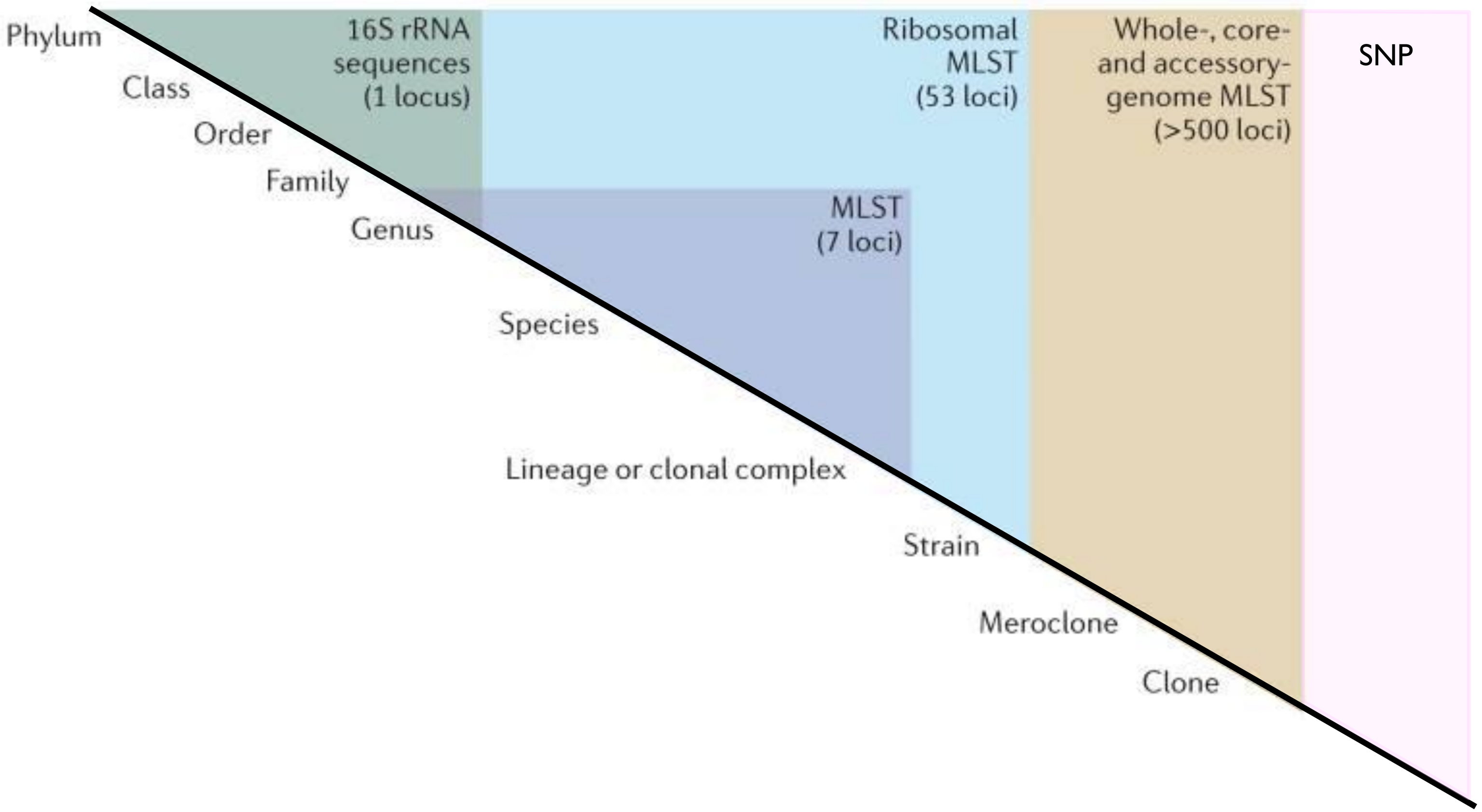



Overlap - Layout - Consensus
Amplified DNA
Shear DNA
Sequenced reads
Overlaps
Layout
Consensus → "Contigs"
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA

Torstern Seeman

Now what?

Choosing a Method

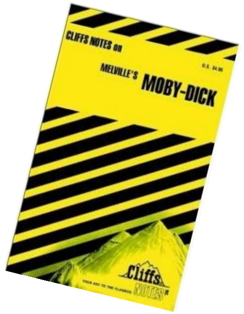

Phylum
Class
Order
Family
Genus
Species
Lineage or clonal complex
Strain
Meroclone
Clone
16S rRNA sequences (1 locus)
MLST (7 loci)
Ribosomal MLST (53 loci)
Whole-, core- and accessory-genome MLST (>500 loci)
SNP

- MLST- Multi Locus Sequencing Typing
- SNP- Single Nucleotide Polymorphism

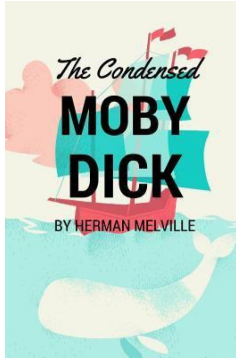
Maiden MC *et al* (2013); Nat Rev Microbiol

