Molecular epidemiology in current times

Burkhard Tümmler

Clinical Research Group, Clinic for Paediatric Pneumology, Allergology and Neonatology, Hannover Medical School, Hannover, Germany

Correspondence: Clinic for Paediatric Pneumology, Allergology and Neonatology, Hannover Medical School, Carl-Neuberg-Str. 1, D-30625 Hannover, Germany

Phone: 0049-511-5322920; FAX: 0049-511-5326723

Email: Tuemmler.burkhard@mh-hannover.de

Running title: Molecular epidemiology of infectious disease

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1462-2920.15238

Abstract

Motivated to find options for prevention or intervention, molecular epidemiology aims to identify the host and microbial factors that determine the transmission, manifestation and progression of infectious disease. The genotyping of cultivatable bacterial strains is performed by either anonymous fingerprinting techniques or sequence-based exploration of variable genomic sites. Multilocus sequence typing of housekeeping genes and allele profiling of the core genome have become standard techniques of bacterial strain typing that may be supplemented by whole genome sequencing to explore all single nucleotide variants and/or the composition of the accessory genome. Next, novel protocols to investigate host and microbiome based upon smart third generation sequencing technologies are being developed for an effective surveillance, rapid diagnosis and real-time tracking of infectious diseases.

Key words: cgMLST, genetic susceptibility to infection, metagenome, nanopore sequencing, optical mapping, pulsed-field gel electrophoresis, within-host microevolution

Introduction

The term 'molecular epidemiology' emerged during the 1970s and early 1980s in three separate areas: cancer epidemiology, environmental epidemiology, and infectious disease epidemiology (Schulte and Perera, 1993; Foxman and Riley, 2001). Within the context of this minireview, we will focus on the molecular epidemiology of bacterial infections in humans and livestock (Riley and Blanton, 2018). Like molecular taxonomy, phylogeny, or population genetics, molecular epidemiology applies biochemical techniques to identify microbes, but the objectives are different. Motivated to find options for prevention or intervention, molecular epidemiology aims to identify the factors that determine the transmission, manifestation and progression of infectious disease. Based on the methodological progress of molecular biology during the last thirty years, molecular epidemiology has primarily applied nucleic-acid based techniques to monitor outbreaks, infection chains and patterns of disease transmission in populations. The item to be tracked could be a taxon, a clone or subclonal lineage, a strain, mobile genetic elements such as genome islands, transposons, phage or plasmids or individual genes that encode e.g. virulence or antimicrobial resistance determinants. Thereby the molecular typing scheme should provide sufficient discriminatory power, be reproducible among different laboratories, and be easily performed and standardized (Wang et al., 2015). Ideally the data should be portable, i.e. they should be able to be digitized, to be stored in a publicly accessible database and be easily transmitted between laboratories. The time lag between the discovery of a novel feature or mechanism, development or refinement of a technique and its application to infection epidemiology has

thereby become shorter and shorter, namely because the very same teams develop methodology and protocols and apply them to both basic research and real-life problems.

Molecular typing exploits genetic diversity

Molecular typing protocols assess either the presence or the diversity of genetic elements (Sabat et al., 2013; Riley, 2018) (Table 1). Genetic diversity within a bacterial taxon is caused by single nucleotide variants (SNVs), the variable composition of the accessory genome and the spatial distribution of repeats and oligonucleotides within the genome. Robust and reliable exploitation of diversity is performed by sequencing, restriction endonuclease digestion or detection of singular genetic elements. Dichotomous yes/no approaches typically examine genetic markers of phenotypically relevant traits such as pathogenicity factors or antimicrobial resistance genes. Restriction fragment patterns are a measure of the physical genome organization. If the bacterial genome is cleaved with a restriction endonuclease, the distribution of genomic restriction fragment size represents the spatial distribution of recognition sites in the genome (Grothues and Tümmler, 1991). The physical distance between sites primarily varies between strains due to insertions, deletions or large rearrangements. Variations of restriction fragment patterns mostly arise from the incorporation or release of elements of the accessory genome and are thus a measure of genome evolution in the time frame from days to years. In contrast, de novo point mutations in the core genome occur orders of magnitude less frequently than changes in the accessory genome. SNVs are conserved in the core genome and represent ancient landmarks of the separation of lineages. The sequencing of multiple loci of the core genome accordingly

identifies SNVs at defined genome positions that can be used to designate the sequence types of a taxon (Maiden et al., 1998) that are stable within a period of years to decades.

Bacterial strain typing: Anonymous fingerprinting techniques

Anonymous PCR-based fingerprinting techniques either explore copy number variations of repeats or perform PCR at relatively low stringency or with low selectivity primers (Table 1). PCR products are separated by polyacrylamide, agarose or capillary electrophoresis generating a fingerprint of bands. The methods are globally applicable to bacteria, do not require specific equipment, are inexpensive and are easy to set up (Table 2). Comparative genotyping by PCR-based fingerprinting is feasible if strains are processed in parallel under standardized conditions at a single site. However, the methods are poorly portable and cannot assess the clonal relatedness of strains.

Macrorestriction fragment pattern analysis is a genotyping method that is globally applicable to all bacteria and hence has been and still is the reference method for strain typing in bacteriology (Table 2). Agarose-embedded chromosomal DNA is cleaved with a rare-cutting restriction endonuclease and the generated 20 – 70 fragments are then separated by pulsed-field gel electrophoresis (PFGE) (Birren and Lai, 1996). Since the recognition sites of most restriction endonucleases are randomly distributed in the chromosome, fragment patterns represent unbiased genomic fingerprints and can thus be exploited to assign clonal relationships between strains (Grothues and Tümmler, 1991). A single enzyme restriction pattern is less discriminatory than genome sequencing (Martak et al., 2020), but PFGE analysis can be refined by the addition of more single and double digests with further rare-

cutting restriction endonucleases. Agarose-embedded high-molecular weight DNA is stable when stored in Tris-EDTA buffer at 4°C. In our hands 20-year old agarose plugs still produced identical fragment fingerprints. However, the method is slow and demands manual skills and extensive experience.

