

Contents

<i>About the authors</i>	xv
<i>Contributors</i>	xvii
<i>Foreword</i>	xix
<i>Preface</i>	xxi
<i>List of websites and resources</i>	xxv
1. Forensic genetics: the basics	1
1.1. Short tandem repeat (STR) analysis	2
1.1.1. Historical development of multiplexed systems	3
1.2. Development and harmonization of European National DNA databases	4
1.2.1. Development of the European set of standard (ESS) markers	5
1.3. Hardy–Weinberg equilibrium	6
1.3.1. Measuring deviation from Hardy–Weinberg equilibrium	8
1.3.2. Extension to multiple alleles in STRs	10
1.4. Quality assurance of data	11
1.5. Recap: the laws of statistics	12
1.5.1. Effect of conditioning on probabilities	13
1.6. The likelihood ratio	14
1.7. Simple mixtures interpretation: the basics	15
1.7.1. Nomenclature	15
1.7.2. Minimum number of contributors	15
1.7.3. Interpretation of two-contributor mixtures	17
1.7.4. Step 1: condition the number of contributors	18
1.7.5. Step 2: state the alternative propositions	18
1.7.6. Step 3: evaluate the probability of the evidence under the defense proposition	18
1.7.7. Step 4: evaluate the probability of the crime-stain evidence under the prosecution proposition	19
1.7.8. Step 5: calculate the likelihood ratio	20
1.7.9. Step 6: report the likelihood ratio	20
1.8. Three allele mixtures	21
1.8.1. Calculation of the likelihood ratio	21
1.9. Two allele mixtures	23
1.10. Multiple contributors	24
1.11. Automating calculations with a computer program	24
1.11.1. Two-contributors	24
1.11.2. Recap	27
1.11.3. Extension to three or more contributors	27
1.12. The effect of population sub-structuring	29
1.12.1. What is a population?	29
1.12.2. Genetic drift	31
1.12.3. Common ancestry	32

1.12.4.	A computer simulation to illustrate effects of genetic drift	35
1.12.5.	Extension to multi-allelic systems (STRs)	36
1.12.6.	Balding and Nichols F_{ST}	37
1.13.	Simulation studies to show the effect of genetic drift and F_{ST} corrections	39
1.13.1.	Details of the experiment	39
1.14.	Sampling error correction	41
1.15.	European Network of Forensic Science Institutes (ENFSI) evaluation of STR markers	42
1.15.1.	Hardy–Weinberg tests	43
1.15.2.	F_{ST} estimation	44
1.15.3.	What F_{ST} to apply in casework?	45
1.16.	Relatedness: kinship analysis	46
1.16.1.	Effect of relatedness on the likelihood ratio calculation	46
1.16.2.	Formal treatment and extension to mixtures	47
1.17.	A brief introduction to the software approaches used to interpret mixtures	49
1.18.	Summary	52
	Notes	53
2.	Empirical characterization of DNA profiles	55
2.1.	Heterozygote balance	55
2.1.1.	Peak height or peak area	55
2.1.2.	Definitions	55
2.1.3.	Methods to visualize data	56
2.1.4.	The use of guidelines	56
2.1.5.	Heterozygote balance summary	60
2.1.6.	Characterization of heterozygote balance with <i>STR-validator</i>	60
2.2.	Stutters	62
2.2.1.	Characterization of stutters	62
2.2.2.	Effect of number of STR repeats on the stutter size	63
2.2.3.	Interpretation challenges with stutters	65
2.2.4.	Stutters summary	66
2.2.5.	Characterization of stutters with <i>STR-validator</i>	66
2.3.	The Clayton guideline: a binary model to interpret simple mixtures	70
2.3.1.	Step 1: identify the presence of a mixture	70
2.3.2.	Step 2: identify the potential number of contributors to a sample	73
2.3.3.	Step 3: determine the approximate ratio of the components in the mixture	74
2.3.4.	Step 4: determine the possible pairwise combinations for the components of the mixture	76
2.3.5.	Step 5: compare with reference samples	82
2.4.	Characterization of massively parallel sequencing (MPS) kits using <i>STR-validator</i>	86
2.5.	Summary	88
3.	Allele drop-out and the stochastic threshold	89
3.1.	What causes allele drop-out?	89
3.1.1.	A mathematical model showing the relationship of heterozygote balance and drop-out	92

