Table of Contents

Preface xiii

1

Python and the Surrounding Software Ecology 1

Installing the required basic software with Anaconda 2
Getting ready 2
How to do it... 4
There's more... 5

Installing the required software with Docker 7
Getting ready 7
How to do it... 8
See also 8

Interfacing with R via rpy2 9
Getting ready 9
How to do it... 10
There's more... 15
See also 16

Performing R magic with Jupyter 16
Getting ready 16
How to do it... 17
There's more... 18
See also 18

2

Getting to Know NumPy, pandas, Arrow, and Matplotlib 19

Using pandas to process vaccine-adverse events 20
Getting ready 20
How to do it... 20
There's more... 25
See also 26

Dealing with the pitfalls of joining pandas DataFrames 26
Getting ready 26
How to do it... 27
There's more... 29

Reducing the memory usage of pandas DataFrames 29
Getting ready 29
How to do it… 29
See also 32

Accelerating pandas processing with Apache Arrow 32

Getting ready 33

How to do it... 33

There's more... 35

Understanding NumPy as the engine behind Python data science and bioinformatics 36

Getting ready 36

How to do it… 36

See also 39

Introducing Matplotlib for chart generation 39

Getting ready 40

How to do it... 40

There's more... 47

See also 47

3

Next-Generation Sequencing 49

Accessing GenBank and moving around NCBI databases 50

Getting ready 50

How to do it... 51

There's more... 53

See also 54

Performing basic sequence analysis 55

Getting ready 55

How to do it... 55

There's more... 56

See also 57

Working with modern sequence formats 57

Getting ready 57

How to do it... 58

There's more... 64

See also 65

Working with alignment data 66

Getting ready 66

How to do it... 66

There's more... 72

See also 72

Extracting data from VCF files 73

Getting ready 73

How to do it... 74

There's more... 75

See also 76

Studying genome accessibility and filtering SNP data 76

Getting ready 76

How to do it... 78

There's more... 88

See also 88

Processing NGS data with HTSeq 88

Getting ready 89

How to do it... 90

There's more... 92

4

Advanced NGS Data Processing 93

Preparing a dataset for analysis 93
Getting ready 94
How to do it… 94
Using Mendelian error information for quality control 101
How to do it… 101
There's more… 105
Exploring the data with standard statistics 106
How to do it… 106
There's more… 111
Finding genomic features from sequencing annotations 111
How to do it… 111
There's more… 114
Doing metagenomics with QIIME 2 Python API 114
Getting ready 114
How to do it... 116
There's more... 119

5

Working with Genomes 121

Technical requirements 121
Working with high-quality reference genomes 122
Getting ready 122
How to do it... 123
There's more... 127
See also 128
Dealing with low-quality genome references 128
Getting ready 128
How to do it... 129
There's more... 133
See also 134
Traversing genome annotations 134
Getting ready 134
How to do it... 134
There's more... 136
See also 137
Extracting genes from a reference using annotations 137
Getting ready 137
How to do it... 138
There's more... 140
See also 140
Finding orthologues with the Ensembl REST API 141
Getting ready 141
How to do it... 141
There's more... 144
Retrieving gene ontology information from Ensembl 144
Getting ready 144
How to do it... 145
There's more... 149
See also 149

6

Population Genetics 151

Managing datasets with PLINK 152
Getting ready 152
How to do it... 154
There's more... 158
See also 158
Using sgkit for population genetics analysis with xarray 158
Getting ready 159
How to do it... 159
There's more... 163
Exploring a dataset with sgkit 163
Getting ready 163
How to do it... 163
There's more... 167
See also 167
Analyzing population structure 167
Getting ready 168
How to do it... 168
See also 174
Performing a PCA 174
Getting ready 174
How to do it... 175
There's more... 177
See also 177
Investigating population structure with admixture 177
Getting ready 177
How to do it... 178
There's more... 183

7

Phylogenetics 185

Preparing a dataset for phylogenetic analysis 185
Getting ready 186
How to do it... 186
There's more... 192
See also 192
Aligning genetic and genomic data 192
Getting ready 192
How to do it... 193
Comparing sequences 195
Getting ready 195
How to do it... 195
There's more... 200
Reconstructing phylogenetic trees 200
Getting ready 200
How to do it... 201
There's more... 204
Playing recursively with trees 205
Getting ready 205
How to do it... 205
There's more... 209
Visualizing phylogenetic data 210
Getting ready 210
How to do it... 210
There's more... 215

8

Using the Protein Data Bank 217

Finding a protein in multiple databases 218
Getting ready 218
How to do it... 218
There's more 222

Introducing Bio.PDB 222
Getting ready 223
How to do it... 223
There's more 228

Extracting more information from a PDB file 228
Getting ready 228
How to do it... 228

Computing molecular distances on a PDB file 232
Getting ready 233
How to do it... 233

Performing geometric operations 237
Getting ready 237
How to do it... 237
There's more 240

Animating with PyMOL 241
Getting ready 241
How to do it... 241
There's more 247

Parsing mmCIF files using Biopython 247
Getting ready 247
How to do it... 247
There's more 248

9

Bioinformatics Pipelines 249

Introducing Galaxy servers 250
Getting ready 250
How to do it… 250
There's more 252

Accessing Galaxy using the API 252
Getting ready 252
How to do it… 254

Deploying a variant analysis pipeline with Snakemake 260
Getting ready 260
How to do it… 261
There's more 266

Deploying a variant analysis pipeline with Nextflow 267
Getting ready 267
How to do it… 268
There's more 272

10

Machine Learning for Bioinformatics 273

Introducing scikit-learn with a PCA example 273

Getting ready 274

How to do it... 274

There's more... 276

Using clustering over PCA to classify samples 276

Getting ready 277

How to do it... 277

There's more... 282

Exploring breast cancer traits using Decision Trees 282

Getting ready 283

How to do it... 283

Predicting breast cancer outcomes using Random Forests 286

Getting ready 286

How to do it… 286

There's more... 289

11

Parallel Processing with Dask and Zarr 291

Reading genomics data with Zarr 292

Getting ready 292

How to do it... 292

There's more... 297

See also 297

Parallel processing of data using Python multiprocessing 297

Getting ready 297

How to do it... 298

There's more... 299

See also 300

Using Dask to process genomic data based on NumPy arrays 300

Getting ready 300

How to do it... 301

There's more... 305

See also 305

Scheduling tasks with dask. distributed 305

Getting ready 305

How to do it... 307

There's more... 311

See also 311

12

Functional Programming for Bioinformatics 313

Understanding pure functions 314

Getting ready 314

How to do it... 314

There's more... 316

Understanding immutability 316

Getting ready 317

How to do it... 317

There's more... 318

Avoiding mutability as a robust development pattern 318

Getting ready 319

How to do it... 319

There's more... 320

Using lazy programming for pipelining 321

Getting ready 321

How to do it... 321

There's more... 323

The limits of recursion with Python 323

Getting ready 324

How to do it... 324

There's more... 326

A showcase of Python's functools module 326

Getting ready 326

How to do it... 326

There's more... 328

See also... 328

Index 329

Other Books You May Enjoy 338