EuroReference
journal of Reference
anses
French agency for food, environmental and occupational health & safety

Journal No. 10
Summer 2013

Point of view

Whole genome sequencing, an efficient investment for developing diagnostic and epidemiology tools: the case of contagious equine metritis

Sandrine Petry (sandrine.petry@anses.fr)
ANSES, Dozulé Laboratory for Equine Diseases, Goustranville, France

The number of available whole genomes has been regularly increasing while costs and sequencing times continue to decrease. Many laboratories — both public and private — participate in whole genome sequencing and this investment has radically transformed basic research and all fields that deal with biological processes. In the near future, whole genome sequencing will likely become a routine tool for microbiology diagnostic and reference laboratories.

Introduction

Genomics has seen great technological progress since the development of the first DNA sequencing methods and the publication of the first whole genome — that of the bacteriophage ΦX174 — in 1977. Today, sequencing an individual's genome only takes a few hours and costs little more than $1000, but the first whole human genome was obtained in 2003 after 13 years of work and cost roughly $2.7 billion. In bacteria alone, and only 20 years after the sequencing of the first genome of a living organism (*Haemophilus influenzae*, with a genome size of 1.83 Mb), nearly 4000 whole genomes are now available and four times this number are incomplete or being completed. Rapid progress in whole genome sequencing has been driven by the quest to understand, treat and predict diseases; to innovate in the fields of biotechnology, environment, agronomy, etc.; and also to better comprehend, on a basic science level, life and biological evolution.

A brief description of contagious equine metritis (CEM)

CEM is an economic threat for the horse industry (breeding, export, sales). This sexually transmitted and contagious bacterial infection appeared in the late 1970s in several regions around the world where horse dealing is particularly active. Today, the disease is still found worldwide, but mandatory testing has limited outbreaks to only a few cases. The clinical signs include inflammation of the endometrium in mares, which is typically accompanied by temporary infertility. Without treatment, horses (males and females) may become carriers of the disease for several years. Treatment failure seems to be frequent although none of the strains appear to show resistance to the administered antimicrobials. First classified in genus *Haemophilus*, the causative agent of CEM was reclassified in a new bacterial genus in 1985 named *Taylorella*, now made up of two species: *Taylorella equigenitalis*, which leads to an outbreak of a case of CEM, and *Taylorella asinigenitalis*, considered as non-pathogenic despite clinical signs of metritis following experimental intra-uterine infections in several mares. Despite

