

Table of Contents

Foreword.....	xiii
Preface.....	xvii
1. Introduction.....	1
The Promises and Challenges of Big Data in Biology and Life Sciences	3
Infrastructure Challenges	3
Toward a Cloud-Based Ecosystem for Data Sharing and Analysis	4
Cloud-Hosted Data and Compute	5
Platforms for Research in the Life Sciences	6
Standardization and Reuse of Infrastructure	8
Being FAIR	10
Wrap-Up and Next Steps	11
2. Genomics in a Nutshell: A Primer for Newcomers to the Field.....	13
Introduction to Genomics	13
The Gene as a Discrete Unit of Inheritance (Sort Of)	14
The Central Dogma of Biology: DNA to RNA to Protein	16
The Origins and Consequences of DNA Mutations	18
Genomics as an Inventory of Variation in and Among Genomes	19
The Challenge of Genomic Scale, by the Numbers	20
Genomic Variation	21
The Reference Genome as Common Framework	21
Physical Classification of Variants	24
Germline Variants Versus Somatic Alterations	29
High-Throughput Sequencing Data Generation	32
From Biological Sample to Huge Pile of Read Data	33
Types of DNA Libraries: Choosing the Right Experimental Design	37

Data Processing and Analysis	41
Mapping Reads to the Reference Genome	41
Variant Calling	43
Data Quality and Sources of Error	48
Functional Equivalence Pipeline Specification	51
Wrap-Up and Next Steps	52
3. Computing Technology Basics for Life Scientists.....	53
Basic Infrastructure Components and Performance Bottlenecks	54
Types of Processor Hardware: CPU, GPU, TPU, FPGA, OMG	54
Levels of Compute Organization: Core, Node, Cluster, and Cloud	55
Addressing Performance Bottlenecks	56
Parallel Computing	60
Parallelizing a Simple Analysis	60
From Cores to Clusters and Clouds: Many Levels of Parallelism.....	61
Trade-Offs of Parallelism: Speed, Efficiency, and Cost	63
Pipelining for Parallelization and Automation.....	64
Workflow Languages	66
Popular Pipelining Languages for Genomics.....	66
Workflow Management Systems	67
Virtualization and the Cloud	68
VMs and Containers	69
Introducing the Cloud	72
Categories of Research Use Cases for Cloud Services	74
Wrap-Up and Next Steps	77
4. First Steps in the Cloud.....	79
Setting Up Your Google Cloud Account and First Project	79
Creating a Project	80
Checking Your Billing Account and Activating Free Credits	81
Running Basic Commands in Google Cloud Shell	84
Logging in to the Cloud Shell VM	84
Using gsutil to Access and Manage Files	85
Pulling a Docker Image and Spinning Up the Container	89
Mounting a Volume to Access the Filesystem from Within the Container	92
Setting Up Your Own Custom VM	94
Creating and Configuring Your VM Instance	94
Logging into Your VM by Using SSH	100
Checking Your Authentication	102
Copying the Book Materials to Your VM	103
Installing Docker on Your VM	105
Setting Up the GATK Container Image	106

Stopping Your VM...to Stop It from Costing You Money	108
Configuring IGV to Read Data from GCS Buckets	109
Wrap-Up and Next Steps	113
5. First Steps with GATK.....	115
Getting Started with GATK	115
Operating Requirements	116
Command-Line Syntax	117
Multithreading with Spark	118
Running GATK in Practice	121
Getting Started with Variant Discovery	125
Calling Germline SNPs and Indels with HaplotypeCaller	125
Filtering Based on Variant Context Annotations	135
Introducing the GATK Best Practices	143
Best Practices Workflows Covered in This Book	145
Other Major Use Cases	145
Wrap-Up and Next Steps	145
6. GATK Best Practices for Germline Short Variant Discovery.....	147
Data Preprocessing	147
Mapping Reads to the Genome Reference	149
Marking Duplicates	151
Recalibrating Base Quality Scores	153
Joint Discovery Analysis	155
Overview of the Joint Calling Workflow	156
Calling Variants per Sample to Generate GVCFs	160
Consolidating GVCFs	162
Applying Joint Genotyping to Multiple Samples	164
Filtering the Joint Callset with Variant Quality Score Recalibration	166
Refining Genotype Assignments and Adjusting Genotype Confidence	171
Next Steps and Further Reading	172
Single-Sample Calling with CNN Filtering	173
Overview of the CNN Single-Sample Workflow	175
Applying 1D CNN to Filter a Single-Sample WGS Callset	176
Applying 2D CNN to Include Read Data in the Modeling	178
Wrap-Up and Next Steps	180
7. GATK Best Practices for Somatic Variant Discovery.....	183
Challenges in Cancer Genomics	183
Somatic Short Variants (SNVs and Indels)	185
Overview of the Tumor-Normal Pair Analysis Workflow	186
Creating a Mutect2 PoN	188

