

全ゲノム・エクソーム遺伝統計解析

Python, Rで実践して身につく, 未知の遺伝要因の探索と疾患リスク予測

CONTENTS

はじめに.....	3
執筆者一覧.....	9
本書の使い方.....	10

第1章 NGS解析 ～ バリエント検出 14

1 研究目的に応じたデータ取得 —— WGSかWESか？ 必要なデータ量は？	牧野悟士 14
--	---------

基礎知識と背景.....	14
リシークエンシングか <i>de novo</i> シークエンシングか？ 15／委託とインハウスのコスト 15／シークエンシング方法は？ 16／適切なリード長は？ 16／どのプラットフォームを選ぶのか？ 16／データ取得量は？ 17／カバレッジは？ 17／WESかWGSか？ 18／データQCは？ 19	
おわりに.....	21

2 データ解析環境に関する基礎知識 —— ラップトップかスパコンか？	船山貴光 22
--	---------

基礎知識と背景.....	22
一般的なスパコンの構成 23／NGS解析に適しているのは、ラップトップとスパコンどちらか？ 24	
練習環境の整備.....	25
おわりに.....	27
Column ① NGS技術による「WES + diversity SNP」パネルのデータ取得.....	28

3 全ゲノムシークエンシングのバリエントコール —— トリオ家系のゲノム解析	仁宮洗太, 高山 順 29
--	---------------

基礎知識と背景.....	29
全ゲノム解析の背景 29／全ゲノム解析手順の全体像 30／解析対象データについて 30	
全ゲノム解析における各手順の解説.....	31
基準ゲノムの選択 32／プリプロセス 32／バリエントコール 33	
解析の準備.....	34
解析の実行.....	40
おわりに.....	52

4 構造多型の検出 — smoooveを用いたSVコール 高山 順 54

基礎知識と背景	54
構造多型 54 / smoooveの機能概要 55	
解析の準備	56
テストデータの解析実行とソフトウェア実行方法の理解	57
おわりに	60

5 コピー数バリエーションの検出 — GATK-gCNVを用いたCNVコール 高山 順 62

基礎知識と背景	62
CNVについて 62 / GATK-gCNVのモデル 62 / GATK-gCNVの機能概要 63	
解析の準備	64
テストデータの解析実行とソフトウェア実行方法の理解	65
おわりに	69
Column ② 集団の祖先性を見直す	70

第2章 一次処理 ～ アノテーション 71

1 bcftoolsを用いたVCFの整形 早坂将聖, 高山 順 71

基礎知識と背景	71
bcftoolsを用いてVCFファイルを編集する理由 71	
解析の準備	72
解析の実行	74
おわりに	80
Column ③ コアレスセント理論と遺伝的祖先性	81

2 SnpEffとSnpSiftによるVCFのアノテーションと操作 仁宮洸太, 高山 順 82

基礎知識と背景	82
アノテーションとは 82 / SnpEffとSnpSift 83 / pathogenicityに注目したバリエーションの効果予測 83 / pathogenicityの判定におけるエビデンス 84	
解析の準備	86
解析の実行	89
おわりに	93

3 vcfannoによるVCFのアノテーションと操作 高山 順 95

基礎知識と背景	95
アノテーションの統合が必要な理由 95 / vcfannoの機能概要 96	
解析の準備	96
テストデータの解析実行とソフトウェア実行方法の理解	100
おわりに	101

1 slivarによる希少難病家系のVCF分析

高山 順 102

基礎知識と背景	102
疾患原因バリエント同定のための絞り込み 102 / slivarの機能概要 103	
解析の準備	103
テストデータの解析実行とソフトウェア実行方法の理解	105
おわりに	108
Column ④ NGSの重要な応用：NICU / PICUでの迅速ゲノム検査	108

2 Exomiserによる希少難病家系のVCF分析

高山 順 109

基礎知識と背景	109
ゲノム以外の情報を用いた病因バリエントの絞り込み 109 / 表現型の機械可読な表現 109 / Exomiserの特徴 111	
解析の準備	111
テストデータの解析実行とソフトウェア実行方法の理解	112
おわりに	118