Minor differences between fragment band patterns are typical for members of the same clonal complex. These shifts are most often not caused by de novo point mutations in the restriction enzyme recognition site but represent genome rearrangements by insertions, deletions or inversions or a strain-specific different composition of the flexible accessory genome (Römling et al., 1997). Point source outbreak strains typically produce indistinguishable fragment patterns, but during the time course of an outbreak horizontal gene transfer within the clonal community may give rise to subtle shifts of the genome fingerprint.

Gel-to-gel comparisons of fragment patterns are not easily portable and require rigorous standardization of PFGE conditions, controls and software. The PulseNet International network (Table 3) has successfully met these challenges (Gerner-Smidt et al., 2006). PulseNet is the first foodborne disease surveillance system in the world based on genotypic stratification of bacterial pathogens. The PFGE database allows the intra- or inter-laboratory comparison of PFGE patterns of strains from different time and geographic locations. According to an economic impact study nearly 270,000 illnesses from *Salmonella*, 9,500 illnesses from *Escherichia coli*, and 60 illnesses from *Listeria monocytogenes* are avoided annually because of this PFGE-based surveillance system (Scharff et al., 2016).

Combinatorial PFGE of large restriction fragments can be used to assemble low-resolution physical genome maps of bacterial strains (Birren and Lai, 1996) that will highlight inter- and intraclonal structural variations like inversions, deletions or insertions. This goal can nowadays be more conveniently achieved by optical mapping (Bocklandt et al., 2019; Rice and Green, 2019). Starting with high molecular weight bacterial DNA, fluorescent labels are attached to a 6 bp sequence motif. The fluorescently labeled DNA fragments are electrophoretically fed through a nanochannel array and imaged to determine the sizes of the molecules and the locations of fluorescent labels. This information is assembled into a genome map, which can then be used to scaffold contigs or to discover structural variations starting at 500 bp.

All anonymous fingerprinting techniques detect unique strains and may therefore still be used in the future to trace local outbreaks, but the generated band patterns (with the exception of PFGE) do not resolve clonal relatedness and are poorly portable because they index variation that is difficult to compare among laboratories. In contrast, sequence-based typing methods generate portable data that are suitable for storage in databases.

Bacterial strain typing: Sequence-based typing methods

Bacterial isolates of taxa of interest may be genotyped by multiplex hybridization onto arrays (Monecke and Ehricht, 2005; Wiehlmann et al., 2007; Dunne et al., 2018) (Tables 1, 2). Such a low resolution GeneChip scans single nucleotide variants of the core genome and variable elements of the accessory genome by hybridization. Targets are amplified directly from bacterial colony DNA by cycles of multiplex primer extension reactions, whereby the nascent

strands are randomly labelled by incorporation of biotin-16-dUTP. The multiplex amplificate is hybridized under high stringency with the oligonucleotide microarray of target sequences. Hybridisation signals visualized with streptavidin conjugate are converted to the electronically portable multilocus genotype (Wiehlmann et al., 2015).

Since its inception in 1998 multilocus sequence typing (MLST) has been and probably for many years to come will remain the standard method for the genotyping of strains (Maiden et al., 1998) (Tables 1-3). MLST utilizes nucleotide sequence data of about 500-bp long internal fragments of six to nine housekeeping genes. For each gene fragment, isolates with identical sequences are assigned the same allele number. For each strain, the combination of alleles at each locus defines its sequence type (ST). MLST is an electronically portable, universal and definitive method for genotyping. The PubMLST database (Table 3) hosts MLST schemes for more than a hundred species and stores MLST data of more than 700,000 isolates. In the past the target genes of one (or more) strains were amplified by (multiplex) PCR and the (barcoded) amplicons were sequenced. With next generation sequencing becoming a more and more affordable low-cost method, the MLST can nowadays be easily extracted in silico from whole-genome sequencing (WGS) data (Kimura, 2018) (Table 3). In other words, the acquisition of ST data is a welcome by-product of any WGS project.

Genome-wide gene-by-gene allele calling of hundreds/thousands of genes (core genome MLST, cgMLST) (Mellmann et al., 2011) is an obvious extension of the traditional MLST concept (Tables 1-3). cgMLST retains the intuitive nature of traditional MLST but offers much greater resolution by utilizing significantly larger portions of the genome. cgMLST is species-

specific, can be standardized and is electronically portable. A set of conserved core genes is extracted from a comprehensive dataset of (hopefully high-quality) genome sequences. Then a sublineage threshold of allelic differences needs to be defined that creates clusters nearly identical to traditional MLST types, providing backwards compatibility to new cgMLST classifications (Liang et al., 2020). A ring trial demonstrated the high reproducibility of cgMLST allele profiles, only 3 of 183,927 (0.0016%) cgMLST allele calls were wrong (Mellmann et al., 2017).

cgMLST is a timely field. For example, within one month in 2020 cgMLST schemes have been published for *Campylobacter jejuni* (Hsu et al., 2020), *Leptospira spp.* (Grillová and Picardeau, 2020), *Pseudomonas aeruginosa* (de Sales et al., 2020; Martak et al., 2020; Stanton et al., 2020), *Streptococcus mutans* (Liu et al., 2020) and *Vibrio cholera* (Liang et al., 2020). cgMLST schemes consist of a fixed set of conserved genome-wide genes. The three cgMLST schemes for *P. aeruginosa* consist of 3,831 (Martak et al., 2020); 2,653 (de Sales et al., 2020) or 4,400 core genes (Stanton et al., 2020) which probably will give rise to confusion in the next years when *P. aeruginosa* genotypes will be published that are based on different typing schemes. A cgMLST scheme is fixed and agreed upon number of genes for each species or group of closely related species. A commercial cgMLST.org Nomenclature Server (Table 3) is available for 16 species. The *P. aeruginosa* case tells us that we need consensus schemes for many more species to avoid chaos in the future.

cgMLST performs allele profiling and provides a detailed overview of the inter-clonal diversity of a bacterial species. Alternatively, WGS could explore all single nucleotide variants

and/or the composition of the accessory genome (Schürch et al., 2018). These more comprehensive approaches are useful to trace intra-clonal diversity in order to understand the microevolution of a clone or to follow the acquisition and loss of traits of interest such as virulence or antimicrobial resistance determinants.