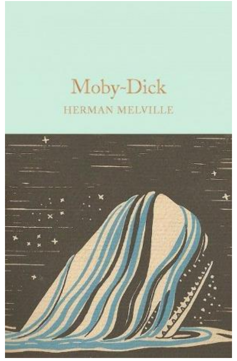
WGS Methods

- **MLST**
  - CliffNotes version of a book

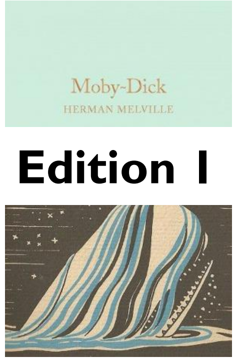


- **cgMLST**
  - Compare abridged versions of a book with other versions and only focus on the chapters that are the same in all of them

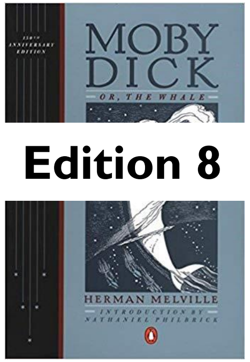




- **wgMLST analysis**
  - Compare two books on a chapter by chapter basis



Edition 1



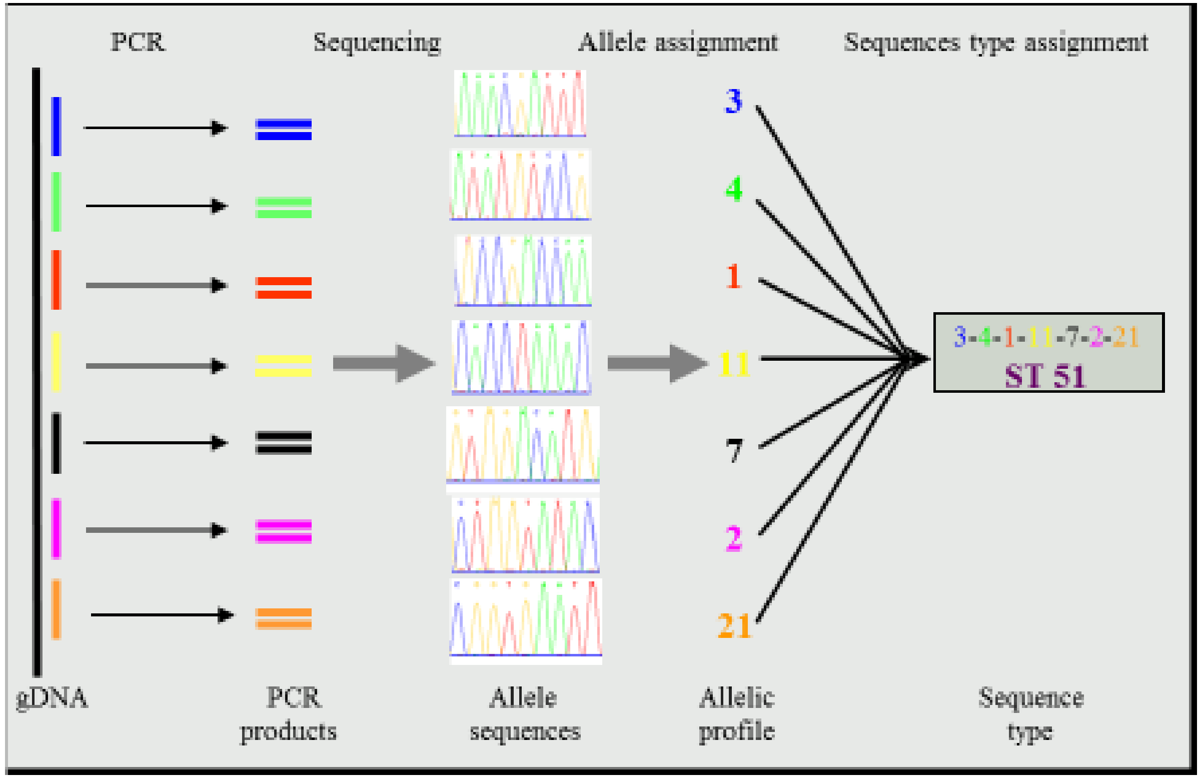
Edition 8

- **SNP analysis**
  - Compare two books character by character

The dog and cat watched the bird sing.
The hog and cat matched the bird song.

