Index

Note: page numbers in *italics* refer to figures and tables

AAUAAA polyadenylation signal, 261
*ab initio* gene prediction, 65, 66, 71, 91–2, 101
abacavir, hypersensitivity reaction, 520
ABC transporters, 518
accession numbers
primary, 98
secondary, 98–9
Acembly structure prediction program, 71, 77
adaptive pressures, 110
ADME genes, 12
ADRB2 receptor polymorphism, 510–13, 514
adverse drug reactions (ADRs), 508
bioinformatics used in understanding, 519–21
factors affecting, 509
genetical associations, 519–21
Affymetrix, 240, 372
microarrays, 166, 351, 372
SNP chips/arrays, 150, 151, 158, 426, 432, 460–1
compared with Illumina panels, 461, 466, 467, 469
Alagille syndrome, 273
alanine, 319, 320, 321
alcohol-related phenotypes, 408
aliases, checking, 99
alkaptonuria, gene responsible, 508
ALLASS software, 230
alleles
frequency, estimation using Perl, 26–8
identical by descent (IBD), 219
identical by state (IBS), 219
sharing, 219
Alzheimer’s disease, 9, 166, 251, 507
amino acids
amphipathic, 319, 328–9
aromatic ‘stacking’ interactions, 320
characteristics (of individual amino acids), 321–34
classifications, 316–18
by chemical/physical/structural properties, 317–18
by mutation/substitution matrices, 316–17
codes, 530
disulphide bond formation, 273–4, 312, 316, 331
glycosylation of, 316, 329, 330, 331
hydrophobic, 319–20
with aliphatic side-chains, 319, 321–3, 333–4
with aromatic side-chains, 319–20, 323–5
pairs, effect on protein structures, 337
phosphorylation of, 316, 324, 330, 331
polar, 320, 324–31
properties, 318–21
evaluation tools, 270
sequences, alteration by DNA variants, 10
small, 320–1
Taylor classification, 317–18
variants, impact on protein structures, 272, 334–9
amino-acid substitutions
factors affecting consequences, 268–74
functional context, 273–4
matrices, 533–5
physicochemical environment affecting, 269, 511
protein structure affected by, 270–2
site-specific, 335–6

ANALYZE software, 224–5
ancient repeats, 127–8, 132
aneuploidy, in tumour genomes, 425
Angelman syndrome, 157
anonychia, homozygosity mapping, 151–2
ApoE/APOE gene, 9, 166, 251
Apolipoprotein E, 9, 166
Appied Biosystem, 148, 150
Arg184Cys mutation, 273
arginine, 319, 320, 326–7
interaction with phosphates, 326, 327
in salt bridges, 326
Argonaute database, 345, 363
ARLEQUIN software, 172, 230
array comparative genomic hybridization (acCGH)
analysis, 428–31
publicly available data resources, 427–8
technologies, 425–7
ArrayExpress data repository, 71, 75, 377, 432
asparagine, 316, 320, 329
aspartate, 317, 328
aspartic acid, 317
association analysis, 223–9
Bayesian approaches, 457–9
compared with linkage analysis, 9, 10, 448
complex disease susceptibility genes identified using 448–9
cross-validation approach, 455
direct approach, 43
gene loci defined using, 190–1
genome-wide, 11, 447
HapMap data used, 42–53, 176–8
indirect approach, 43
investigating positive findings, 454–5
replication aspects, 455–6, 475–6
sample sizes, 450–1, 452
thresholds of significance, 451–2
type I errors (false-positives), 11, 167, 187, 451, 454
asthma, 510
AVID alignment tool, 119, 121–2
BACE1 gene, 97
accession numbers, 98, 99
 aliases/synonyms, 99
ECR comparative genome browser, 92, 93
UCSC genome browser, 89, 90, 98, 100, 101
Bacillus steaothermophilus, 338
backcross (BC), 235
bacterial artificial chromosome (BAC) clones, 60
basal cell carcinoma, 158
BASE microarray database package, 379–80
compared with other tools, 378
Bayesian analysis, 457
in genome-wide association analysis, 457–9
BayesMiRNAfind tool, 345, 350
BiDil (nitrous oxide enhancer), 521
BioCarta pathway tool, 478
and DAVID annotation tool, 483, 486, 487
BioMart software, 71, 80
BLAST, 71, 87, 97
applications, 62, 97–8, 120
limitations, 8
BLASTN, 65, 120
BLASTP, 67
BLASTX, 201
BlastZ alignment tool, 119, 121
BLAT, 65, 71, 76, 87, 97, 119, 124
homology searching, 87, 192, 193
BLOSUM matrices, 317one mineral density (BMD)
factors affecting, 167, 227
lowering stringency, 452, 474
Bonferroni correction, 180, 451
BRAC1/BRAC2 genes, 9, 148, 254, 416, 433
BRAF locus, 416
breast cancer, 9, 148, 254, 416
BXD recombinant inbred (mouse) strains, 398–9
C. programming language, 473
cancer
breast cancer, 9, 148, 254
colon cancer, 439
epigeneics, 204, 438
lung cancer, 424
miRNA/ncRNA in, 356–8
neuroblastomas, 416, 425, 439
ovarian cancer, 9, 416
prostate cancer, 166, 175, 300, 437–8
cancer cell lines, 417