3.2.	Stochastic thresholds and drop-out	93
3.3.	National DNA databases and low-template DNA crime stains	94
3.4.	Definition of drop-out and extreme drop-out	95
3.5.	Towards a meaningful definition of the low-template-DNA (T) threshold	96
3.6.	Historical development of probabilistic evaluation of drop-out	98
3.7.	Introducing probability of drop-out	99
3.7.1.	The defense proposition	100
3.7.2.	The prosecution proposition	100
3.8.	A method to calculate probability of drop-out and the stochastic threshold	102
3.8.1.	Experimental design and data collection	102
3.8.2.	An explanation of logistic regression	103
3.9.	Using <i>STR-validator</i> to plot drop-out characteristics	105
3.10.	Summary	107
	Notes	110
4.	Low-template DNA	111
4.1.	Historical	111
4.2.	The consensus method explained	112
4.2.1.	Contamination vs. drop-in	115
4.3.	Statistical analysis	118
4.4.	Working with replicates	118
4.4.1.	An example to assess the strength of the evidence of three replicates	119
4.5.	Degradation	122
4.5.1.	Pre-PCR quantification and characterization of degradation	123
4.6.	Summary	127
	Note	128
5.	<i>LRmix</i> model theory	129
5.1.	Background	129
5.2.	Model description	130
5.3.	Theoretical considerations	131
5.4.	Replicate probability	133
5.4.1.	Drop-out and drop-in: definitions	133
5.4.2.	Drop-out and drop-in: formalization	134
5.4.3.	Example 1	135
5.4.4.	Example 2	136
5.5.	Genotype probability	137
5.5.1.	Determining the genotypes for the unknowns	137
5.5.2.	Correcting for population sub-structuring	141
5.5.3.	Example 3	142
5.5.4.	Calculating genotype probabilities for relatedness	143
5.5.5.	Example with relatedness in mixtures	146
5.6.	Full example for calculating LR	146
5.7.	Summary	149
	Notes	150

6. A qualitative (semi-continuous) model: <i>LRmix Studio</i>	153
6.1. Interpretation of a major/minor DNA mixture, where the minor contributor is evidential	153
6.1.1. Case circumstances	153
6.1.2. Exploratory data analysis	154
6.1.3. The propositions	154
6.1.4. Sensitivity analysis	157
6.1.5. Calculation of levels of drop-out using an empirical method	158
6.1.6. Non-contributor analysis	163
6.1.7. How the <i>LR</i> is used in a statement	166
6.2. A case with two suspects	167
6.2.1. Case circumstances	167
6.2.2. Exploratory data analysis	167
6.2.3. Minimum number of contributors	167
6.2.4. Non-contributor tests	168
6.2.5. Formulation of propositions and exploratory data analysis	169
6.3. Considerations of the ISFG DNA commission	174
6.3.1. Formulation of propositions: summary	174
6.4. An example of a replicate analysis	175
6.5. Relatedness	177
6.6. Summary	178
Notes	179
7. The quantitative (continuous) model theory	181
7.1. Introduction	181
7.2. Towards a quantitative model	181
7.2.1. An example using the "MasterMix" Excel spreadsheet program	185
7.2.2. Using the information across loci	190
7.3. Likelihood ratio estimation using probabilistic models	191
7.3.1. The normal distribution model	192
7.3.2. Step 1: calculation of expected peak heights	192
7.3.3. Step 2: calculate the weighting	193
7.3.4. Step 3: combine weightings with probability of genotype given the proposition	194
7.3.5. Estimating the parameters of the normal distribution: maximum likelihood estimation (MLE)	196
7.4. Recap	198
7.5. The Gamma model	199
7.5.1. Gamma distribution	199
7.5.2. Model for one contributor	200
7.5.3. Reparameterization	201
7.5.4. An example	201
7.6. Drop-out for the gamma model explained	205
7.6.1. The <i>Q</i> allele	206
7.6.2. Comparison of the Excel spreadsheet with <i>EuroForMix</i> : a summary	210
7.7. Deconvolution	212
7.8. Degradation	215

7.8.1.	Demonstration with a spreadsheet	219
7.8.2.	Recap	219
7.9.	Stutter	220
7.9.1.	Recap: taking account of peak height expectations using mixture proportions M_x	222
7.9.2.	The full stutter model	223
7.9.3.	A worked example showing the impact of stutter in <i>EuroForMix</i>	223
7.9.4.	Combining stutter and degradation in the same model	226
7.9.5.	Using a spreadsheet model to illustrate stutter	226
7.9.6.	Dealing with forward and complex stutters	234
7.10.	The drop-in model	235
7.10.1.	Estimating lambda using <i>EuroForMix</i>	236
7.10.2.	Characterization of drop-in	236
7.11.	Summary	237
	Notes	238
8.	<i>EuroForMix</i>	239
8.1.	<i>EuroForMix</i> theory	239
8.2.	Interpretation using <i>EuroForMix</i>	239
8.3.	Theoretical considerations	240
8.4.	Model features	240
8.5.	An example	241
8.5.1.	ENFSI exercise 1: a major/minor profile, where the minor component is evidential	241
8.5.2.	Reference samples	243
8.5.3.	Preliminary assessment of the evidence	243
8.5.4.	Model specification	244
8.5.5.	The MLE fit tab	246
8.5.6.	Model selection	248
8.5.7.	Model validation	251
8.5.8.	Probability-probability, PP-plots	251
8.5.9.	<i>LR</i> sensitivity	254
8.5.10.	Model fitted peak height tab	255
8.5.11.	Non-contributor analysis	255
8.5.12.	Deconvolution	257
8.6.	ENFSI exercise 2: illustration of the advantage of using the quantitative model	258
8.6.1.	Case circumstances	258
8.6.2.	Effect of removing the conditioned suspect from the analysis	261
8.7.	A complex case: robbery	263
8.7.1.	Case circumstances	263
8.7.2.	Analysis	263
8.7.3.	<i>EuroForMix</i> analysis	266
8.8.	Relatedness calculations	268
8.8.1.	Results and discussion	271