Running Mutect2 on the Tumor-Normal Pair	190
Estimating Cross-Sample Contamination	191
Filtering Mutect2 Calls	193
Annotating Predicted Functional Effects with Funcotator	195
Somatic Copy-Number Alterations	197
Overview of the Tumor-Only Analysis Workflow	198
Creating a Somatic CNA PoN	201
Applying Denoising	202
Performing Segmentation and Call CNAs	204
Additional Analysis Options	207
Wrap-Up and Next Steps	208
8. Automating Analysis Execution with Workflows	209
Introducing WDL and Cromwell	210
Installing and Setting Up Cromwell	212
Your First WDL: Hello World	216
Learning Basic WDL Syntax Through a Minimalist Example	216
Running a Simple WDL with Cromwell on Your Google VM	218
Interpreting the Important Parts of Cromwell's Logging Output	219
Adding a Variable and Providing Inputs via JSON	222
Adding Another Task to Make It a Proper Workflow	223
Your First GATK Workflow: Hello HaplotypeCaller	226
Exploring the WDL	226
Generating the Inputs JSON	229
Running the Workflow	231
Breaking the Workflow to Test Syntax Validation and Error Messaging	233
Introducing Scatter-Gather Parallelism	236
Exploring the WDL	237
Generating a Graph Diagram for Visualization	242
Wrap-Up and Next Steps	244
9. Deciphering Real Genomics Workflows	245
Mystery Workflow #1: Flexibility Through Conditionals	245
Mapping Out the Workflow	246
Reverse Engineering the Conditional Switch	251
Mystery Workflow #2: Modularity and Code Reuse	257
Mapping Out the Workflow	257
Unpacking the Nesting Dolls	262
Wrap-Up and Next Steps	268
10. Running Single Workflows at Scale with Pipelines API	269
Introducing the GCP Genomics Pipelines API Service	269

Enabling Genomics API and Related APIs in Your Google Cloud Project	270
Directly Dispatching Cromwell Jobs to PAPI	271
Configuring Cromwell to Communicate with PAPI	272
Running Scattered HaplotypCaller via PAPI	275
Monitoring Workflow Execution on Google Compute Engine	277
Understanding and Optimizing Workflow Efficiency	280
Granularity of Operations	281
Balance of Time Versus Money	282
Suggested Cost-Saving Optimizations	284
Platform-Specific Optimization Versus Portability	286
Wrapping Cromwell and PAPI Execution with WDL Runner	287
Setting Up WDL Runner	288
Running the Scattered HaplotypCaller Workflow with WDL Runner	288
Monitoring WDL Runner Execution	290
Wrap-Up and Next Steps	293
11. Running Many Workflows Conveniently in Terra	295
Getting Started with Terra	295
Creating an Account	296
Creating a Billing Project	298
Cloning the Preconfigured Workspace	301
Running Workflows with the Cromwell Server in Terra	302
Running a Workflow on a Single Sample	302
Running a Workflow on Multiple Samples in a Data Table	305
Monitoring Workflow Execution	311
Locating Workflow Outputs in the Data Table	316
Running the Same Workflow Again to Demonstrate Call Caching	318
Running a Real GATK Best Practices Pipeline at Full Scale	320
Finding and Cloning the GATK Best Practices Workspace for Germline	
Short Variant Discovery	320
Examining the Preloaded Data	321
Selecting Data and Configuring the Full-Scale Workflow	323
Launching the Full-Scale Workflow and Monitoring Execution	324
Options for Downloading Output Data—or Not	327
Wrap-Up and Next Steps	328
12. Interactive Analysis in Jupyter Notebook	331
Introduction to Jupyter in Terra	332
Jupyter Notebooks in General	332
How Jupyter Notebooks Work in Terra	334
Getting Started with Jupyter in Terra	340
Inspecting and Customizing the Notebook Runtime Configuration	341

