

MOLECULAR EPIDEMIOLOGY

Dr Charlene Kahler

OBJECTIVES

- ❖ By the end of this lecture, you should be able to answer questions relating to:
 - ❖ Understand the principle of bacterial variation and the processes that drive this
 - ❖ Revise terms and objectives for infectious diseases public health epidemiology
 - ❖ Understand the process of typing strains for outbreak investigations
 - ❖ Describe typing schemes based on multi-locus sequence typing (MLST)
 - ❖ Understand the terminology used to describe phylogenetic trees
 - ❖ Explain how a phylogeny is obtained using a simple dataset of alleles from a gene

THE POWER OF BACTERIAL VARIATION

DEFINITION:

A bacterial species is a collection of strains with a conserved core of genes and phenotypes

- For example, all *Escherichia coli* are gram negative rods, facultative anaerobe, non-sporulating

DEFINITION:

A strain is a subvariant of a bacterial species.

It can be defined by genetic content and sometimes phenotype

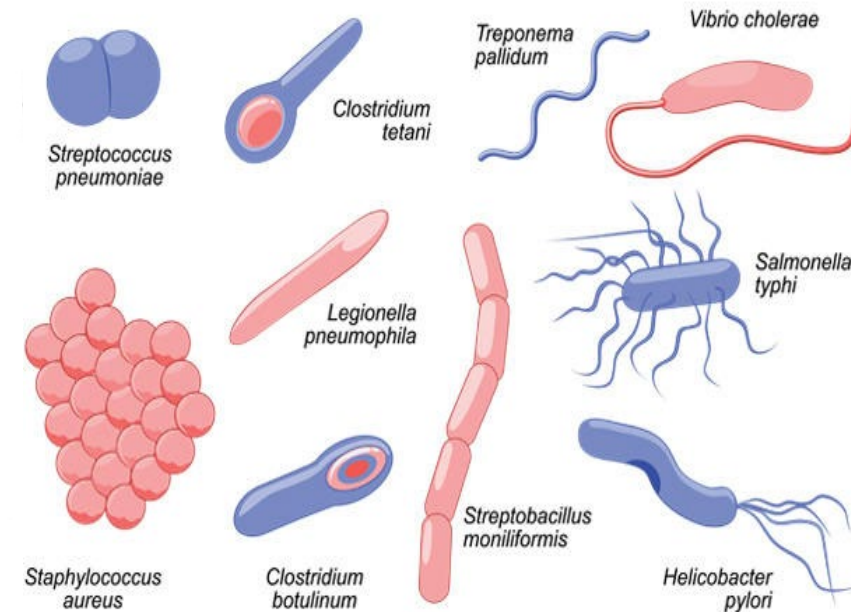
Isolates (individual pure cultures from different sources) of a strain may share 99% of gene sequence identity

Strains of the same species may have a completely different phenotype

- *E. coli* is generally a commensal
- *E. coli* have many pathotypes (EPEC, APEC, UPEC etc)- these are different strains causing different disease syndromes

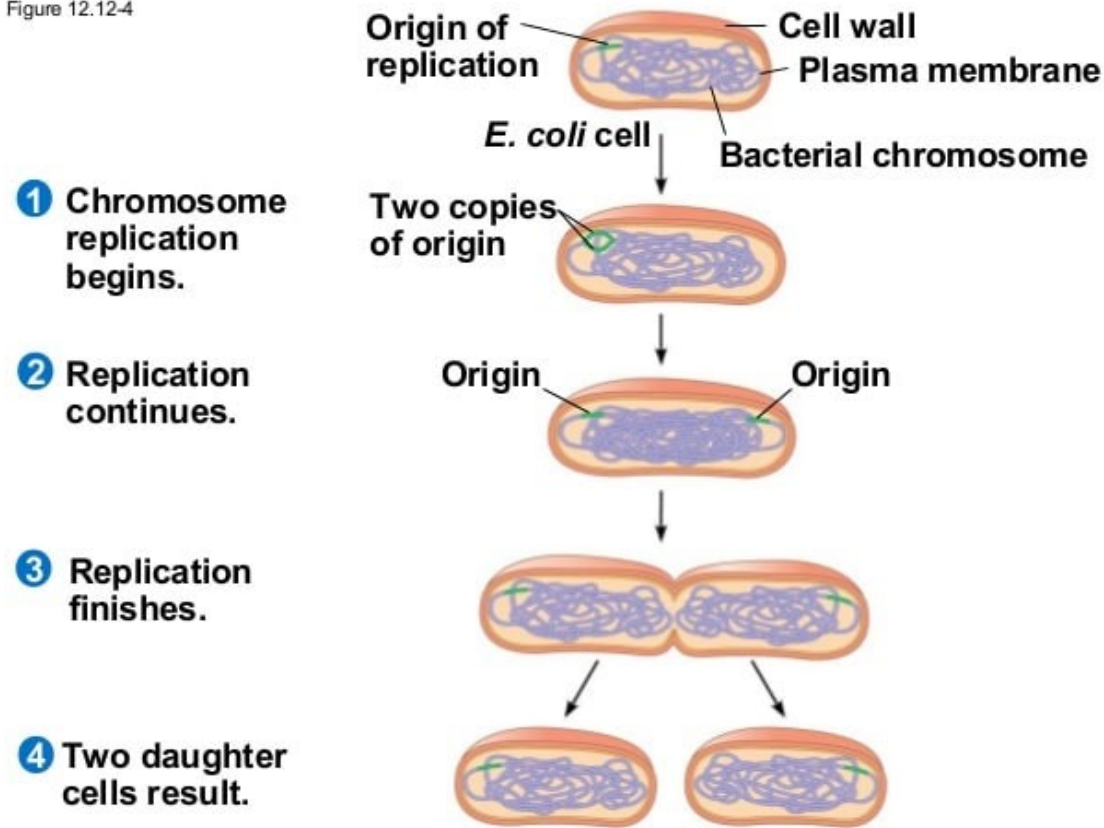
Genotypes and Phenotypes of strains change based upon

- Mutation of the core genome
- Acquisition of foreign DNA from other sources



REVISION: BACTERIAL CELL DIVISION

Figure 12.12-4



© 2011 Pearson Education, Inc.

If replication is faithful, both bacterial cells are identical = same isolate
But if replication has errors, the bacterial cells are not identical = different strains

See MICR5831 lectures

REVISION: BACTERIAL VARIATION:

DNA DAMAGE AND REPAIR

- A permanent, heritable change to the base sequence of the DNA
- **Spontaneous**
 - Error in DNA replication is the main cause: rate of 1 in 10^8 to 10^{11} nucleotides is copied incorrectly by the DNA polymerase
- **Induced**
- DNA is damaged by chemical alteration
 - Alkylation
 - Electrophiles add alkyl groups to phosphates, stalls replication
 - eg. carcinogens, ethylmethane sulphonate (EMS)
 - UV-induced thymine dimers
 - DNA absorbs UV at 260nm
 - Forms intra-strand pyrimidine dimers, mainly T-T
 - Distortion of double helix prevents DNA replication, thus is lethal
 - Oxygen radicals
 - Cause single and double stranded breaks
 - eg. Gamma and X-rays
- Initial damage, if not repaired, will result in mutation of the DNA which will be inherited by daughter cells via cell division