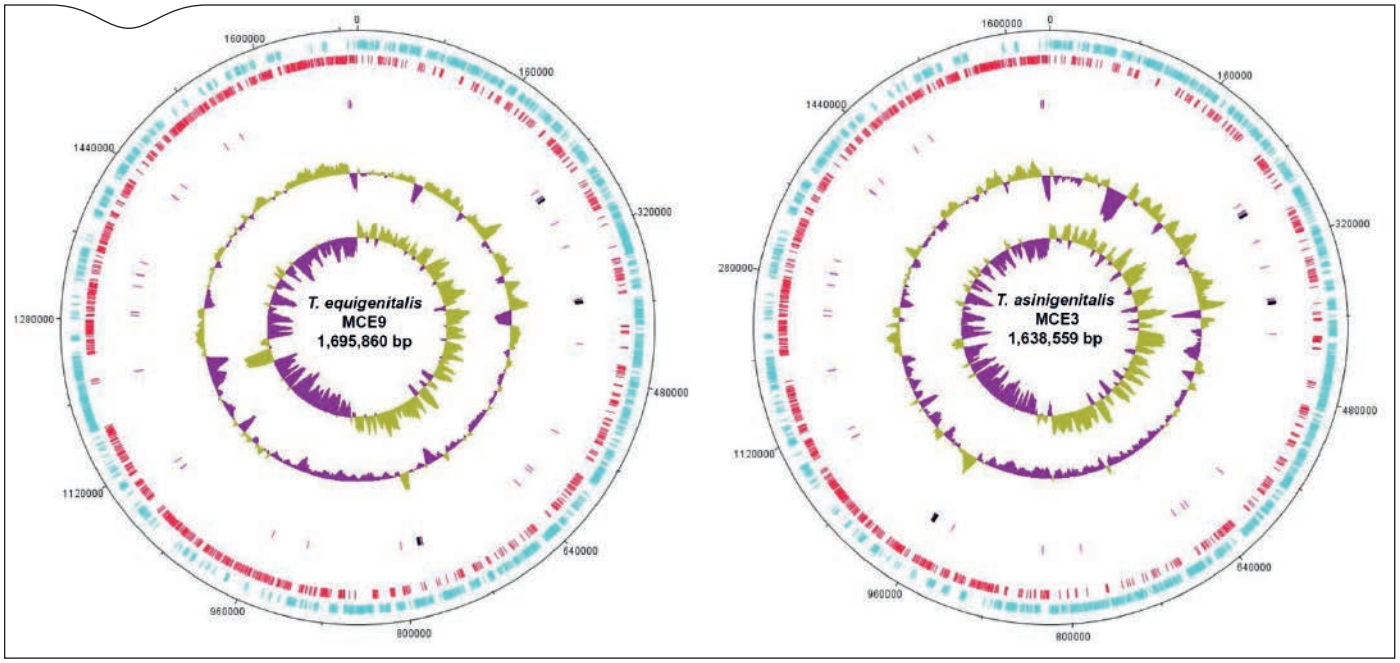

Figure 1. Genetic maps of the *T. equigenitalis* MCE9 and *T. asinigenitalis* MCE3 genomes.

Point of view

the slow growth of *Taylorella* bacteria on current culture media compared to the commensal flora of the genital tract, the official diagnosis of CEM is based on the isolation and biochemical identification of the pathogen. Moreover, the species must be confirmed by PCR because *T. equigenitalis* and *T. asinigenitalis* are phenotypically very similar.
The national reference laboratory (NRL) and European reference activities (carried out as part of the European Union Reference Laboratory for equine diseases) for CEM are based at the Bacteriology and Parasitology Unit of the ANSES Dozulé Laboratory for Equine Diseases.

Whole genome sequencing for reference activities, the case of CEM

For a reference laboratory whose role is to develop reliable tools for detecting and characterising pathogens as well as new typing methods, the lack of genetic information can drastically hamper progress. For example, in CEM, the theoretically close genetic relationship between genus *Taylorella* and other bacterial genera (e.g. *Bordetella* and *Haemophilus*) and the absence of sequence data other than ribosomal operons has limited the development of efficient diagnostic tools. To address this limitation, we undertook the *de novo* sequencing of the genomes of both *Taylorella* species in collaboration with the UMR1319 Micalis at INRA (Jouy-en-Josas, France); two high-throughput sequencing technologies were combined: 454 technology (Margulies *et al.,* 2005) coupled with "paired-end" reads (Fullwood *et al.,* 2013) to facilitate genomic assembly of both genomes, and Illumina (Solexa) technology (Bentley *et al.,* 2008). Both genomes were then annotated and compared (Hébert *et al.,* 2011; Hébert *et al.,* 2012; **Figure 1**).
The availability of genetic sequences has greatly aided the development of molecular diagnostic tools. To improve the reliability of CEM testing, we are currently collaborating to optimise a real-time multiplex PCR protocol that will improve upon the 16S ribosomal DNA sequences and therefore increase specificity. Knowledge of genome sequences also benefits the development of bacteriological or serological diagnostic methods. For example, the *in silico* reconstruction of the metabolic pathways of *Taylorella* species helps to better understand the bacteria's nutritional needs and develop a more appropriate culture medium to increase the sensitivity of the current official diagnosis method for CEM. The optimisation of this culture medium began in January 2013 in collaboration with the AES CHEMUNEX company (bioMérieux Industry, France).
There are several available molecular epidemiology tools, depending on the pathogen in question and the intended application (Sabat *et al.,* 2013). For genus *Taylorella*, we chose multilocus sequence typing (MLST), which is not only a robust, easy to perform and portable method, but also a standard method for global epidemiological surveillance of a disease and the study of the evolution of bacterial populations. Based on the annotated genomes of *T. equigenitalis* and *T. asinigenitalis*, we selected several genes that are shared between the two species and whose products are essential for cell survival. Seven of these genes were validated in 113 *T. equigenitalis* and 50 *T. asinigenitalis* strains collected in six countries over 35 years. This molecular epidemiology tool for CEM will eventually be shared with other NRLs in the European Union to determine the status of the *Taylorella* strains that circulate in Europe and to conduct retrospective studies on the available strain collections. MLST data (epidemiology of strains and DNA sequences from seven MLST markers) will be shared and made accessible in a *Taylorella*-specific database (http://pubmlst.org/taylorella/) hosted at the Department of Zoology, University of Oxford, UK (Jolley and Maiden, 2010).
Along with the development of diagnostic and molecular epidemiology tools, reference laboratories conduct research to enhance basic knowledge on pathogen biology and their infectivity in interaction with their host and the environment. The whole genome sequence will greatly help take this research forward over the long term. For example, from the comparison of the annotated genomes of *T. equigenitalis* and *T. asinigenitalis*, we have listed the potential colonisation and virulence factors that are common to both species and those that are specific to one of the two species. To shed further light on the genomic diversity of *Taylorella* species, we plan to sequence the genomes of some 10 strains in collaboration with ANSES' genomic/transcriptomic facility (at the Viral Genetics and Biosafety Unit of the ANSES Ploufragan-Plouzané Laboratory).

