

# Contents

0.1	Preface	8
0.1.1	Chapter outline	8
0.1.2	Acknowledgments	9
<b>1</b>	<b>The Basic Coalescent</b>	<b>11</b>
1.1	Introduction	11
1.2	A Y-chromosome data set	15
1.3	Data and Theory	20
1.4	The Wright-Fisher Model	22
1.4.1	Assumptions of the Wright-Fisher Model	24
1.4.2	The number of descendants of a gene in one generation	25
1.4.3	An Example	26
1.5	The geometric distribution	28
1.6	The exponential distribution	30
1.7	The Discrete-Time Coalescent	32
1.7.1	Coalescence of a sample of two genes	32
1.7.2	Coalescence of a sample of $n$ genes	33
1.7.3	Example: Effect of Approximations	34
1.8	The continuous time coalescent	34
1.9	Calculating Simple Quantities on the Coalescent Tree	36
1.9.1	The Height of the Tree	36
1.9.2	The Total Branch Length of the Tree	38
1.9.3	The Effect of Sampling More Sequences	39
1.10	The Effective Population Size	39
1.11	The Moran model	41
1.12	Robustness of the Coalescent	42
1.13	Recommended readings	43
<b>2</b>	<b>From genealogies to sequences</b>	<b>45</b>
2.1	Mathematical models of alleles	46
2.1.1	The infinite alleles model	46
2.1.2	The infinite sites model	47
2.1.3	Finite sites model	50
2.2	The Wright-Fisher Model with mutation	52
2.3	Algorithms for simulating sequence evolution	54
2.4	The probability of a sample configuration	58
2.4.1	Infinite Alleles Model	59
2.4.2	Infinite Sites Model	63
2.4.3	Impossible ancestral states	70
2.5	Quantities related to the infinite sites model	73
2.5.1	The number of segregating sites	73
2.5.2	Haplotypes	75
2.5.3	Pairwise mismatch distribution	76
2.5.4	Estimators of $\theta$	76
2.6	Evolutionary versus Sampling Variance	77
2.6.1	Example 1: The variable $S_n$	78
2.6.2	Example 2: Tajima's estimator $\hat{\pi}$	79
2.7	Recommended readings	79
<b>3</b>	<b>Trees and Topologies</b>	<b>81</b>
3.1	Some terminology	81
3.1.1	The jump process and the waiting time process	81
3.1.2	The coalescent and phylogenetic trees	81
3.2	Counting trees and topologies	85
3.3	Gene Trees	88
3.3.1	How to build a gene tree	90
3.4	Nested sub-samples	91
3.5	Hanging sub-trees	93
3.5.1	Unbalanced trees	95
3.5.2	Example: Neanderthal sequences	97
3.6	A Single Lineage	97
3.7	Disjoint Subsamples	98
3.7.1	Examples	101
3.8	A sample partitioned by a mutation	102
3.8.1	Unknown ancestral state	104
3.8.2	The Age of the MRCA for two sequences	105
3.9	The probability of going from $n$ ancestors to $k$ ancestors	105
<b>4</b>	<b>Extensions to the basic coalescent</b>	<b>109</b>
4.1	Introduction	109
4.2	The coalescent with fluctuating population size	110
4.2.1	Stochastic and systematic changes	110
4.2.2	How to model population changes in the coalescent	110
4.3	Exponential growth	113
4.3.1	The genealogy under exponential growth	113
4.4	Population bottlenecks	119
4.4.1	Genealogical effect of bottlenecks	120
4.5	Effective population size revisited	121

4.6	The coalescent with population structure . . . . .	122
4.6.1	The finite island model . . . . .	123
4.6.2	The coalescent tree in the finite island model . . . . .	124
4.6.3	General models of subdivision . . . . .	129
4.6.4	Non-equilibrium models . . . . .	130
4.7	Coalescent with balancing selection . . . . .	131
4.7.1	Two allele balancing selection . . . . .	133
4.7.2	Multi-allelic balancing selection . . . . .	135
4.8	Coalescent with directional selection . . . . .	138
4.8.1	The ancestral selection graph . . . . .	138
4.9	Summary . . . . .	141
4.10	Recommended reading . . . . .	141
<b>5</b>	<b>The Coalescent with Recombination.</b>	<b>143</b>
5.1	Introduction . . . . .	143
5.2	Data example with recombination . . . . .	144
5.3	Modelling recombination . . . . .	147
5.3.1	Hudson's model of recombination . . . . .	147
5.3.2	Biological features of recombination . . . . .	149
5.4	The Wright-Fisher Model with Recombination . . . . .	154
5.5	Algorithms . . . . .	157
5.5.1	The Ancestral Recombination Graph . . . . .	157
5.5.2	Sampling ARGs: not back in time, but along sequences.162	
5.5.3	Efficiency of different algorithms . . . . .	164
5.6	The effect of a single recombination event . . . . .	167
5.7	The number of recombination events . . . . .	170
5.8	The probability of a data set. . . . .	172
5.9	The number of segregating sites . . . . .	174
5.10	The coalescent with gene conversion . . . . .	174
5.11	Gene Trees with Recombination - from incompatibilities to minimal ARGs. . . . .	176
5.11.1	Recombination as subtree transfer . . . . .	177
5.11.2	Recombination Inferred from Haplotypes . . . . .	184
5.11.3	From local to global bounds. . . . .	185
5.11.4	Minimal ARGs . . . . .	186
5.11.5	Topologies, Recombination and Compatibility . . . . .	187
5.12	Recommended reading . . . . .	190
<b>6</b>	<b>Getting Parameters from Data</b>	<b>193</b>
6.1	Introduction . . . . .	193
6.2	Estimators of $\theta$ . . . . .	194
6.2.1	Watterson's Estimator . . . . .	195
6.2.2	Tajima's Estimator . . . . .	196
6.2.3	Fu's Two Estimators . . . . .	197