cancer genetics, 413–45
approaches to studying, 415–18
cancer cell lines, 417
recent technologies, 418
study design, 416–17
general resources, 418–20
Cancer Genome Anatomy Project (CGAP), 418–20, 419
Cancer Genome Project (CGP), 418, 419
cancer-related gene list, 422
genome copy number data, 428
cancer genomes, 414–15
candidate gene(s)
bioinformatics approaches to identify, 524–5
biological rationale, 201–3
Crohn's disease, 204, 210, 212, 251
interactions, 5–6
mapping, 152
Parkinson's disease, 479
candidate polymorphisms, 251–2
CARD15 gene
association with Crohn's disease, 192, 251, 468
see also NOD2 gene
case-control cohorts, 167, 176
Catalogue of Somatic Mutations in Cancer (COSMIC), 419, 421–4
query interface, 423
catechol drug pharmacogenetics, gene(s)
associated, 251
cathepsin K, 495
CCDS project, 87
CCND1 gene, 433
CCR6 gene, 189
CDK4 gene, 435, 436
CDKN2A gene, 415, 424, 425
Celera Genomics (CG)
genomic sequence assembly method, 60–1, 61–2
compared with other assembly methods, 62–3
Center for Information Biology Gene Expression (CIBEX) data repository, 377
CFTR gene, 8
CGHAnalyzer software, 429–30, 436, 437
CHAOS alignment tool, 119, 120
ChARM software, 435

