From theory to practice: molecular strain typing for the clinical and public health setting

R V Goering (rgoeri@creighton.edu)[1], R Köck[2], H Grundmann[3], G Werner[4], A W Friedrich[3], on behalf of the ESCMID Study Group for Epidemiological Markers (ESGEM)[5]

1. Department of Medical Microbiology and Immunology, Creighton University School of Medicine, Omaha, USA
2. Institute of Hygiene, University Hospital Münster, Münster, Germany
3. Department of Medical Microbiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
4. National Reference Centre for Staphyloccoci and Enterococci, Robert Koch Institute, Wernigerode Branch, Wernigerode, Germany
5. European Society for Clinical Microbiology and Infectious Diseases, Basel, Switzerland

The persistence and transmission of infectious disease is one of the most enduring and daunting concerns in healthcare. Over the years, epidemiological analysis especially of bacterial etiological agents has undergone a remarkable evolutionary metamorphosis. While initially relying on purely phenotypic characterisation, advances in molecular biology have found translational application in a number of approaches to strain typing which commonly centre either on 'epityping' (molecular epidemiology) to characterise outbreaks, perform surveillance, and trace evolutionary pathways, or 'pathotyping' to compare strains based on the presence or absence of specific virulence or resistance genes. A perspective overview of strain typing is presented here considering the issues surrounding analyses which are employed in the localised clinical setting as well as at a more regional/national public health level. The discussion especially considers the shortcomings inherent in epidemiological analysis: less than full isolate characterisation by the typing method and limitations imposed by the available data, context, and time constraints of the epidemiological investigation (i.e. the available epidemiological window). However, the promises outweigh the pitfalls as one considers the potential for advances in genomic characterisation and information technology to provide an unprecedented aggregate of epidemiological information and analysis.

Introduction

Since the time of Semmelweis and Koch's Postulates, medical science has recognised the cause-and-effect relationship between the transmission of etiological agents and the persistence and spread of infectious disease. In this context, routine clinical and infection control interests commonly centre on the detection of multifocal patient infection or dissemination within a defined patient population (e.g. outbreak identification, control, or other rather short-term epidemiological issues). Conversely, public health concerns include local, regional, national, and international emergence and spread of pathogens, global microbiological and molecular surveillance, as well as longer term evolutionary interrelationships. Classical epidemiology uses the three parameters (time, place, person) to find epidemiological links. However, in both healthcare and community-associated infections today, those three parameters do not necessarily provide the desired resolution to identify an outbreak event or the causing pathogen. Clinical microbiology provides species-level isolate identification and molecular analysis provides the strain type or subtype fingerprint. Bringing these five parameters together provides the greatest hope of associating outbreaks of infectious disease with certain types of the same bacterial species. This perspective overview considers the epidemiological analysis of infectious diseases in both the clinical and public health setting, focusing on bacterial etiologies to illustrate issues associated with moving molecular strain typing from theory to practical application. Regardless of the setting, the interrelationships that strain typing seeks to clarify are generally in the context of epityping (i.e. transmission investigation (e.g. outbreak)) or pathotyping to compare strains based on the presence or absence of specific virulence genes. The former is emphasised here and discussed in the context of two principal challenges independent of the methods employed: isolate characterisation and the available data, context, and time constraints of the epidemiological investigation (i.e. the available epidemiological window).

The challenge of isolate characterisation

In both the clinical and public health setting, the assessment of potential interrelationships between isolates is based on a comparison of specific characteristics which ideally will identify (i.e. fingerprint) transmitted strains as the same type while not overlooking

epidemiologically relevant variants (subtypes) or mistakenly including unrelated isolates (i.e. issues of sensitivity and specificity). Isolate characterisation has been historically based on phenotypic assessment which is most certainly still of value (e.g. antibiograms, serotyping). However, recognition of the bacterial chromosome as the fundamental molecule of cellular identity has firmly established the importance of molecular (genomic) epidemiological evaluation. Thus, molecular approaches to isolate characterisation are considered here. In general, historical review reveals a consistent 'translational' trend of genotypic methods moving from the basic science laboratory to clinical application. These approaches to molecular epidemiology are reviewed more completely elsewhere [1,2] and are only summarised here to note the challenges faced in terms of providing definitive isolate characterisation for epidemiological purposes.

