



Genomics in the AWS[®] Cloud

Analyzing Genetic Code
Using Amazon Web Services

Catherine Vacher • David Wall

WILEY

สำนักหอสมุด มหาวิทยาลัยเชียงใหม่

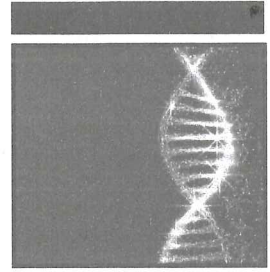
๒ 16703959

๐ 12576716

i 226 84396

Genomics in the AWS[®] Cloud





Genomics in the **AWS®** Cloud

Analyzing Genetic Code Using
Amazon Web Services

Catherine Vacher
David Wall

WILEY



Contents at a Glance

Introduction	xix
Chapter 1 Why Do Genome Analysis Yourself When Commercial Offerings Exist?	1
Chapter 2 A Crash Course in Molecular Biology	9
Chapter 3 Obtaining Your Genome	25
Chapter 4 The Bioinformatics Workflow	39
Chapter 5 AWS Services for Genome Analysis	59
Chapter 6 Building Your Environment in the AWS Cloud	77
Chapter 7 Linux and AWS Command-Line Basics for Genomics	115
Chapter 8 Processing the Sequencing Data	143
Chapter 9 Visualizing the Genome	211
Chapter 10 Containerizing Your Workflow on the Desktop	235
Chapter 11 Variants and Applications	249
Chapter 12 Cancer Genomics	267
Index	291

Contents

Introduction	xix
Chapter 1 Why Do Genome Analysis Yourself When Commercial Offerings Exist?	1
Commercial Sequencing Services	2
Typical Results	3
Summary	8
Chapter 2 A Crash Course in Molecular Biology	9
DNA	9
DNA at Work: RNA and Proteins	13
Inheritance	20
Summary	23
Chapter 3 Obtaining Your Genome	25
Preparing to Have Your Genome Sequenced	25
Can It Affect My Insurance?	25
Privacy	26
Humility and Levelheadedness	26
Validation with a Clinically Accredited Test	26
Alternatives to Using Your Own Genome	27
Specifying Lab Work	27
Depth	27
Sample Type	28
Type of Output Files	28
Sequencing Technology	28
Genome vs. Exome vs. SNP Arrays	30
Engaging a Laboratory	30
Getting a Tissue Sample for DNA Extraction	31
Rules and Regulations	32

	Do-It-Yourself Phlebotomy	33
	Legal Considerations	34
	Shipping the Sample	35
	Receiving the Results	36
	Sequences and Quality Control Information	36
	Alignment Information	37
	Variation Information	38
	Summary	38
Chapter 4	The Bioinformatics Workflow	39
	Extraction of DNA	40
	Deriving Nucleated Cells from Whole Blood	40
	Processing Nucleated Cells	41
	FASTA Files	41
	FASTQ Files	42
	Phred Scores	44
	ASCII Encoding of Phred Scores	44
	Alignment to a Reference Genome	46
	Reference Genomes	48
	Quality Control	49
	Trimming	50
	The Alignment Process	51
	Marking Duplicates	53
	Recalibrating Base Quality Score	53
	Calling SNVs and Indel Variants	54
	Annotating SNVs and Indel Variants	55
	Prioritizing Variants	56
	Inheritance Analysis	56
	Identifying SVs and CNVs	57
	Bioinformatics Workflow	58
	Summary	58
Chapter 5	AWS Services for Genome Analysis	59
	General Concepts	61
	Networking	61
	AWS Functionalities	61
	AWS Accounts	61
	Virtual Private Cloud	62
	Subnets	63
	Elastic IP Addresses	65
	Custom Environments	65
	Storage	66
	S3	67
	Glacier	67
	Computing	68
	Elastic Compute Cloud	68
	Containers	70
	Lambda Functions	73