Routine use of whole genome sequencing

The routine use of whole genome sequencing has been addressed in several papers and review articles with, among others, a pilot study on *Staphylococcus aureus* and *Clostridium difficile* (Eyre *et al.,* 2012). All have demonstrated that the ever-decreasing costs and sequencing time have resulted in a dramatic change in the capability of microbiology diagnostic and reference laboratories. The most obvious applications are the identification of microorganisms (particularly those that are non-culturable, difficult-to-culture or highly pathogenic) and genotyping strains with optimal resolution in "real-time". The enterohemorrhagic *Escherichia coli* O104:H4 outbreak in Germany in 2011 (Mellmann *et al.,* 2011) and the recent identification of the causative agent of Theiler's disease (Chandriani *et al.,* 2013), which conventional identification techniques had failed to identify for more than a century, are good examples of these applications. Data from genome sequencing and new sequencing technologies can also be used to analyse specific target genes alone (e.g. antibiotic resistance genes) and thus reduce the cost of preparation, analysis and data storage.
It is however clear that, even though the technical constraints and costs incurred by generating sequence data are no longer limiting factors, as shown by the advent of 'push-button' sequencing, it seems difficult to envisage the routine use of whole genome sequencing as a method of diagnosis and epidemiology in the near future. Indeed, several technological advances have yet to be made to facilitate the analysis of data using software that can automatically interpret results and can be run by microbiologists who are not bioinformatics specialists, and to share data over common interfaces that are user-friendly and efficient enough to allow data incrementation and comparison with existing data. This last point implies the standardization of the formats and software from different sequencing platforms, and the suitability of infrastructures for the storage and transport of the large amounts of data generated.
In conclusion, although it is unlikely that whole genome sequencing will completely replace current diagnostic and epidemiology methods, it nevertheless represents a promising method for reference laboratories and offers greater possibilities than at present.

Point of view

Acknowledgments
I would like to thank Marie-France Breuil, Fabien Duquesne and Laurent Hébert for their daily commitment to CEM reference and research activities, the Administrative and Financial Department at the Dozulé Laboratory for Equine Diseases and the Laboratory Director, Claire Laugier. I would also like to thank L. Hébert for his active participation and his comments on this article.
List of funding sources for the research on CEM presented in this article: ANSES, the Basse-Normandie Regional Council, the European Regional Development Fund (FEDER), the French Horse and Riding Institute (*Institut Français du Cheval et de l'Equitation*), the European Union.

References
Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218): 53-9.

Chandriani S, Skewes-Cox P, Zhong W, Ganem DE, Divers TJ, Van Blaricum AJ, Tennant BC, Kistler AL. 2013. Identification of a previously undescribed divergent virus from the Flaviviridae family in an outbreak of equine serum hepatitis. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15): 1407-1415.

Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, Ip CL, Wilson DJ, Didelot X, O'Connor L, Lay R, Buck D, Kearns AM, Shaw A, Paul J, Wilcox MH, Donnelly PJ, Peto TE, Walker AS, Crook DW. 2012. A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. *BMJ Open*. 2012, 2(3): e001124.

Fullwood MJ, Lee J, Lin L, Li G, Huss M, Ng P, Sung WK, Shenolikar S. 2011. Next-generation sequencing of apoptotic DNA breakpoints reveals association with actively transcribed genes and gene translocations. *PLoS One*, 6(11): e26054.

Hébert L, Moumen B, Duquesne F, Breuil MF, Laugier C, Batto JM, Renault P, Petry S. 2011. Genome sequence of *Taylorella equigenitalis* MCE9, the causative agent of contagious equine metritis. *Journal of Bacteriology*, 193(7): 1785.

Hébert L, Moumen B, Pons N, Duquesne F, Breuil MF, Goux D, Batto JM, Laugier C, Renault P, Petry S. 2012. Genomic characterization of the *Taylorella* genus. *PLoS One*, 7(1): e29953.

Jolley KA., Maiden MC. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC bioinformatics*, 11, 595.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in open microfabricated high density picoliter reactors. *Nature*, 437(7057): 376-80.

Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, Rothberg JM, Karch H. 2011. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One*, 6(7): e22751

Sabat AJ, Budimir A, Nashev D, Sá-Leão R, van Dijl Jm, Laurent F, Grundmann H, Friedrich AW, on behalf of the ESCMID Study Group of Epidemiological Markers (ESGEM). 2013. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Eurosurveillance, 18(4): 20380.