Simply stated, when it comes to epidemiological sensitivity and specificity the key methodological issues are: (i) the degree to which the targets/markers being analysed provide epidemiologically relevant information and (ii) the precision with which the queried characteristic(s) are identified and analysed. The former relates to epidemiological validation which has been considered elsewhere [3] and is beyond the scope of this discussion. However, by way of summary it is important to note that, regardless of analytical precision, other than whole genome sequencing (WGS) all methods strive to assess isolate interrelatedness based on a subset of targets that represent a genomically incomplete, but epidemiologically relevant, dataset. Thus, for these approaches, additional data is more informative than less (e.g. see [4]). In terms of precise data output, while newer methods employ instrumentation (e.g. capillary electrophoresis using an automated DNA sequencer [5]), a significant number of currently used protocols rely on visual inspection of data output generated by agarose gel electrophoresis (Table). While such analysis can be accurate for protocols involving the presence or absence of end point polymerase chain reaction (PCR) products, visual assessment of fragment-size comparisons (e.g. by agarose gel electrophoresis) can be problematic. For example, digestion of total cellular DNA by common restriction enzymes (restriction endonuclease analysis (REA)) can generate greater than 600 fragments from a typical 2 to 3 Mb bacterial chromosome. In addition, there is an element of imprecision in the visual comparison of DNA banding patterns in electrophoresis gels since DNA fragments differing by ±10% may be seen as identical [6]. This could amount to a 70 kb discrepancy, for example, in a pulsed-field gel with bands ca. 700 kb in size.

As noted earlier, the chromosome is the most fundamental molecule of identity in the cell. Thus, it is the sequence-based methods that ultimately hold the greatest promise for accurately assessing epidemiological interrelationships in problem pathogens. Reviewed elsewhere [2,7] these methods can be found in three general iterations: single locus sequence typing (SLST), multilocus sequence typing (MLST), and WGS (Table). Of these, the first two have found broad epidemiological application although, as noted above for other methods, both represent a genomically incomplete dataset, while WGS holds clear promise for providing total chromosomal analysis. While WGS was impossible with older dideoxy/chain termination sequencing technology [8], newer (i.e. next generation sequencing, NGS) methods have made this goal a reality. The technology behind NGS is discussed in detail elsewhere [7,9], however, from a strain typing

Table

Characteristics of methods commonly used for molecular epidemiology

Data generation	Chromosomal target(s)	Data output	Method examples
Restriction enzymes	Common restriction sites	DNA fragments visualised after agarose gel electrophoresis (AGE)	Restriction endonuclease analysis (REA)
Restriction enzymes	Common restriction sites	Ordered sequence scaffolds identified via instrument software	Optical mapping
Restriction enzymes	Rare restriction sites	DNA fragments visualised after AGE	Pulsed-field gel electrophoresis (PFGE)
Polymerase chain reaction (PCR)	Repetitive element or variable-number tandem repeat (VNTR) sequences	Amplified DNA fragments either visualised after AGE or via instrument software	Repetitive-element PCR (rep-PCR); VNTR typing; PCR ribotyping
DNA probes	Multiple genes	Hybridisation signal either identified visually or via instrument software	Microarray
DNA sequencing	Single or multiple genes	DNA sequence obtained via instrument software	*Staphylococcus aureus* protein A gene (spa) typing; multilocus sequence typing (MLST)
DNA sequencing	Whole genome	DNA sequence obtained via instrument software	Whole genome sequencing (WGS); next generation sequencing (NGS)

A full description of methods is reported elsewhere [1,2].

