

Contents

1	The basic coalescent	1
---	----------------------	---

1.1	Introduction	1
1.2	A Y-chromosome data set	5
1.3	Data and theory	10
1.4	The Wright–Fisher model	11
1.4.1	Assumptions of the Wright–Fisher model	13
1.4.2	The number of descendants of a gene in one generation	14
1.4.3	An example	15
1.5	The geometric distribution	17
1.6	The exponential distribution	19
1.7	The discrete-time coalescent	21
1.7.1	Coalescence of a sample of two genes	21
1.7.2	Coalescence of a sample of n genes	22
1.7.3	Example: Effect of approximations	23
1.8	The continuous time coalescent	24
1.9	Calculating simple quantities on a coalescent tree	25
1.9.1	The height of a tree	25
1.9.2	The total branch length of a tree	27
1.9.3	The effect of sampling more sequences	28
1.10	The effective population size	29
1.11	The Moran model	31
1.12	Robustness of the coalescent	32

2	From genealogies to sequences	33
---	-------------------------------	----

2.1	Mathematical models of alleles	33
2.1.1	The infinite alleles model	33
2.1.2	The infinite sites model	35
2.1.3	Finite sites model	37
2.2	The Wright–Fisher model with mutation	39
2.3	Algorithms for simulating sequence evolution	41
2.4	The probability of a sample configuration	45
2.4.1	Infinite alleles model	46
2.4.2	Infinite sites model	50

2.4.3	Impossible ancestral states	55
2.5	Quantities related to the infinite sites model	58
2.5.1	The number of segregating sites	58
2.5.2	Haplotypes	60
2.5.3	Pairwise mismatch distribution	61
2.5.4	Estimators of θ and Tajima's D	62
2.6	Evolutionary versus sampling variance	63
2.6.1	Example 1: The variable S_n	64
2.6.2	Example 2: Tajima's estimator $\hat{\pi}$	65
3	Trees and topologies	67
3.1	Some terminology	67
3.1.1	The jump process and the waiting time process	67
3.1.2	The coalescent and phylogenetic trees	67
3.2	Counting trees and topologies	70
3.3	Gene trees	72
3.3.1	How to build a gene tree	75
3.4	Nested subsamples	76
3.5	Hanging subtrees	78
3.5.1	Unbalanced trees	81
3.5.2	Example: Neanderthal sequences	81
3.6	A single lineage	82
3.7	Disjoint subsamples	83
3.7.1	Examples	86
3.8	A sample partitioned by a mutation	87
3.8.1	Unknown ancestral state	89
3.8.2	The age of the MRCA for two sequences	90
3.9	The probability of going from n ancestors to k ancestors	91
4	Extensions to the basic coalescent	95
4.1	Introduction	95
4.2	The coalescent with fluctuating population size	95
4.2.1	Stochastic and systematic changes	95
4.2.2	How to model population changes in the coalescent	96
4.3	Exponential growth	99
4.3.1	The genealogy under exponential growth	100
4.4	Population bottlenecks	104
4.4.1	Genealogical effect of bottlenecks	106
4.5	Effective population size revisited	107

4.6	The coalescent with population structure	108
4.6.1	The finite island model	108
4.6.2	The coalescent tree in the finite island model	110
4.6.3	General models of subdivision	114
4.6.4	Non-equilibrium models	116
4.7	Coalescent with balancing selection	118
4.7.1	Two allele balancing selection	118
4.7.2	Multiallelic balancing selection	120
4.8	Coalescent with directional selection	123
4.8.1	The ancestral selection graph	123
4.9	Summary	126
5	The coalescent with recombination	127
5.1	Introduction	127
5.2	Data example with recombination	128
5.3	Modelling recombination	130
5.3.1	Hudson's model of recombination	130
5.3.2	Biological features of recombination	132
5.4	The Wright-Fisher model with recombination	137
5.5	Algorithms	139
5.5.1	The ancestral recombination graph	139
5.5.2	Sampling ARGs: Not back in time, but along sequences	144
5.5.3	Efficiency of different algorithms	147
5.6	The effect of a single recombination event	148
5.7	The number of recombination events	152
5.8	The probability of a data set	153
5.9	The number of segregating sites	155
5.10	The coalescent with gene conversion	156
5.11	Gene trees with recombination—from incompatibilities to minimal ARGs	158
5.11.1	Recombination as subtree transfer	159
5.11.2	Recombination inferred from haplotypes	165
5.11.3	From local to global bounds	166
5.11.4	Minimal ARGs	167
5.11.5	Topologies, recombination, and compatibility	169
6	Getting parameters from data	173
6.1	Introduction	173
6.2	Estimators of θ	174
6.2.1	Watterson's estimator	175
6.2.2	Tajima's estimator	176
6.2.3	Fu's two estimators	178

6.3	Estimators of ρ	181
6.3.1	Estimators based on summary statistics	183
6.3.2	Pseudo-likelihood estimators	185
6.4	Monte Carlo methods	187
6.4.1	The likelihood curve	189
6.4.2	Monte Carlo integration and the coalescent	191
6.4.3	Markov chain Monte Carlo	195
7	LD mapping and the coalescent	199
7.1	The potential of LD mapping	199
7.2	Linkage versus LD mapping	200
7.3	Complex disease aetiology	202
7.4	Formulating the task	205
7.5	A role for the coalescent	206
7.6	Genealogical trees around a disease mutation	208
7.6.1	Qualitative measures	209
7.6.2	An example	210
7.6.3	Quantifying genealogical tree differences	213
7.7	The genealogical process reflected in data	216
7.8	Linkage disequilibrium (LD)	217
7.8.1	Testing for LD	220
7.8.2	Accounting for population admixture	220
7.8.3	Differences between human populations	221
7.9	Measuring association using single markers	223
7.10	Haplotype LD mapping	223
7.11	Model based LD mapping	224
7.11.1	Star shaped genealogy	225
7.11.2	Coalescent based genealogy	225
7.11.3	An example	227
7.11.4	Further challenges	228
8	Human evolution	231
8.1	Introduction	231
8.2	Our phylogenetic position and ancestral population genetics	232
8.2.1	The number of genetic ancestors to a genome	235
8.3	Human migrations and population structure	240
8.3.1	Our relationship to the Neanderthaler	242
8.3.2	Population growth	244
8.3.3	Structure within global modern human populations	244
8.3.4	Specific histories	245
8.3.5	Empirical pedigrees and the coalescent	246

8.3.6 Other genealogical issues	250
8.3.7 Tracing genetic material within the parent genealogy	252

Appendix: Web based tools	255
---------------------------	-----

Bibliography	259
--------------	-----

Index	273
-------	-----