MLST

- Traditional MLST
  - ~6-7 housekeeping genes
  - Allelic variation for each gene is identified
  - Sequence Type (ST) assigned based on allelic profile
  - Databases are internationally available
  - Suitable for examining differences at the population level



- **Allele**: one or more alternative forms of a gene
- **Locus**: gene or region of the gene that is being extracted and compared

Ruppitsch W (2016); J Land Management, Food and Env

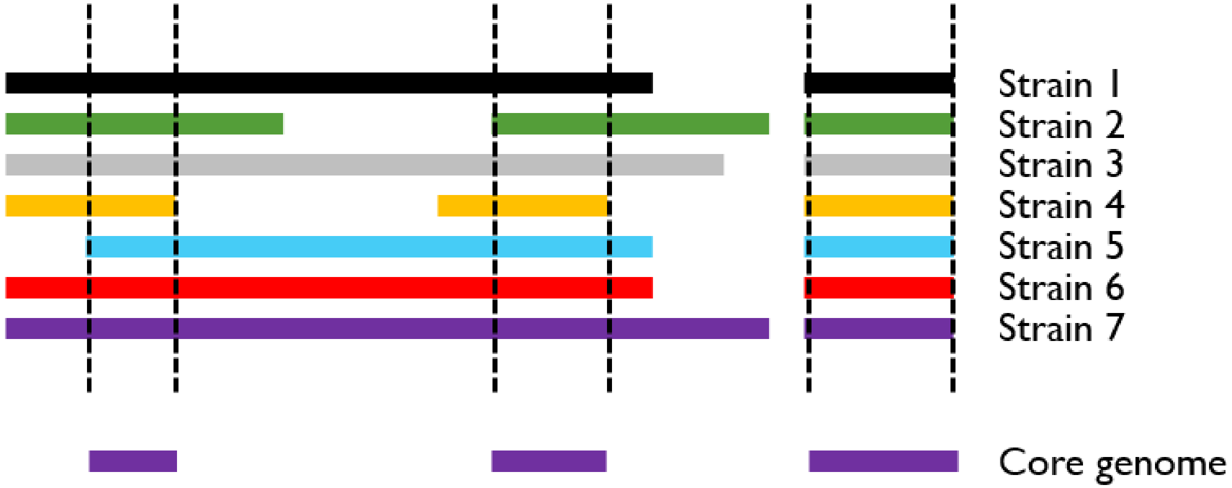
wgMLST (whole genome)

- Compare genomes of interest gene by gene
- Database comprises genes from the genus/species of interest and represents a diverse genetic background within the genus/species
  - Does not have to be the full gene included in the database
  - A new allele is only added as it is encountered
- STANDARDIZED!

	Gene A	
Allele 1	ATGTAGCGCTAGCC	
Allele 2	ATGTAGCCCTAGCC	SNP
Allele 3	ATGTAGATGGCTAGCC	SNP + insertion

cgMLST (core genome)

- Similar to wgMLST, compares on a gene by gene basis
- What genes/portions of genes are common to all within the species
  - Results in a smaller databases
- Still STANDARDIZED



- wgMLST or cgMLST
  - wgMLST – higher resolution
  - cgMLST – more stability

Allele codes/profiles

MLST
Sequence A – 4 . 4 . 5 . 6 . 10 . 1 . 5 – ST30
Sequence B – 4 . 6 . 5 . 6 . 10 . 1 . 5 – ST35
Sequence C – 1 . 4 . 5 . 6 . 10 . 1 . 1 – ST4

wg/cgMLST
Sequence A – 5 . 5 . 21 . 6
Sequence B – 5 . 5 . 21
Sequence C – 1 . 5

- The allele codes/profiles do not change regardless of the isolates that are added to an analysis