standpoint it is important to note that revolutionary developments in NGS have made WGS possible with benchtop instrumentation such as the Ion Torrent PGM (Life Technologies, Guilford), GS Junior (454 Life Sciences/Roche, Branford), and the MiSeq (Illumina, San Diego). Such instrumentation now allows WGS to be completed in hours to days with extensive multi-fold coverage allowing isolates to be compared down to the level of single nucleotide polymorphisms (SNPs). However, as with previous sequencing iterations, the critical issues for NGS are throughput, quality, read length and cost. All of these are currently in a state of flux as commercial technology improves and positions itself in the scientific marketplace. In addition, it must be noted that the present state of WGS has not reached accurate base-by-base total origin-to-termini output. For example, the assembly and analysis of the relatively short read lengths from current NGS platforms are problematic for repeat sequences (e.g. clustered regularly interspaced short palindromic repeats (CRISPRs), homopolymers, and variable-number tandem repeats (VNTRs) [10]). An additional bottleneck is the bioinformatics requirement for proper WGS annotation and analysis which at present is far from routine, with costs (in time and money) that may exceed that of the sequencing itself [11,12]. Nevertheless, these are exciting 'problems' to have, confirming that the scientific stage is clearly set for remarkable developments in this most fundamental approach to determining isolate epidemiological interrelationships.

The challenge of the epidemiological window and detecting significant difference

Regardless of the epidemiological approach, the focus ultimately becomes data interpretation. Thus, it is important to note that while the term 'molecular' epidemiology implies a precise process, this is not always the case regardless of the method employed since epidemiological analysis always has an unavoidable context and time-driven component. A variety of environmental factors as well as interaction between the host and infectious agent may all influence the course of disease transmission. In addition, the time leading up to, as well as that required for, the epidemiological investigation provides opportunity for the outbreak strain to evolve. Whether in a clinical or public health setting, infectious disease scenarios benefiting from epidemiological evaluation do not typically give advance warning. Hence, in many investigations where the starting point of the epidemiological scenario (e.g. the source case or the outbreak source) is not identified, the process of data analysis attempts to work backward in time which, depending on the available information, may necessitate drawing conclusions based on probabilities rather than absolute certainty [13]. However, as with classical epidemiological approaches, molecular epidemiological analysis may to some extent implicate the source 'beyond a reasonable doubt'.

In the absence of a source isolate, all strain typing methods are challenged as the opportunity for chromosomal change over time increases the potential for genetic distance between epidemiologically related isolates (i.e. confounding the recognition of interrelationships in the isolates being analysed). This can be illustrated (Figure) considering a simple example of six epidemiologically-relevant characters ('A') in a reference genome (e.g. the characters could be restriction sites, specific genes, other chromosomal loci). Evolution through two generations, with sequential genetic events of unknown complexity (e.g. insertions, deletions, rearrangements, recombination) designated as changes from 'A' to 'B', results in second-generation genomes varying from each other by four differences. As the process continues through subsequent generations additional complexity in the population dramatically increases. This scenario illustrates the issue central to the interpretation of any bacterial strain typing data, the definition and detection of significant difference. This relates to the issues of sensitivity and specificity previously addressed, in particular specificity, which is important to insure adequate case definitions for outbreak investigations, in order to avoid inclusion of non-cases and detect maximum epidemiological associations between the isolates. Thus, for optimum epidemiological outcome, proper analysis of strain typing data requires knowledge of: (i) the genetics of the microbial pathogen (e.g. clock speed/rate of change of the characteristics being analysed), (ii) the limitations of the typing method, (iii) the degree of concordance between different typing methods, if more than one technique is applied in parallel, and (iv) the setting within which the issue is being studied. Regardless of the typing approach, these details must be considered in attempting to discern the relatedness and transmission patterns of infectious agents in both the clinical and public health setting.

Figure

Diagrammatic illustration of interrelationships between a reference genome and two subsequent generations each of which differs from the previous by a single genetic event
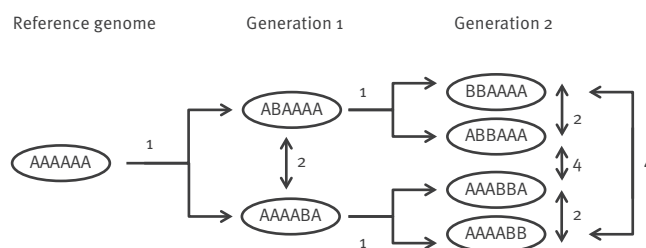
The reference genome has six epidemiologically-relevant characteristics (designated 'A'). Each generation differs from the previous by a single genetic event (indicated by the number 1 above the horizontal arrows) changing characteristics from A to B. For each generation, the numbers of genetic differences between members are indicated by figures on the side of the vertical arrows (adapted from [13]).

