Index

Page references followed by f denote figures. Page references followed by t denote tables.

A
Abeel, Thomas, 103
ABI. See Applied Biosystems Inc.
Ab initio genome annotation, 172, 178, 180t–181t
ab1PeakReporter software, 52–53
A-Brujin graph, 133–134
ABySS (Assembly by Short Sequencing), 134, 142, 147–153
effect of k-mer size and minimum pair number on assembly, 148–149, 149f
overview of, 147–148
quality of assembly, 149–153, 150t, 151f–152f
transcriptome assembly (Trans-ABySS), 158t, 160–161, 166
AceView database, 294, 295f
Acrylamide gels
capillary tube, 4
Sanger sequencing and, 2, 3–4
ACT, 179t
Adapter removal, 37–39, 39f, 43
Adapter Removal program, 38
Af
fine gaps, 42, 110, 111–112
Algorithms
database searching, 113–115
development, 364
DNA fragment/genome assembly, 127–129, 133–134, 142
dynamic programming, 110–124
file compression, 79
Golay error-correcting, 31
heuristic, 113–115
numerical optimization, 285
peak-finding, 224
protein sequence database searching, 325–326
transcriptome assembly software, 155–166, 158t
variant detection, 90
Aligners, 80
Alignment, 29, 74, 314, 362. See also Alignment algorithms;
Sequence alignment
nanopore sequencing and, 346, 347f
pileup alignment format, 194
RNA-seq, 281–283
Alignment algorithms, 90, 129, 338, 344
BLAST (Basic Local Alignment Search Tool), 114–115
BLAT, 182
ChIP-seq and, 223
FASTA (Fast Alignment), 113–114
Needleman–Wunsch (NW) algorithm, 49, 54, 110–113
overview, 109–110
Smith–Waterman (SW) algorithm, 38, 49, 62–63, 111–113
Splign, 182
TopHat, 43, 182
Alignment score, FASTA, 64–65
Allele, 52, 354
Allele frequency, 76, 94, 193
Allele-specific expression, 155, 298
ALLPATHS, 134
ALN format, 92
α-diversity indices, 319
Alternative splicing, 182, 293–296, 294f–295f
Altschul, Stephen, 65
Amino acids, pairwise comparisons, 48–49
Implicons, 8, 30, 89, 204, 309, 312
Implicon Variant Analyzer, 101
AmpliSeq Cancer Panel (Ion Torrent), 206
Annotation, 75. See also Genome annotation
ChIP-seq peak, 240–242, 255, 259, 262–263, 262f–263f
proteogenomics and, 327–328, 328f
of variants, 208–212
ANNOVAR, 211
Anthrax, 141
Anti-sense RNA, 281
Application programming interface (API), 368
Applied Biosystems Inc. (ABI)
ABI 3700 machine, 17
ab1PeakReporter software, 52–53
desktop sequence assembly and editing software, 58–59, 59f
electropherograms, 51–52, 51f
fluorescent dye use, 3–4
Human Genome Project and, 4, 60
Macintosh computer use, 58
Phred program and, 60–61
Sequence Scanner, 52
SOLiD, 17–20, 18f–20f
Arabidopsis Information Resource, 104
Archival data storage, 73–74
Argo, 179t
ArrayExpress, 178t
Arrays. See also Microarrays
array comparative genomic hybridization (aCGH), 199
variant detection and, 199

© 2015 by Cold Spring Harbor Laboratory Press. All rights reserved.
Best Match Tagger (BMTagger), 42
Benjamini Hochberg FDR method, 291, 297
Bar coding, 30
Best Practice Guidelines, dbGaP approved, 85
BEDTools, 209
BED format, 92, 99
ChIP-seq and, 240, 252, 255, 259
HTSeq-count, 284
proteogenomics and, 331, 332
bedGraph file, 333
BAM file, 74, 76, 80–81, 83, 94, 102
fields in file format, 1961
Genome Analysis Toolkit (GATK), 197
quality assessment, 33
RNA-seq and, 283–284
SAM file conversion to, 256, 258, 283
short read alignment software and, 223–224
variant detection and, 193, 195
Bar Code Index Sequence, Illumina, 31, 32f
Bar coding, 30–33, 38
Base-calling software, 52, 55–56, 59–62
BaseSpace, 254, 368–370
Basic4Seq, 249
Basic Local Alignment Search Tool, 115
BayesPeak, 235, 237f
baySeq package, 288, 292, 293t
BEADS, 239f
BEAST, 316
BED format, 92, 99
ChIP-seq and, 240, 252, 255, 259
HTSeq-count, 284
proteogenomics and, 331, 332
bedGraph file, 333
BEDTools, 209
ChIP-seq and, 240, 254, 255, 259
coverageBed, 286
Benjamini–Hochberg FDR method, 291, 297
Best Match Tagger (BMTagger), 42–43
Best Practice Guidelines, dbGaP approved, 85–86
β-diversity, 319
BETA-minus tool, 240
BFAST, 281
BGZF (block compressed gzip), 80
Biblospec, 326
Binary search, 119
BioCarta, 296
Bioinformatics, defined, 9
Bioinformatics Open Source Conference (BOSC), 366
Biomarkers, 155
Biostrings, 241
BitSeq, 239f
BLAST, 56, 114–115
BLAST (Basic Local Alignment Search Tool), 56, 114–115
e-value score, 66–67
Gapped BLAST, 67–68
gene annotation process and, 172, 181
as heuristic method, 67
maximal segment pair (MSP), 66
metagenomics and, 314
ortholog hit ratio (OHR), 165
overview, 65–68
speed of searches, 67–68
translated, 179
variant detection and, 195
Blast2GO, 172
BLASTN, 114–115, 315
BLASTX, 315
BLAT, 182
Block compressed gzip (BGZF), 80
BMNTagger (Best Match Tagger), 42–43
Bolger, Anthony, 38
Bonferroni correction, 319
BOSC (Bioinformatics Open Source Conference), 366
Bowtie, 80, 81, 116, 121, 124
ChIP-seq and, 223
FASTQ file preprocessing, 41
RNA-seq and, 281, 282–283
Brenda, 178f
Bridge amplification, 14f, 15
Broad Institute, 102, 103f, 368
Burrows–Wheeler matrix, 119–120, 120f
Burrows–Wheeler Transformation (BWT), 116–124, 120f, 195, 281
Butterfly, 161
BWA (Burrows–Wheeler Alignment), 80, 81
ChIP-seq and, 223, 256
DeconSeq human contaminant filter, 42–43
FASTQ file preprocessing, 41
metagenomics and, 314
MiSeq software and, 30
RNA-seq and, 281
trimming algorithm, 41
variant detection, 195
Bzip, 81
Bzip2, 121

Cancer
“driver” vs. “passenger” mutations, 77
gene expression studies, 77
sequence databases, 77–78
The Cancer Genome Atlas (TCGA), 77
European Genome-Phenome Archive (EGA), 78
International Cancer Genome Consortium (ICGC), 78
somatic variant discovery, 201–204