- Allows comparison with other labs/states/outbreaks etc

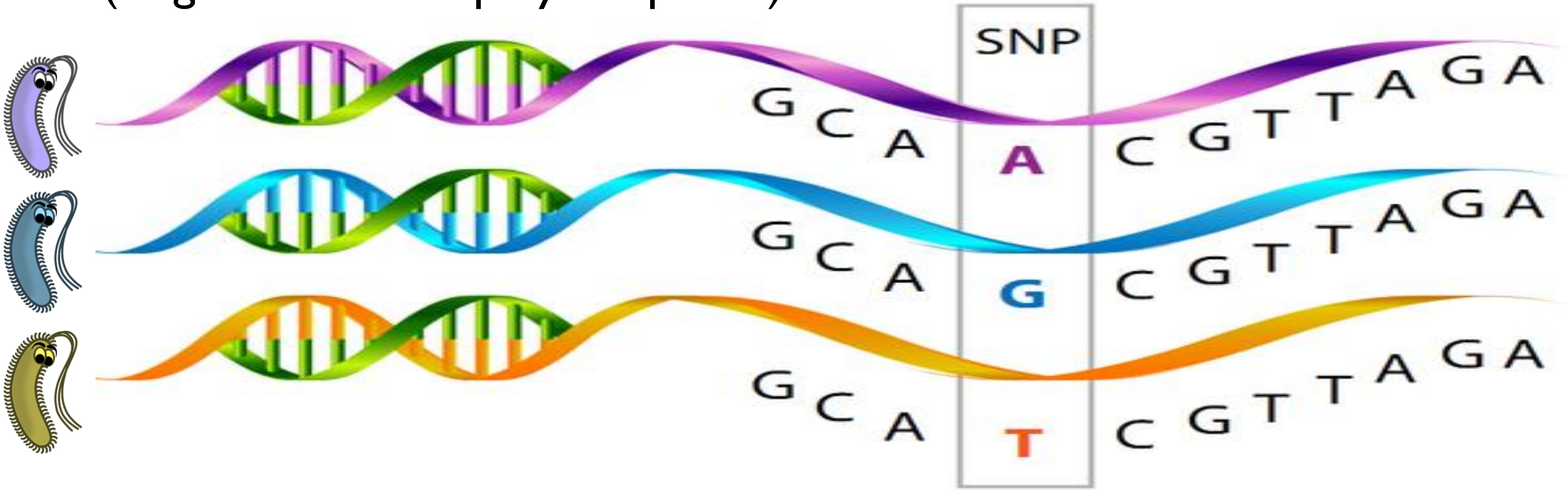
Benefits/Disadvantages of MLST methods

- Disadvantage
  - A lot of upfront development to develop and continually curate the databases
  - SNPs, insertions, deletions are all treated the same, an allele may have multiple evolutionary events but it is not evident by looking at the allele number
  - Comparison is based on the allele numbers and not the genetic sequences
  - Requires genome assembly
  - Does not include non-coding regions
- Advantage
  - Provides high enough differentiation of strains
  - Allele call is stable
  - Low computational power
  - Standardized data

Best for surveillance

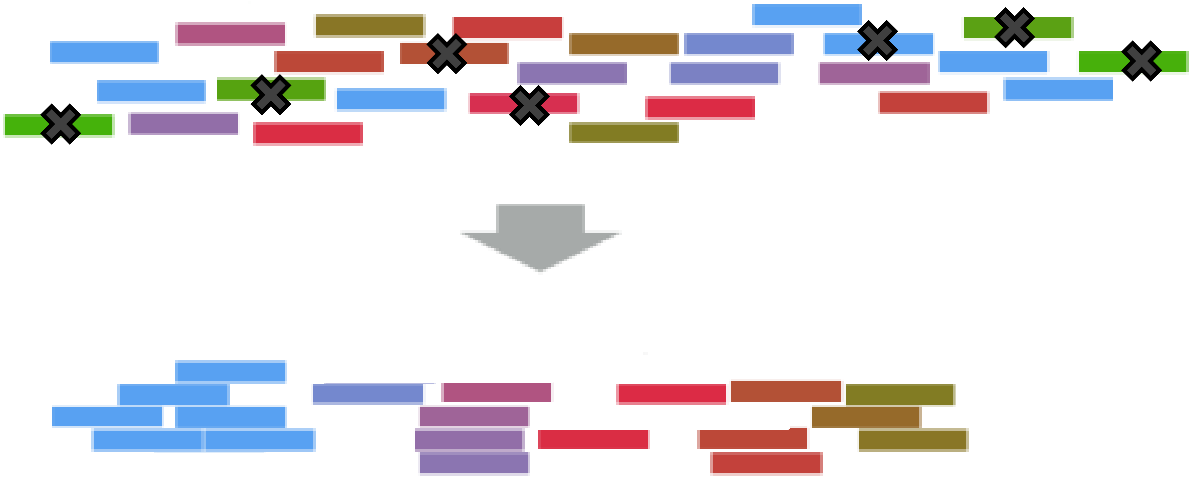
SNP typing - Basics

- Map reads against a reference genome
- Identify SNPs between reference and reads
  - Quality, Coverage, Frequency
- Build phylogenetic tree based on concatenated SNPs
- **SNP** (single nucleotide polymorphism)



hqSNP (high quality)

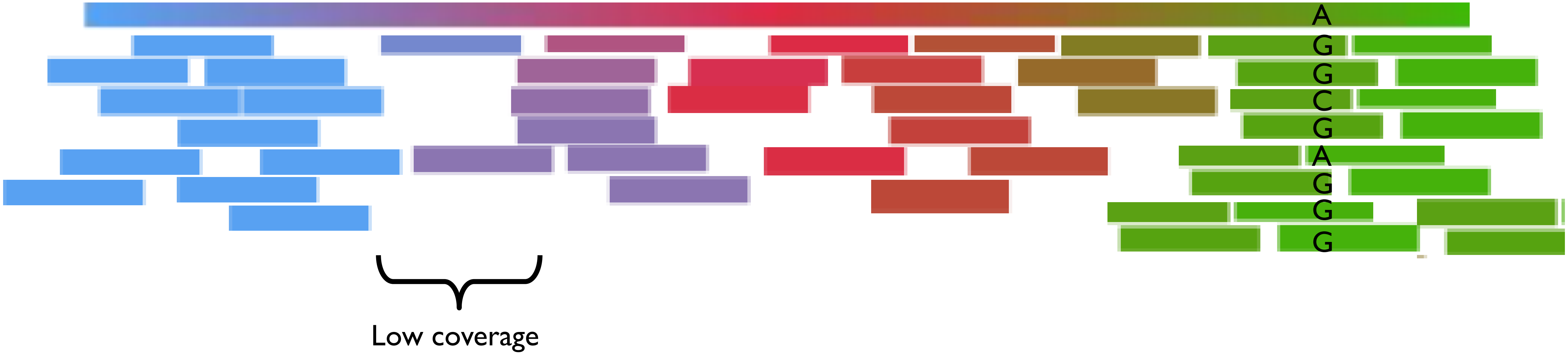
- Takes into account the quality, coverage and frequency of the SNPs
- Quality: reads are filtered to ensure they pass a certain threshold