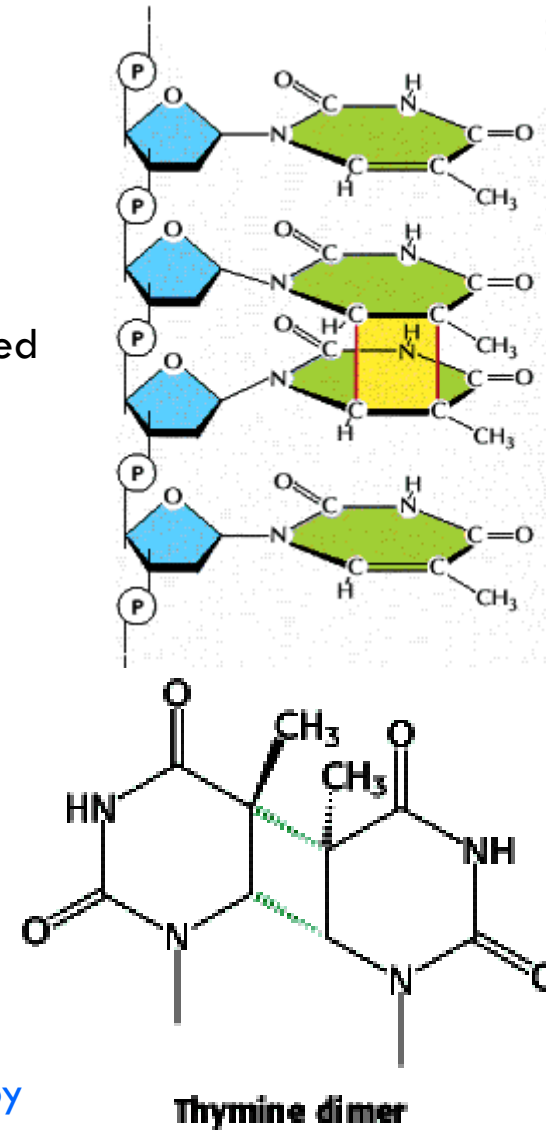
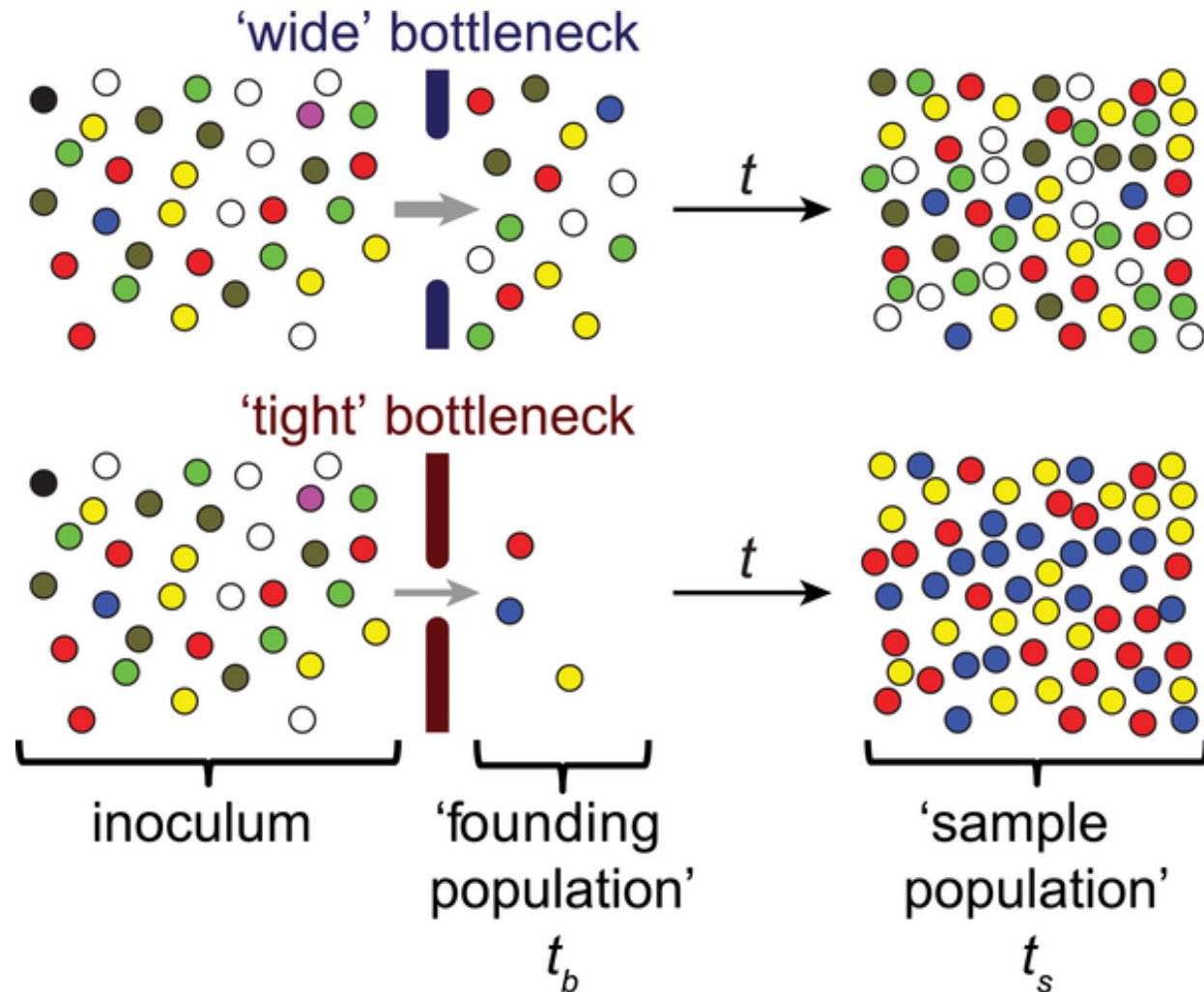


Fig 1. Schematic representation of the effect of bottlenecks on genetic diversity.

Imagine every isolate (every dot) has a different genetic makeup of single nucleotide polymorphisms or acquired other genetic elements



Abel S, Abel zur Wiesch P, Davis BM, Waldor MK (2015) Analysis of Bottlenecks in Experimental Models of Infection. PLOS Pathogens 11(6): e1004823. <https://doi.org/10.1371/journal.ppat.1004823>
<https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1004823>

INFECTIOUS DISEASE EPIDEMIOLOGY

Definition of Epidemiology

- **Study of distribution & determinants of disease and health related events and its application in control and prevention.**

Epidemiology

Deals with one population

Risk → case

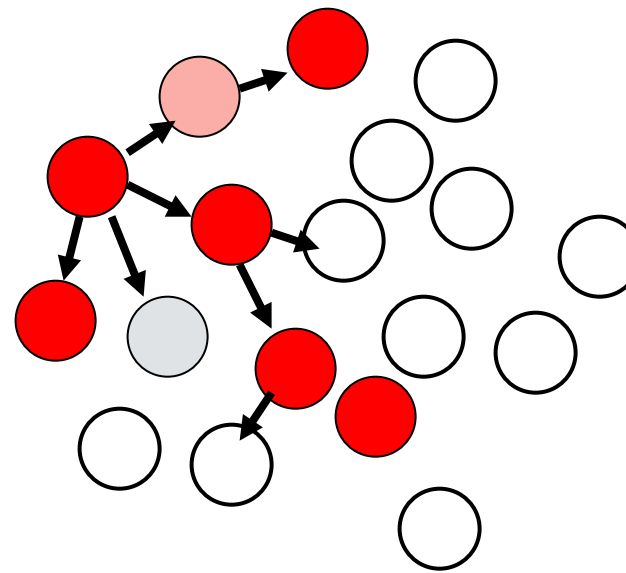
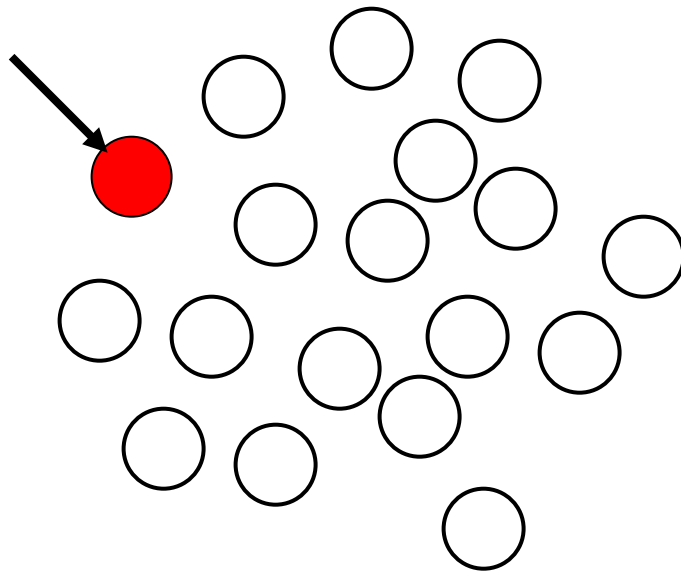
Identifies causes




Infectious disease epidemiology

- ❖ Two or more populations
- ❖ A case is a risk factor
- ❖ The cause often known