The 'typing Esperanto'

It is of utmost importance, that typing methods produce data that can be compared not only within the same laboratory or clinical setting, but also between different facilities. Therefore, the 'typing Esperanto' or language should produce data that are clear, reproducible, and include strain nomenclature which allows for the independent identification of specific types. However, it is important to note that the probability of an outbreak due to a certain strain type depends on its frequency in the associated environment (e.g. both within and outside of the healthcare setting, the community). The less frequent a strain type is, the more probable it becomes that multiple isolates (a cluster) of a certain strain type represent a true outbreak. Thus, epidemiological analysis must recognise the nuances associated with disease transmission such as distinguishing outbreaks from pseudo-outbreaks [14]. The latter occur frequently in environments associated with an endemic prevalence of antibiotic-resistant microorganisms. For example, in a clinical setting, patients on the same hospital ward may carry similar but distinct problem pathogens which could superficially mimic an outbreak. Useful typing should properly identify such a pseudo-outbreak thus helping to avoid inappropriate escalation of 'outbreak' management. This kind of 'de-compromising' and 'de-escalating' is one of the major reasons why local hospitals and their laboratories perform strain typing for outbreak analysis. Thus, whether in a clinical or public health setting, the discriminatory or resolving power of a given epidemiological analysis is not solely dependent on a method or a method-pathogen combination but may be also be influenced by the pathogens' diversity (i.e. the more or less frequent appearance/epidemicity or endemicity of a specific type).

Choosing the 'best' method for typing

Whether considering strain typing from the clinical or public health perspective, the logical question is: what is the best method procedurally to use? However, there are a number of reasons why a 'one size fits all' answer to this question is impractical.

Considering first the clinical environment, as noted earlier, strain typing is commonly of value in assessing therapeutic concerns such as multisite infection or emergence of antimicrobial resistance in the individual patient, and transmission of problem pathogens within a limited patient population (e.g. a healthcare or family unit). In this context the key issues include: (i) having the required technical expertise, (ii) potential for automation/routine applicability, (iii) cost, (iv) required time-to-answer, (v) equipment maintenance and footprint size, (vi) intuitive data output and objective, standardisable, or automated interpretation, (vii) relevance of the typing result for further investigations (e.g. screening of staff) or for reporting to public health authorities.

It is logical to aspire to the most recently published cutting-edge method. However, the newest iteration of the most sophisticated and advanced technology is of little value if one does not have physical room for it, cannot afford it, properly operate it, or readily achieve clinically or epidemiologically relevant outcomes from the data generated. While one would never recommend gravitating to the lowest technological denominator for strain typing, to a large extent the 'best' method in a given clinical environment depends on the available resources addressing the issues noted above. In this context, as stated earlier, it is important to recognise that, regardless of sophistication, molecular strain typing commonly operates from an incomplete data set since all relevant clinical isolates may not be available and all isolate characteristics may not have been analysed, although the latter issue will be less of a concern in the future as WGS becomes more refined and widespread. In addition, communication between appropriate clinical interests (e.g. physician, laboratory, nursing, infection control) is vital to putting the 'incomplete' typing data into the fullest context for a meaningful outcome in terms of infection prevention and control.

Taken together, in addition to routine and real time strain typing, key elements for successful strain typing in the clinical setting most certainly include [3,15]: (i) initiation of strain typing by the hospital epidemiologist in consultation with infection control, infectious disease, and microbiology personnel, (ii) targeting of strain typing to investigate specific infectious disease issues such as an unusual increase in the rate of isolation of a pathogen, a cluster of infections in a particular healthcare unit, and multiple isolates with unusual (e.g. antibiotic susceptibility) characteristics, (iii) understanding that strain typing in the absence of epidemiological context and follow-up is an inefficient use of laboratory resources. Strain typing should supplement, not replace, careful epidemiological investigation.