- Reads are cleaned to minimize the amount of erroneous data and improve average quality by removing read duplicates, reads with high frequency of ambiguous bases and adapter dimers

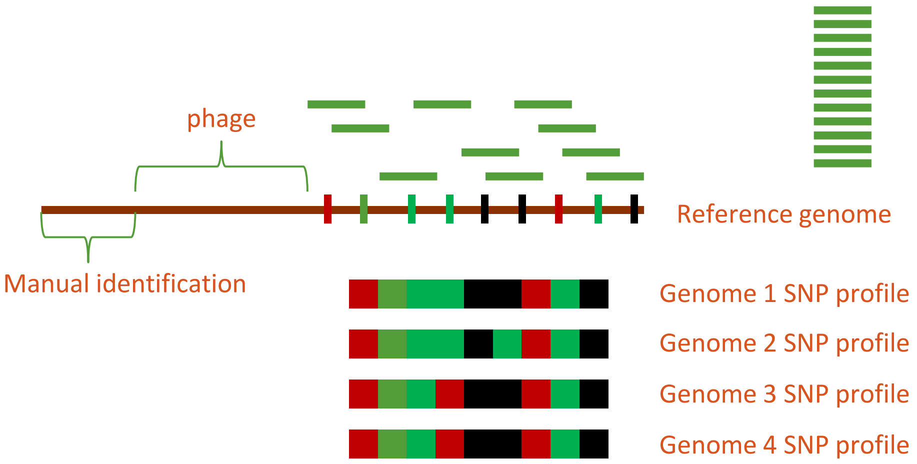
hqSNP

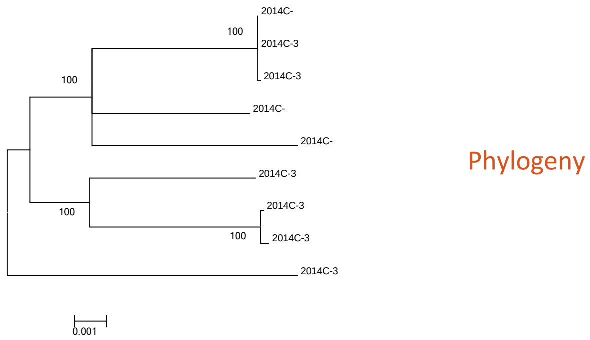
- Coverage: how many reads do we have at a particular nucleotide
- Frequency: how many of the reads support the new SNP



Lyve-Set (hqSNP) Process

- Pre-Processing
  - Phage discovery/masking
  - Manual identification of troublesome regions
  - Read cleaning
- Mapping –SMALT
  - 95% read identity
  - Unambiguous mapping
- SNP calling – VarScan
  - 75% consensus
  - 10X depth
- Phylogeny inferring – RAxML
  - Removal of clustered SNPs
  - Ascertainment bias model
  - Maximum likelihood