8.9. Important updates to <i>EuroForMix</i>	272
8.10. A summary of the interpretation process	274
Notes	275
9. Validation	277
9.1. Context	277
9.2. Avoiding the black-box	278
9.3. Model validation	279
9.3.1. Conceptual validation	279
9.3.2. Operational validation	280
9.4. Defining benchmarks for <i>LR</i> -based models	281
9.4.1. Software validation	281
9.5. Validation study of <i>LRmix Studio</i> and <i>EuroForMix</i>	282
9.5.1. NGM DNA profiles	283
9.5.2. PPF6C DNA profiles	283
9.5.3. Replicates	284
9.5.4. Stutter filter	284
9.5.5. Allele frequency database	285
9.6. Design of experiments	285
9.6.1. Experiments using NGM profiles	285
9.6.2. Experiments using PPF6C profiles	285
9.7. Method	286
9.7.1. Defining characteristics of the models	286
9.8. Comparison of <i>LRmix Studio</i> and <i>EuroForMix</i> when the POI was the true contributor	287
9.9. Comparison of <i>LRmix</i> and <i>EuroForMix</i> when the POI was not a true contributor	288
9.10. Characterization of false positive results	290
9.10.1. NGM results	290
9.10.2. PPF6C results	291
9.11. Characterization of false negative results	293
9.12. LR values as a function of allele drop-out	294
9.12.1. NGM results	294
9.12.2. PPF6C results	296
9.13. Comparing the performance of two or more different models using ROC plots	296
9.14. Comparison of the stutter model in <i>EuroForMix</i> versus GeneMapper stutter filter	297
9.15. Calibration of the likelihood ratio	299
9.16. Further comparative studies between different software	302
9.17. Defining guidelines for using probabilistic genotyping software in forensic casework	304
9.18. Summary	306
Notes	308
10. Development and implementation of DNAXs	309
The DNAXs Team	
10.1. DNA expert system	309
10.1.1. Functionalities of DNAXs	310

10.2. <i>LoCIM-tool</i>	310
10.3. <i>DNAStatistX</i>	315
10.3.1. Similarities and differences between <i>EuroForMix</i> and <i>DNAStatistX</i>	316
10.3.2. Optimizer to estimate parameters	317
10.3.3. Number of optimizer iterations	319
10.3.4. Overall versus dye-specific detection threshold	320
10.3.5. Model validation	321
10.3.6. <i>DNAXs</i> software testing	323
10.4. An exemplar case analyzed with <i>DNAXs-DNAStatistX</i> and <i>EuroForMix</i>	324
10.4.1. Analysis using <i>DNAXs-DNAStatistX</i>	324
10.4.2. Comparison of the exemplar using <i>EuroForMix</i>	332
10.5. Summary	337
11. Investigative forensic genetics: <i>SmartRank</i>, <i>CaseSolver</i> and <i>DNAMatch2</i>	339
11.1. National DNA databases	339
11.2. When is evaluative reporting appropriate?	341
11.3. A cautionary tale	342
11.4. Current methods used to compare crime stains with national DNA databases	342
11.4.1. Prüm inclusion rules	343
11.4.2. Prüm matching rules	343
11.4.3. CODIS inclusion and matching rules	344
11.5. Limitations of traditional national DNA databases: introducing " <i>SmartRank</i> "	344
11.6. Experimental details to test efficiency of <i>SmartRank</i>	345
11.6.1. <i>SmartRank</i> performance	346
11.7. <i>SmartRank</i> versus CODIS	347
11.7.1. Specifying a likelihood ratio threshold and a top ranked candidate list	349
11.7.2. Limitations	350
11.8. <i>SmartRank</i> exercise	352
11.9. Using <i>EuroForMix</i> for database searching	354
11.10. <i>CaseSolver</i> : an expert system based on <i>EuroForMix</i>	356
11.10.1. Method	358
11.10.2. Mixture comparison	359
11.10.3. Model parameters	362
11.11. Demonstration using a real case example	362
11.11.1. Case circumstances	362
11.11.2. Importing data to CS	364
11.11.3. Comparing references to mixtures	365
11.11.4. Viewing the results	366
11.11.5. Deconvolution	367
11.11.6. Relatedness searches	368
11.11.7. Advanced options	369
11.11.8. More detailed summary of results	370
11.11.9. Searching a large fictive national DNA database	373
11.12. The need for new digital tools in forensic genetics: improving efficiency of searching large DNA databases	374