	Workflow Management	74
	AWS Batch	74
	AWS Step Functions	74
	Simple Workflow Service	75
	Third-Party Solutions	75
	Summary	75
Chapter 6	Building Your Environment in the AWS Cloud	77
	Setting Up a Virtual Private Cloud	77
	Setting Up and Launching an EC2 Instance	82
	Shutting Down an Instance to Save Money	91
	Setting Up S3 Buckets	91
	Configuring Your Account Securely	95
	Turning On Multifactor Authentication	97
	Establishing an AWS IAM Password Policy	101
	Creating Groups	102
	Creating Users	105
	Setting Up Your Client Environment	106
	Connecting to an EC2 Instance	106
	Connecting from macOS or Unix/Linux	108
	Connecting from Windows	109
	Making S3 Buckets Available Locally	110
	Mounting an S3 Bucket as a Windows Drive	111
	Mounting an S3 Bucket Under macOS and Linux	111
	Summary	113
Chapter 7	Linux and AWS Command-Line Basics for Genomics	115
	Selecting a Linux Distribution	115
	Accessing Your AWS Linux Instance from Your	
	Local Computer	118
	From Windows	118
	From macOS	120
	Options for Setting Up Linux on Your Personal Computer	120
	Getting Familiar with the Command Line	123
	Absolute and Relative References	124
	Manipulating Files	126
	Transferring Files to and from Your AWS Instance	127
	Keyboard Shortcuts	128
	Running Programs in the Background	128
	Understanding File Permissions	129
	Compressing and Archiving Files	130
	Compression	131
	Grep	132
	Pipes and Redirection Operators	132
	Text Processing Utilities: awk and sed	133
	Managing Linux	135
	Package Management Systems	135

	The AWS Command-Line Interface	135
	Installing the AWS CLI Environment	136
	Windows	136
	macOS and Linux	137
	Configuring the AWS CLI	137
	Setting the Configuration at the Command Line	138
	Storing the Configuration in the Configuration File	138
	Testing Your Installation	139
	AWS CLI Essentials	139
	An Alternative Approach: AWS Systems Manager	140
	Summary	141
Chapter 8	Processing the Sequencing Data	143
	Getting from Data to Information	143
	Aligning to the Reference Genome	145
	Making Adjustments and Refinements to the Aligned Reads in the BAM File	150
	Identifying the Small Differences and Recording Them in the VCF File	155
	Making Adjustments and Refinements to the Variants in the VCF File	157
	Annotating the SNVs and Indels	160
	Prioritizing the Variants to Identify the Most Consequential Ones	162
	Trio Analysis and Inheritance Analysis	164
	Identifying and Annotating SVs and CNVs	167
	Setting Up AWS Services and Data Storage	172
	Copying the FASTQ Files	196
	Installing Docker and Containers	197
	Summary	210
Chapter 9	Visualizing the Genome	211
	Introducing Genome Visualizers	211
	Installing the IGV Desktop Visualizer	214
	Connecting the IGV Visualizer to Our AWS Data	216
	Loading Data into the IGV Visualizer	220
	Visualizing Aligned Sequencing Reads in IGV	226
	Have a CIGAR	229
	Analyzing Variants in IGV	230
	Summary	233
Chapter 10	Containerizing Your Workflow on the Desktop	235
	Introducing Containerization	235
	Understanding and Using Docker	239
	Installing Docker on Your Local Machine	240
	Downloading a Docker Image	241
	Viewing Available Docker Images	242
	Running a Docker Container Interactively	242

Removing a Docker Image	243
More on Using the Docker Hub	244
Containers for Genomics Work	244
Summary	248
Chapter 11 Variants and Applications	249
Polygenic Risk Scores	249
Genome-wide Association Studies	249
Calculating a Polygenic Score	251
Metagenomics	254
AlphaFold	255
Predicting Protein Structure from Protein Sequence—A 50-Year Puzzle	256
Installing and Running AlphaFold	258
Viewing and Comparing AlphaFold Results	261
Summary	266
Chapter 12 Cancer Genomics	267
Somatic Genomes	267
Cancer	268
Oncogenes	268
Tumor Suppressors	269
The Promise and Reality of Cancer Precision Medicine	270
Somatic or Germline? Cancer Predisposition	273
Chromothripsis	274
Epigenetics of Cancer	275
Mechanisms of Cancer	276
Samples	279
Somatic Variant Analysis	279
Copy Number Changes	284
Measuring Tumor Genomic Instability	287
Summary	288
Notes	289
Index	291