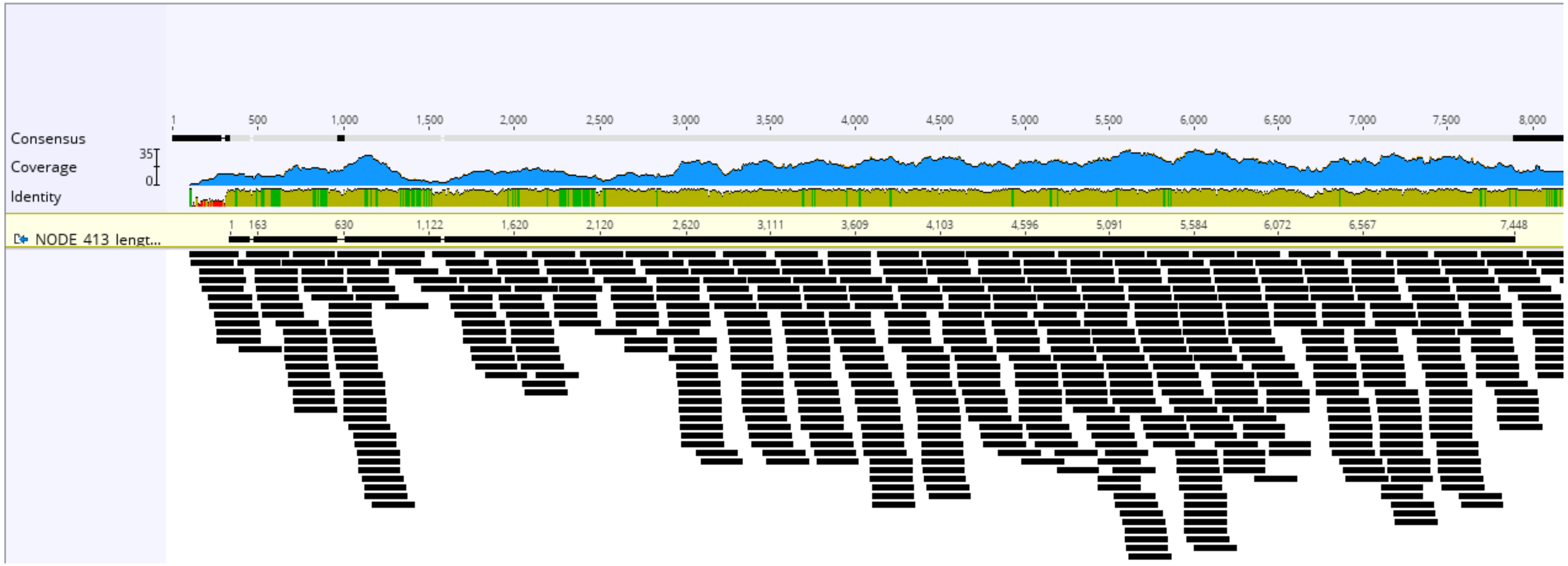




https://github.com/lskatz/lyve-SET/blob/master/docs/FAQ.md
https://www.sanger.ac.uk/resources/software/smalt
Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L., & Wilson, R. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing *Genome Research* DOI: 10.1101/gr.129684.111
Stamatakis, A. (2014) RAxML version8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. doi:10.1093/bioinformatics/btu033