WGS of a bacterial strain by second-generation sequencing-by-synthesis platforms generates an avalanche of short reads that may be aligned to a high-quality reference genome to perform allele profiling. The assembly of short reads to large contigs or a complete genome, however, is hampered by gene duplications and copy number variations of long perfect or imperfect repeats where contigs compiled from short reads typically terminate. Moreover structural variants such as large insertions or inversions are often not detected by short read sequencing on an Illumina platform because breakpoints typically reside in a repeat region. Third-generation Single-Molecule Real-Time sequencing (SMRT) (Ardui et al., 2018) or Nanopore sequencing (Runtuwene et al., 2019) generate reads that are long enough to span any repeat known in bacterial genomes. Since both the SMRT and Nanopore technologies are still compromised by a comparably high sequencing error particularly in homo-oligomers, a hybrid approach of mapping short Illumina reads onto a scaffold of long reads is currently the most rapid and economic approach to assemble a complete high-quality genomic sequence of a strain of interest. The third-generation sequencing technologies moreover have the advantage that structural variants and breakpoints are resolved by the base and that the clonal assignment of hypermutable strains carrying multiple de novo

mutations is feasible from the inspection of the long-range organization of the core and accessory genome.

Real-world applications of strain typing by genome sequencing

In 2020 hundreds of papers have been monthly published which apply molecular techniques to resolve the epidemiology of infectious diseases. To illustrate the increasing impact of whole genome sequencing in the surveillance of infectious disease, I arbitrarily picked the themes of recently published papers of clinically important pathogens co-authored by infectious disease specialists and genome researchers from the Wellcome Sanger Institute:

- Definition of a quality metrics for the sequencing of methicillin-resistant *Staphylococcus aureus* (MRSA) in clinical practice (Raven et al., 2020)
- Assessment of the feasibility and utility of adding genomics to epidemiological surveillance of MRSA associated with bloodstream infection to determine the national population structure of MRSA, contextualise previous outbreaks, and detect high-risk lineages (Toleman et al., 2019)
- Rapid sequencing of MRSA direct from clinical plates in a routine microbiology laboratory (Blane et al., 2019)
- Evaluation of a fully automated bioinformatics system for analysis of MRSA genomes and detection of outbreaks (Brown et al., 2019)
- Reconstruction of transmission trees for MRSA in a high-transmission hospital setting including tracing of an individual at different anatomical sites (Hall et al., 2019)

- Longitudinal genomic surveillance of community- and hospital-associated MRSA lineages in the UK (Coll et al., 2017)
- Evolutionary history and genomic epidemiology of the origins and global spread of a multidrug-resistant, community-associated *S. aureus* lineage from the Indian subcontinent, the Bengal Bay clone (ST772) (Steinig et al., 2019)
- Emergence of a new *Streptococcus pyogenes* lineage in England associated with an increased incidence of scarlet fever and invasive disease (Lynskey et al., 2019)
- Longitudinal population dynamics and antimicrobial resistance mechanisms in *Klebsiella pneumoniae* and *Enterobacter cloacae* between 2007 and 2012 in a major UK hospital (Ellington et al., 2019)
- Population genomics of *K. pneumoniae* in patients, livestock and environment in the East of England (Ludden et al., 2020)
- Prevalence of vancomycin-resistant *Enterococcus faecium* lineages in humans, livestock, hospitals and municipal wastewater treatment plants (Gouliouris et al., 2018; 2019)
- Genomic history of the *Vibrio cholera* epidemics in Africa and the Americas (Domman et al., 2017; Weill et al., 2017)

These examples demonstrate the thematic versatility of applications of WGS to hospital infection control, epidemiology and public health and the synergy gains of collaborations between genome centers, microbiologists and infection disease specialists. The phylogenomic studies mentioned above stand out by the high-quality metadata of their source. In contrast metadata are typically scarce for completely sequenced bacterial genomes in the databases which unfortunately limits the informative value of phylogenomic intraspecies comparisons.

Within-host bacterial microevolution

Genome sequencing of sequential bacterial isolates resolves the within-host microevolution of a pathogen during the infection. The genomic history of chronic infections in humans has so far mainly been studied in patients with cystic fibrosis (Elborn, 2016). Unlike *Helicobacter pylori* (Moodley et al., 2012) or *Mycobacterium tuberculosis* (Achtman, 2016) which were already prevalent in man in prehistoric times the airway infections in cystic fibrosis emerged just 50 years ago. The bacterial microevolution in cystic fibrosis lungs has been monitored by WGS for *Burkholderia dolosa* (Lieberman et al., 2011)*, Mycobacterum abscessus* (Bryant et al., 2016) and *Pseudomonas aeruginosa* (Cramer et al., 2011; Yang et al., 2011; Marvig et al., 2015; Klockgether et al., 2018). The studies consistently observed convergent molecular evolution in multiple individuals. Non-neutral mutations predominantly emerge in genes relevant for protection against and communication with signals from the lung environment,

but these mutations rarely fix in a patient's pathogen population--instead, diversifying lineages coexist for many years. Most *P. aeruginosa* lineages found in the aquatic inanimate environment can also colonize and persist in cystic fibrosis airways (Wiehlmann et al., 2015), but the infections with *M. abscessus* are caused by recently emerged dominant circulating clones that have spread globally and are associated with worse clinical outcomes (Bryant et al., 2016).

Susceptibility of the human host to infection

The molecular epidemiology of infectious disease is much more than strain typing to monitor the spread of pathogens or traits at the local, domestic or international level. A molecular epidemiology study will also search for biomarkers that reflect host susceptibility to infection. The inter-individual variability of clinical outcome during infection can be accounted for by the variability of the microbes themselves, by the variability of the hosts, and by factors with no effect on the intrinsic capacities of the host and microbe, such as the numbers of invading microbes and their route of infection and global features such as climate, living conditions, socioeconomic status and the health care system (Casanova, 2015; Casanova and Abel, 2013).