Opening Notebook in Edit Mode and Checking the Kernel	346
Running the Hello World Cells	347
Using gsutil to Interact with Google Cloud Storage Buckets	350
Setting Up a Variable Pointing to the Germline Data in the Book Bucket	351
Setting Up a Sandbox and Saving Output Files to the Workspace Bucket	352
Visualizing Genomic Data in an Embedded IGV Window	353
Setting Up the Embedded IGV Browser	354
Adding Data to the IGV Browser	355
Setting Up an Access Token to View Private Data	357
Running GATK Commands to Learn, Test, or Troubleshoot	358
Running a Basic GATK Command: HaplotypeCaller	359
Loading the Data (BAM and VCF) into IGV	360
Troubleshooting a Questionable Variant Call in the Embedded IGV Browser	363
Visualizing Variant Context Annotation Data	365
Exporting Annotations of Interest with VariantsToTable	365
Loading R Script to Make Plotting Functions Available	366
Making Density Plots for QUAL by Using makeDensityPlot	367
Making a Scatter Plot of QUAL Versus DP	370
Making a Scatter Plot Flanked by Marginal Density Plots	371
Wrap-Up and Next Steps	372
13. Assembling Your Own Workspace in Terra.....	373
Managing Data Inside and Outside of Workspaces	373
The Workspace Bucket as Data Repository	374
Accessing Private Data That You Manage Outside of Terra	374
Accessing Data in the Terra Data Library	377
Re-Creating the Tutorial Workspace from Base Components	378
Creating a New Workspace	378
Adding the Workflow to the Methods Repository and Importing It into the Workspace	380
Creating a Configuration Quickly with a JSON File	382
Adding the Data Table	384
Filling in the Workspace Resource Data Table	386
Creating a Workflow Configuration That Uses the Data Tables	387
Adding the Notebook and Checking the Runtime Environment	389
Documenting Your Workspace and Sharing It	390
Starting from a GATK Best Practices Workspace	390
Cloning a GATK Best Practices Workspace	391
Examining GATK Workspace Data Tables to Understand How the Data Is Structured	391
Getting to Know the 1000 Genomes High Coverage Dataset	394

Copying Data Tables from the 1000 Genomes Workspace	396
Using TSV Load Files to Import Data from the 1000 Genomes Workspace	398
Running a Joint-Calling Analysis on the Federated Dataset	400
Building a Workspace Around a Dataset	407
Cloning the 1000 Genomes Data Workspace	407
Importing a Workflow from Dockstore	408
Configuring the Workflow to Use the Data Tables	411
Wrap-Up and Next Steps	412
14. Making a Fully Reproducible Paper.....	413
Overview of the Case Study	413
Computational Reproducibility and the FAIR Framework	414
Original Research Study and History of the Case Study	416
Assessing the Available Information and Key Challenges	417
Designing a Reproducible Implementation	419
Generating a Synthetic Dataset as a Stand-In for the Private Data	421
Overall Methodology	422
Retrieving the Variant Data from 1000 Genomes Participants	424
Creating Fake Exomes Based on Real People	425
Mutating the Fake Exomes	430
Generating the Definitive Dataset	432
Re-Creating the Data Processing and Analysis Methodology	432
Mapping and Variant Discovery	433
Variant Effect Prediction, Prioritization, and Variant Load Analysis	435
Analytical Performance of the New Implementation	436
The Long, Winding Road to FAIRness	438
Final Conclusions	440
Glossary.....	441
Index.....	445