A CASE IS A RISK FACTOR

Infection in one person can be transmitted to others

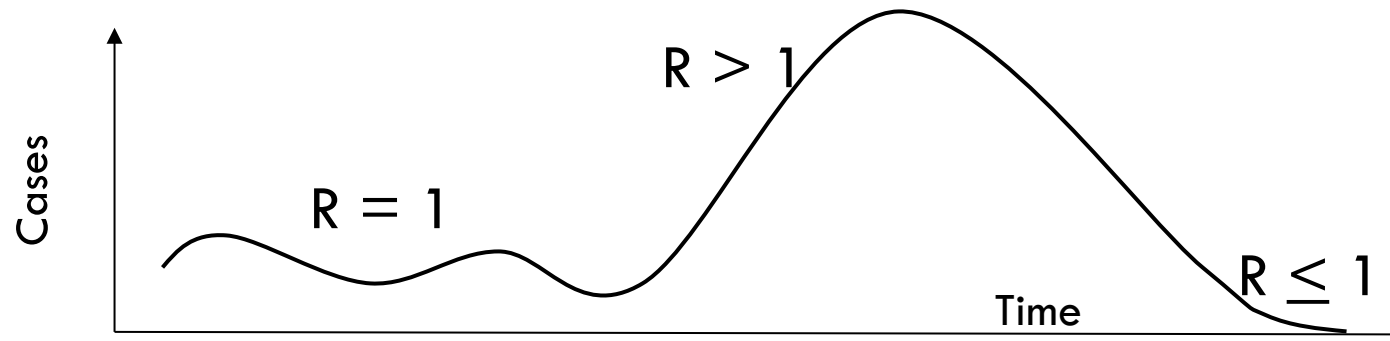


-  Carrier – no signs of disease, but colonised, will transmit
-  Case
-  Immune, cannot be colonised, will not transmit

TERMS IN INFECTIOUS DISEASE EPIDEMIOLOGY

Term	Disease incidence	Infectious agent	Time span (rolling average)
Sporadic disease	occasional cases occurring at irregular intervals	Caused by unrelated strains of the same infectious agent	Months- years
Endemic disease	persistent occurrence with a low to moderate level	Caused by unrelated strains of the same infectious agent	Months- years
Hyper-endemic disease	persistently high level of occurrence	Appearance of small clusters of disease in the population caused by highly related strains	Month- years
Epidemic outbreak	occurrence clearly in excess of the expected level for a given time period	Appearance of large clusters of disease in the population caused by highly related strains	Weeks to months
Pandemic	epidemic spread over several countries or continents, affecting a large number of people	Appearance of large clusters of disease in the population caused by highly related strains	Months to years

ENDEMIC-EPIDEMIC-PANDEMIC



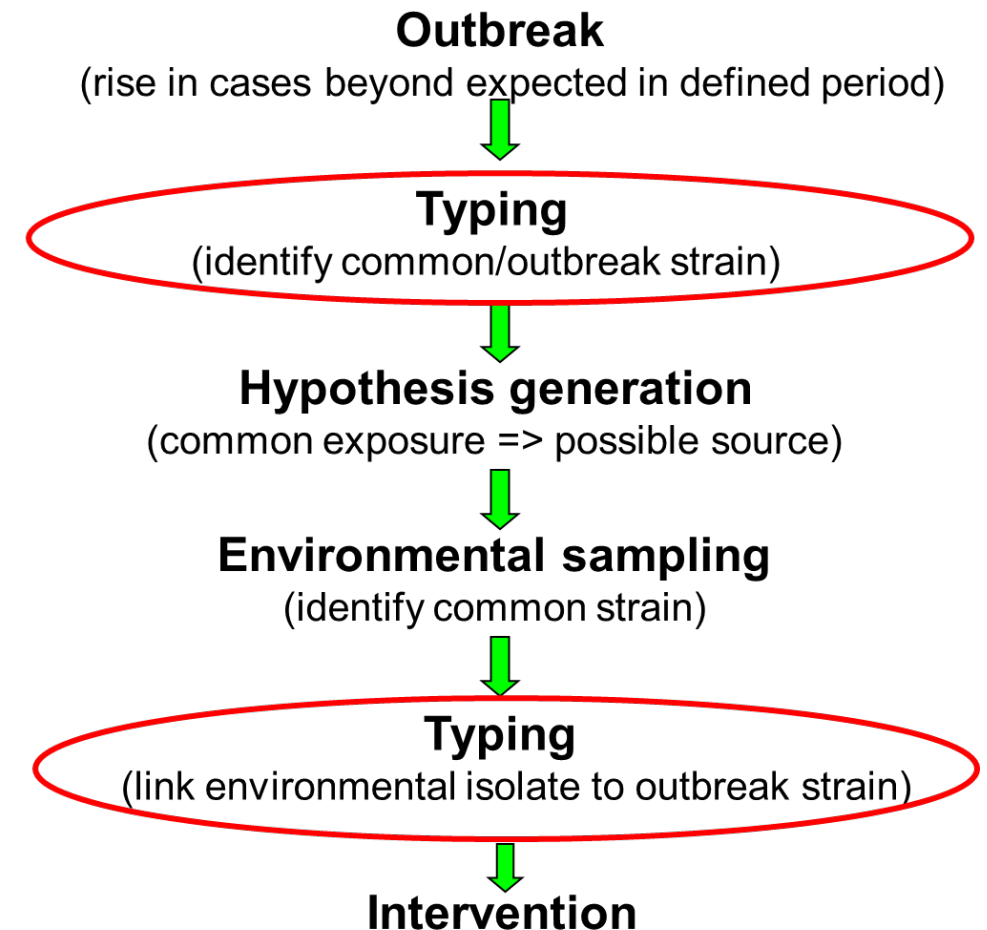
- ❖ **R =** *The basic reproductive number, R_0 ,*
 - ❖ the mean number of individuals directly infected by an *infectious* case through the total infectious period, when introduced to a susceptible population
- ❖ **Endemic ($R=1$)**
 - ❖ Transmission occur, but the number of cases remains constant
- ❖ **Epidemic ($R>1$)**
 - ❖ The number of cases increases
- ❖ **Pandemic ($R>1$)**
 - ❖ When epidemics occur at several continents – global epidemic

THE USE OF MOLECULAR TYPING METHODS IN OUTBREAK INVESTIGATIONS

The identification of a causative agent of an outbreak is confounded by

- Presence of other strains in the environment that do not cause disease
- Presence of hosts which carry virulent organisms but do not show signs of disease

Need to discriminate between these strains genotypically to determine which strains are endemic (not associated with the outbreak) or epidemic (associated with the outbreak)



MOLECULAR TYPING

Strengths

Do not need a live culture to make an identification

Therefore, more time efficient

Also very sensitive as amplification technology will detect very low numbers of organisms in a sample

Weaknesses

Antibiotyping is very important and requires culture

Need to be aware that detection of an organism does not always equate with causation of disease

Change in staff training and education

Initial cost of set up eg. Specialist machines

At least initially, this work will be done in central reference laboratories but as costs come down, it could be performed in a decentralised network which would improve information sharing/communication regarding outbreaks

TYPING QUESTIONS & SUITABLE METHODS

Questions	Suitable methods	Discriminatory power	Time span
Outbreak investigations Short term/local surveillance Control of hygiene measures	PCR-based methods Mostly on bacteria and fungi	Medium	Weeks-Months
Long term/global epidemiological studies Population genetics Analysis of population-based interventions eg. Vaccination	WGS- bacteria (MLST) WGS- virus (must have small genome) Not used on fungi yet	*High *High *Needs large central public databases	Weeks- Months Days (with automation-hours)

Discriminatory power- a method's ability to assign **a different type to two unrelated strains** sampled randomly from the population of a given species.