To a large extent, the issues affecting approaches to strain typing for public health purposes are similar to those previously noted for local clinical efforts. However, there are important differences. The concerns of public health, while clinical in nature, are much broader in scope especially focusing on the transmission of problem pathogens on a local, regional, national, and international scale. Therefore, while financial and technical resources are generally more abundant at the regional/national level, the complexity of the necessary outcomes is greater as well. Effective communication to insure that the typing method's results are comparable between all laboratories involved is at the heart of a proper large-scale understanding of infectious disease occurrence and transmission. Everything from choice of typing method to data output and interpretation revolves around this issue. Thus, from a methodological standpoint the strain typing approach should: (i) be as standardised

as possible to be performed with similar efficiency, accuracy, and reproducibility in different participating laboratories, (ii) generate output that can be efficiently databased and shared, with interpretative criteria as objective as possible and a common terminology for strain type and subtype designations.

In this regard, sequence-based approaches hold the greatest promise. For example, SLST of the staphylococcal protein A gene (spa-typing) is effectively used in the epidemiological monitoring of specific *Staphylococcus aureus* strains (i.e. SeqNet; www.seqnet.org) with 540 laboratories from 51 countries submitting strains from 90 countries worldwide using the Ridom spa server as a common platform [16]. As noted earlier, approaches to WGS are rapidly being developed and refined with the potential to ultimately provide strain typing data ranging from key gene subsets [17] to total chromosomal comparison [18]. However, the success of the Pulse-Net System, designed by the United States Centers for Disease Control to investigate food-borne outbreaks [19], as well as refinements in VNTR-based analysis of pathogens such as meticillin-resistant *S. aureus* [5,20], illustrate that older molecular typing approaches also have potential for effective public health application.

Clinical and public health strain typing in perspective

Whether performed in a local clinical or more regional/national public health setting, the effective use of strain typing requires an understanding of both the pitfalls and the promises of the process. While the pitfalls can certainly be methodological, perhaps the most fundamental caveat, as noted above, is that strain typing is not a standalone method. Therefore, more information and communication is better than less. The scenario is not unlike an unfolding mystery story where one needs as much evidence as possible to figure out who 'did it.' For both local and larger-scale regional settings, the promise is a better understanding of the dynamics of infectious disease transmission with the hope of effective intervention (prevention, infection control, and treatment). Remarkable possibilities are on the horizon when one considers advances in genomic characterisation and the power of the Internet to facilitate the linking of strain typing analysis and databasing to other previously disparate data such as antimicrobial resistance (e.g. European Antimicrobial Resistance Surveillance Network (EARS-Net); www.ecdc.europa.eu/en/activities/surveillance/EARS-Net/Pages/index.aspx) and geographic information systems (GIS) as elegantly shown by the European Staphylococcal Reference Laboratory (SRL) working group (www.spatialepidemiology.net/srl-maps)[21] EpiScanGIS (www.episcangis.org), Global Network for Geospatial Health (GnosisGIS) (www.gnosisgis.org), and the World Health Organization (WHO)'s Public Health Mapping GIS effort (www.who.int/health_mapping/en). Most recently, during the *Escherichia coli* O104:H4 outbreak in Germany, open-source genomic analysis, available hardware/software resources and

international expertise contributed tremendously to the rapid understanding of the pathogens' evolution, dissemination, and pathology [22]. Thus, for the future, the promises outweigh the pitfalls as molecular strain typing seeks to address enduring infectious disease issues with important morbidity, mortality, economic, and general quality of life implications.

References

1. Goering RV. Molecular typing techniques: state of the art. In: Tang YW, Stratton CW, editors. Advanced techniques in diagnostic microbiology. 2nd ed. New York (NY): Springer; 2013. p. 239-61.