© 2015 by Cold Spring Harbor Laboratory Press. All rights reserved.
<table>
<thead>
<tr>
<th>Term</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChIP-on-chip (ChIP-chip)</td>
<td>219, 220</td>
</tr>
<tr>
<td>Chimera Slayer</td>
<td>314</td>
</tr>
<tr>
<td>Chimeras</td>
<td>316</td>
</tr>
<tr>
<td>ChIP-seq Peak Finder</td>
<td>231, 233, 233t, 237t</td>
</tr>
<tr>
<td>Chemiluminescence</td>
<td>10</td>
</tr>
<tr>
<td>ChEBI</td>
<td>178t</td>
</tr>
<tr>
<td>Chain terminators</td>
<td>3, 4</td>
</tr>
<tr>
<td>ChEBI, 178t</td>
<td></td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>25–26, 74, 89, 95, 217–264, 348</td>
</tr>
<tr>
<td>ChIP-seq tool RCSB PDB</td>
<td>240177t</td>
</tr>
<tr>
<td>CLIP-chip</td>
<td>218, 220, 231</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>238, 239t</td>
</tr>
<tr>
<td>ChIP-on-chip</td>
<td>218, 220, 231</td>
</tr>
<tr>
<td>ChIP-seq tool</td>
<td>238, 239t</td>
</tr>
<tr>
<td>Cloud bioinformatics platforms</td>
<td>369</td>
</tr>
<tr>
<td>contaminate sequences</td>
<td>41</td>
</tr>
<tr>
<td>cost</td>
<td>220</td>
</tr>
<tr>
<td>data analysis</td>
<td></td>
</tr>
<tr>
<td>BWA alignment</td>
<td>256</td>
</tr>
<tr>
<td>file format conversion</td>
<td>256, 258</td>
</tr>
<tr>
<td>peak annotation</td>
<td>255, 259, 262–263, 262f–263f</td>
</tr>
<tr>
<td>practical guide</td>
<td>255–264</td>
</tr>
<tr>
<td>reference genomes preparation</td>
<td>256</td>
</tr>
<tr>
<td>visualization of alignment</td>
<td>256, 258</td>
</tr>
<tr>
<td>data visualization</td>
<td>95, 223–224, 256, 258</td>
</tr>
<tr>
<td>differential analysis</td>
<td>236–240, 239t</td>
</tr>
<tr>
<td>ENCODE project</td>
<td>252–254</td>
</tr>
<tr>
<td>history of</td>
<td>218</td>
</tr>
<tr>
<td>motif discovery</td>
<td>240–242</td>
</tr>
<tr>
<td>peak calling</td>
<td>225–236, 227f, 230f, 237f, 255, 258–259, 260f–261f</td>
</tr>
<tr>
<td>approaches for</td>
<td>225–230</td>
</tr>
<tr>
<td>software</td>
<td>230–236</td>
</tr>
<tr>
<td>quality control</td>
<td>223, 255</td>
</tr>
<tr>
<td>RNA-seq and</td>
<td>238, 298, 299f</td>
</tr>
<tr>
<td>sequencing depth</td>
<td>220–223, 222f</td>
</tr>
<tr>
<td>transcription factor binding</td>
<td>217–218, 221–222, 226, 228–229, 233–235, 239t, 240</td>
</tr>
<tr>
<td>workflow</td>
<td>219f, 223</td>
</tr>
<tr>
<td>ChIP-seq Peak Finder</td>
<td>231, 233, 233t, 237t</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td></td>
</tr>
<tr>
<td>Chromatin</td>
<td>52</td>
</tr>
<tr>
<td>Chromatin. See also ChIP-seq</td>
<td></td>
</tr>
<tr>
<td>Assay for Transposase-Accessible Chromatin (ATAC-seq)</td>
<td>243–244</td>
</tr>
<tr>
<td>DNase-seq</td>
<td>242–245</td>
</tr>
<tr>
<td>FAIRE-seq</td>
<td>243–245</td>
</tr>
<tr>
<td>three-dimensional structure</td>
<td>245–249</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td></td>
</tr>
<tr>
<td>Chromatin conformation capture (3C)</td>
<td>245–246</td>
</tr>
<tr>
<td>Chromatin immunoprecipitation (ChIP)</td>
<td>217–218. See also ChIP-seq</td>
</tr>
<tr>
<td>Chromatogram Explorer</td>
<td>52</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td></td>
</tr>
<tr>
<td>Cloud bio system</td>
<td>315</td>
</tr>
<tr>
<td>Cloud Computing</td>
<td>78–79</td>
</tr>
<tr>
<td>Cloud Virtual Machine (CloVR)</td>
<td>43, 314–315</td>
</tr>
<tr>
<td>Code of conduct, dbGaP approved</td>
<td>85</td>
</tr>
<tr>
<td>CodonCode Aligner</td>
<td>52</td>
</tr>
<tr>
<td>COG database</td>
<td>315</td>
</tr>
<tr>
<td>Cloud BioLinux project</td>
<td>365–368, 365f, 367f</td>
</tr>
<tr>
<td>Cloud computing</td>
<td>78–79</td>
</tr>
<tr>
<td>Cloud computing</td>
<td>78–79</td>
</tr>
<tr>
<td>Cloud Computing</td>
<td>78–79</td>
</tr>
<tr>
<td>Cloud Virtual Machine (CloVR)</td>
<td>43, 314–315</td>
</tr>
<tr>
<td>Code of conduct, dbGaP approved</td>
<td>85</td>
</tr>
<tr>
<td>Color space, SOLID sequencing and</td>
<td>19–20, 19f–20f</td>
</tr>
<tr>
<td>Comparative ChIP-seq</td>
<td>239t</td>
</tr>
<tr>
<td>Comparative genomic hybridization (CGH)</td>
<td>199–200</td>
</tr>
<tr>
<td>Complete Genomics</td>
<td>76, 354</td>
</tr>
<tr>
<td>Compressed Sequence Read Archive (cSRA)</td>
<td>74, 83</td>
</tr>
<tr>
<td>Compute clusters</td>
<td>364–365</td>
</tr>
<tr>
<td>Consensus Assessment of Sequence and Variation. See CASAVA</td>
<td></td>
</tr>
<tr>
<td>Consensus by plurality algorithm</td>
<td>59</td>
</tr>
<tr>
<td>Consensus sequence</td>
<td>59, 170, 198, 240</td>
</tr>
<tr>
<td>nanopore sequencing and</td>
<td>345f</td>
</tr>
<tr>
<td>PacBio sequencing</td>
<td>22</td>
</tr>
<tr>
<td>in shotgun sequencing strategy</td>
<td>54</td>
</tr>
<tr>
<td>splice junction</td>
<td>283</td>
</tr>
<tr>
<td>Contaminating sequences, filtering</td>
<td>41–42</td>
</tr>
<tr>
<td>Contig (contiguous sequence)</td>
<td></td>
</tr>
<tr>
<td>adapter sequence blocking of assembly</td>
<td>37</td>
</tr>
<tr>
<td>assembling consensus sequence</td>
<td>4</td>
</tr>
<tr>
<td>assembling with Phrap, 62–63</td>
<td></td>
</tr>
<tr>
<td>Celera Assembler and, 351</td>
<td></td>
</tr>
<tr>
<td>internal joins</td>
<td>55</td>
</tr>
</tbody>
</table>
Contig (contiguous sequence) (Continued)
joining in Staden package, 55
mapping to genome sequence, 142
MetAMOS tool and, 315
primer walking strategy, 7
properties of, 170
in reference genome, 102
RNA, 156, 158–160, 162–166, 182–183
in shotgun sequencing strategy, 54
viewed with Consed editor, 63
Contig ordering, 133, 150
CONTRAST, 180t
Copy number variation, 52, 191, 198–201, 199f
Coverage, 7, 30, 81, 95, 313
cSRA (compressed Sequence Read Archive), 74, 83
CSA, 175t
cDNA
coverage of, 102
primer walking strategy, 7
mapping to genome sequence, 142
rule-based digraph, 135
for de novo RNA-seq assembly, 155–158, 160
generalized de Bruijn digraph, 136–138, 131f, 135f–137f,
142, 351
de Bruijn–based digraph, 135–137, 135f, 137f
for de novo RNA-seq assembly, 155–158, 160
generalized de Bruijn digraph, 136–138, 136f–137f
de Bruijn sequence, 130, 131–132
DeconSeq human contaminant filter, 42–43
Deep sequencing, 11, 202, 206, 276
Deletion/insertion index, 49
Deletions. See Indels (insertions and deletions)
Demultiplexing
e-utils, 33, 318
FASTX-Toolkit, 32–33
Illumina, 31–32, 38
metagenomics and, 318
Quantitative Insights into Microbial Ecology (QIIME), 32
Sabre, 33
de novo assemblers
ABySS (Assembly by Short Sequencing), 134, 142, 147–153,
158t, 160–161
Velvet, 134, 142–147
de novo assembly, 8, 89, 127, 130, 169, 361. See also DNA frag-
ment assembly
adapter removal, 37
DbSNP (Database of Single-Nucleotide Polymorphisms), 192
dbGAP, 74, 84f
Database of Genomic Variants (DGV), 198
Database searching, 64–68, 113–115
Database sequence, 110
Data compression, 79–84, 82f–83f
Data formats, 90–95, 91t. See also specific formats
ALN, 92
SAM, 93
Useq, 94–95
VCF (variant call format), 94
WIG (wiggle), 92–93
Data preprocessing, 37–43
adapter removal, 37–39, 39f, 43
DeconSeq human contaminant filter, 42–43
filtering contaminating sequences, 41–42
quality trimming, 37, 40–41, 43–44
Data privacy, 84–86
Data representation, 90. See also Data visualization
Data visualization, 89–106, 96f, 99f–106f
ChIP-seq, 95, 223–224, 256, 258
choosing visualization tools, 95–98
data formats, 90–95, 91t
ALN, 92
BAM, 43, 223–224
BED, 92
FASTA, 90–91
FASTQ, 90–91
GFF3, 93
GFF/GTF, 93
NGS data file types, 91
SAM, 93
Useq, 94–95
VCF (variant call format), 94
WIG (wiggle), 92–93
overview, 89
GBrowse (Generic Genome Browser), 104–106, 105f–106f, 224
GenomeView, 102–103, 104f, 224
Illumina Genome Studio, 100, 101f
Integrative Genomics Viewer (IGV), 101–102,
103f, 224
MAUVE, 100
Newbler, 101, 102f
Staden, 98–99
UCSC Genome Browser, 98–99, 100f
Dayhoff, Margaret, 48
DBChIP, 239t
dbSNP, 174t, 192, 207, 210
dbVar, 198
dCLIP tool, 251
DDBJ (DNA Data Bank of Japan), 73
de Bruijn digraphs/paths, 40, 127, 130–138, 131f, 135f–137f,
142, 351
de Bruijn–based digraph, 135–137, 135f, 137f
for de novo RNA-seq assembly, 155–158, 160
generalized de Bruijn digraph, 136–138, 136f–137f
de Bruijn sequence, 130, 131–132
DeconSeq human contaminant filter, 42–43
Deep sequencing, 11, 202, 206, 276
Deletion/insertion index, 49
Deletions. See Indels (insertions and deletions)
Demultiplexing
e-utils, 33, 318
FASTX-Toolkit, 32–33
Illumina, 31–32, 38
metagenomics and, 318
Quantitative Insights into Microbial Ecology (QIIME), 32
Sabre, 33
de novo assemblers
ABySS (Assembly by Short Sequencing), 134, 142, 147–153,
158t, 160–161
Velvet, 134, 142–147
de novo assembly, 8, 89, 127, 130, 169, 361. See also DNA frag-
ment assembly
adapter removal, 37
© 2015 by Cold Spring Harbor Laboratory Press. All rights reserved.
of bacterial genomes from short reads, 141–153