- should be calculated using a **test population** that includes **epidemiologically unrelated strains**

SEQUENCING BASED METHODS: MULTI-LOCUS SEQUENCE TYPING (MLST)

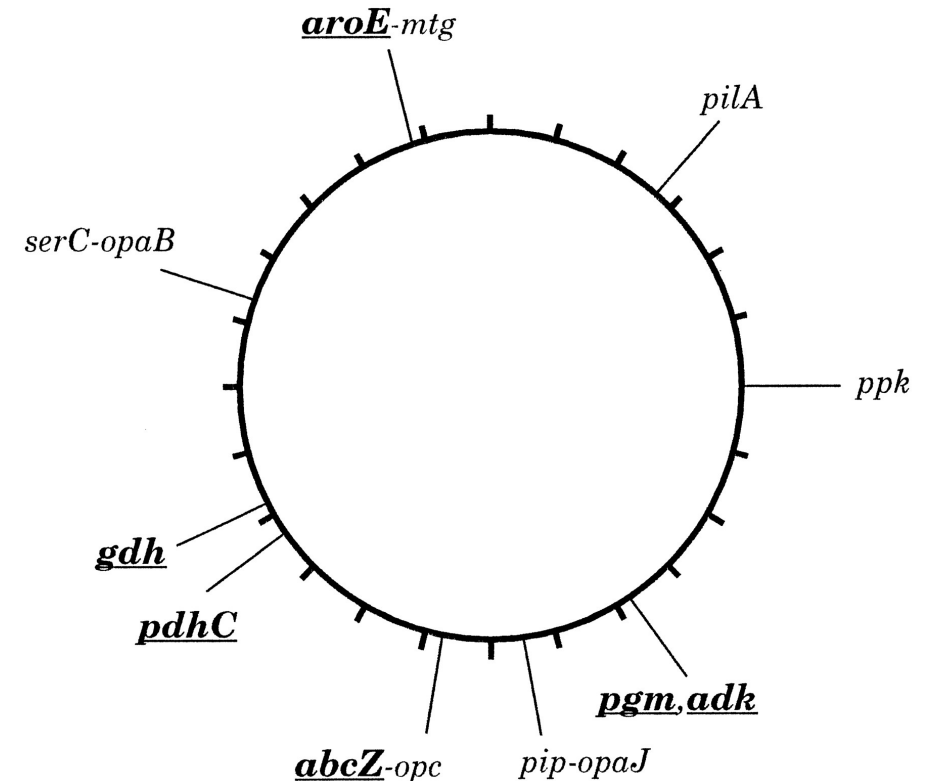
Sequencing is the most accurate method for assessing divergence and therefore relatedness

MLST involves the PCR amplification and sequencing of 7 housekeeping genes which are highly conserved relative to other genes

- Gene targets are under neutral selection and change is considered to be random
- These gene are evolving very slowly over time and represent the strain
- Fragment sizes 400-500bp

Convenient as can be efficiently sequenced using a single primer extension reaction in each direction

Each sequence is assigned an allele number based on single nucleotide polymorphisms (SNPs)



MLST ANALYSIS

Alleles are given a unique integer as a label

A combination of 7 alleles is given a unique integer = sequence type (ST)

STs can be clustered and analysed to create phylogenetic trees showing relatedness

An average of 30 alleles per locus allows about 20 billion genotypes to be resolved

ADVANTAGES: data is unambiguous, reproducible between labs, easily transferred and compared between labs, scalable and automated using high throughput sequencing

DISADVANTAGES: expensive, data bases are only available for some pathogens

Gene 1

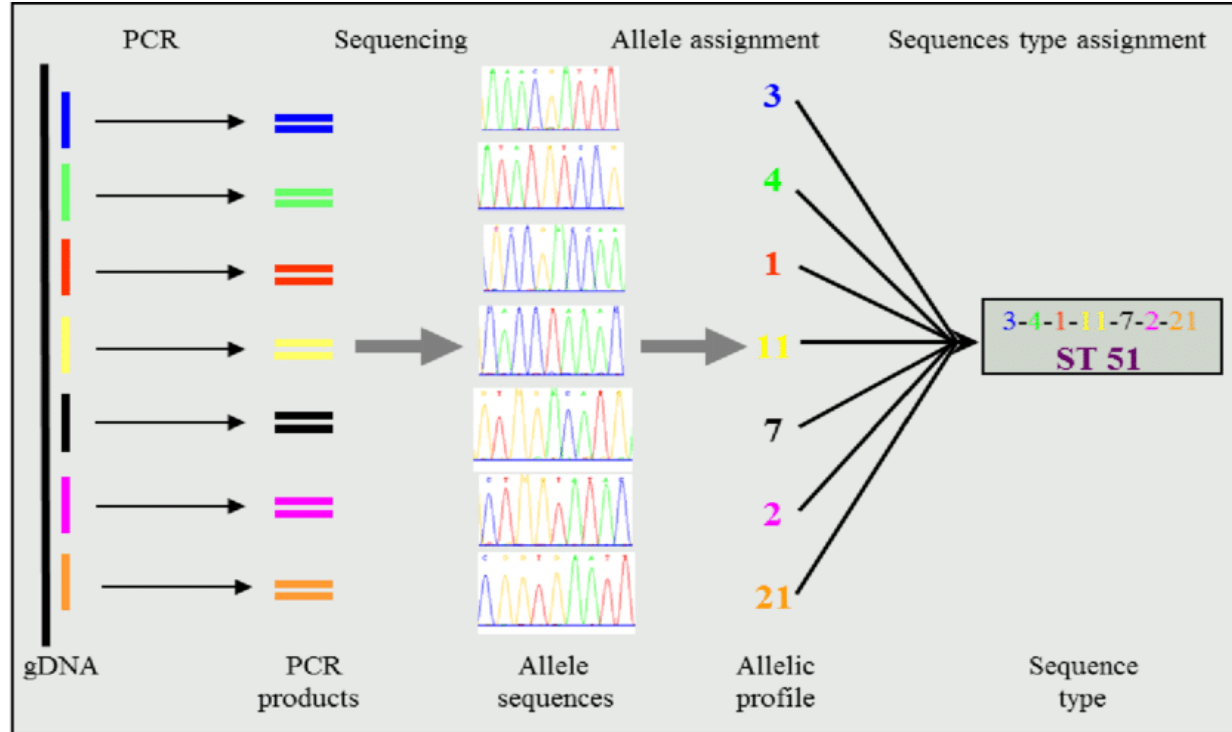
ATTCGGCGCGAAATCGCA allele 1

ITTCGGCGCGAAATCGCA allele 2

TTTGGGCGCGAAATCGCA allele 3

Etc.

NAMING SCHEME USING MLST



To understand the relationships of sequence Types (strains) to one another, we need to draw a phylogenetic tree.

WHAT IS A PHYLOGENETIC TREE

Phylogeny is a model of the relationships between organisms, genes, proteins or other structures based on common ancestry

Lamarck (1809) first used this to demonstrate taxonomic groups for all forms of life

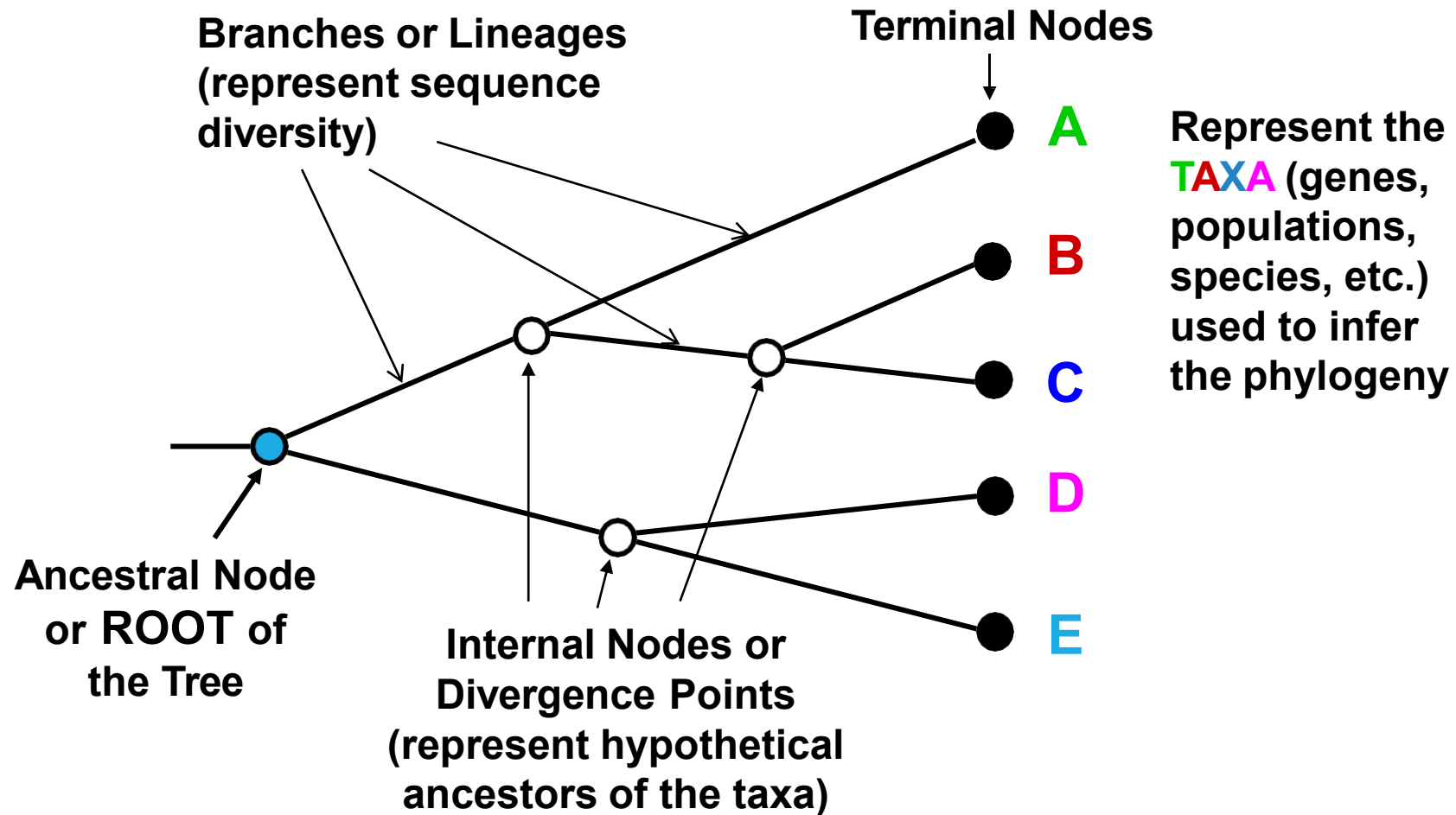
Darwin and others have continually refined the concepts

