

Titre : Génomique comparée à grande échelle de souches d'*Escherichia coli* responsables de bactériémies chez l'Homme : implications cliniques et analyse des réseaux métaboliques

Résumé: *Escherichia coli* est la bactérie aéro-anaérobie facultative majoritaire du tube digestif de l'Homme et, également, l'espèce la plus fréquemment isolée au cours de bactériémies dans les pays industrialisés. La population d'*E. coli* présente une diversité importante mais néanmoins structurée, avec certains groupes phylogénétiques préférentiellement associés à un mode de vie (e.g. A/B1 et commensal, B2/D et pathogène extraintestinal). De nombreux facteurs de virulence ont été décrits chez les souches pathogènes extraintestinales (ExPEC), mais le pronostic des bactériémies à *E. coli* semble dépendre des facteurs associés à l'hôte. Cependant, certains clones virulents et multirésistants aux antibiotiques ont récemment émergé, modifiant drastiquement l'épidémiologie de ces infections. En parallèle, la démocratisation des méthodes de séquençage offre aujourd'hui une granularité encore jamais atteinte dans l'analyse des génomes bactériens. L'objectif de ce travail de thèse est de tirer profit des méthodes de génomique comparée pour améliorer notre compréhension de la physiopathologie des bactériémies à *E. coli*, tout en prenant en compte les modifications épidémiologiques majeures qui sont observées. La réalisation de telles comparaisons génomiques a nécessité, tout d'abord, la mise en place d'une stratégie d'analyse, incluant notamment le développement d'une approche ciblée pour l'identification des séquences plasmidiques au sein d'assemblages de génomes. Cette stratégie, appliquée à 545 souches recueillies au cours de l'étude Septicoli en 2016-2017, a permis de confirmer le rôle mineur des déterminants bactériens dans l'issue des bactériémies à *E. coli*. De plus, par la comparaison de ces souches à celles de l'étude Colibafi, nous avons étudié la dynamique de la population sur 12 ans. Si celle-ci apparaît globalement stable, l'exploration plus fine des principaux clones montre d'importants remaniements, parfois associés à des facteurs de virulence typiques des ExPEC. D'autre part, la diversité antigénique de ces clones est variable et suggère des pressions de sélection différentes en fonction de leur niche écologique respective. Enfin, dans une dernière partie nous avons réalisé la reconstruction des réseaux métaboliques de plus de 1400 souches d'*E. coli* afin d'étudier les liens entre métabolisme et mode de vie. Les résultats soulignent la conservation du métabolisme chez *E. coli* et son association forte avec la phylogénie. Par ailleurs, une analyse détaillée des principaux clones retrouvés dans les bactériémies, dont notamment le ST131, met en évidence une voie métabolique impliquée dans la dégradation de composés aromatiques dérivés de la lignine. Cette voie, habituellement absente des souches de phylogroupe B2, pourrait procurer un avantage sélectif à ce clone pandémique mondial d'émergence récente.

Mots clefs : *Escherichia coli*, bactériémie, génomique comparée, réseaux métaboliques

Title:

Abstract: *Escherichia coli* is the predominant aero-anaerobic bacterium of the human gut and also the leading cause of bacteremia in industrialized countries. The *E. coli* population presents a high but structured diversity, with certain phylogenetic groups preferentially associated with a given lifestyle (e.g. A/B1 and commensal, B2/D and extraintestinal pathogen). Many virulence factors have been described in extraintestinal pathogenic strains (ExPEC), but the prognosis of *E. coli* bacteremia appears to be linked mainly to host-associated factors. However, some virulent and multi-resistant clones have recently emerged and dramatically modified the epidemiology of these infections. At the same time, the democratization of sequencing methods now offers a granularity yet never reached in the analysis of bacterial genomes. The objective of this thesis is to take advantage of comparative genomic methods to improve our understanding of the physiopathology of *E. coli* bacteremia, while taking into account the major epidemiological changes that are observed. Carrying out such genomic comparisons required, first of all, the implementation of an analysis strategy, including in particular the development of a targeted approach for the identification of plasmid sequences within genome assemblies. This strategy, applied on 545 strains collected during the Septicoli study in 2016-2017, confirmed the minor role of bacterial determinants in the outcome of *E. coli* bacteremia. In addition, by comparing these strains to those of the Colibafi study, we studied the population dynamics over 12 years. While the population appears to be stable overall, further exploration of the main clones shows significant changes, sometimes associated with virulence factors typical of ExPEC. On the other hand, the antigenic diversity of these clones is variable and suggests different selective pressures according to their respective ecological niches. Finally, in a last part we reconstructed the metabolic networks of more than 1400 *E. coli* strains in order to study the links between metabolism and lifestyle. The results highlight the conservation of metabolism in *E. coli* and its strong association with phylogeny. In addition, a detailed analysis of the main clones found in bacteremia, including the ST131, highlights a metabolic pathway involved in the degradation of aromatic compounds derived from lignin. This pathway, usually absent in phylogroup B2 strains, could provide a selective advantage to this recently emerging global pandemic clone.

Key words: *Escherichia coli*, bacteremia, comparative genomic, metabolic networks