6.3	Estimators of $\rho$ . . . . .	200
6.3.1	Estimators based on summary statistics . . . . .	202
6.3.2	Pseudo-Likelihood Estimators . . . . .	204
6.4	Monte Carlo Methods . . . . .	206
6.4.1	The Likelihood Surface . . . . .	208
6.4.2	Monte Carlo integration and the Coalescent . . . . .	210
6.4.3	Markov Chain Monte Carlo . . . . .	212
6.5	Recommended reading . . . . .	213
<b>7</b>	<b>LD mapping and the coalescent</b>	<b>215</b>
7.1	The potential of LD mapping . . . . .	215
7.1.1	Complex disease aetiology . . . . .	216
7.2	Linkage versus LD mapping . . . . .	218
7.3	Simulation of genealogies and LD mapping . . . . .	222
7.4	Genealogical trees around a disease mutation . . . . .	222
7.4.1	Different types of trees . . . . .	224
7.4.2	An example . . . . .	224
7.4.3	Quantifying genealogical tree differences . . . . .	225
7.5	The genealogical process reflected in data . . . . .	232
7.5.1	Linkage disequilibrium . . . . .	234
7.5.2	Measures of LD . . . . .	234
7.5.3	Which measure of LD to use? . . . . .	237
7.5.4	Testing LD . . . . .	238
7.5.5	Accounting for population admixture . . . . .	241
7.5.6	Differences between human populations . . . . .	241
7.6	Measuring association of single markers . . . . .	242
7.7	Haplotype LD mapping . . . . .	243
7.7.1	Haplotype blocks and the HapMap project . . . . .	244
7.8	Bayesian multipoint LD mapping . . . . .	247
7.8.1	Star shaped genealogy . . . . .	249
7.8.2	Coalescent based genealogy . . . . .	249
7.9	HapMap or Multipoint LD? . . . . .	252
<b>8</b>	<b>Human Evolution</b>	<b>255</b>
8.1	Our phylogenetic position and ancestral population genetics. . . . .	256
8.1.1	The number of genetic ancestors to a genome . . . . .	259
8.2	Human migrations and population structure . . . . .	265
8.2.1	Our relationship to the Neanderthaler. . . . .	266
8.2.2	Population Growth . . . . .	269
8.2.3	Structure within global modern human populations. . . . .	270
8.2.4	Specific histories . . . . .	271
8.2.5	Empirical pedigrees and The Coalescent . . . . .	272
8.2.6	Other Genealogical Issues . . . . .	276
8.2.7	Tracing genetic material within the parent genealogy. . . . .	279

<b>A WWW tools</b>	<b>283</b>
A.1 Illustrations of the coalescent . . . . .	283
A.1.1 The discrete coalescent under the Wright-Fisher model: 283	
A.1.2 The continuous time coalescent (the Hudson-animator)	283
A.2 Simulations of the coalescent . . . . .	286
A.2.1 The infinite sites model . . . . .	286
A.2.2 The finite sites model . . . . .	287
A.3 Inference . . . . .	287
A.3.1 The DataAnalyzer . . . . .	287
A.4 Illustration of the coalescent and LD mapping . . . . .	289
A.4.1 Qualitative measures . . . . .	289
A.4.2 Quantitative measures . . . . .	289

## 0.1 Preface

Coalescent Theory has gone from an obscure corner of population genetics to a central concept for anybody that studies variation at the sequence level.

Besides filling the obvious need for such a book, it is also our wish to present this theory in a straightforward and elementary manner that could dispel the misconception that Coalescent Theory is inherently very difficult and needs a strong mathematical background to understand. The key issues needed for data analysis only needs basic combinatorics. Despite the present prominence of Coalescent Theory, it also belongs to the future. From an application point of view, human evolution and association mapping/fine scale mapping are two areas that are bound to grow enormously in the next few years. And to make optimal use of the coming flood of data, theoretical advances will be needed. There are areas, where present theory fails (or is impractically slow) in presence of real data and if empirical researcher are to use Coalescent based method, there are plenty of challenges for the theoretician both in modelling and in improvement of simulation algorithms.

The present book is definitely not exhaustive, but is only meant to provide a good basis for further study.

### 0.1.1 Chapter outline

The book consists of eight Chapters:

Chapter 1 provides the basics for understanding the assumptions behind and derivation of the basic coalescent model, and some simple properties of the resulting genealogies.

Chapter 2 introduces the models of alleles and sequences and associated mutation processes. Prominent models are: Infinite alleles, infinite sites and finite sites models. When these models and mutation processes are combined with genealogies, data can either be simulated or the probability of data can be evaluated.

Chapter 3 gives some more examples of statistics that can be calculated on coalescent genealogies and mutations on such genealogies.

Chapter 4 relaxes some of the assumptions of the basic coalescent model by introducing extensions necessary in the analysis of real data. These include population size changes, population subdivision, bottlenecks, balancing selection, and directional selection.

Chapter 5 extends the coalescent model to include genetic exchange in the form of recombination and gene conversion. An introduction to the biological features of the process, the model and algorithms used and