Only very few severe infectious diseases are known to segregate as Mendelian traits (Casanova and Abel, 2020a), but numerous infections are monogenic showing extensive

genetic heterogeneity (Casanova et al., 2020b) (Table 4). For example, Mendelian susceptibility to infection with mycobacteria is caused by mutations in genes of the interferon gamma signaling pathway such as *Tyk2* (Table 4). Autosomal recessive deficiency of non-receptor tyrosine-protein kinase Tyk2 is a very rare inborn error of immunity predisposing to primary tuberculosis (TB) in otherwise healthy patients. However, a catalytically inactive *Tyk2* missense variant, P1104A is common in human populations. Homozygosity for P1104A was enriched in patients with TB who participate in the UK biobank and in cohorts of patients with TB from non-European countries in which TB is endemic (Boisson-Dupuis et al., 2018; Kerner et al., 2019). This currently most striking example demonstrates that the search for the rare inborn errors of infectious disease has identified candidate genes that may carry the genetic modifiers of susceptibility to infectious disease in the global population.

What next – the challenge of the microbiome

The genotyping of single cultivatable strains from one source may be sufficient for the surveillance of a local outbreak, but in real life patients and livestock may carry numerous lineages of a species. Within-host diversity of a cultivatable pathogen can be represented by multi-locus sequence types whereby the allele distribution and their proportions are derived from WGS data (Gan et al., 2019). Alternatively one may determine the clonal composition and intra-clonal variation of the species of interest from the hypergeometric frequency distribution of species-specific single nucleotide variants extracted from shotgun

metagenome sequencing data sets yielding quantitative estimates of the proportion of clones and subclonal variants (Moran Losada et al., 2016).

Metagenomics provides the next layer of information coming closer to the real world of pathogens thriving within polymicrobial communities. WGS metagenomics is taxonomically agnostic and exploits the cultivatable and uncultivatable microbial diversity of fungi, archaea, eubacteria and DNA viruses (Quince et al., 2017). Untargeted shotgun sequencing has moreover the principal advantage to avoid any PCR-generated amplification biases and any skews introduced by divergent gene copy numbers. Hence this approach should be optimal for any microbiome study of the molecular epidemiology of infectious diseases. Deep sequencing is required to determine the proportion of clones and subclonal variants. This approach is only applicable to the dominant species in the habitat of interest. The microbes of the rare biosphere are identified from the distribution of pairwise genomic distances for all reads aligned to one reference genome, but the number of clones cannot be determined due to the inherently low number of reads. According to calculus and our experience twenty to thirty reads are sufficient to identify a rare species in a metagenome with a statistical error of less than 1%.

High-throughput sequencing on Illumina platforms is widely used for metagenome sequencing, but it requires the sequencing run to be complete before analysis can begin. However, microbial epidemiology investigations of outbreaks or epidemics often require low turnaround times that should be even shorter than the 2 – 7 days required for classical culturing and susceptibility testing. Nanopore sequencing has the advantage of rapid-library

preparation and real-time data acquisition. Pilot studies have demonstrated that the diagnosis of lower respiratory tract infections by nanopore sequencing is indeed feasible in six hours from sample to result (Pendleton et al., 2017; Charalampous et al., 2019). The small size of the portable equipment facilitates sequencing at the sampling site even in remote environments (Quick et al., 2017; Giovanetti et al., 2020). Provided that there is sufficient time for base calling and bioinformatic work-up, the generated long reads moreover provide reliable information about the presence of rare species and the abundance of taxa, clones, clonal variants and mobile genetic elements in real-life proportions (Nicholls et al., 2019).

Molecular epidemiology is a field which heavily relies on the exchange of data between laboratories. Metagenome-based discoveries of yet uncultivated microbes need to be properly communicated. The International Code of Nomenclature of Prokaryotes only recognizes cultures as 'type material', thereby preventing the naming of uncultivated organisms of which the genome has been assembled from sequencing data. However, we urgently need consistent rules for nomenclature of uncultivated taxa in order to provide day-to-day communication between laboratories during outbreaks and epidemics. Hopefully the on-going initiatives (Bowers et al., 2017; Konstantinidis et al., 2020; Murray et al., 2020) will be successful that DNA sequences are accepted as type material because it will facilitate to set standards for the exchange of metagenomics data.

Conclusion

Starting with poorly portable PCR-based genomic fingerprinting methods in the late 1980s the wet-lab and in silico methods have since then been continuously improved. Protocols were standardized, the generated data became electronically portable and genotypes and sequences were being stored in databases. There is only very limited evidence for co-speciation of mammalian hosts and microbes (Groussin et al., 2020), but the sequencing of host genome and microbiomes will nevertheless transform the holobiome concept into reality in near future (Theis et al., 2016; Klimovich and Bosch, 2018). The surveillance, rapid diagnosis and real-time tracking of infectious diseases has become feasible by novel protocols based upon smart third generation sequencing technologies. As stated some time ago by Jennifer Gardy and Nicholas Loman (2018), “Coupling genomic diagnostics and epidemiology to innovative digital disease detection platforms raises the possibility of an open, global, real-time digital pathogen surveillance system”. This concept of molecular epidemiology of infectious disease could globally improve public health, particularly in settings and areas of the world lacking robust laboratory capacity (Gardy and Loman, 2018).

Acknowledgment

Work about the molecular epidemiology of infectious disease in the author’s laboratory is funded by Niedersächsisches Ministerium für Wissenschaft und Kultur under the initiative ‘Big Data in den Lebenswissenschaften der Zukunft’, Projektnummer ZN 3432 and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s

Excellence Strategy – EXC 2155 – Projektnummer 390874280 and the Collaborative Research Center 900 - Projektnummer 158989968. The authors declares no conflict of interest.

References

Achtman, M. (2016) How old are bacterial pathogens? Proc Biol Sci 17: 283.

Ardui, S., Ameur, A., Vermeesch, J.R., and Hestand, M.S. (2018) Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res 46:2159-2168. doi: 10.1093/nar/gky066.

Birren, B., and Lai, E., eds. (1996) Nonmammalian genomic analysis. San Diego, USA: Academic Press.

Blane, B., Raven, K.E., Leek, D., Brown, N., Parkhill, J., and Peacock, S.J. (2019) Rapid sequencing of MRSA direct from clinical plates in a routine microbiology laboratory. J Antimicrob Chemother 74:2153-2156. doi: 10.1093/jac/dkz170.