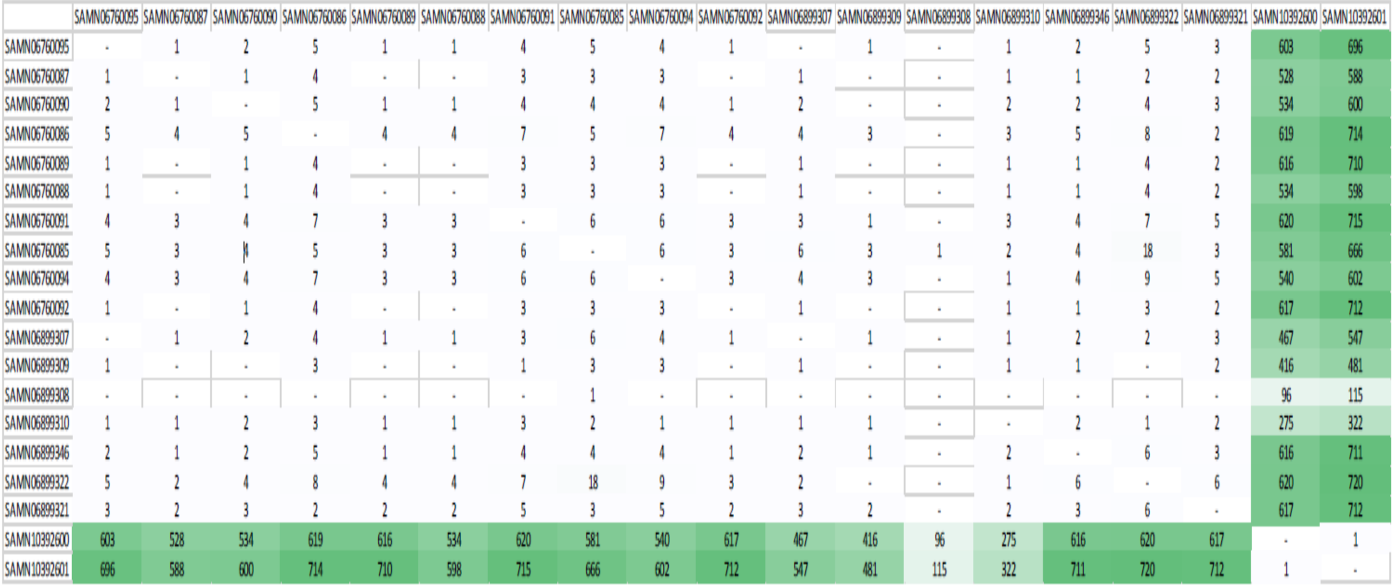
Lee Katz

SNP typing

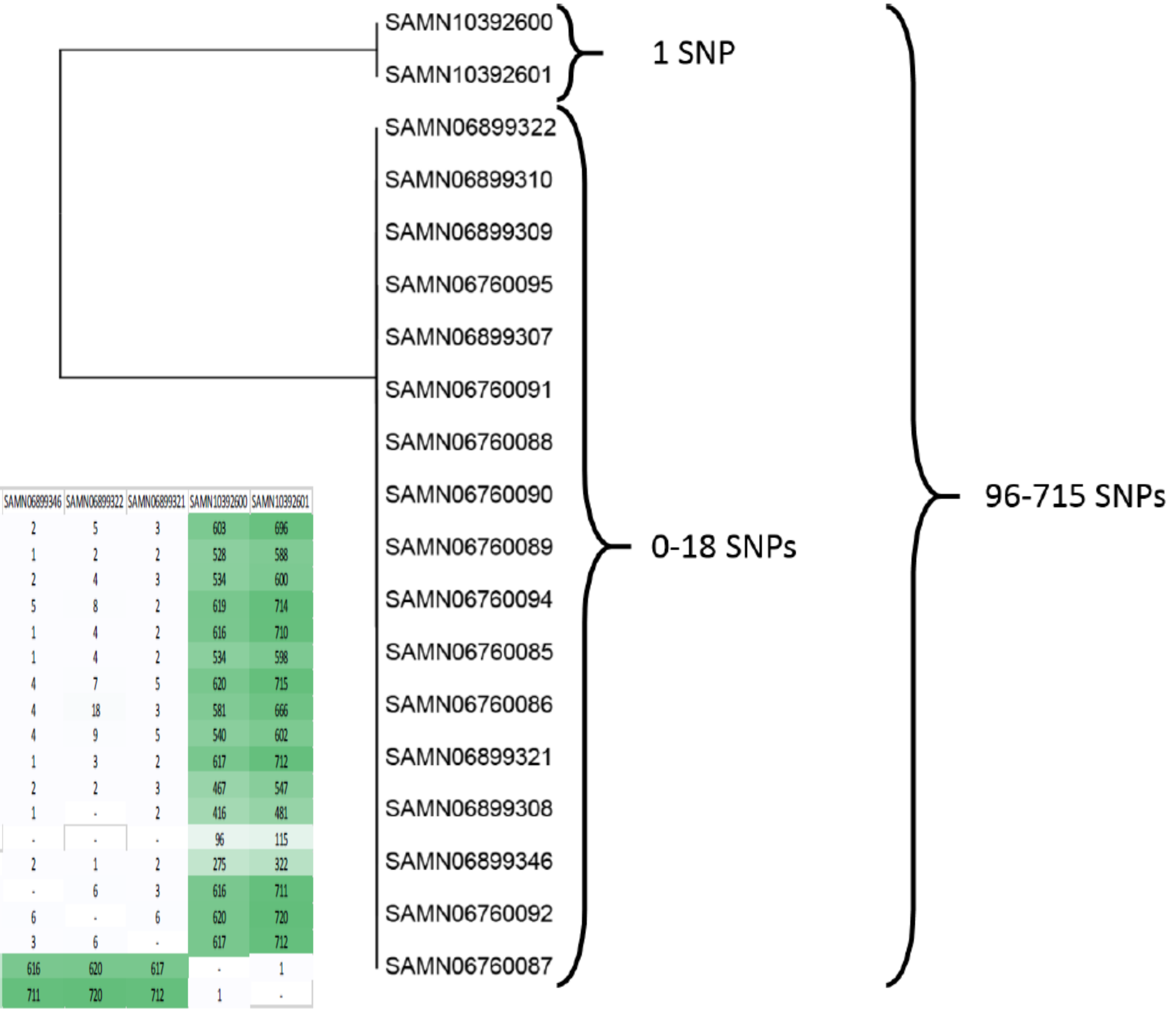


- Allows for every nucleotide in the genome to be analyzed
- Very dependent upon the reference genome that is chosen
- All mobile elements (ie phages) are masked and not included in the analysis

SNP visualization



	SAMN06760095	SAMN06760087	SAMN06760090	SAMN06760086	SAMN06760089	SAMN06760088	SAMN06760091	SAMN06760085	SAMN06760094	SAMN06760092	SAMN06899307	SAMN06899309	SAMN06899308	SAMN06899310	SAMN06899346	SAMN06899322	SAMN06899321	SAMN10392600	SAMN10392601
SAMN06760095	-	1	2	5	1	1	4	5	4	1	-	1	-	1	2	5	3	603	696
SAMN06760087	1	-	1	4	-	-	3	3	3	-	1	-	-	1	1	2	2	528	588
SAMN06760090	2	1	-	5	1	1	4	4	4	1	2	-	-	2	2	4	3	534	600
SAMN06760086	5	4	5	-	4	4	7	5	7	4	4	3	-	3	5	8	2	619	714
SAMN06760089	1	-	1	4	-	-	3	3	3	-	1	-	-	1	1	4	2	616	710
SAMN06760088	1	-	1	4	-	-	3	3	3	-	1	-	-	1	1	4	2	534	598
SAMN06760091	4	3	4	7	3	3	-	6	6	3	3	1	-	3	4	7	5	620	715
SAMN06760085	5	3	4	5	3	3	6	-	6	3	6	3	1	2	4	18	3	581	666
SAMN06760094	4	3	4	7	3	3	6	6	-	3	4	3	-	1	4	9	5	540	602
SAMN06760092	1	-	1	4	-	-	3	3	3	-	1	-	-	1	1	3	2	617	712
SAMN06899307	-	1	2	4	1	1	3	6	4	1	-	1	-	1	2	2	3	467	547
SAMN06899309	1	-	-	3	-	-	1	3	3	-	1	-	-	1	1	-	2	416	481
SAMN06899308	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	96	115
SAMN06899310	1	1	2	3	1	1	3	2	1	1	1	1	-	-	2	1	2	275	322
SAMN06899346	2	1	2	5	1	1	4	4	4	1	2	1	-	2	-	6	3	616	711
SAMN06899322	5	2	4	8	4	4	7	18	9	3	2	-	-	1	6	-	6	620	720
SAMN06899321	3	2	3	2	2	2	5	3	5	2	3	2	-	2	3	6	-	617	712
SAMN10392600	603	528	534	619	616	534	620	581	540	617	467	416	96	275	616	620	617	-	1
SAMN10392601	696	588	600	714	710	598	715	666	602	712	547	481	115	322	711	720	712	1	-



