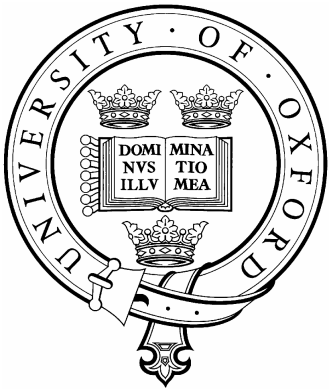
University of Oxford



Multilocus sequence analysis of the pathogen *Neisseria meningitidis*

Daniel J. Wilson

St. John's College

Mathematical Genetics and Bioinformatics

Department of Statistics

A thesis submitted for the degree of Doctor of Philosophy

September 2005

Multilocus sequence analysis of the pathogen *Neisseria meningitidis*

Daniel J. Wilson, St. John's College
D.Phil. thesis, Trinity Term 2005

ABSTRACT

Neisseria meningitidis is the bacterium responsible for meningococcal meningitis and septicaemia in humans. Meningococcal disease is primarily a disease of young children, characterized by rapid deterioration from first symptoms to death, with an 11% fatality rate and a global distribution. Patterns of genetic diversity in meningococcal populations provide an account of their evolutionary history and structure, which can be inferred by population genetics modelling. Understanding these phenomena can inform control and prevention strategies, and provides interesting case studies in evolution. The aim of this thesis is to develop population genetics techniques for inferring the evolutionary history of meningococci.

I begin by reviewing the field, and justifying the use of coalescent methods in modelling microparasite populations. Inference on carriage populations of meningococci under the standard neutral model and the neutral microepidemic model is performed using a modification to approximate Bayesian computation. AMOVA and Mantel tests are used to quantify the differentiation between carriage and disease populations, and the extent to which geography and host age structure carriage populations. The results are used to propose revised coalescent models for meningococcal evolution.

The role of natural selection in shaping meningococcal diversity is investigated using a novel method that utilises an approximation to the coalescent and reversible-jump Markov chain Monte Carlo to detect sites under selection in the presence of recombination. Having performed a simulation study to assess the statistical properties of the method, I apply it to the *porB* antigen locus and seven housekeeping loci in *N. meningitidis*. There is strong evidence for selection imposed by the host immune system in the antigen locus, but not the housekeeping loci which are functionally constrained. Finally I discuss the future direction of population genetic approaches to understanding infectious disease.

Acknowledgements

Thanks go to members of the Mathematical Genetics and Bioinformatics Group of the Department of Statistics and the Bacterial Population Structure and Public Health Group of the Department of Zoology. In particular I would like to thank Adam Auton, Ella Chase, Daniel Falush, Bob Griffiths, Rosalind Harding, Chris Holmes, Keith Jolley, Stephen Leslie, Jonathan Marchini, Noel McCarthy, Chris Spencer and Rachel Urwin. I would also like to thank Graham Coop and Don Conrad of the Human Genetics Department at the University of Chicago. Ziheng Yang kindly provided C computer code and Jeremy Derrick the molecular structure of the PorB molecule. Special thanks to Gillian Kay and my parents Brian and Janet, who in addition to support also kindly helped to proof-read the thesis. Finally I would like to thank my co-supervisors, Gil McVean and Martin Maiden.

This thesis was supported by a research studentship from the Biotechnology and Biological Sciences Research Council. Much of the computational work presented here was conducted on a multi-node AMD compute cluster that was bought with a grant awarded by the Wolfson Foundation to Peter Donnelly. Part of the work in this thesis was presented at the Society for Molecular Biology and Evolution, Newport Beach, California, June 2003, the London Mathematical Society Symposium on Mathematical Genetics, Durham, July 2004, the 14th International Pathogenic Neisseria Conference, Milwaukee, Wisconsin, September 2004, and the 10th Congress of the European Society for Evolutionary Biology, Krakow, Poland, August 2005. I would like to thank the BBSRC, the London Mathematical Society and St. John's College, Oxford for financial support in attending these conferences.

Thanks also to my examiners, Brian Charlesworth and Jotun Hein.

Table of Contents

Abstract i
Acknowledgements ii
Table of Contents iii

Chapter 1
Epidemiology of *Neisseria meningitidis* 1
1.1 Overview of *Neisseria meningitidis* 4
1.1.1 Epidemiology 4
1.1.1.1 Pathology 4
1.1.1.2 Epidemiology of meningococcal disease 6
1.1.1.3 Epidemiology of carriage 8
1.1.2 Typing 10
1.1.2.1 Immunological typing 11
1.1.2.2 Electrophoretic typing 13
1.1.2.3 Sequence typing 15
1.1.3 Control and prevention 18
1.1.3.1 Polysaccharide vaccines 19
1.1.3.2 Polysaccharide-protein conjugate vaccines 20
1.1.3.3 Outer membrane protein vesicle vaccines 22
1.2 Population biology of *Neisseria meningitidis* 24
1.2.1 The clonal complex 24
1.2.1.1 Serogroup A lineages 25
1.2.1.2 Serogroup B and C lineages 29
1.2.2 How clonal are bacteria? 30
1.2.2.1 Epidemic clone model 31
1.2.2.2 Relative contribution of recombination and mutation 33
1.2.2.3 BURST 37
1.2.3 Strain theory 42
1.2.3.1 Immune selection can structure the pathogen population 42