Genome Assembly Program 4 (GAP4), 56

QUAKE read correction method, 40

quality filters, 40

RNA, 179

using short reads, 133–134, 142
de novo sequencing, 127, 169

contaminating sequences, 41
derror-correcting PacBio long sequences, 349

expressed sequence tags (ESTs), 273
de novo transcriptome assembly, 155–166, 182–183

metrics for evaluation of assemblies, 164–165

Oases for Velvet, 158–160, 158t, 166

overview, 155–158

Soapdenovo-Trans, 158t, 162–164, 166

SOP (standard operating procedure), 157

Trans-ABYSS, 158t, 160–161, 166

Trinity, 158, 158t, 161–162, 166

Density function curve, 225

DESeq, 289, 291, 293t

Desktop sequence assembly and editing software, 58–59, 59f

DETONATE, 165

DIAG cloud, 43, 315

Diagonal runs, 113–114

DiffBind, 239t

differential analysis of ChIP-seq data, 238–240, 239t

differential expression, 275–276, 278–279, 281, 287–289, 293–297,

Differentially methylated regions, 348, 354, 356f

diffReps, 239t

Digenorm, 157, 161

DIME, 239t

diploid organism, 52, 192, 351

Directed reads, 127, 130f

Diversity, metagenomics and, 319

DNA Baser, 13

DNA Data Bank of Japan (DDBJ), 73

DNA fragment assembly, 170. See also Genome assembly algorithms, 59, 127–129, 133–134

coverage, 128

de Bruijn digraphs, 127, 130–138, 131f, 135f–137f

multiple alignments and consensus sequence, 128

Overlap-Layout-Consensus (OLC) model, 128–129

overview, 4–7

reference genome creation by, 198

repetitive DNA sequences and, 127, 128

sequencing by hybridization and, 129–130, 133

shotgun sequencing and, 127–128, 133

DNA fragments

ChIP isolation of, 217–218

ChIP-seq and, 95, 218, 221f, 224–226, 228, 233–234, 236, 243,

246–247, 254

electrophorograms of sequenced, 51

DNA fragment synthesis, in Sanger sequencing, 1–2

DNAnexus, 368–369

DNA polymerase

PacBio sequencing, 22

in pyrosequencing, 10

in Sanger sequencing, 1, 4, 13

Taq polymerase, 52

DNA sequence reads compression (DSRC), 81

DNA sequencing

ABI SOLID, 17–20, 18f–20f

cloning for, 3, 4–7, 5f–7f

errors, 9, 22, 128, 133

experimental applications, 24–26

454 system, 9–13, 12f

history of, 1–4, 2f

Illumina Genome Analyzer, 13–17, 14f–15f

Ion Torrent, 20–22

massively parallel, 9

mate-pair sequencing, 24, 25f

next-generation sequencing (NGS), 7–9

Pacific Biosciences (PacBio), 22–23, 23f

paired-end sequencing strategy, 24

DNaseR, 245

DNase-seq, 242–245

Doolittle, Russell, 48

Double principal coordinate analysis (DPCoA), 320

dPeak, 237t

DREME, 242, 263

duplex-specific nuclease (DSN) normalization, 274

duplicate sequences, 36

dynamic programming algorithms, 110–124

E

EASE, 297

ea-utils, 33, 318

EBA (Estimate Base Accuracy) program, 56

EBI (European Bioinformatics Institute), 73

EC2. See Amazon Elastic Compute Cloud

Edena (Exact De Novo Assembler), 134

dgeK, 239t, 291, 293t

Edman, Pehr, 48

EGA (European Genome-Phenome Archive), 78

ELAND (Efficient Large-Scale Alignment of Nucleotide Databases), 16, 116, 233, 281–282

electropherograms, 51–53, 51f, 53f

electrophoresis

base-calling and, 61

capillary, 4, 5f, 17

Sanger sequencing and, 2, 3–4, 61

EMBL, 57–58, 73, 75, 83, 84, 91, 366

EMBOSS (European Molecular Biology Open Software Suite), 58, 240, 263

Emerging technologies and applications, 337–357

ENA (European Nucleotide Archive), 74–76, 83

ENCODE, 181, 211, 221, 244, 252–254, 283, 348, 362

ENCODEnCODE, 354

Encyclopedia of DNA Elements. See ENCODE

Enrichment score, 298

ENSEMBL, 91, 172, 174t, 184, 259

alternate transcripts in, 294

as reference proteome, 329, 331

Entrez system, 74

Epigenetically modified cytosine, nanopore sequencing detection of, 346, 348–349

E-RANGE, 231

Erlich, Yaniv, 86

Error

base-calling, 61–62

cloning artifact, 192

454 system base calling errors, 316

PCR, 9, 224, 228, 280

sequencing

base position in read, effect of, 192

ChIP-seq and, 224

coverage and, 193

dinucleotides and, 192–193

error-correcting bar codes, 31–32

genome assembly and, 128, 133, 165

Illumina sequencing, 14, 29–30
Index

Error (Continued)
multiplexing and, 31
nanopore, 344–346
PacBio sequencing, 22
pyrosequencing, 9
Sanger sequencing, 9
single-molecule read correction and assembly method, 349, 350, 351
Escherichia coli, genome sequencing, 21
ESTs (expressed sequence tags), 74, 273
Eucalyptus, 364
EUGENE, 180t
EULER-DB algorithm, 133
Eulerian path, 121, 130, 132
EULER (software tool), 133–134
EULER-SR algorithm, 134
European Bioinformatics Institute (EBI), 73
European Genome-Phenome Archive (EGA), 78
European Nucleotide Archive (ENA), 74–76, 83
E-value score, BLAST, 66–67
EvidenceModeler, 183
Exact mapping, BWT, 121–122
Excavator software, 200
Exome capture, 204–206
ExomeDepth, 200
Exome sequencing, GC bias in, 200
Exon capture, 36
Exonereate sequence alignment tool, 42
ExonHunter, 180t
Exons, 240, 354
alternative splicing, 293–296, 294f–295f
genome annotation, 171, 327
intron/exon boundaries, 329–330, 330f
RNA-seq and, 279–280
splice junctions, 281–283
Exonuclease III, 4
Expressed sequence tags (ESTs), 74, 273