A phylogenic analysis only makes sense when the characters being compared have a high degree of homology

Four common uses are;

- Classification (taxonomy)
- Grouping of genes, proteins, and other molecular sequences including non-coding sequences
- Epidemiological investigations
- Analysis of parallel evolution between host and parasite

TERMINOLOGY FOR TREES



CONSTRUCTING A PHYLOGENETIC TREE FROM MLST DATA

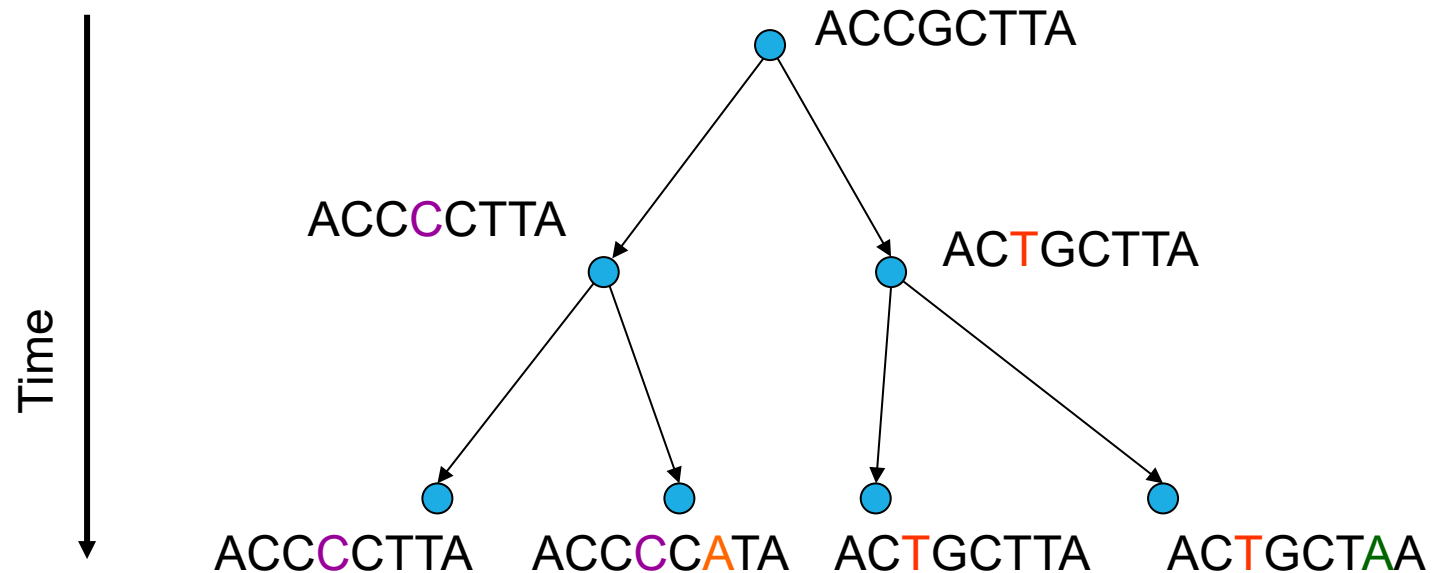
The sequence is known for the 7 housekeeping genes

These sequences are aligned and then clustered based on changes in the DNA sequence

Heirachical clustering algorithms are used to create a hypothetical tree which shows the number of changes that have occurred in the sequence

Hypothetical ancestors are included to create the clade

(see next lecture on phylogenetic trees)



We see the aligned modern day sequences

...ACCCCTTA...
...ACCCATA...
...ACTGCTTA...
...ACTGCTAA...

And want to recover the underlying evolutionary tree.

WHOLE GENOMIC MULTILOCUS SEQUENCE TYPING (WGMLST)

This scheme uses more than 7 housekeeping genes

As databases have become larger and algorithms faster, larger sets of genes can be used for typing

wgMLST is most useful for recombining populations eg.

- *N. gonorrhoeae* which is naturally transformable which leads to a population structure that is non-clonal (panmictic)
- This means that the rate of recombination is higher than mutation resulting in an index of association (I_A) = 0 (in other words, it is random)
- wgMLST for *N. gonorrhoeae* uses 1600 genes

See Christensen 11.1.5, 11.3.1, 11.3.2

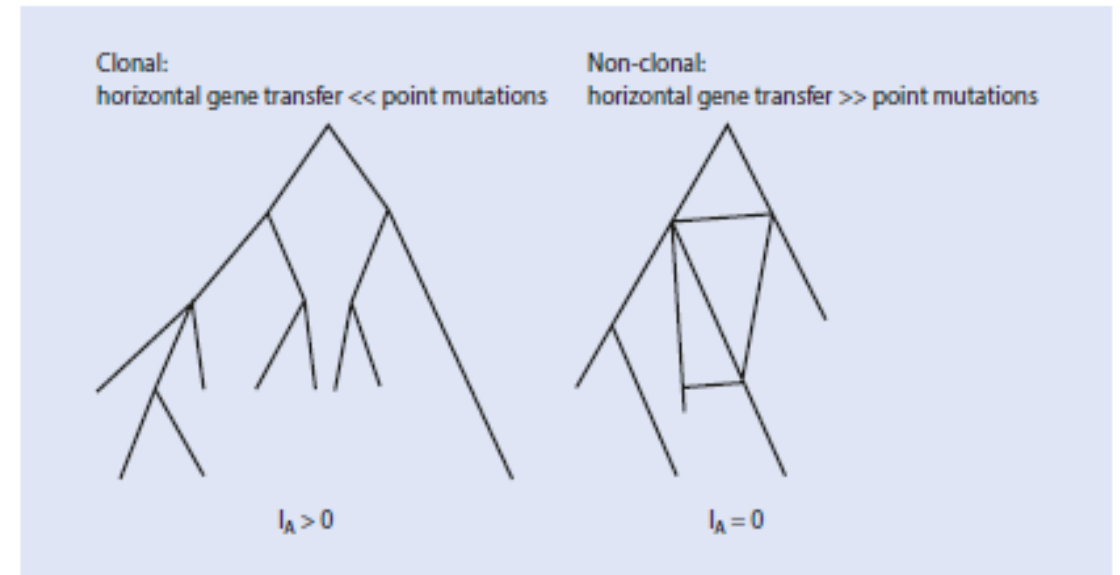


Fig. 11.4 Population structures are related to the ratio between recombination and point mutations. The expected variance of allele frequency, V_E should equal the observed variance V_O . If the populations have evolved like clones, the alleles will be identical or highly related within clones and very different between clones, and V_O will be higher than V_E . The index of association is calculated as $I_A = ((V_O/V_E) - 1)$