1.2.3.2 Evidence for meningococcal strain structure 44
1.2.4 Neutral models 45
1.2.4.1 Standard neutral model 46
1.2.4.2 Neutral microepidemic model 48
1.3 Population genetics in epidemiology 50
1.3.1 Pathogen biology 50
1.3.2 The origin and history of pathogens 52
1.3.3 Immune-mediated selection on pathogen genomes 54
1.3.4 The relevance of recombination 56
1.3.5 Phylogenetic and population genetic approaches to inference 59
1.3.6 Advantages and disadvantages of population genetics 61
1.4 Coalescent models of *Neisseria meningitidis* 63
1.4.1 Epidemiological models 63
1.4.1.1 SIS 64
1.4.1.2 SIRS 66
1.4.2 Metapopulations and the coalescent 67
1.4.2.1 The coalescent 67
1.4.2.2 The coalescent with recombination 68
1.4.2.3 Coalescence in a metapopulation 69
1.4.3 Epidemiology and the coalescent 74
1.4.3.1 SIRS with superinfection 74
1.4.3.2 Metapopulation with SIRS 77

Chapter 2
Population genetics of *Neisseria meningitidis* 81
2.1 Description of a carriage population 82
2.1.1 Diversity 83
2.1.2 Frequency distributions 87
2.1.3 Recombination 90
2.2 Fitting the standard neutral model 95
2.2.1 Composite likelihood inference 96
2.2.2 Parameter estimates 99

2.2.3 Simulating under the coalescent 103
2.2.4 Goodness-of-fit testing 105
2.3 Approximate Bayesian inference 110
2.3.1 MCMC without likelihoods 112
2.3.2 Fitting the standard neutral model 116
2.3.2.1 Update θ 118
2.3.2.2 Update κ 119
2.3.2.3 Update ρ 119
2.3.3 Parameter estimates 120
2.3.4 Bayesian cross-validation 125
2.4 Refining the model 131

Chapter 3
Genetic structuring in *Neisseria meningitidis* 134
3.1 Neutral microepidemic model 134
3.1.1 Coalescent formulation of the microepidemic model 137
3.1.2 Approximate Bayesian inference 138
3.2 Analysing population structure 140
3.2.1 Analysis of molecular variance 140
3.2.1.1 Two-way AMOVA 143
3.2.2 Mantel test 145
3.3 Geographic structuring in Europe 145
3.3.1 Structuring within the Czech Republic 147
3.3.2 Differentiation between European countries 149
3.4 Meningococcal population structure in Bavaria 152
3.4.1 Role of host age-structure 153
3.4.2 Geographic differentiation 155
3.4.2.1 Evidence for population structure 156
3.4.2.2 Evidence for isolation by distance 159
3.4.3 Institution type and genetic structure 163
3.5 Relationship between disease and carriage 165
3.6 Summary 172

3.6.1 Causes of structure in meningococcal populations 172

Chapter 4

Evolutionary Model of Immune Selection 176

4.1 The dN/dS ratio 177

4.1.1 Models that incorporate the dN/dS ratio 177

4.1.1.1 Purifying selection and dN/dS 179

4.1.1.2 Positive selection and dN/dS 182

4.1.2 Inferring immune selection using dN/dS 183

4.1.2.1 CODEML 184

4.1.2.2 MrBayes 186

4.1.2.3 SLR 187

4.1.2.4 Problems with current methods 188

4.2 Modelling selection with recombination 189

4.2.1 Population genetics inference 189

4.2.2 An approximation to the coalescent 191

4.2.2.1 Sampling formula with recombination 192

4.2.2.2 Mutation model 193

4.2.2.3 Recombination model 195

4.2.2.4 Computing the likelihood 196

4.2.3 NY98 in the coalescent approximation 197

4.2.4 An indel model for NY98 200

4.2.5 Variation in ω and ρ along a gene 202

4.3 Bayesian inference 204

4.3.1 Type A. Change ω within a block 206

4.3.2 Type B. Extend an ω block 5′ or 3′ 207

4.3.3 Types C and D. Split and Merge an ω block 207

4.3.3.1 Ratio of priors 208

4.3.3.2 Ratio of proposal probabilities 208

4.3.3.3 Ratio of density functions 209

4.3.3.4 Jacobian 210

4.3.3.5 Acceptance probabilities 211

4.3.4 Implementation 213
4.4 Simulation study 215
4.4.1 Permutation test for recombination 215
4.4.2 Simulation study A 217
4.4.3 Mixing properties of reversible jump moves 220
4.4.4 Simulation study B 223
4.5 Summary 226

Chapter 5
Evidence for Immune Selection in an Antigen of *Neisseria meningitidis* 228
5.1 Analysis of the *porB* locus 228
5.1.1 Previous analyses 230
5.1.2 Isolates 233
5.1.3 Test for recombination 234
5.1.4 Codon frequencies 235
5.1.5 Priors 236
5.1.6 Results 238
5.2 Model criticism 243
5.2.1 Prior sensitivity analysis 243
5.2.2 Posterior predictive *p*-values 246
5.2.3 Simulating under a PAC model 247
5.2.4 Combining *p*-values 248
5.2.5 Choice of statistics and results 250
5.2.6 Analysis of the global study 252
5.3 Evidence for false positives 254
5.4 Analysis of housekeeping loci 258
5.5 Summary 264

Chapter 6
Further Developments 265
6.1 Meningococcal population structure 267

6.1.1 Advantages of explicit evolutionary models....267
6.1.2 Bayesian inference in the structured coalescent....270
6.2 Detecting selection in *Neisseria meningitidis*....271
6.2.1 Comparison of PorB3 analyses....272
6.2.2 Aspects of the Bayesian approach....274
6.2.3 Limitations of the method....276
6.2.4 Extensions to the method....279
6.2.5 Implications for vaccine research....280
6.2.6 Separation of timescales in microparasite evolution....282
6.3 Summary....284

Glossary of Acronyms....286
Literature Cited....287