F
FAIRE-seq, 243–245
FAKII, 56
False discovery rate, 319
Benjamini-Hochberg FDR method, 291, 297
ChIP-seq and, 232–234, 258
RNA-seq and, 276–292, 296
FASTA (Fast Alignment), 13, 74, 90–91, 113–114
adapter sequences, 38
demultiplexing, 32
in de novo RNA assembly process, 182
protein sequences database, 254, 333
Quantitative Insights into Microbial Ecology (QIIME), 32
reference genomes, 256
searching sequence databases, 64–65
SOAPdenovo-Trans and, 162–163
variant detection and, 194
FASTP, 64
FastQC program, 33–37, 34f–35f, 43, 255
duplicate sequences, 36
overrepresented sequences, 36–37
per base “N” content, 36
%GC, 35
Q score, 34–35, 34f
RNA-seq data, 280
sequence length distribution, 36
FASTQ format, 74, 76, 79–81, 90–91
ChIP-seq and, 223, 256
demultiplexing, 32–33, 38
in genome assembly process, 170
Illumina sequencing and, 13, 16–17, 32, 142, 259
preprocessing, 41
quality assessment, 33–34, 36–37, 280
SOAPdenovo-Trans and, 162
variant detection and, 193, 194
FASTQ_Groomer, 366
FastqMultlx, 33
FastTree, 316
Fastx_clipper, 38
FASTX Quality Filter, 40
FASTX Quality Trimmer, 40
FASTX-Toolkit, 32–33, 35, 38
Ferragina and Manzini algorithm, 121
FGENESH, 180t
Field programmable gate arrays (FPGAs), 10, 109, 112
Fiers, Walter, 1
File formats, 51. See also specific formats
Filtering contaminating sequences, 41–42
FindTV, 52
FindPeaks, 231, 233, 237f
FireDB/Firestar, 175t
Fisher, R.A., 47
FISH (fluorescent in situ hybridization), 170
FisHiCal, 249
Fitch, Walter, 48
5C, 246
FLASH, 313
Flow cell
Illumina Genome Analyzer, 15, 15f
SOLiD system, 17
Flower, 13
Fluorescent dyes/labels
ABI sequencers, 3–4, 51, 60
Illumina sequencers, 13–14, 15f
SOLiD system, 18–19
Fluorescent in situ hybridization (FISH), 170
FlyBase, 104, 174t
Fly Factor Survey, 241
FM index, 119
Foodborne pathogens, 141
FoundationHeme panel, 206
Foundation Medicine, 206
FoundationOne cancer panel, 206
4C (circular chromosome conformation capture), 246
454 system
base calling errors, 316
Genome Sequencer 20 (GS20), 10
GS FLX system (Titanium), 10
GS Junior, 361
Human Microbiome Project, 312–313
Newbler program use for assembling reads, 164
overview, 9–13, 12f
sequence read length, 36
sequence variants and, 11
SFF files, 11, 13
Fourier series, 61
Fourier transform, 225
4Peaks, 52
FPGAs. See Field programmable gate arrays
FPKM (fragments per kilobase per million), 285, 288, 304
Fragment assembly. See DNA fragment assembly
F-seq, 245
Functional annotation of features, 179, 181
Fungi, 18S genes of, 313

© 2015 by Cold Spring Harbor Laboratory Press. All rights reserved.
Index 393

G

Galaxy web server
ChIP-seq and, 240, 254, 255, 259, 262f–264f
cloud-based informatics and, 364, 365f, 366, 367f
RNA-seq analysis exercise, 300–305, 301f–302f, 304f
GAP4, 56
GAP5, 56
Gap_extension, 112
GAP (Genome Assembly Program), 55–56, 55f
Gap_open, 112
Gapped BLAST, 67–68
Gap penalty, 49, 62, 64
GATK (Genome Analysis Toolkit), 192–193, 197–198
GBrowse (Generic Genome Browser), 93, 104–106, 105f–106f, 183, 223
GCG (Genetics Computer Group), 57–58
GELEN program, 54
Gel reading, 49–50, 50f, 54
GEL system, 57
Genapsys, 341t, 343
GenBank, 73, 74, 79, 91
as annotation data resource, 174t, 184
Database of Single-Nucleotide Polymorphisms (dbSNPs), 192
GCG package compatibility with, 57–58
searching, 67, 113
16S sequences, 316
GENCODE Consortium, 183
GeneCodes Corporation, 58
Gene3D, 175t
Gene expression, quantification of, 284–286
GeneID, 178, 180t
GeneMark, 180t
Gene Ontology Consortium, 172, 179, 181, 296
Gene ontology (GO) terms, 296
GenePattern, 284, 298
Generic Genome Browser. See GBrowse
GENESIS program, 57
GeneTorrent, 77
Genia Nano-Tag system, 341t
Genome Analysis Toolkit (GATK), 192–193, 197–198
Genome annotation, 169–184
combining evidence and expert curation, 183–184
community-based projects and software, 179f
de novo transcriptome assembly, 182–183
GFF file, 171–172, 173t
importance of, 170–171
overview, 169–172, 171f
proteogenomics and, 327–328, 328f
source of information, 171–172, 173t–179t
strategies and methods, 172, 178–181
ab initio approach, 172, 178, 180–181t
functional annotation of features, 179, 181
reference-based/evidence-based, 179, 180t–181t
trends and advanced practice, 182–184
Genome Assembly Program (GAP), 55–56, 55f
Genome dictionary, 123, 125, 170
GenomeSpace, 368
Genome Studio, Illumina, 100, 101f, 231, 233, 233t
GenomeView, 102–103, 104f, 223
Genome-wide association studies (GWASs), variant detection and, 192, 207–208, 209f
Genomic Threading Database, 177t
GENSCAN, 178, 180t
GENUIS, 341t, 343
GEO, 84
GEPAS, 178t
GERP++ (genomic evolutionary rate profiling) scores, 211
Gerstein, Mark, 16
GFF3 (general feature format, version 3), 93, 285, 333
GFF (general feature format) file, 93
annotation and, 171–172, 173t, 209
ChIP-seq and, 240, 259
GBrowse and, 104–105
RNA-seq and, 296
GHeneWorks, 59
Gilbert, Don, 51
Gilbert, Walter, 1, 3
GimmeMotifs, 242, 263
GitHub, 39, 366
GLIMMER, 180t
GMOD (General Model Organism Database) package, 93, 104, 183
Gnomon, 180t, 183–184
Golay error-correcting algorithm, 31, 313
GOLD, 174t
Gordon, Assaf, 38
Gordon, David, 60
GPS/GEM, 237t
Graphical user interface (GUI), 364
Graphics processing unit (GPU), 109, 112
GraphProt, 251
Green, Phil, 60
Greengenes, 316, 317, 318
GSCAN, 178t
GS de novo Assembler, 101
GSEA, 297–298
GTF (general transfer format) file, 82, 93
Cloud BioLinux and, 366–367
genome annotation, 171
RNA-seq and, 285, 287, 296
variant detection, 209–210
GWASs (genome-wide association studies), variant detection and, 192, 207–208, 209f
H