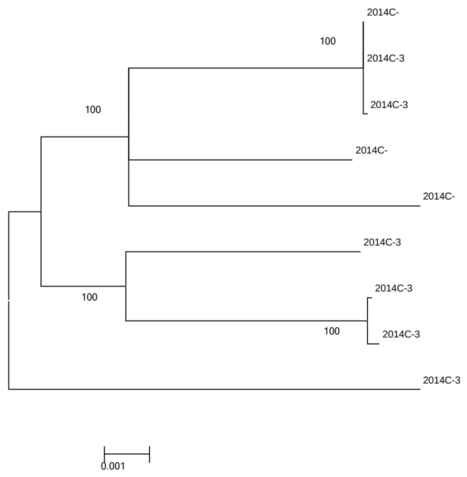
Benefits/Disadvantages of SNP methods

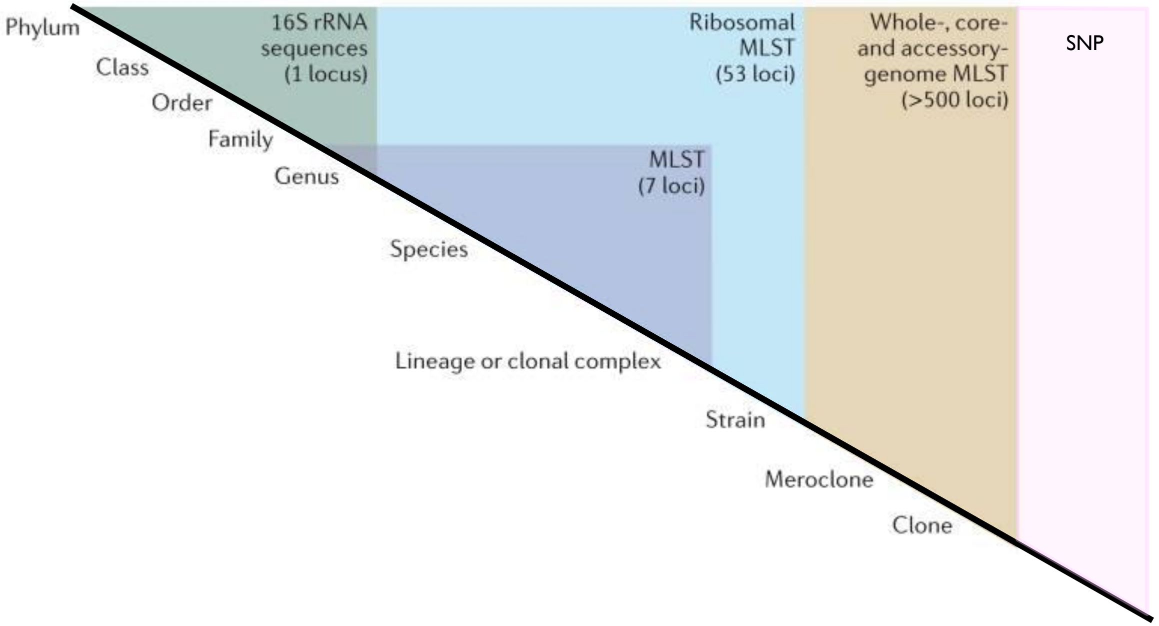
- Disadvantage
  - Computationally intensive
  - Reliant on the reference – not standardized
  - Time and computationally intensive
  - Not stable and relies on the genomes present
- Advantage
  - Does not require assembly of the genomes
  - Very high discriminative power
  - Each mutation is taken into account
  - Non-coding regions are utilized
  - Adaptable for all organisms

Best for high discriminatory power and outbreak investigation

Utilization of WGS Analysis

- Analysis will give us phylogenetic trees = <u>hypothesis</u>
- Strains that are genetically related based on analysis **<u>MAY</u>** share an epidemiological association
  - What defines a cluster?
  - Are cutoffs absolute?
  - Does WGS "match" mean association?





Phylum
Class
Order
Family
Genus
Species
Lineage or clonal complex
Strain
Meroclone
Clone
16S rRNA sequences (1 locus)
MLST (7 loci)
Ribosomal MLST (53 loci)
Whole-, core- and accessory-genome MLST (>500 loci)
SNP

Maiden MC *et al* (2013); Nat Rev Microbiol

QUESTIONS?

- blankenshiph@michigan.gov
- selheime@msu.edu