Bocklandt, S., Hastie, A., and Cao, H. (2019) Bionano genome mapping: High-throughput, ultra-long molecule genome analysis system for precision genome assembly and haploid-resolved structural variation discovery. Adv Exp Med Biol 1129:97-118. doi: 10.1007/978-981-13-6037-4_7.

Boisson-Dupuis, S., Ramirez-Alejo, N., Li, Z., Patin, E., Rao, G., Kerner, G., et al. (2018) Tuberculosis and impaired IL-23-dependent IFN-γ immunity in humans homozygous for a

common TYK2 missense variant. Sci Immunol 3:eaau8714. doi: 10.1126/sciimmunol.aau8714.

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., et al. (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Version 2. Nat Biotechnol 35:725-731. doi: 10.1038/nbt.3893.

Brown, N.M., Blane, B., Raven, K.E., Kumar, N., Leek, D., Bragin, E., et al. (2019) Pilot evaluation of a fully automated bioinformatics system for analysis of methicillin-resistant *Staphylococcus aureus* genomes and detection of outbreaks. J Clin Microbiol 57:e00858-19. doi: 10.1128/JCM.00858-19.

Bryant, J.M., Grogono, D.M., Rodriguez-Rincon, D., Everall, I., Brown, K.P., Moreno, P., et al. (2016) Emergence and spread of a human-transmissible multidrug-resistant nontuberculous mycobacterium. Science 354:751-757. doi: 10.1126/science.aaf8156.

Casanova, J.L. (2015) Human genetic basis of interindividual variability in the course of infection. Proc Natl Acad Sci USA 112:E7118–E7127. https ://doi.org/10.1073/pnas.15216 44112

Casanova, J.L., and Abel, L. (2013) The genetic theory of infectious diseases: a brief history and selected illustrations. Annu Rev Genomics Hum Genet 14:215–243. https :// doi.org/10.1146/annur ev-genom-09121 2-15344 8

Casanova, J.L., and Abel, L. (2020a) Lethal infectious diseases as inborn errors of immunity: toward a synthesis of the germ and genetic theories. Annu Rev Pathol, in press. https ://doi.org/10.1146/annur evpathol-03192 0-10142 9

Casanova, J.L., and Abel, L. (2020b) The human genetic determinism of life-threatening infectious diseases: genetic heterogeneity and physiological homogeneity? Hum Genet 139:681-694. doi: 10.1007/s00439-020-02184-w.

Charalampous, T., Kay, G.L., Richardson, H., Aydin, A., Baldan, R., Jeanes, C., et al. (2019) Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol 37:783-792. doi: 10.1038/s41587-019-0156-5.

Coll, F., Harrison, E.M., Toleman, M.S., Reuter, S., Raven, K.E., Blane, B., et al. (2017) Longitudinal genomic surveillance of MRSA in the UK reveals transmission patterns in hospitals and the community. Sci Transl Med 9:eaak9745. doi: 10.1126/scitranslmed.aak9745.

Cramer, N., Klockgether, J., Wrasman, K., Schmidt, M., Davenport, C.F., and Tümmler, B. (2011) Microevolution of the major common *Pseudomonas aeruginosa* clones C and PA14 in cystic fibrosis lungs. Environ Microbiol 13:1690-704. doi: 10.1111/j.1462-2920.2011.02483.x.

De Sales, R.O., Migliorini, L.B., Puga, R., Kocsis, B., and Severino, P. (2020) A core genome multilocus sequence typing scheme for *Pseudomonas aeruginosa*. Front Microbiol 11:1049. doi: 10.3389/fmicb.2020.01049.

Domman, D., Quilici, M.L., Dorman, M.J., Njamkepo, E., Mutreja, A., Mather, A.E., et al. (2017) Integrated view of *Vibrio cholerae* in the Americas. Science 358:789-793. doi: 10.1126/science.aao2136.

Dunne W. M, Jr., Pouseele, H., Monecke, S., Ehricht, R., and van Belkum, A. (2018) Epidemiology of transmissible diseases: Array hybridization and next generation sequencing as universal nucleic acid-mediated typing tools. Infect Genet Evol 63:332-345. doi: 10.1016/j.meegid.2017.09.019.

Elborn, J.S. (2016) Cystic fibrosis. Lancet 388:2519-2531.

Foxman B, and Riley L. (2001) Molecular epidemiology: focus on infection. Am J Epidemiol 153:1135-41. doi: 10.1093/aje/153.12.1135.

Gan, G.L., Willie, E., Chauve, C., and Chindelevitch, L. (2019) Deconvoluting the diversity of within-host pathogen strains in a multi-locus sequence typing framework. BMC Bioinformatics 20(Suppl 20):637. doi: 10.1186/s12859-019-3204-8.

Gardy, J.L., and Loman N.J. (2018) Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet 19:9-20. doi: 10.1038/nrg.2017.88.

Gerner-Smidt, P., Hise, K., Kincaid, J., Hunter, S., Rolando, S., Hyytiä-Trees, E., et al. (2006) PulseNet USA: a five-year update. Foodborne Pathog Dis 3:9-19. doi: 10.1089/fpd.2006.3.9.

Giovanetti, M., Faria, N.R., Lourenço, J., Goes de Jesus, J., Xavier, J., Claro, I.M., et al. 2020. Genomic and epidemiological surveillance of Zika Virus in the Amazon region. Cell Rep 30: 2275-2283.e7. doi: 10.1016/j.celrep.2020.01.085.

Gouliouris, T., Raven, K.E., Ludden, C., Blane, B., Corander, J., Horner, C.S., et al. (2018) Genomic surveillance of *Enterococcus faecium* reveals limited sharing of strains and resistance genes between livestock and humans in the United Kingdom. mBio 9:e01780-18. doi: 10.1128/mBio.01780-18.

Gouliouris, T., Raven, K.E., Moradigaravand, D., Ludden, C., Coll, F., Blane, B., et al. (2019) Detection of vancomycin-resistant *Enterococcus faecium* hospital-adapted lineages in municipal wastewater treatment plants indicates widespread distribution and release into the environment. Genome Res 29:626-634. doi: 10.1101/gr.232629.117.

