

Table of Contents

Preface

xiii

1

Python and the Surrounding Software Ecology 1

Installing the required basic software with Anaconda	2	Interfacing with R via rpy2	9
Getting ready	2	Getting ready	9
How to do it...	4	How to do it...	10
There's more...	5	There's more...	15
		See also	16
Installing the required software with Docker	7	Performing R magic with Jupyter	16
Getting ready	7	Getting ready	16
How to do it...	8	How to do it...	17
See also	8	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	26
Getting ready	20	How to do it...	27
How to do it...	20	There's more...	29
There's more...	25		
See also	26	Reducing the memory usage of pandas DataFrames	29
Dealing with the pitfalls of joining pandas DataFrames	26	Getting ready	29
		How to do it...	29
		See also	32

Accelerating pandas processing with Apache Arrow	32	Getting ready	36
Getting ready	33	How to do it...	36
How to do it...	33	See also	39
There's more...	35	Introducing Matplotlib for chart generation	39
Understanding NumPy as the engine behind Python data science and bioinformatics	36	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	How to do it...	66
Getting ready	50	There's more...	72
How to do it...	51	See also	72
There's more...	53	Extracting data from VCF files	73
See also	54	Getting ready	73
Performing basic sequence analysis	55	How to do it...	74
Getting ready	55	There's more...	75
How to do it...	55	See also	76
There's more...	56	Studying genome accessibility and filtering SNP data	76
See also	57	Getting ready	76
Working with modern sequence formats	57	How to do it...	78
Getting ready	57	There's more...	88
How to do it...	58	See also	88
There's more...	64	Processing NGS data with HTSeq	88
See also	65	Getting ready	89
Working with alignment data	66	How to do it...	90
Getting ready	66	There's more...	92

4

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

5

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

6

Population Genetics **151**

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

7

Phylogenetics **185**

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

8

Using the Protein Data Bank 217

Finding a protein in multiple databases	218	Getting ready	233
Getting ready	218	How to do it...	233
How to do it...	218	Performing geometric operations	237
There's more	222	Getting ready	237
Introducing Bio.PDB	222	How to do it...	237
Getting ready	223	There's more	240
How to do it...	223	Animating with PyMOL	241
There's more	228	Getting ready	241
Extracting more information from a PDB file	228	How to do it...	241
Getting ready	228	There's more	247
How to do it...	228	Parsing mmCIF files using Biopython	247
Computing molecular distances on a PDB file	232	Getting ready	247
		How to do it...	247
		There's more	248

9

Bioinformatics Pipelines 249

Introducing Galaxy servers	250	Deploying a variant analysis pipeline with Snakemake	260
Getting ready	250	Getting ready	260
How to do it...	250	How to do it...	261
There's more	252	There's more	266
Accessing Galaxy using the API	252	Deploying a variant analysis pipeline with Nextflow	267
Getting ready	252	Getting ready	267
How to do it...	254	How to do it...	268
		There's more	272

10

Machine Learning for Bioinformatics 273

Introducing scikit-learn with a PCA example	273	Exploring breast cancer traits using Decision Trees	282
Getting ready	274	Getting ready	283
How to do it...	274	How to do it...	283
There's more...	276	Predicting breast cancer outcomes using Random Forests	286
Using clustering over PCA to classify samples	276	Getting ready	286
Getting ready	277	How to do it...	286
How to do it...	277	There's more...	289
There's more...	282		

11

Parallel Processing with Dask and Zarr 291

Reading genomics data with Zarr	292	Using Dask to process genomic data based on NumPy arrays	300
Getting ready	292	Getting ready	300
How to do it...	292	How to do it...	301
There's more...	297	There's more...	305
See also	297	See also	305
Parallel processing of data using Python multiprocessing	297	Scheduling tasks with dask. distributed	305
Getting ready	297	Getting ready	305
How to do it...	298	How to do it...	307
There's more...	299	There's more...	311
See also	300	See also	311

12

Functional Programming for Bioinformatics		313
Understanding pure functions	314	Using lazy programming for
Getting ready	314	pipelining
How to do it...	314	Getting ready
There's more...	316	How to do it...
		There's more...
Understanding immutability	316	The limits of recursion with Python
Getting ready	317	323
How to do it...	317	Getting ready
There's more...	318	How to do it...
		There's more...
Avoiding mutability as a robust	318	A showcase of Python's functools
development pattern		module
Getting ready	319	326
How to do it...	319	Getting ready
There's more...	320	How to do it...
		There's more...
		See also...
		328
Index		329