Table of Contents

Preface 1

Chapter 1: Python and the Surrounding Software Ecology 9
Introduction 9
Installing the required software with Anaconda 10
Getting ready 10
How to do it... 13
There's more... 14
Installing the required software with Docker 14
Getting ready 14
How to do it... 15
See also 16
Interfacing with R via rpy2 16
Getting ready 16
How to do it... 17
There's more... 23
See also 23
Performing R magic with Jupyter Notebook 24
Getting ready 24
How to do it... 24
There's more... 26
See also 26

Chapter 2: Next-Generation Sequencing 27
Introduction 27
Accessing GenBank and moving around NCBI databases 28
Getting ready 29
How to do it... 29
There's more... 33
See also 34
Performing basic sequence analysis 34
Getting ready 34
How to do it... 35
There's more... 36
See also 37
Working with modern sequence formats 37
Getting ready 37
How to do it... 38
There's more... 44
See also 45

Working with alignment data 46
Getting ready 46
How to do it... 47
There's more... 52
See also 53
Analyzing data in VCF 54
Getting ready 54
How to do it... 55
There's more... 56
See also 57
Studying genome accessibility and filtering SNP data 57
Getting ready 58
How to do it... 59
There's more... 69
See also 70
Processing NGS data with HTSeq 70
Getting ready 71
How to do it... 71
There's more... 74

Chapter 3: Working with Genomes 75
Introduction 75
Working with high-quality reference genomes 76
Getting ready 76
How to do it... 77
There's more... 82
See also 82
Dealing with low-quality genome references 82
Getting ready 83
How to do it... 83
There's more... 87
See also 88
Traversing genome annotations 88
Getting ready 88
How to do it... 89
There's more... 90
See also 91
Extracting genes from a reference using annotations 91
Getting ready 92
How to do it... 92
There's more... 94
See also 95
Finding orthologues with the Ensembl REST API 95
Getting ready 95
How to do it... 96

There's more... 99
Retrieving gene ontology information from Ensembl 99
Getting ready 99
How to do it... 100
There's more... 103
See also 104

Chapter 4: Population Genetics 105
Introduction 105
Managing datasets with PLINK 106
Getting ready 107
How to do it... 108
There's more... 112
See also 113
Introducing the Genepop format 113
Getting ready 114
How to do it... 114
See also 118
Exploring a dataset with Bio.PopGen 118
Getting ready 119
How to do it... 119
There's more... 124
See also 124
Computing F-statistics 124
Getting ready 124
How to do it... 125
See also 130
Performing Principal Components Analysis 131
Getting ready 131
How to do it... 131
There's more... 135
See also 136
Investigating population structure with admixture 136
Getting ready 136
How to do it... 137
There's more... 142

Chapter 5: Population Genetics Simulation 143
Introduction 143
Introducing forward-time simulations 144
Getting ready 144
How to do it... 144
There's more... 150
Simulating selection 150
Getting ready 151
How to do it... 151

There's more... 157
Simulating population structure using island and stepping-stone models 157
Getting ready 157
How to do it... 158
Modeling complex demographic scenarios 163
Getting ready 163
How to do it... 164
Chapter 6: Phylogenetics 171
Introduction 171
Preparing a dataset for phylogenetic analysis 172
Getting ready 172
How to do it... 172
There's more... 177
See also 178
Aligning genetic and genomic data 178
Getting ready 178
How to do it... 178
Comparing sequences 180
Getting ready 180
How to do it... 180
There's more... 185
Reconstructing phylogenetic trees 185
Getting ready 185
How to do it... 186
There's more... 190
Playing recursively with trees 190
Getting ready 190
How to do it... 191
There's more... 195
Visualizing phylogenetic data 195
Getting ready 196
How to do it... 196
There's more... 202
Chapter 7: Using the Protein Data Bank 203
Introduction 203
Finding a protein in multiple databases 204
Getting ready 204
How to do it... 205
There's more... 208
Introducing Bio.PDB 208
Getting ready 209
How to do it... 209

There's more... 213
Extracting more information from a PDB file 213
Getting ready 213
How to do it... 214
Computing molecular distances on a PDB file 217
Getting ready 218
How to do it... 218
Performing geometric operations 222
Getting ready 222
How to do it... 222
There's more... 225
Animating with PyMOL 225
Getting ready 225
How to do it... 226
There's more... 231
Parsing mmCIF files using Biopython 232
Getting ready 232
How to do it... 232
There's more... 233

Chapter 8: Bioinformatics Pipelines 235
Introduction 235
Introducing Galaxy servers 236
Getting ready 236
How to do it... 237
There's more... 239
Accessing Galaxy using the API 239
Getting ready 239
How to do it... 241
Developing a Galaxy tool 247
Getting ready 247
How to do it... 248
There's more... 250
Using generic pipelines with bioinformatics data 251
Getting ready 251
How to do it... 251
Deploying a variant analysis pipeline with Airflow 253
Getting ready 254
How to do it... 254
There's more... 260

Chapter 9: Python for Big Genomics Datasets 261
Introduction 261
Using high-performance data formats – HDF5 262
Getting ready 262

How to do it... 263
There's more... 267
Doing parallel computing with Dask 267
Getting ready 268
How to do it... 268
There's more... 271
Using high-performance data formats – Parquet 272
Getting ready 272
How to do it... 272
There's more... 273
Computing sequencing statistics using Spark 274
Getting ready 274
How to do it... 275
There's more... 276
Optimizing code with Cython and Numba 277
Getting ready 277
How to do it... 277
There's more... 281
Chapter 10: Other Topics in Bioinformatics 283
Introduction 283
Doing metagenomics with the QIIME 2 Python API 284
Getting ready 284
How to do it... 286
There's more... 289
Inferring shared chromosomal segments with Germline 289
Getting ready 289
How to do it... 291
There's more... 294
Accessing the Global Biodiversity Information Facility via REST 294
How to do it... 295
There's more... 300
Georeferencing GBIF datasets 301
Getting ready 301
How to do it... 301
There's more... 306
Plotting protein interactions with Cytoscape the hard way 307
Getting ready 307
How to do it... 308
There's more... 313
Chapter 11: Advanced NGS Processing 315
Introduction 315
Preparing the dataset for analysis 316
Getting ready 316

How to do it... 317
Using Mendelian error information for quality control 322
How to do it... 322
There's more... 326
Using decision trees to explore the data 326
How to do it... 327
Exploring the data with standard statistics 329
How to do it... 329
There's more... 334
Finding genomic features from sequencing annotations 334
How to do it... 335
There's more... 337
Index 339