Grillová, L., and Picardeau, M. (2020) Core genome multi-locus sequence typing analyses of *Leptospira* spp. using the bacterial isolate genome sequence database. Methods Mol Biol 2134:11-21. doi: 10.1007/978-1-0716-0459-5_2.

Grothues, D, and Tümmler, B. (1991) New approaches in genome analysis by pulsed-field gel electrophoresis: application to the analysis of Pseudomonas species. Mol Microbiol 5:2763-76. doi: 10.1111/j.1365-2958.1991.tb01985.x.

Groussin, M., Mazel, F., and Alm, E.J. (2020) Co-evolution and co-speciation of host-gut bacteria systems. Cell Host Microbe 28:12-22. doi: 10.1016/j.chom.2020.06.013.

Hall, M.D., Holden, M.T., Srisomang, P., Mahavanakul, W., Wuthiekanun, V., Limmathurotsakul, D., et al. (2019) Improved characterisation of MRSA transmission using within-host bacterial sequence diversity. Elife. 2019 Oct 8;8:e46402. doi: 10.7554/eLife.46402.

Hsu, C.H., Harrison, L., Mukherjee, S., Strain, E., McDermott, P., Zhang, Q., et al. (2020) Core genome multilocus sequence typing for food animal source attribution of human *Campylobacter jejuni* infections. Pathogens 9(7):E532. doi:10.3390/pathogens9070532.

Kerner, G., Ramirez-Alejo, N., Seeleuthner, Y., Yang, R., Ogishi, M., Cobat, A., et al. (2019) Homozygosity for *TYK2* P1104A underlies tuberculosis in about 1% of patients in a cohort of European ancestry. Proc Natl Acad Sci U S A 116:10430-10434. doi: 10.1073/pnas. 1903561116.

Kimura, B. (2018) Will the emergence of core genome MLST end the role of in silico MLST? Food Microbiol 75:28-36. doi: 10.1016/j.fm.2017.09.003

Klimovich, A.V., and Bosch, T.C.G. (2018) Rethinking the role of the nervous system: Lessons from the Hydra holobiont. Bioessays 40:e1800060. doi: 10.1002/bies.201800060.

Klockgether, J., Cramer, N., Fischer, S., Wiehlmann, L., and Tümmler, B. (2018) Long-term microevolution of *Pseudomonas aeruginosa* differs between mildly and severely affected cystic fibrosis lungs. Am J Respir Cell Mol Biol 59:246-256. doi: 10.1165/rcmb.2017-0356OC.

Konstantinidis, K.T., Rosselló-Móra, R., and Amann, R. (202) Advantages outweigh concerns about using genome sequence as type material for prokaryotic taxonomy. Environ Microbiol 22:819-822. doi: 10.1111/1462-2920.14934.

Liang, K.Y.H., Orata, F.D., Islam, M.T., Nasreen, T., Alam, M., Tarr, C.L., et al. (2020) A *Vibrio cholerae* core genome multilocus sequence typing scheme to facilitate the epidemiological study of cholera. J Bacteriol JB.00086-20. doi: 10.1128/JB.00086-20.

Lieberman, T.D., Michel, J.B., Aingaran, M., Potter-Bynoe, G., Roux, D., Davis, M.R. Jr., et al. (2011) Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat Genet 43:1275-80. doi: 10.1038/ng.997.

Liu, S., Li, X., Guo, Z., Liu, H., Sun, Y., Liu, Y., et al. (2020) A core genome multilocus sequence typing scheme for *Streptococcus mutans*. mSphere 5(4):e00348-20. doi: 10.1128/mSphere.00348-20.

Ludden, C., Moradigaravand, D., Jamrozy, D., Gouliouris, T., Blane, B., Naydenova, P., et al. (2029) A one health study of the genetic relatedness of *Klebsiella pneumoniae* and their mobile elements in the East of England. Clin Infect Dis 70:219-226. doi: 10.1093/cid/ciz174.

Lynskey, N.N., Jauneikaite, E., Li, H.K., Zhi, X., Turner, C.E., Mosavie, M., et al. (2019) Emergence of dominant toxigenic M1T1 *Streptococcus pyogenes* clone during increased scarlet fever activity in England: a population-based molecular epidemiological study. Lancet Infect Dis 19:1209-1218. doi: 10.1016/S1473-3099(19)30446-3.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., et al. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95:3140-5. doi: 10.1073/pnas.95.6.3140.

Martak, D., Meunier, A., Sauget, M., Cholley, P., Thouverez, M., Bertrand, X., et al. (2020) Comparison of pulsed-field gel electrophoresis and whole-genome-sequencing-based typing confirms the accuracy of pulsed-field gel electrophoresis for the investigation of local

P*seudomonas aeruginosa* outbreaks. J Hosp Infect S0195-6701(20)30298-X. doi: 10.1016/j.jhin.2020.06.013.

Marvig, R.L., Sommer, L.M., Molin, S., and Johansen, H.K. (2015) Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. Nat Genet 47:57-64. doi: 10.1038/ng.3148.

Mellmann, A., Harmsen, D., Cummings, C.A., Zentz, E.B., Leopold, S.R., Rico, A., et al. (2011) Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. PLoS One 6:e22751. doi: 10.1371/journal.pone.0022751.

Mellmann, A., Andersen, P.S., Bletz, S., Friedrich, A.W., Kohl, T.A., Lilje, B., et al. (2017) High interlaboratory reproducibility and accuracy of next-generation-sequencing-based bacterial genotyping in a ring trial. J Clin Microbiol 55:908-913. doi: 10.1128/JCM.02242-16.

Monecke, S., and Ehricht, R. (2005) Rapid genotyping of methicillin-resistant *Staphylococcus aureus* (MRSA) isolates using miniaturised oligonucleotide arrays. Clin Microbiol Infect 11:825-33. doi: 10.1111/j.1469-0691.2005.01243.x.

Moodley, Y., Linz, B., Bond, R.P., Nieuwoudt, M., Soodyall, H., Schlebusch, C.M., et al. (2012) Age of the association between *Helicobacter pylori* and man. PLoS Pathog 8:e1002693.