Hamiltonian path, 129, 132
Hamming distances, 31
Hannon, Greg, 38
Haplotyping, 354, 355f, 356
Haplotyping phaseing, 337, 338, 354, 356
HapMap, 198, 207, 354
Hash table lookup, BLAST, 66
Hash tables, 115–116, 156, 158–160, 256, 281
Haskell language, 13
Health Insurance Portability and Accountability Act (HIPAA), 86
Henning, Willi, 47
Heracle Biosoft S.R.L., 13
Heterozygosity
cancer mutations and, 202
PCR sequencing and, 52–53, 53f
Heterozygotes, 194, 354
Heuristic algorithms, 113–115
HiBrowse, 249
Hi-C method, 247, 248f
HiCNorm, 248
Hidden Markov models, 178, 181, 238, 346
Index 395

L
LaserGene, 59
Last-first (LF) mapping, 121, 122f
Limma package, 289, 292, 293t
Linkage mapping, 47
Liquid chromatography (LC) tandem mass spectrometry (MS/MS), 325, 326f
Locally collinear blocks (LCBs), 150, 151f
Long-fragment read technology, 354
Long-read sequencing (LRseq), 349–354, 352f–353f, 355f–356f
Lossless compression, 79, 80, 83
Lossy compression, 79–80, 81, 83–84
LOWESS regression, 289, 292
Luciferase/luciferin, 10
LZ encoding, 81

M
MACGE, 178t
MACS, 239f, 258
MACS (model-based analysis of ChIP-seq), 95, 96f, 231, 232–233, 237f, 235, 259–260
MacVector, 59
Major histocompatibility locus, 193
MAKER annotation pipeline toolkit, 183
Manate, 179t
Mann–Whitney test, 319
MANORM, 239t
Mapping
ChIP-seq reads, 224
discordant, 200
DNA fragments to chromosomes, 170
ortholog, 172
proteomic mapping to genomic coordinates, 332–333
RNA-seq reads to genes, 281–283
MAQ (Mapping and Assemblies with Qualities), 80, 116, 193–195, 223
Markov chain Monte Carlo (MCMC), 235
Martin, Marcel, 39
Mascot, 326
Massively parallel sequencing, 9, 142
Mass spectrometry, proteogenomics and, 325, 326f, 327, 329–332
Mate-pair sequencing, 24, 25f
MAUVE, 100, 150
Mauve Contig Mover (MCM), 150
Maxam, Allan, 1, 3
Maxam-Gilbert sequencing, 3
Maximal alignment length, 283
Maximal Mappable Prefix, 283
Maximal segment pair, 66
MAXIMIZE program, 57
MaxInfo, 38
MCMC (Markov chain Monte Carlo), 235
MDS (multidimensional scaling), 287
Mean sequence quality graph, 35
MEGABLAST, 115
MEME, 242
MEME-ChIP, 242, 263
MEMSAT, 176t
Message-Passing Interface (MPI)-cluster approach, 134
Metagenomics, 26, 37, 74, 309–320
overview, 309–311
polymerase chain reaction technologies and, 311–313
shotgun metagenomic sequencing, 314–316
16 data analysis, 316–318, 317f
tutorial, 318–320, 320f

N
NABsys, 340, 341t
Nanopores, 338–339
Nanopore sequencing, 338–349, 356
alignment, 346, 347f
base calling, 344, 345f
bioinformatics and data-flow control architecture, 348f
challenges, 340

© 2015 by Cold Spring Harbor Laboratory Press. All rights reserved.
Index

Nanopore sequencing (Continued)
detection of epigenetically modified cytosine, 346, 348–349
idealized sequencer, 339–340, 339f
prototypes, 340, 341–342f
NarrowPeaks, 239
National Biomedical Research Foundation (NBRF), 48, 64
AceView database, 294, 295f
COG database, 315
National Center for Biotechnology Information (NCBI), 73, 74
Amazon EC2, availability on, 366
BLAST program creation, 65, 67
BLAST program selection guide, 114–115
data compression, 81–83
data privacy, 84–85
dbVar, 198
downloading reference genomes, 256
eukaryotic genome annotation, 183–184
Gnomon, 180t, 183–184
SRA Toolkit, 81–82
National Heart, Lung, and Blood Institute Exome Sequence
Project (NHHLBI-ESP), 208, 211
NCBI SRA Toolkit, 81–82
NCBI tools, 174
ncRNA (noncoding functional RNA), 181
Neanderthal genome sequence, 11
Needleman–Wunsch (NW) algorithm, 49, 54, 110
Neanderthal genome sequence, 11
Needleman–Wunsch (NW) algorithm, 49, 54, 110–113
Newbler, 101, 102f, 134, 164
New York University Langone Medical Center (NYULMC), 95,
96f, 105
Next-generation sequencing (NGS)
deﬁned, 8
experimental applications, 24–26
history, 7–9
mate-pair sequencing, 24, 25f
paired-end sequencing strategy, 24
platforms
ABI SOLID, 17–20, 18f–20f
454 system, 9–13, 12f
Illumina Genome Analyzer, 13–17, 14f–15f
Ion Torrent, 20–22
Pacific Biosciences (PacBio), 22–23, 23f
NGC compression software, 80
NGC QC Toolkit, 38
NimbleGen SeqCap, 205
Nimbus Informatics, 368
NNPP, 180t
NNSPLICE, 180t
NobleGen, 340, 341f
NOIseq, 292, 293t
Noncoding functional RNA (ncRNA), 181
Normalization, RNA-seq, 274–275, 287–289
Novoalign, 223
NuChart, 249
Nucleosomes, 244, 244f
Numerical optimization algorithm, 285

O
Oases, 156, 158–160, 158t, 166, 182
OMIM, 174t
1000 Genomes Project, 76, 83
cloud servers and, 362, 366
variant detection and, 198, 206–207, 211
Open source, 57, 97–98, 233, 366, 587
Open Source Initiative, 57
OpenStack, 364
Operational taxonomic unit (OTU), 312, 316–319
ORF FINDER, 181t
Ortholog hit ratio (OHR), 165
Ortholog mapping, 172
Otterlace, 179t
Overlap-Layout-Consensus (OLC) model, 128–129, 134
Oxford Nanopore Technologies, 340, 342t

P
Paabo, Svante, 11
PacBio HD, 76
PacBio RSII, 349, 350f, 351
Pacific Biosciences (PacBio), 22–23, 23f
sequence read length, 36
SMRT (single-molecule real-time), 9, 22
Paired-end reads, 13, 39f, 182
ChIP-seq, 224–225
Illumina HiSeq reads, 142
insert size estimation, 284
RNA-seq, 157–159, 161–163, 284
transforming short into long, 356
variant detection, 200–201
Paired-end sequencing, 24, 31, 38, 39f
ChIP, 218
Cufflinks and, 285
PAM-120 matrix, 65
PAM250 scoring matrix, 64
Pan-Cancer project, 78
Panther, 176t
PAR-CLIP, 250
PASA, 183
PRBCr (PacBio corrected reads), 349, 350f, 351
PCA (principal component analysis), 287
PC/GENE, 58, 59
PCoA (principal coordinate analysis), 319, 320f
PCR, See Polymerase chain reaction
PDBsum, 177t
Peak annotation, ChIP-seq, 240–242, 255, 259, 262–263,
262f–263f
Peak calling, ChIP-seq, 225–236, 227f, 230f, 237f, 255, 258–259,
260f–261f
approaches for, 225–230
biases, 228–229
software, 230–236, 237f
BayesPeak, 235, 237t
ChIP-seq Peak Finder, 231, 233, 233t, 237t
CisGenome, 231, 234, 237t
FindPeaks, 231, 233, 237t
GenomeStudio, 231, 233, 233t, 237t
MACS, 231, 232–233, 237t
overview, 231–232
PeakSeq, 233–234, 237t
QuEST, 231, 234, 237t
SICER, 231, 235, 237t
SISSRs, 231, 233t, 234, 237t
SPP, 231, 235, 237t
Useq, 231, 234–235, 237t
ZINBA, 231, 235–236, 237t
PeakFinder, 231, 233, 233t, 237t
Peak-ﬁnding algorithm/software, 224, 226
Peak2qpe tool, 262–263, 262f
PeakKD Eck, 245
PeakRanger, 237t
PeakSeq, 233–234, 237t
PEP program, 57