**3 症例対照研究における
レアバリエント関連解析**

森井 航, 秦 千比呂, 椎橋卓哉 119

基礎知識と背景	119
レアバリエント関連解析の重要性 119 / レアバリエント関連解析について 120	
解析の準備	120
解析データの加工	122
解析の実行	125
応用例	130
おわりに	131
Column ⑤ 混交を定量する統計量	132

**4 STAARpipelineによる多因子遺伝疾患の
レアバリエント関連解析**

小嶋崇史, 岡田随象 133

基礎知識と背景	133
レアバリエント解析の遺伝統計学的な考え方 133 / STAARpipelineの特徴 134	
解析の準備	136
解析の実行	138
結果の解釈方法	147
おわりに	148

第4章

原因／リスク遺伝子同定後の解析

150

1 希少疾患の原因遺伝子同定に向けた
大規模データの利用

浅海 真, 椎橋卓哉 150

基礎知識と背景	150
Genomics Englandとは 151 / データアクセスと解析環境：GEL Research EnvironmentとAirlock process 151 / Genomics Englandデータの内容 152 / データアクセスまでの手続き 158	
おわりに	159

2 PRS / PGS 計算

田端佑介, 三宅顕光, 成田 暁 161

基礎知識と背景	161
PRS / PGS 計算の流れ 163 / 品質管理 (QC) 165 / トレーニング (計算モデルの構築) 167 / バリデーション (ハイパーパラメータの最適化) 169 / テスト (PRS / PGSの精度評価) 169	
PRS / PGS 計算の手法	170
C+T法 170 / LDpred2 172 / PRS-CS 173 / Lassosum 174 / STMGP 175	
解析の準備	177
解析の実行	180
おわりに	188

3 絶対リスク計算

三宅顕光, 櫻井利恵子 189

基礎知識と背景	189
絶対リスクによる集団の層別化の重要性 189 / 絶対リスク 191 / 絶対推定リスクモデルの構築 193 / 研究デザイン 195 / 絶対リスクモデルの性能評価 198 / 臨床的有用性の評価 203	
おわりに	205
Column ⑥ 多様性ゲノムコホート：米国 All of Us	206

4 メンデル無作為化による因果推論

成田 暁 207

基礎知識と背景	207
MRの登場 208 / MRの理論 209 / MRの手法 211 / MRの歴史と現状 212 / MRの限界 214 / 日本分子疫学コンソーシアムの取組み 214	
MRの演習	214
Wald法 214 / 2SLS 215 / MR-Egger 215	
おわりに	216

5 タンパク質立体構造を用いたバリエーションの
臨床的意義の検討

城田松之 217

基礎知識と背景	217
解析の準備	218
解析手法	221
対象とするバリエーション 221	
A) PDBにタンパク質構造がある場合の解析手順 (バリエーション1)	223
B) PDBにタンパク質構造がない場合の解析 (バリエーション2)	234
おわりに	242
Column ⑦ 多様性ゲノムデータ：HGDP / SGDP	243

1	半導体技術による次世代シーケンシングの高速化	原島圭介 244
	基礎知識と背景	244
	解決すべき問題 244 / 現状のソリューション 245 / 通常のCPUで行う演算との相違 245	
	おわりに	248
	Column ⑧ 電子カルテデータの利用	249
2	量子時代におけるゲノム解析データのセキュアな活用	佐藤英昭, 長神風二 250
	基礎知識と背景	250
	量子暗号通信によるゲノム情報の伝送 251	
	今後の展開	253
	おわりに	254
3	スーパーコンピューター富岳での大規模計算 —— スーパーコンピューター富岳へ プログラムを移植するにあたって	松岡 光, 鈴木永久 255
	基礎知識と背景	255
	STMGPとは 255 / Pythonとは 255 / 富岳とは 256 / 富岳の特徴 256 / 富岳を使用するにあたって気を付ける点 258 / STMGPで使ったデータセットおよび分析手法 259 / 分析の結果 260	
	おわりに	260
	索引	262