Moran Losada, P., Chouvarine, P., Dorda, M., Hedtfeld, S., Mielke, S., Schulz, A., et al. (2016) The cystic fibrosis lower airways microbial metagenome. ERJ Open Res 2:00096-2015. doi: 10.1183/23120541.00096-2015.

Murray, A.E., Freudenstein, J., Gribaldo, S., Hatzenpichler, R., Hugenholtz, P., Kämpfer, P., et al. (2020) Roadmap for naming uncultivated Archaea and Bacteria. Nat Microbiol , in press. doi: 10.1038/s41564-020-0733-x.

Nicholls, S.M., Quick, J.C., Tang, S., and Loman, N.J. (2019) Ultra-deep, long-read nanopore sequencing of mock microbial community standards. Gigascience 8:giz043. doi: 10.1093/gigascience/giz043.

Pendleton, K.M., Erb-Downward, J.R., Bao, Y., Branton, W.R., Falkowski, N.R., Newton, D.W., et al. (2017). Rapid pathogen identification in bacterial pneumonia using real-time metagenomics. Am J Respir Crit Care Med 196:1610-1612. doi: 10.1164/rccm.201703-0537LE.

Quick, J., Grubaugh, N.D., Pullan, S.T., Claro, I.M., Smith, A.D., Gangavarapu, K., et al. (2017) Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc 12:1261-1276. doi: 10.1038/nprot.2017.066.

Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., and Segata, N. (2017) Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35:833-844. doi: 10.1038/nbt.3935

Raven, K.E., Blane, B., Kumar, N., Leek, D., Bragin, E., Coll, F., et al. (2020) Defining metrics for whole-genome sequence analysis of MRSA in clinical practice. Microb Genom 6:e000354. doi: 10.1099/mgen.0.000354.

Riley, L.W. (2018) Laboratory methods in molecular epidemiology: Bacterial infections. Microbiol Spectr 6(6). doi: 10.1128/microbiolspec.AME-0004-2018.

Riley, L.W., and Blanton, R.E. (2018) Advances in molecular epidemiology of infectious diseases: Definitions, approaches, and scope of the field. Microbiol Spectr 6:10.1128/microbiolspec.AME-0001-2018. doi: 10.1128/microbiolspec.AME-0001-2018.

Rice, E.S., and Green, R.E. (2019) New approaches for genome assembly and scaffolding. Annu Rev Anim Biosci 7:17-40. doi: 10.1146/annurev-animal-020518-115344.

Römling, U., Schmidt, K.D., and Tümmler, B. (1997) Large genome rearrangements discovered by the detailed analysis of 21 *Pseudomonas aeruginosa* clone C isolates found in environment and disease habitats. J Mol Biol 271:386-404. doi: 10.1006/jmbi.1997.1186.

Runtuwene, L.R., Tuda, J.S.B., Mongan, A.E., and Suzuki, Y. (2019) On-site MinION sequencing. Adv Exp Med Biol 1129:143-150. doi: 10.1007/978-981-13-6037-4_10.

Sabat, A.J., Budimir, A., Nashev, D., Sá-Leão, R., van Dijl, Jm., Laurent, F., et al. (2013) Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro Surveill 18:20380. doi: 10.2807/ese.18.04.20380-en.

Scharff, R.L., Besser, J., Sharp, D.J., Jones, T.F., Peter, G.S., Hedberg, C.W. (2016) An economic evaluation of PulseNet: a network for foodborne disease surveillance. Am J Prev Med 50(Suppl 1):S66–S73. http://dx.doi.org/10.1016/j.amepre.2015.09.018.

Schürch, A.C., Arredondo-Alonso, S., Willems, R.J.L., and Goering, R.V. (2018) Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. Clin Microbiol Infect 24:350-354. doi: 10.1016/j.cmi.2017.12.016.

Schulte, P.A., and Perera, F.P., eds. (1993) Molecular epidemiology: Principles and Practices. San Diego, USA: Academic Press.

Steinig EJ, Duchene S, Robinson DA, Monecke S, Yokoyama M, Laabei M, et al. (2019) Evolution and global transmission of a multidrug-resistant, community-associated methicillin-resistant *Staphylococcus aureus* lineage from the Indian subcontinent. mBio 10:e01105-19. doi: 10.1128/mBio.01105-19.

Theis, K.R., Dheilly, N.M., Klassen, J.L., Brucker, R.M., Baines, J.F., Bosch, T.C., et al. (2016) Getting the hologenome concept right: an eco-evolutionary framework for hosts and their microbiomes. mSystems 1:e00028-16. doi: 10.1128/mSystems.00028-16.

Toleman, M.S., Reuter, S., Jamrozy, D., Wilson, H.J., Blane, B., Harrison, E.M., et al. (2019) Prospective genomic surveillance of methicillin-resistant *Staphylococcus aureus* (MRSA) associated with bloodstream infection, England, 1 October 2012 to 30 September 2013. Euro Surveill 24:1800215. doi: 10.2807/1560-7917.ES.2019.24.4.1800215.

Wang, X., Jordan, I.K., and Mayer, L.W. (2015) A Phylogenetic Perspective on Molecular Epidemiology. In: Tang, Y-W., Sussman, M., Liu, D., Poxton, I., and Schwartzman, J. Molecular Medical Microbiology, 2nd ed. Academic Press, Elsevier, pp. 517-536. DOI: http://dx.doi.org/10.1016/B978-0-12-397169-2.00029-9

Weill, F.X., Domman, D., Njamkepo, E., Tarr, C., Rauzier, J., Fawal, N., et al. (2017) Genomic history of the seventh pandemic of cholera in Africa. Science 358:785-789. doi: 10.1126/science.aad5901.

Wiehlmann, L., Wagner, G., Cramer, N., Siebert, B., Gudowius, P., Morales, G., et al. (2007) Population structure of *Pseudomonas aeruginosa*. Proc Natl Acad Sci U S A 104:8101-6. doi: 10.1073/pnas.0609213104.

Wiehlmann, L., Cramer, N., and Tümmler, B. (2015) Habitat-associated skew of clone abundance in the *Pseudomonas aeruginosa* population. Environ Microbiol Rep 7:955-60. doi: 10.1111/1758-2229.12340.