2. Sabat AJ, Budimir A, Nashev D, Sá-Leão R, van Dijl JM, Laurent F, et al. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro Surveill. 2013;18(4):pii=20380. Available from: www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20380

3. Van Belkum A, Tassios PT, Dijkshoorn L, Haeggman S, Cookson B, Fry NK, et al. Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clin Microbiol Infect. 2007;13 Suppl 3:1-46.

4. Robinson DA, Enright MC. Evolutionary models of the emergence of meticillin-resistant Staphylococcus aureus. Antimicrob Agents Chemother. 2003;47(12):3926-34.

5. Schouls LM, Spalburg EC, van LM, Huijsdens XW, Pluister GN, Van Santen-Verheuvel MG, et al. Multiple-locus variable number tandem repeat analysis of Staphylococcus aureus: comparison with pulsed-field gel electrophoresis and spa-typing. PLoS One. 2009;4(4):e5082.

6. Goering RV, Ribot EM, Gerner-Smidt P. Pulsed-field gel electrophoresis: laboratory and epidemiologic considerations for interpretation of data. In: Persing DH, Tenover FC, Tang YW, Nolte FS, Hayden RT, Belkum A, et al., editors. Molecular microbiology. 2nd ed. Washington (DC): ASM Press; 2011. p. 167-77.

7. Higuchi R, Glyyensten U, Persing DH. Next-generation DNA sequencing and microbiology. In: Persing DH, Tenover FC, Tang YW, Nolte FS, Hayden RT, Belkum A, et al., editors. Molecular microbiology: diagnostic principles and practice. 2nd ed. Washington (DC): ASM Press; 2011. p. 301-12.

8. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463-7.

9. Köser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog. 2012;8(8):e1002824.

10. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012;30(5):434-9.

11. Schatz MC, Langmead B, Salzberg SL. Cloud computing and the DNA data race. Nat Biotechnol. 2010;28(7):691-3.

12. Angiuoli SV, White JR, Matalka M, White O, Fricke WF. Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing. PLoS One. 2011;6(10):e26624.

13. Goering RV. Pulsed field gel electrophoresis: a review of application and interpretation in the molecular epidemiology of infectious disease. Infect Genet Evol. 2010;10(7):866-75.

14. Hallin M, Deplano A, Roisin S, Boyart V, De Ryck R, Nonhoff C, et al. Pseudo-outbreak of extremely drug-resistant Pseudomonas aeruginosa urinary tract infections due to contamination of an automated urine analyzer. J Clin Microbiol. 2012;50(3):580-2.

15. Tenover FC, Arbeit RD, Goering RV. How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: a review for healthcare epidemiologists. Infect Control Hosp Epidemiol. 1997;18(6):426-39.

16. Harmsen D, Claus H, Witte W, Rothganger J, Claus H, Turnwald D, et al. Typing of meticillin-resistant Staphylococcus aureus in a university hospital setting by using novel software for spa repeat determination and database management. J Clin Microbiol. 2003;41(12):5442-8.

17. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, et al. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology. 2012;158(Pt 4):1005-15.

18. Vogel U, Szczepanowski R, Claus H, Junemann S, Prior K, Harmsen D. Ion torrent personal genome machine sequencing for genomic typing of Neisseria meningitidis for rapid determination of multiple layers of typing information. J Clin Microbiol. 2012;50(6):1889-94.

19. Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg Infect Dis. 2001;7(3):382-9.

20. Sabat AJ, Chlebowicz MA, Grundmann H, Arends JP, Kampinga G, Meessen NE, et al. Microfluidic-chip-based multiple-locus variable-number tandem-repeat fingerprinting with new primer sets for meticillin-resistant Staphylococcus aureus. J Clin Microbiol. 2012;50(7):2255-62.

21. Grundmann H, Aanensen DM, van den Wijngaard CC, Spratt BG, Harmsen D, Friedrich AW. Geographic distribution of Staphylococcus aureus causing invasive infections in Europe: a molecular-epidemiological analysis. PLoS Med. 2010;7(1):e1000215.

22. Rohde H, Qin J, Cui Y, Li D, Loman NJ, Hentschke M, et al. Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. N Engl J Med. 2011;365(8):718-24.