© 2015 by Cold Spring Harbor Laboratory Press. All rights reserved.
Index

Peppy, 331
PePr, 239t
Peptide Identification by Unbiased Search (PIUS), 331
Per base “N” content, 36
%GC, 35
Perl, 96, 97
Peptide Identi, 239t
PISA, 177t
piRNA (Piwi-interacting RNA), 274
Piranha, 251
PIR, 176t
PIPE-CLIP, 251
Pileup alignment format, 194
PIECLIP, 251
PIR, 176t
Piranal, 251
piRNA (Piwi-interacting RNA), 274
PISA, 177t
PIUS (Peptide Identification by Unbiased Search), 331
Ploidy, 202
PMut, 176t
Ploidy, 202
PMut, 176t
Poisson distribution, 109
ChIP-seq, 232
of coverage depth, 128, 156
RNA-seq, 290
of shotgun DNA fragments, 6–7
PoissonSeq package, 288, 292, 293t
PolyA Peak, 237t
Poly(A) selection, 274–275, 283
Poly(A) tails, 273
Polymerase chain reaction (PCR)
chimera creation, 316
ChIP isolated DNA fragments and, 217–218, 224
eмулюс, 10, 12f, 17, 20
errors/artifacts, 9, 224, 228, 280
Illuma Genome Analyzer sequencing, 14f
metagenomcis and, 311–313
PCR sequencing and heterozygosity, 32–33, 53f
quantitative PCR (qPCR), 279
16S sequences, 316
POLYPHEMUS, 239t
Polyphen, 210–211
Poly(T) oligonucleotides, 274
Poly(T) primers, 274, m283
Population data, variant detection and, 206–208, 209f
PRIDE, 176t
Primer extension, 2–3, 4
Primer removal, 157
Primer walking strategy, 7
Principal component analysis (PCA), 287
Principal coordinate analysis (PCoA), 319, 320f
PRINSEQ, 41
Prints, 176t
PRISM, 201, 354
Probability score, alignment, 116
ProCognate, 177t
ProDom, 176t
ProFunc, 177t
Prosite, 176t
Protein sequence databases, 328–333
construction of, 328–330
proteomic mapping to genomic coordinates, 332–333
six-frame translation databases, 331–332
tools for creating, 330–333
Protein sequence database searching algorithms, 325–326
Protein sequencing, 48
Proteogenomics, 325–333
gene annotation and, 327–328, 328f
liquid chromatography (LC) tandem mass spectrometry (MS/MS), 325, 326f
proteomic mapping to genomic coordinates, 332–333
sequence driven database construction, 328–330
Proteogenomics Mapping Tool, 333
Proteomic mapping to genomic coordinates, 332–333
ProtoNet, 176t
Pseudogenes, 178
Public sequence databases, 73–86
PubMed, 74
PupaSuite, 176t
pyDNase, 245
Pyrosequencing
insertion and deletion errors, 9
Ion Torrent technology, 20
Newbler program use for assembling reads, 164
overview, 10–11, 12f
sequencing by synthesis, 10
Q
QIIME (Quantitative Insights into Microbial Ecology), 32, 317, 320
eqps, 237t
Q score, FastQC program, 34–35, 34f
QSVanalyzer, 53
Quadromer, 339, 344
QUAKE read correction method, 40
Quality control, 29–44
ChIP-seq, 223, 225
data preprocessing, 37–43
demultiplexing, 30–32, 31f–32f
overview, 29–30
Phred scores and, 62
RNA-seq
postalignment QC of data, 283–284
prealignment QC of data, 280–281
sequence quality, 33–37, 34f–35f
SOP recommendations, 43–44
adapter removal, 43
overall evaluation, 43
quality trimming, 43–44
transcriptome assembly and, 157
Quality score
454 system, 316
Phred, 62, 63, 192
recalibration, 192–193, 197–198
variant detection and, 192–193, 197–198
Quality thresholds, 43
Quality trimming, 37, 40–41, 43–44, 157
Quantapore, 342t
Quantitative Insights into Microbial Ecology (QIIME), 32, 317, 320
Query sequence, 65–66, 110
Query substring, BWT, 119
<table>
<thead>
<tr>
<th><strong>Index</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>QuEST, 231, 234, 237t</td>
</tr>
<tr>
<td>QUILTS, 331</td>
</tr>
</tbody>
</table>