11.12.1.	Why use a stepwise strategy to extract candidate matches?	378
11.12.2.	The flexibility of CS	378
11.12.3.	Searching large national DNA databases	379
11.12.4.	Using <i>DNAmatch2</i> to search large databases and as a contamination search engine	380
11.12.5.	Future improvements for estimating the number of contributors	382
11.13.	Summary	382
	Notes	383
12.	Interpretation, reporting, and communication	385
12.1.	Early methods used to express the value of the evidence and their limitations	385
12.2.	An outline of the important principles to interpret scientific evidence	387
12.3.	Propositions	388
12.4.	Scope of the forensic scientist as an investigator or evaluator	389
12.5.	Hierarchy of propositions	390
12.6.	Formulation of propositions to avoid bias	393
12.7.	Activity level propositions	394
12.8.	The prosecutor's fallacy and prior probabilities	395
12.8.1.	Prior odds can be updated	397
12.9.	Avoidance of the prosecutor's fallacy	397
12.10.	Database searches	399
12.11.	Statement writing	400
12.12.	The use of the verbal scale and the "weak evidence effect"	401
12.13.	Using terms like "inconclusive" and "excluded" as expressions in statements	403
12.14.	Conditioning and the application to DNA mixtures	404
12.15.	The effect of relatedness on the likelihood ratio	406
12.16.	Assigning the number of contributors	407
12.17.	Non-contributor tests	407
12.17.1.	The "two suspect" problem	411
12.18.	Summary	416
	Notes	417
13.	Interpretation of complex DNA profiles generated by massively parallel sequencing	419
	Peter Gill, Rebecca Just, Walther Parson, Chris Phillips, Øyvind Bleka	
13.1.	Introduction	419
13.2.	SNP analysis	420
13.3.	Number of contributors	421
13.4.	Effect of choosing the wrong number of contributors	421
13.5.	Comparison of quantitative and qualitative methods	422
13.6.	Limitations	422
13.7.	The effect of uncertainty of the number of contributors on the likelihood ratio	425
13.8.	Summary	425
13.9.	Short tandem repeats	427

13.9.1. Nomenclature	427
13.10. Historical perspective: lessons of the past	429
13.11. STR-MPS stutters	431
13.11.1. Extension to LUS+ nomenclature	433
13.12. Characterization of stutters with MPS	435
13.13. MPS programming using <i>EuroForMix</i>	437
13.13.1. Automated conversion of sequence into RU/LUS/LUS+ nomenclature	437
13.14. Demonstration of likelihood ratio calculations using <i>EuroForMix</i>	440
13.14.1. Analytical threshold (<i>AT</i>)	440
13.14.2. Rules for stutter filtering	440
13.14.3. Noise	442
13.14.4. Experimental summary	442
13.15. Automating the interpretation strategy	442
13.15.1. <i>EuroForMix</i> analysis	444
13.15.2. Stutter filter effect	446
13.16. Information gain: measuring the effect of the stutter filter and different nomenclatures	446
13.16.1. Information gain: comparing stutter filter vs no stutter filter	449
13.17. Summary	450
Notes	451

A. Formal descriptions of the genotype probabilities 453

A.1. Extending the F_{ST} -formula to mixtures	453
A.2. Relatedness	454
A.2.1. Formulae for relatedness	454
A.2.2. Extension with the F_{ST} -formula	455
A.2.3. Specific relatedness formulae F_{ST} -correction	456

B. Formal description of the probabilistic models 459

B.1. Definitions	459
B.2. Extension to parameterized statistical models	460
B.2.1. The likelihood for parameterized statistical models	460
B.2.2. LR calculations for parameterized statistical models	461
B.3. Mathematical details of the probabilistic models	462
B.3.1. The contribution from assumed genotypes	462
B.3.2. Mathematical details of the qualitative model	463
B.3.3. Inference of the drop-out parameter	464
B.3.4. Mathematical details of the quantitative model	465
B.3.5. The gamma distribution	466
B.3.6. Model for a single contributor	466
B.3.7. Model for multiple contributors	468
B.3.8. Model for allele drop-out	470
B.3.9. Model for allele drop-in	471
B.3.10. Model for multiple replicates	472

B.3.11. Model for backward-stutters

472

B.3.12. Estimating the *LR* in *EuroForMix*

473

B.4. Deconvolution

475

Bibliography

477

Index

497