V_E = expected allele frequency

V_O = observed allele frequency

EXAMPLE: NEISSERIA MENINGITIDIS IN WESTERN AUSTRALIA

- Caused by *Neisseria meningitidis*
- A contagious bacterial disease that causes cerebrospinal **meningitis** and/or **septicaemia**
- Symptoms onset is sudden and death can follow within hours
- Spread by person-to-person contact through respiratory droplets
- Rates of disease are highest among infants, with a secondary peak during adolescence and early adulthood
- Currently, the rate of disease is 0.7 per 100,000 in Australia
- Mortality rate is 5-10% and ~15% of survivors suffer from neurologic disorders and amputations



THE BEXSERO® VACCINE FOR SEROGROUP B MENINGOCOCCAL DISEASE

A new vaccine against serogroup B licenced in 2013

Contains 4 meningococcal surface antigens

- **FHbp** (factor H-binding protein) – **fHbp-1**, fHbp-2, fHbp-3
- **NadA** (neisserial adhesin A) – **NadA-1**, **NadA-2/3**, NadA-4/5, NadA-6
- **PorA** (**P1.4 family**)
- **NHBA** (neisserial heparin-binding antigen)

Cross-protective immunity elicited against strains expressing one or more suitable antigens

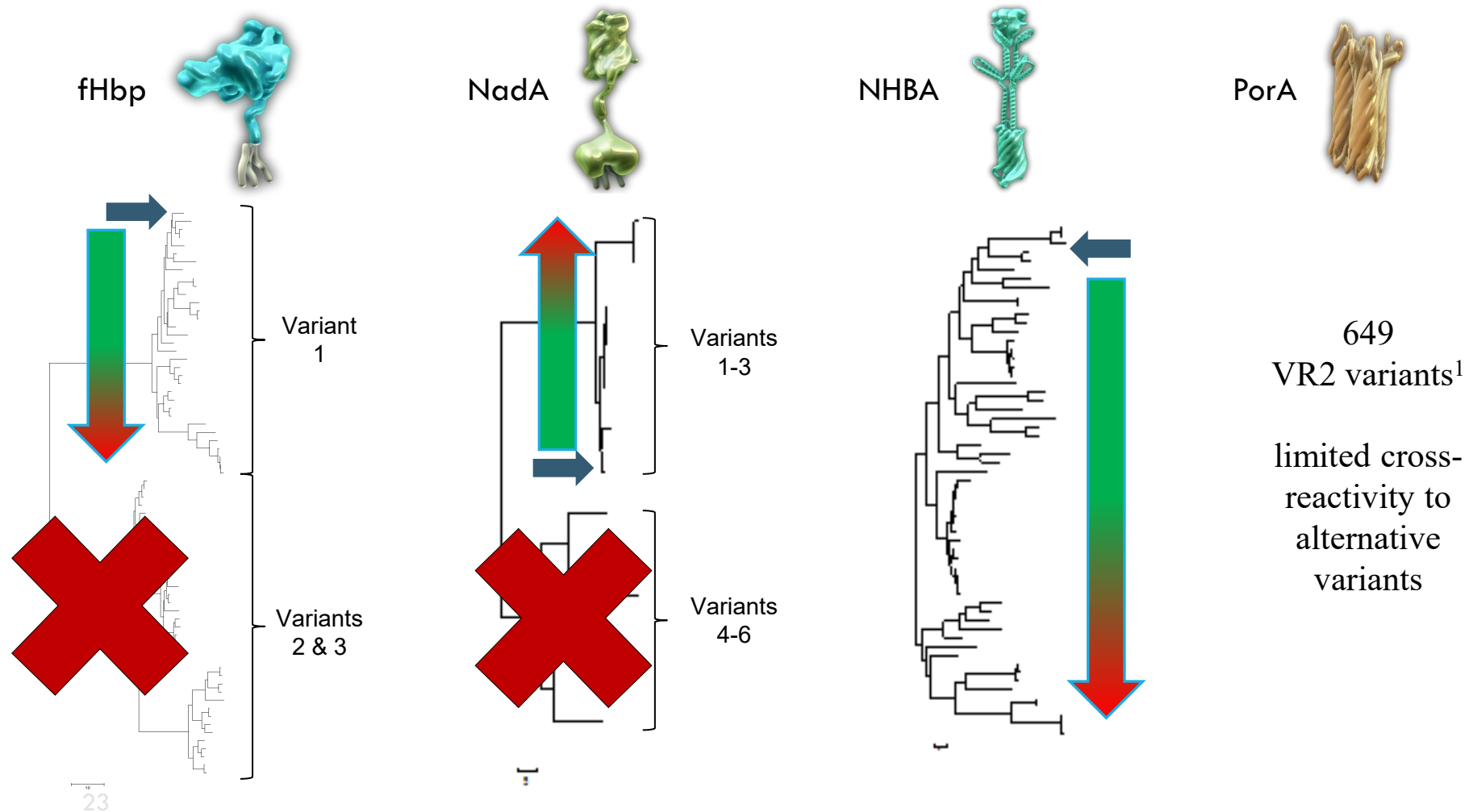
- Measured using an ELISA method termed MATS (Meningococcal Antigen Typing System)

Strain coverage varies by jurisdiction (65-91%)

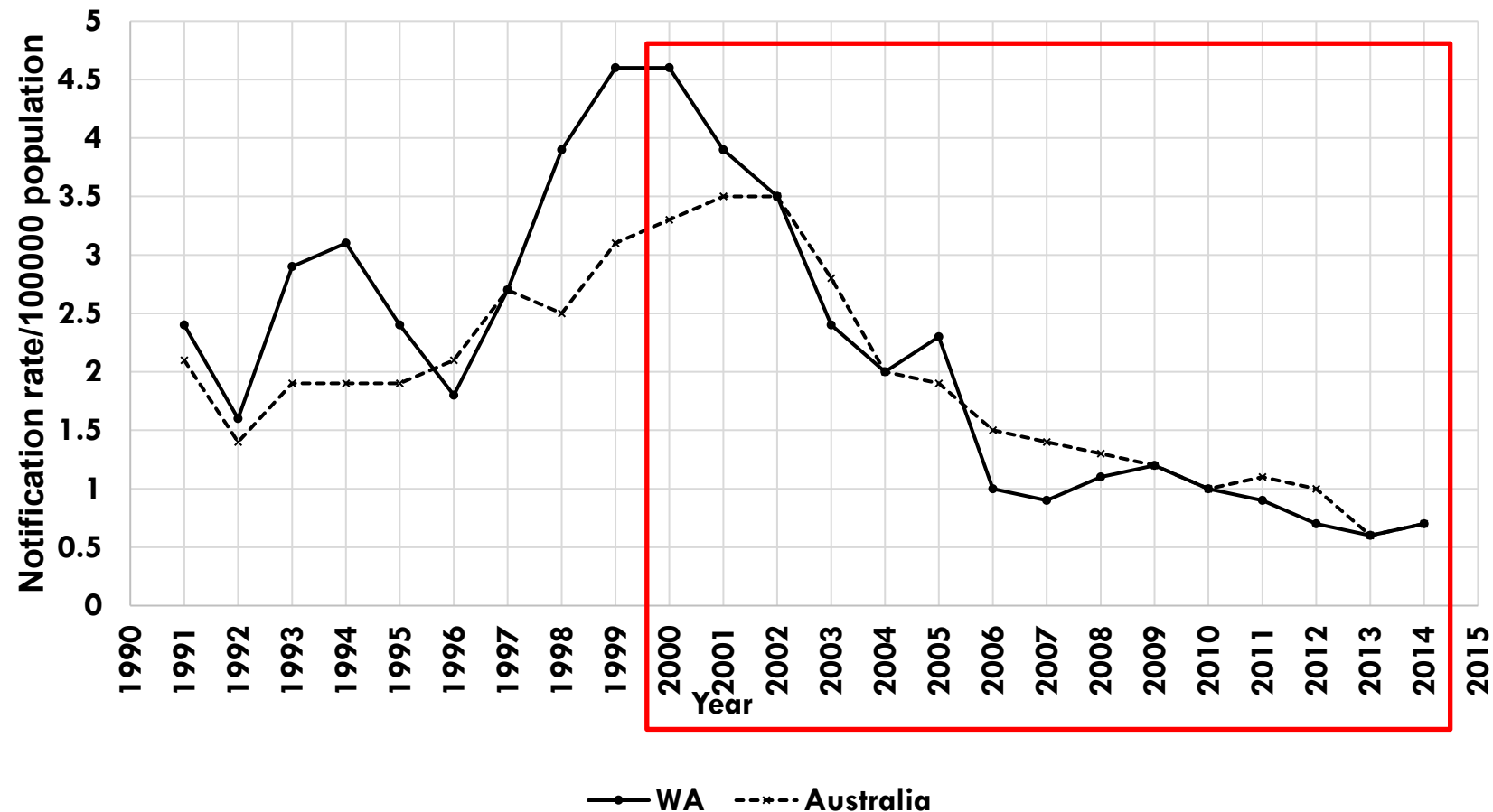
Prediction of **76% in Australia** using strains from 2006-2011