### R

<table>
<thead>
<tr>
<th><strong>RAM, 97</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>R/Bioconductor software suite</strong></td>
</tr>
<tr>
<td>ChIP-seq and, 235, 240–241, 249</td>
</tr>
<tr>
<td>edgeR package, 291</td>
</tr>
<tr>
<td>limma package, 289, 292, 293t</td>
</tr>
<tr>
<td>peak annotation, 240–241</td>
</tr>
<tr>
<td>SPP, 235</td>
</tr>
<tr>
<td>r3Cseq, 249</td>
</tr>
<tr>
<td>RDF2 program, 64–65</td>
</tr>
<tr>
<td>RDP Classifier, 318</td>
</tr>
<tr>
<td>Read length, 36</td>
</tr>
<tr>
<td>Reads, 75, 170. See also Sequence reads</td>
</tr>
<tr>
<td>alignment of, 115–116</td>
</tr>
<tr>
<td>Sanger sequencing, 115</td>
</tr>
<tr>
<td>Readseq, 51</td>
</tr>
<tr>
<td>Reads per kilobase per million (RPKM), 284, 288–289, 292</td>
</tr>
<tr>
<td>Real-Time Analysis (RTA), 29, 38, 254, 255</td>
</tr>
<tr>
<td>Reference genomes, 8, 37, 74, 89, 109, 361</td>
</tr>
<tr>
<td>ChIP-seq use of, 218</td>
</tr>
<tr>
<td>creation by assembly, 198</td>
</tr>
<tr>
<td>Mauve Contig Mover (MCM) use of, 150</td>
</tr>
<tr>
<td>nanopore sequencing, 340, 344</td>
</tr>
<tr>
<td>preparation for ChIP-seq, 256</td>
</tr>
<tr>
<td>proteogenomics and, 329, 331</td>
</tr>
<tr>
<td>RNA-seq mapping reads to, 155, 170–171, 274, 281–283</td>
</tr>
<tr>
<td>variant detection and, 191</td>
</tr>
<tr>
<td>Reference sequences, 19, 30, 52, 81, 157</td>
</tr>
<tr>
<td>single-nucleotide polymorphisms (SNPs) in, 192</td>
</tr>
<tr>
<td>16S ribosomal DNA, 316</td>
</tr>
<tr>
<td>Reference string, BWT, 117</td>
</tr>
<tr>
<td>RefSeq, 172</td>
</tr>
<tr>
<td>alternate transcripts in, 294</td>
</tr>
<tr>
<td>as annotation data resource, 174t</td>
</tr>
<tr>
<td>ChIP-seq and, 231, 259, 262–263</td>
</tr>
<tr>
<td>as reference proteome, 329, 331</td>
</tr>
<tr>
<td>RepeatMasker, 56</td>
</tr>
<tr>
<td>Repetitive DNA sequences, 109, 127, 128</td>
</tr>
<tr>
<td>Residual substitution score table, 110–111, 111t</td>
</tr>
<tr>
<td>Reverse transcription, 274</td>
</tr>
<tr>
<td>Reversible terminators, 13, 14f</td>
</tr>
<tr>
<td>Rfam database, 181</td>
</tr>
<tr>
<td>Ribonucleoprotein (RNP) complexes, 250, 251</td>
</tr>
<tr>
<td>Ribosomal Database Project, 316</td>
</tr>
<tr>
<td>Ribosomal RNA, 312</td>
</tr>
<tr>
<td>gene annotation and, 181</td>
</tr>
<tr>
<td>quality assessment of RNA-seq data, 280</td>
</tr>
<tr>
<td>removal for RNA-seq, 274–275</td>
</tr>
<tr>
<td>RIP-seq (RNA-immunoprecipitation sequencing), 250</td>
</tr>
<tr>
<td>R language, 96</td>
</tr>
<tr>
<td>RLE (run length encoding), 81</td>
</tr>
<tr>
<td>RMAP, 223</td>
</tr>
<tr>
<td>RNA assembly, de novo, 179, 182–183</td>
</tr>
<tr>
<td>RNA editing, 281</td>
</tr>
<tr>
<td>RNA polymerase II, 229–230</td>
</tr>
<tr>
<td>RNA-protein interactions, 249–250</td>
</tr>
<tr>
<td>RNase, 249</td>
</tr>
<tr>
<td>RNA-seq, 25, 33, 77, 89, 169, 273–305</td>
</tr>
<tr>
<td>alignment, 281–283</td>
</tr>
<tr>
<td>alternative splicing, 293–296, 294f–295f</td>
</tr>
<tr>
<td>applications of, 274</td>
</tr>
<tr>
<td>biased distribution of fragments, 156</td>
</tr>
<tr>
<td>ChIP-seq and, 238, 298, 299f</td>
</tr>
<tr>
<td>Cloud BioLinux desktop virtual machine, 366–367, 367f</td>
</tr>
<tr>
<td>contaminating sequences, 41</td>
</tr>
<tr>
<td>de novo transcriptome assembly, 155–166, 182–183</td>
</tr>
<tr>
<td>depth of coverage, 275–276, 277f, 278–280, 279f</td>
</tr>
<tr>
<td>differential expression, 275–276, 278–279, 281, 287–293, 293t, 297</td>
</tr>
<tr>
<td>downstream analysis, 296–298</td>
</tr>
<tr>
<td>duplication in libraries, 36, 281</td>
</tr>
<tr>
<td>filtering data, 286–287</td>
</tr>
<tr>
<td>mapping around ADM gene, 279–280, 280f</td>
</tr>
<tr>
<td>mapping reads to genes, 281–283</td>
</tr>
<tr>
<td>mapping reads to reference genome, 170–171</td>
</tr>
<tr>
<td>mRNA isolation/purification, 274–275</td>
</tr>
<tr>
<td>normalization, 274–275, 287–289</td>
</tr>
<tr>
<td>overview, 273–275</td>
</tr>
<tr>
<td>postalignment quality control of data, 283–284</td>
</tr>
<tr>
<td>prealignment quality control of data, 280–281</td>
</tr>
<tr>
<td>protein sequence database construction, 329, 331</td>
</tr>
<tr>
<td>quality trimming, 40</td>
</tr>
<tr>
<td>quantification of gene expression, 284–286</td>
</tr>
<tr>
<td>replicate number, 276, 278–280, 278f–279f, 278t</td>
</tr>
<tr>
<td>tutorial, 298, 300–305, 301f–302f, 304f</td>
</tr>
<tr>
<td>variant detection and, 191</td>
</tr>
<tr>
<td>RNA-SeQC, 284</td>
</tr>
<tr>
<td>RNP (ribonucleoprotein) complexes, 250, 251</td>
</tr>
<tr>
<td>Roberts, Richard, 50</td>
</tr>
<tr>
<td>Roche</td>
</tr>
<tr>
<td>Genia Nano-Tag system, 341</td>
</tr>
<tr>
<td>454 Genome Sequencer, 101</td>
</tr>
<tr>
<td>Newbler assembler, 101, 102f, 134</td>
</tr>
<tr>
<td>NimbleGen SeqCap, 205</td>
</tr>
<tr>
<td>RPKM_count tool, 284</td>
</tr>
<tr>
<td>RPKM (reads per kilobase per million), 284, 288–289, 292</td>
</tr>
<tr>
<td>RSEQ, 237t</td>
</tr>
<tr>
<td>RSEM-EVAL score, 165</td>
</tr>
<tr>
<td>RSEG, 237t</td>
</tr>
<tr>
<td>RPKM, count tool, 284</td>
</tr>
<tr>
<td>RPKM (reads per kilobase per million), 284, 288–289, 292</td>
</tr>
<tr>
<td>SAGE, 237t</td>
</tr>
<tr>
<td>RSEM-EVAL score, 165–166</td>
</tr>
<tr>
<td>RSeQC, 284</td>
</tr>
<tr>
<td>RTA (Real-Time Analysis), 29, 38, 254, 255</td>
</tr>
<tr>
<td>Run length encoding (RLE), 81</td>
</tr>
</tbody>
</table>

---

© 2015 by Cold Spring Harbor Laboratory Press. All rights reserved.
## Index

Schatz, Michael, 349  
Science as a service (SciaaS), 368–370  
SCOP, 177t  
Score matrix, 110–112, 112t  
Scripting language, 96  
Scythe, 39  
Searching DNA sequence databases, 64–68, 113–115  
FASTA program, 64–65  
Smith–Waterman method, 64  
SeattleSNP, 207  
Seed-and-extend, 115  
SeqPos motif tool, 263  
SeqMap, 116  
SEQFIT program, 54  
Seqed, 57  
Seed-and-extend, 115  
SeattleSNP, 207  
Scythe, 39  
Scripting language, 96  
Score matrix, 110–112, 111t–112t  
alignment algorithms, 109–124  
BLAST (Basic Local Alignment Search Tool), 56, 65–68, 114–115  
Bowie, 80, 81, 116, 121, 124  
Burrows–Wheeler Transformation (BWT), 116–124, 120f  
ChIP-seq, 223–225  
database searching, 113–115  
ELAND (Efficient Large-Scale Alignment of Nucleotide Databases), 16, 116, 233  
FASTP program, 64  
delops (insertions and deletions), 110, 111–112  
Mauve Contig Mover (MCM) and, 150  
nanopore sequencing and, 343  
Needleman–Wunsch (NW) algorithm, 49, 54, 110–113  
of next-generation sequencing reads, 115–116  
overview, 109–110  
Phrap program and, 62–63  
Short Oligonucleotide Alignment Program (SOAP), 115–116  
similarity to sequence assembly, 6  
Smith–Waterman (SW) algorithm, 38, 49, 62–63, 111–113  
SOAP2, 116, 124  
statistical significance of, 64–65  
Sequence analysis tools, 50  
Sequence alignment, 90, 109–124, 117f, 120f, 122f–123f  
affine gaps, 110, 111–112  
alignment algorithms, 109–124  
BLAST (Basic Local Alignment Search Tool), 56, 65–68, 114–115  
Bowie, 80, 81, 116, 121, 124  
Burrows–Wheeler Transformation (BWT), 116–124, 120f  
ChIP-seq, 223–225  
database searching, 113–115  
ELAND (Efficient Large-Scale Alignment of Nucleotide Databases), 16, 116, 233  
FASTP program, 64  
delops (insertions and deletions), 110, 111–112  
Mauve Contig Mover (MCM) and, 150  
nanopore sequencing and, 343  
Needleman–Wunsch (NW) algorithm, 49, 54, 110–113  
of next-generation sequencing reads, 115–116  
overview, 109–110  
Phrap program and, 62–63  
Short Oligonucleotide Alignment Program (SOAP), 115–116  
similarity to sequence assembly, 6  
Smith–Waterman (SW) algorithm, 38, 49, 62–63, 111–113  
SOAP2, 116, 124  
statistical significance of, 64–65  
Sequence analysis tools, 50  
Sequence assembly. See Assembly; DNA fragment assembly  
Sequence databases, 73–86  
arhival data storage, 73–74  
cancer genomes, 77  
cloud computing, 78–79  
data compression, 79–84  
data privacy, 84–86  
European Genome-Phenome Archive (EGA), 78  
European Nucleotide Archive (ENA), 74–76  
GenBank, 73, 74  
identifiability of sequence data, 86  
International Cancer Genome Consortium (ICGC), 78  
National Center for Biotechnology Information (NCBI), 73, 74  
1000 Genomes Project, 76  
protein sequence, 328–333  
searching, 64–68, 113–115  
Sequence editing  
desktop software, 58–59  
Wisconsin/GCG package, 57  
Sequence enrichment methods, 242–244  
Sequence file formats, 51. See also specific formats  
Sequence fragments, 37, 54, 101  
Sequence length distribution, 36  
Sequence Ontology Project, 93  
Sequence quality, 33–37, 34f–35f  
duplicate sequences, 36  
FastQC program, 33–37, 34f–35f  
overrepresented sequences, 36–37  
per base “N” content, 36  
%GC, 35  
sequence length distribution, 36  
Sequence Read Archive (SRA), 73–76  
Sequence reads, 4, 31, 73, 90, 93. See also Reads  
assembling shotgun with GAP program, 55  
base position in, 192  
ChIP, 218, 220, 221f, 223–225  
prefiltering RNA-seq, 275  
Sequence Scanner, ABI, 52  
SequenceSqueeze, 81, 82  
Sequence variants, 11, 37, 90, 109, 351. See also Single-nucleotide polymorphisms (SNPs); Variant detection  
genome annotation, 171  
PCR sequencing and heterozygosity, 52–53  
in RNA, 280  
single-nucleotide variants (SNVs), 191–198, 329–331, 330f, 332f  
transcript, 331  
Sequencer software, 51f, 52, 58–59, 59f, 152f  
Sequencing Analysis Viewer (SVA), 29–30, 30f  
Sequencing by hybridization, 129–130, 133  
Sequencing by synthesis, 10, 13, 254  
Sequencing chip, 129  
Sequencing depth  
ChIP-seq, 220–223, 222f  
deep sequencing, 11, 202, 206, 276  
Sequencing informatics, history of  
desktop sequence assembly and editing software, 58–59, 59f  
early methods, 48–49  
electropherograms, 51–52, 51f  
gel reading software, 49–50, 50f  
IntelliGenetics, 57  
overview, 47–48  
PCR sequencing and heterozygosity, 52–53, 53f  
Phred/Phrap, 60–64, 63f  
searching DNA sequence databases, 64–68  
sequencing analysis tools, 50  
sequence file formats, 51  
Staden package, 53–56, 55f  
Wisconsin/GCG package, 57–58  
Sequencing primers, 4, 31  
design with Consed, 63  
transcriptome assembly process and, 157  
SeqWare, 368–369  
SFF (standard flow formats), 11, 13, 76, 90  
SFF Workbench, 13  
SGP, 181t  
Shannon, Claude, 47  
Shannon entropy/evenness, 319  
SHARCGS, 134  
Short Oligonucleotide Alignment Program (SOAP), 115–116, 223  
Short read archive, 337, 338f  
Short reads, 109–110, 115, 119  
alignment software for, 223–224  
assembling and editing in GAP5 program, 56  
CLCbio system and, 315  
de novo assembly of bacterial genomes from, 141–153  
de novo DNA assembly using, 128, 133–134  
454 system, 11  
Hi-C and, 248