Yang, L., Jelsbak, L., Marvig, R.L., Damkiær, S., Workman, C.T., Rau, M.H., et al. (2011) Evolutionary dynamics of bacteria in a human host environment. Proc Natl Acad Sci U S A 108:7481-6. doi: 10.1073/pnas.1018249108.

Table 1. Nucleic – acid based methods for the typing of cultivatable bacterial strains

Method	Principle
Fragment fingerprints	
AP-PCR; RAPD	Amplicon patterns generated by low-stringency PCR of genomic DNA
REP; ERIC; BOX	PCR amplicon patterns of repetitive elements
MLVA	Multilocus analysis of copy number variations of tandem repeats
PFGE	Whole genome macrorestriction fragment length polymorphisms
Imaging	Optical physical genome mapping of structural variants
Hybridization	
Arrays	Multilocus SNP detection by chip hybridization of amplified gene fragments
Sequencing	
MLST	Allele profiling of a few (< 10) conserved loci
cgMLST	Allele profiling of (a large part of) the genes of the conserved core genome
wgMLST	Whole genome gene-by-gene profiling or whole genome SNV analysis
WGS	Complete genome represented by one closed contig or few scaffolds

Table 2. Features of genotyping methods for the typing of cultivatable bacterial strains

Method	Time spent per assay	Turnover of specimens	Reproduci-bility	Porta-bility	Manual skills	Hands-on experience	Bioinformatic knowledge	Online database	Cost
AP-PCR; RAPD	½ day	high	poor	no	low	desirable	low	no	low
REP; ERIC; BOX	½ day	high	low	no	low	desirable	low	no	low
MLVA	1-2 day	moderate	high	yes	desirable	desirable	low	yes	low
PFGE	3-7 day	moderate	moderate	yes	necessary	necessary	low	yes	moderate
Imaging	2-3 day	moderate	high	yes	desirable	necessary	intermediate	no	high
Arrays	2-4 h	high	high	yes	low	desirable	intermediate	no	moderate
MLST	2-7 day	high	high	yes	low	low	intermediate	yes	moderate
cgMLST	2-3 day	high	high	yes	low	low	intermediate	yes	high
wgMLST	2-7 day	high	high	yes	low	low	high	(yes)	high

WGS	2-30 day	moderate	high	yes	low	low	high	(yes)	high

Table 3. Web-based bioinformatic tools and databases for molecular epidemiology of infectious disease

Host	Software package, contents and modules	Web address
Bacterial Isolate Genome Sequence Database (BIGSdb)	Software platform for wide range of applications of genome-based molecular epidemiology	https://pubmlst.org/software/database/bigsdb/
PulseNet International	BioNumerics Server and Client Software (for PFGE): Fingerprint and character types, comparison & cluster analysis, Identification and libraries, database (sharing tools)	https://pulsenetinternational.org/protocols/bionumerics/
	MLVA protocols and Bionumerics analysis	https://pulsenetinternational.org/protocols/mlva/
PubMLST	MLST schemes and isolate data	https://pubmlst.org/databases/

Center for Genomic Epidemiology (CGE)	MLST extracted from total genome sequenced bacteria	https://cge.cbs.dtu.dk/services/MLST/
Applied Maths	BioNumerics software; applications: AFLP, ITS, MIRU-VNTR, MVLA, PFGE, wgSNP, wgMLST	http://www.applied-maths.com/bionumerics
Integrated Rapid Infectious Disease Analysis (IRIDA)	Analytics and visualizations for WGS-based microbial pathogen investigations	https://www.irida.ca/
Ridom	SeqSphere software: cgMLST	https://www.ridom.de/
Ridom	cgMLST.org nomenclature server	https://www.cgmlst.org/ncs
PHYLOViZ Online	cgMLST web-based visualization software	https://online.phyloviz.net/index
GenGis	Phylogeography and statistical analysis platform	https://beikolab.cs.dal.ca/gengis/Main_Page
ExaBayes	Bayesian tree inference for phylogenetic reconstruction	https://doi.org/10.1093/molbev/msu236

Phyloscanner	Inference of the direction of transmission from sequence data	https://github.com/BDI-pathogens/phyloscanner
RAxML-NG	Phylogenetic tree inference tool which uses maximum-likelihood optimality criterion	https://github.com/amkozlov/raxml-ng
Bioconductor	ggtree: R package for visualization and annotation of a phylogenetic tree	https://bioconductor.org/packages/release/bioc/html/ggtree.html
plotTree	R package for plotting and annotation of a phylogenetic tree	https://github.com/katholt/plotTree
interactive Tree of Life (iTOL)	Web-based online tool for display, annotation and management of phylogenetic trees	http://itol.embl.de

Table 4. Genetic susceptibility of humans to bacterial infection (Casanova and Abel, 2020b)

Pathogen	Clinical phenotype	Immunological phenotype	Gene[a]
Broad spectrum	Invasive disease	core NF-κB / TLR / IL-1R pathways	*NEMO, NFKBIA, IKBKA, IKBKB, RBCK1, RNF31, IRAK1, IRAK4, MYD88, HOIL1, HOIP*
		Complement deficiency	*C1QA, C1QB, C1QC, C1S, C2, C3, C4A, C4B, CFH, CFI*
Staphylococcus aureus	Recurrent disease	TLR2 response deficiency or IL-6 deficiency	*TIRAP, IL6RA, IL6ST, ZNF341, STAT3*
Neisseria	Invasive disease	Complement deficiency	*C5*, *C6*, *C7*, *C8A, C8B*, *C9, CFB, CFD, CFP*
Environmental mycobacteria	MSMD[b]	IFN-γ deficiency	*IFNGR1, IFNGR2, IL12RB1, IL12B, NEMO, STAT1, CYBB, TYK2, IRF8, ISG15, RORC, IL12RB2, IL23R, PPL2A, JAK1*
Mycobacterium tuberculosis	Tuberculosis (TB)	IFN-γ deficiency	*IL12RB1, TYK2*

[a]Monogenic disorders: Single locus mutations cause infectious disease with complete or incomplete clinical penetrance

[b]MSMD, Mendelian susceptibility to mycobacterial disease