

Contents

<i>Preface</i>	v
Chapter 1 Preparing Your Computing Environment	1
1.1 Buying Your Own Computer	1
1.2 Setting up a Computing Server	4
1.3 Establishing a Remote Connection to a Server	6
Chapter 2 Learning Basic Linux Commands	17
2.1 No Need to be a Linux Guru to use Linux Effectively	17
2.2 Folder (Directory) Operations	18
<i>Controlling your command prompt</i>	25
2.3 File Operations	32
2.4 Assignment of Permissions	39
<i>The path</i>	46
2.5 Understanding System Status	47
<i>UNIX redirection and pipes</i>	52
2.6 Other Useful Commands	54
Chapter 3 Checking Sequence Quality	59
3.1 Basic High-throughput Sequencing	59
3.2 Challenges of High Throughput Genome Sequencing	61

3.3 Standards of Quality Score	62
3.4 Quality Check	65
<i>FastQC</i>	65
<i>FASTX-Toolkit</i>	72
Chapter 4 Sequence Alignment	81
4.1 The Purpose of Sequence Alignment	81
<i>Sequence assembly</i>	83
4.2 Selection of the Sequence Alignment Tools	83
<i>Burrows Wheeler</i>	85
<i>The BWT encoding-decoding algorithm</i>	88
4.3 Actual Operation of the Sequence Alignment	90
<i>Download and installation of Bowtie</i>	90
<i>Executing sequence alignment</i>	96
4.4 Sequence Alignment Results File Conversion	99
<i>Downloading SAMtools</i>	99
4.5 Using the Genome Browser	108
Chapter 5 Speeding-up with GPUs	117
5.1 Computational Advantages of the Graphics Card	117
5.2 Industry Standards and Usage of GPU Computing	119
5.3 Practical CUDA Applications in Bioinformatics	138
<i>Preparing the reference sequence</i>	140
<i>Alignment with CUSHAW2-GPU</i>	143
5.4 The Reason for the Limited Success of GPUs	145
Chapter 6 Establishing a Research Workflow Pipeline	147
6.1 Automating Your Computational Workflow	147
6.2 Scripting Language	148
<i>Script command</i>	150
6.3 Testing and Debugging	157
<i>Keeping track of the current project</i>	158
<i>Complementing tests of code blocks</i>	159
<i>Calculating the execution time</i>	160
6.4 Implementation Case Studies	162
6.5 Case Study of Common Mistakes	170
<i>Mistake 1: Confusing mess of relative paths</i>	170
<i>Mistake 2: Failure to change the necessary permissions</i>	172

<i>Mistake 3: The disk becomes full during execution</i>	172
<i>Mistake 4: Ignoring cross-platform shell portability considerations</i>	174
Chapter 7 Using a Bioinformatics Cloud Computing Platform	177
7.1 Simple Introduction to the Cloud Computing Platform	177
7.2 Amazon Web Service	178
7.3 Bioinformatics Cloud Computing Platforms	182
<i>Logging in to use Galaxy services</i>	184
<i>Uploading sequence data</i>	187
<i>Sequence quality testing</i>	195
<i>Execution of sequence alignment</i>	202
<i>Selecting other Galaxy servers</i>	205
<i>Design and use of research workflows</i>	207
<i>Establishing new research workflows</i>	207
<i>Sharing and publishing process</i>	209
<i>Execution of research workflows</i>	212
<i>Downloading or exporting research workflows</i>	212
<i>Importing research workflows</i>	214
7.4 Installing and Setting up your Own Galaxy Server	215
<i>Downloading the latest version of the Galaxy</i>	216
<i>Starting your Galaxy server</i>	217
<i>Allowing external execution</i>	220
<i>Installation of bioinformatics tools</i>	220
<i>Adding new reference sequences</i>	229
Appendix Learning Regular Expressions through Practising Simple Data Processing	235
Regular Expressions	236
<i>One character pattern match</i>	236
<i>Numbering a file and printing line number of a hit</i>	236
<i>Counting number of grepped hits</i>	237
<i>UNIX redirection using pipes</i>	237
<i>Grep and output several lines of context around the hit</i>	237
<i>Grepping for non-matching lines</i>	238
<i>Grepping for unwanted characters</i>	238
<i>Mistake of logic</i>	238
<i>Egrep or grep -E extended regular expression grep</i>	239
<i>Egrep and the character class</i>	239
<i>Egrep character class negation</i>	240

<i>Regular expression: Beginning of line anchor ^</i>	241
<i>Case-sensitive and case-insensitive grep</i>	241
<i>Regular expression: End of line anchor</i>	242
<i>More about regular expressions</i>	242
<i>Even more regular expression</i>	244
<i>Substitution with SED Awk and Perl</i>	244
<i>Using Excel to do data processing</i>	250
<i>Index</i>	259