© 2015 by Cold Spring Harbor Laboratory Press. All rights reserved.
Short reads (Continued)
   Illumina Genome Analyzer, 142
   k-mers, 130
   MAQ and, 194
   next-generation sequencing and, 8
   RNA, 182
   shotgun sequencing generation of, 127
   Velvet as short read assembler, 158
Short tandem repeats on Y chromosome (Y-STR), 86
Shotgun sequencing, 127–128, 133
   metagenomic sequencing, 314–316
   overview, 6–7
   strategy, 54
   SHRIMP, 281
   SICER, 231, 235, 237t, 255
Sickle, 40
SIFT (sorting intolerant from tolerant) algorithm, 210
SILicos, 150
SMRT (single-molecule real-time) technology, 349
SnpEff/SnpSift software package, 210
SNPeffect, 174t
snoRNA (small nucleolar RNA), 274
Small nuclear RNA (snRNA), 250
SMRT, 176t
Smith–Waterman–Gotoh algorithm, 42
Smith–Waterman (SW) algorithm, 38, 49, 62–63, 111–113, 113f
   BWA compared, 195
   nanopore sequencing and, 344
   searching databases, 64
SMRT (single-molecule real-time) technology, 349
SNAP, 181t
snRNA (small nuclear RNA), 274
SNPctect, 174t
SnpEff/SnpSift software package, 210–211
SNPs. See Single-nucleotide polymorphisms
snRNA (small nuclear RNA), 250, 274
SNVs. See Single-nucleotide variants
Snyder, Michael, 16
SOAP2, 116, 124, 223
SOAPdenovo, 158t, 182
SOAPdenovo-Trans, 158t, 162–164, 166, 182
SOAP (Short Oligonucleotide Alignment Program), 115–116, 223
Software. See also specific applications; specific programs
   analysis of DNase-seq/FAIRE-seq experimental data,
   244–245
   base-calling, 52, 59–62
   ChIP-seq differential analysis, 238–240, 239t
   ChIP-seq peak calling, 230–236
   chromatin three-dimensional structure analysis,
   247–249
   CLIP-seq data analysis, 251–252
   desktop sequence assembly and editing, 58–59, 59f
   differential expression detection in RNA-seq data, 291–293, 293t
   gel reading, 49–50, 54
   motif discovery, ChIP-seq and, 240–242
   open source, 57, 58
   transcriptome assembly, 155–166, 158t
Solexa, 13–15
SolexaQA, 40
SOLiD sequencing, 313
   European Nucleotide Archive and, 76
   overview, 17–20, 18f–20f
   Somatic mutations, cancer and, 201–204
   Somatic Sniper, 202–203
SONS, 317
Sorting intolerant from tolerant (SIFT) algorithm, 210–211
SourceForge, 194
Spectral library, 325–326
SpectraST, 326
Spectrum Mill, 326
Splice junctions, 281–283
Splicing, alternative, 293–296, 294f–295f
Splign, 182
Split read methods, 200–201
SPP, 231, 235, 237t
SRA Toolkit, NCBI, 81
SPP, 231, 235, 237t
SIRIUS, 142
SraRuninfo, 111f
SRA statistics, 229
SRA Toolkit, 81
Staple, 40
Staphylococcus aureus, 142, 145, 150, 151f, 153
Staphylococcus pneumoniae genome, 9, 11
Staden package
   base calling, 55–56
   capabilities of, 53–54
   ERA (Estimate Base Accuracy) program, 56
   GELIN program, 54
   Genome Assembly Program (GAP), 55–56, 55f
   overview, 53–56
   SEQFIT program, 54
   sequence assembly, 49, 55–56, 98, 99f, 128
StamLab, 241
Staphylococcus aureus, 141, 351
STAR (Spliced Transcripts Alignment to a Reference), 282
   Statistically aided long-read haplotyping (SLRH), 354
   Strelka, 202–203
Streptococcus, Group A, 142
Streptococcus mutans genome, 142, 145, 150, 151f, 153
Streptococcus pneumoniae genome, 9, 11
Structural variants, 191, 198–201
Sturtevant, 47
Suffix array, BWT, 120, 121, 123
Sun Grid Engine resource management system, 162
Superfamily, 176t
Supported Oligo Ligation Detection. See SOLID sequencing
   Support vector machine, 344, 346, 348f
   SureSelect, 205
SVAp (Sequencing Analysis Viewer), 29–30, 30f
   SW algorithm. See Smith–Waterman (SW) algorithm
   SwissModel, 177t
   Synapse, 78–79
Index

VNTRs (variable number tandem repeats), 199
Voom function, 289, 292

W
Walk-left algorithm, 121
Watson, James, 10
Webb, Watt W., 22
WGSQuikr, 315
Whole-genome sequencing
  cost, 204
  nanopore sequencing, 340
  variant detection and, 206–208
WIG (wiggle) format, 92–93, 99
Wisconsin/GCG package, 57–58
Word hashing, 62
WormBase, 104, 174t
Wu, Ray, 1
wwPDB, 178t

X
X11, 60, 364
X!Hunter, 326
X! Tandem, 326

Y
Y-STR (short tandem repeats on Y chromosome), 86

Z
Zero-mode waveguide (ZMW) nanostructure arrays, 22
ZINBA (zero-inflated negative binomial algorithm), 231,
  235–236, 237t, 244
Zmap, 179t
ZOOM, 223