Alternative methods such as whole genome sequencing has been used

THE BEXSERO® ANTIGEN DIVERSITY AND CROSS REACTIVITY OF INDUCED ANTIBODY



NOTIFICATION RATE OF MENINGOCOCCAL DISEASE IN AUSTRALIA AND WESTERN AUSTRALIA 1991-2014

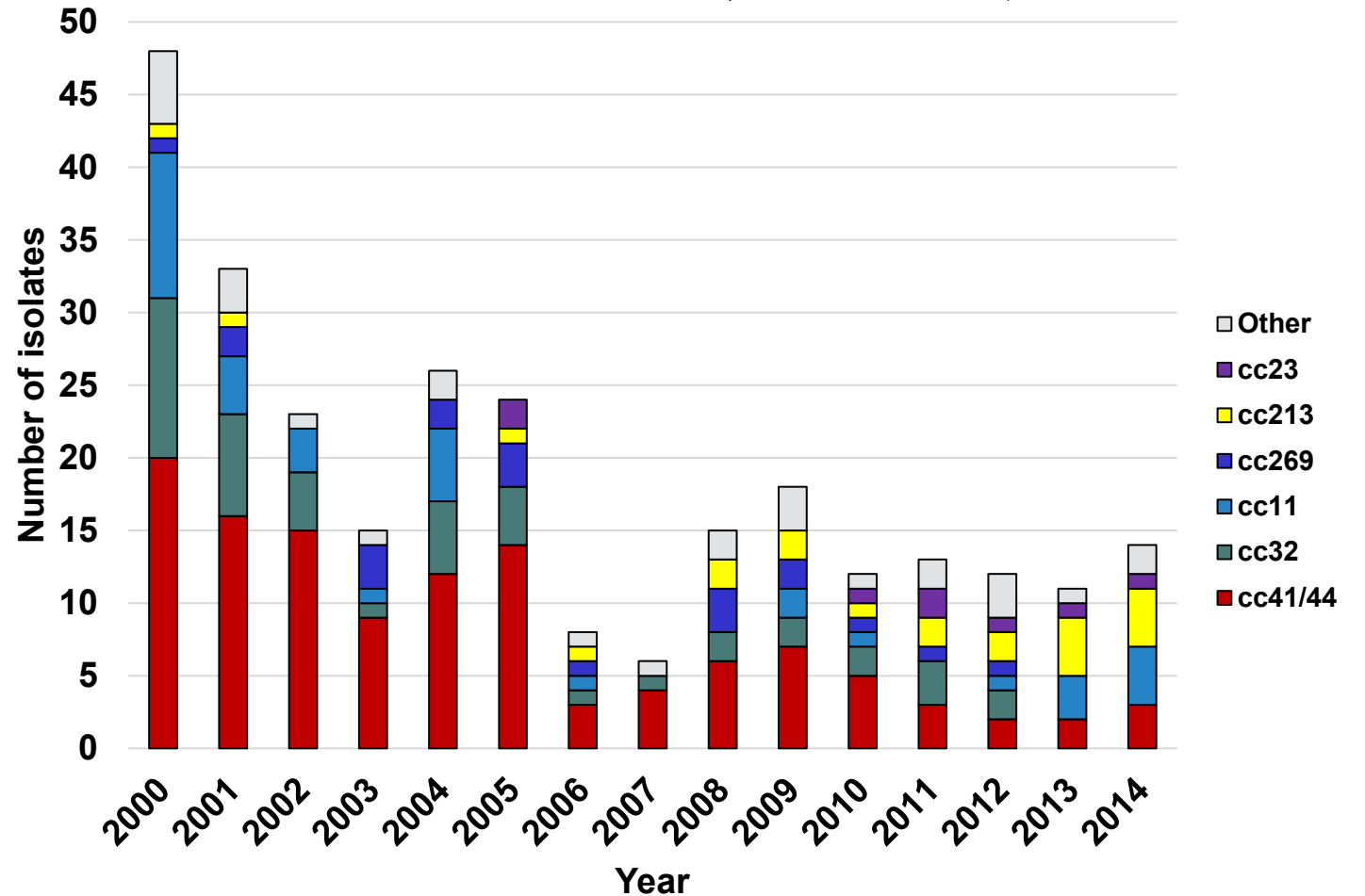


**68% of strains in
this time frame were
collected and
sequenced**

DISTRIBUTION OF CLONAL COMPLEXES IN WA (2000-14)

MLST is used to identify sequence types and group them into clusters called clonal complexes

The incidence of clonal clusters changes over time



PREVALENCE AND DIVERSITY OF VACCINE ANTIGENS AFFECTS VACCINE COVERAGE

	fHbp	NadA	NHBA	PorA
Presence of gene	100%	33%	100%	100%
Mutations**	3%	10%	-	-
Number of alleles	58	15	37	51
Number of peptides (subvariants)	55	11	34	49
Most common antigen	fHbp-1.19 (21%)	NadA-1.1 (48%)	NHBA-43 (20%)	P1.22,14-6 (19%)

Notes:

Each antigen in the vaccine is variable- look at the numbers of alleles! These alleles are under positive selection!

** represents Frameshift mutations or Insertion of a transposable element (*IS1301*)

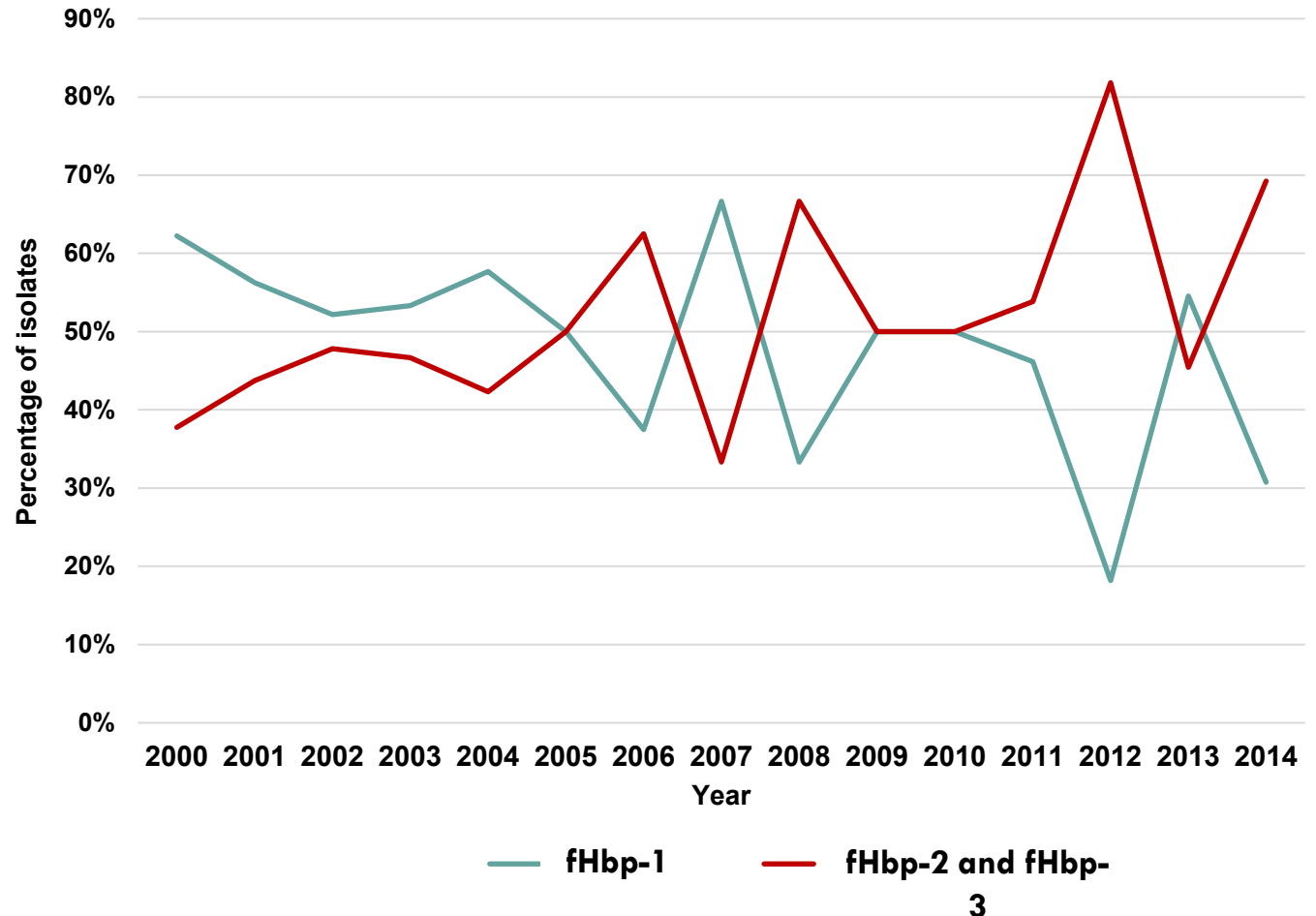
VACCINE ANTIGENS DRIFT OVER TIME IN THE ABSENCE OF A VACCINE

Previous studies from other jurisdictions showed that fHbp-1 was the predominant variant

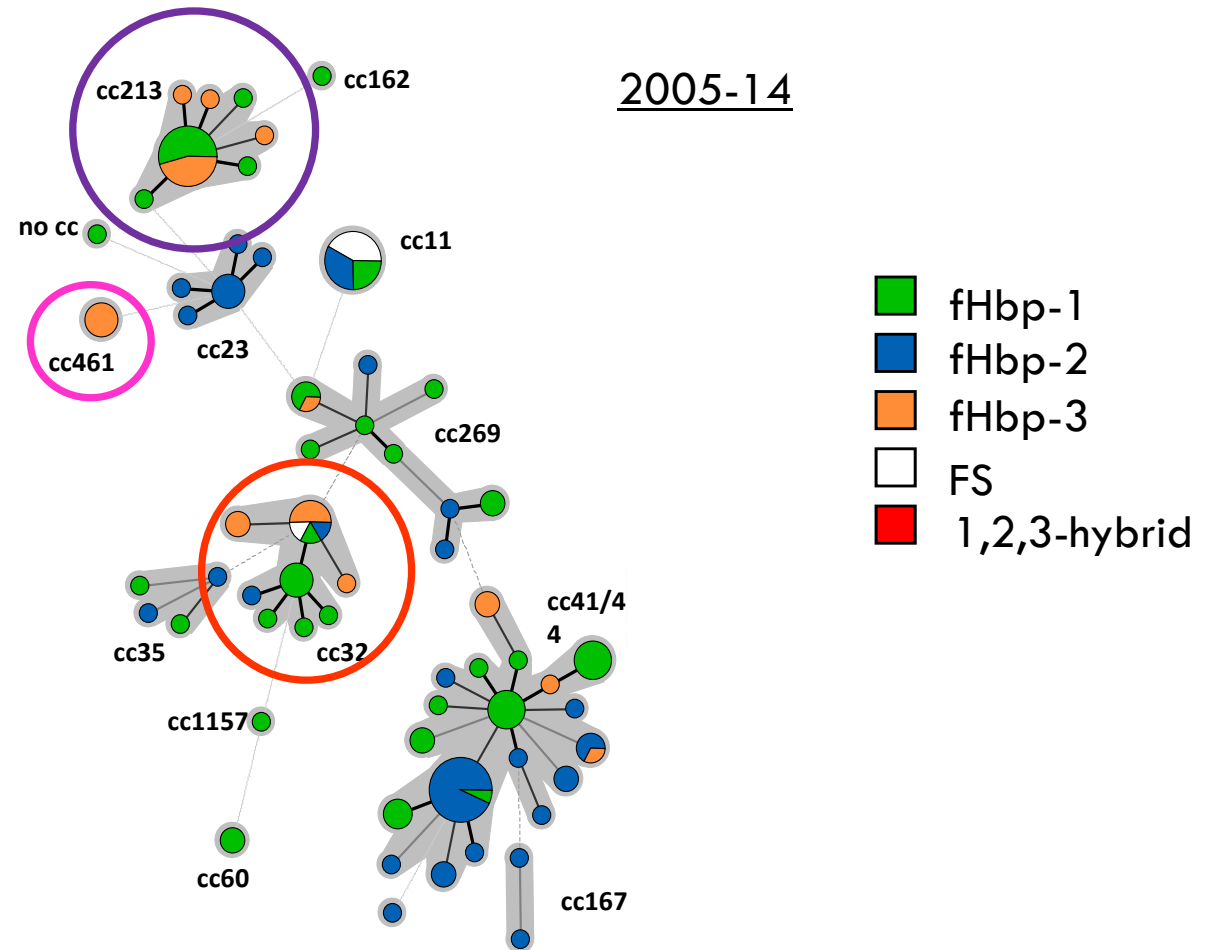
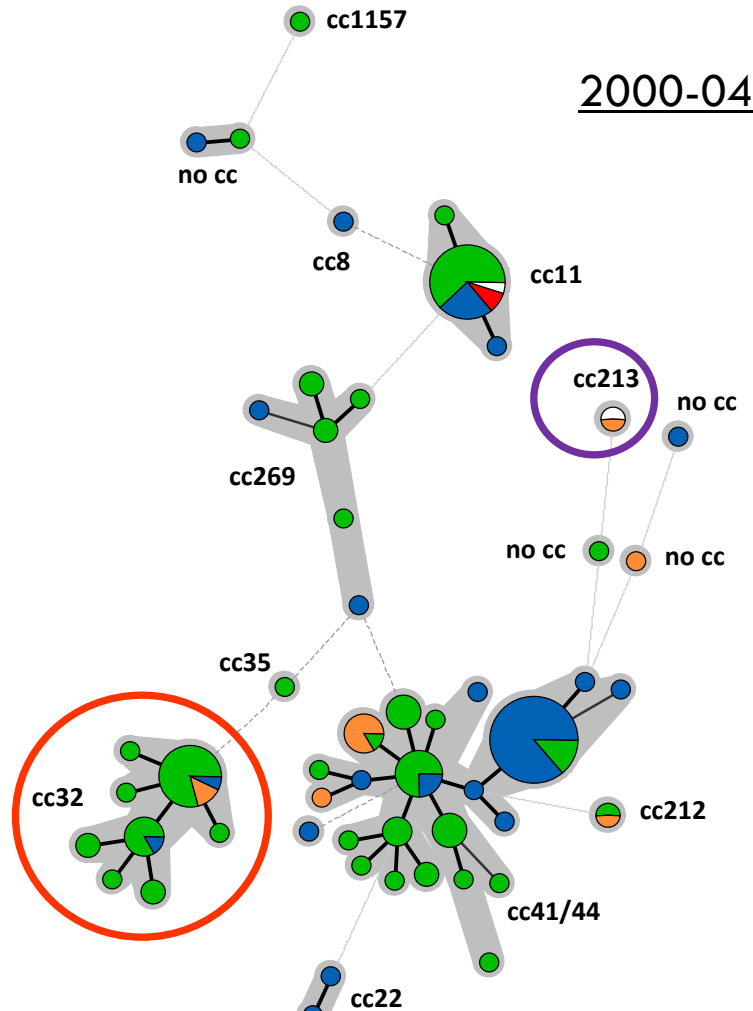
fHbp-1 predominated pre-2004

Post-2004, temporal shifts in prevalence of fHbp-2/3 was observed after every 2 years

Minimal Spanning Tree (MST) of cc was generated for the 2 periods to further investigate



MINIMUM SPANNING TREE OF MLST CLONAL COMPLEXES SHOWING ASSOCIATION WITH FHBP VARIANTS



SUMMARY

WGS and bioinformatics has revolutionized our understanding of the microbial world

Application to clinical microbiology and infectious disease public health will revolutionize infection control strategies both locally and globally

Large consortia are being organized in many countries to collect and manage this surveillance data

In the follow up labs, you will be taken through the process of drawing a simple phylogenetic tree using a set of genes.

You will be asked to complete a written report to be completed as you go through the tutorials which will be handed in after the last session.