Chemical Effects in Biological Systems (CEBS) database, 515, 525
'chemical individuality', 508, 509
chi-square test, 175, 223, 224
chromatin immunoprecipitation (ChIP) technique, 207, 286, 295
data, 206, 295
chromosomal aberration(s), databases, 357, 419, 427
circular binary segmentation (CBS) algorithm, 430, 431
cis-acting expression quantitative trait loci, 14
CisModule algorithm, 283, 293
cis-regulatory elements, 106
location, 136–7
cis-regulatory modules (CRMs), 257–8
constructing, 293
databases, 294
predicting, 294
Clusters of Orthologous Groups of proteins (COGs), database, 67, 71
coregulated genes, identification of, 295
codon usage, 531
colon cancer, 439
combinatorial network-analysis methods, for system genetic analysis, 400
common disease/common variant (CD/CV) hypothesis, 251, 449, 469
alleles supporting, 251
comparative genomic hybridization (CGH), detection of, 157–60
comparative genomics, 8, 92–4, 105–44
applications, 132–7
disease-related studies, 135–7
specific locus studies, 135–7
ultra-conserved regions, 133–5
concepts, 109–13
branch length/distance, 111–12
divergence time, 111
phylogenetic scope, 111
future directions, 137–8
integration with genetic variation data, 137–8
and population genetics, 131–2
practicalities, 113–18
available genomic sequences, 113, 115–16
defining and obtaining genomic sequences, 116–18
comparative genomics (cont.)
technology, 118–32
alignment production, 118–25
detecting selection, 126–31
visualizing genomic alignments, 114, 125–6
‘compartmentalized shotgun assembler’, 61
complex disease, 9
multigenetic causes, 250
non-coding RNAs in, 359–62
role of epigenetics, 205
complex disease susceptibility alleles, likely nature, 450
composite interval mapping (CIM), 235, 396
computer programming
documentation for, 19–20
modularity, 19
problem solution by, 20
COMT gene, 251
consensus coding sequence (CCDS) project, 89
CONSENSUS program, 283, 291
conserved non-coding sequences (CNS), 267
conserved non-genic sequences (CNGs), 133
copy number alterations, in cancer, 425–31
copy number polymorphisms (CNPs), 157–60, 186, 209, 415
tools for study, 209, 213
CpG islands, 255, 285
Crohn's disease, 9, 166, 174, 191
candidate genes, 9, 166, 174, 192, 204, 210, 212, 251, 468
biological rationale, 201–2
susceptibility locus (IBDS), 191–209
cross-species genome comparison, 8
CSHLmpd database, 283, 285–6
CTSK gene, 227
CYP2D6 enzyme, 516, 517
CYPAlleles database, 515, 516
cysteine, 320, 331–2
cystic fibrosis, 8, 153
cytochrome (CYP) enzymes, 516, 517
cytoplasmic sulfotransferase (SULT) enzymes, 517
data entry, 20
data integration, 7
data management, 7, 17–31
data manipulation, 17–31
basic principles, 18–21
elements, 21–30
practical details, 21–2
see also Perl coding language
data mining, 7
factors affecting, 7
data repositories for microarray data, 377
compared with research databases, 377
data storage, 20–1
Database of Genomic Variants, 209, 213
database software, 21
Database of Transcription Start Sites (DBTSS), 256, 283, 285
databases
cancer mutation, 419, 421–52
chromosomal aberrations, 357, CHAP 17
cis-regulatory modules, 294
disease-specific, 424–5
drug–gene interactions, 515, 523–4
drug-metabolizing enzymes, 515, 516, 517
gene-expression data, 71, 75, 377, 427, 432
microRNA, 345, 351, 354, 363
Mouse Genome, 69, 71, 236
mRNA, 89–91, 265
non-coding RNAs, 72, 345
promoter regions, 256, 283, 285
proteins, 87, 91, 271, 340, 345
SNPs, 7, 8, 71, 72, 150
etnic group specific, 522
transcription factor binding sites, 283, 290
see also individual named databases
DAVID tool, 478, 483–6
annotation by
BioCarta pathway, 486, 487
GO terms, 484–5, 486
KEGG pathway, 486, 487
dbSNP database, 7, 8, 71, 72, 150
HapMap data available, 39
dChip software, 434–5
deep vein thrombosis, gene(s) associated, 251
DeepView tool, 270, 272, 511
examples of use, 512, 514
diabetes
type 1, 180
type 2, 9, 251
gene(s) associated, 9, 251, 448, 452
DiGeorge syndrome, 157
disease-associated genes, 9, 103, 251
INDEX 541

Distributed Annotation System (DAS), 71, 80, 87
disulphide bonds, formation of, 273–4, 312
divergence time, 111
DMAP program, 230
DNA copy number alterations, 425–31
databases, 427–8
DNA methylation sites, 207
DNA mutations
diversifying selection, 110
purifying selection, 109
DNA polymorphism, 252, 260
dominant disease, mutation, 152
dotplots, genomic sequence alignments visualized by, 114, 125–6
Down's syndrome, 157
Dragon PF and GSF programs, 283, 288
drug absorption, distribution, metabolism and excretion (ADME), 513, 515–18
genes involved in, 515, 516–18
routes, 516
drug discovery and development
causes of failures, 496
genetics in, 11, 497–8
project timescale, 498
drug-metabolizing enzymes, 516–17
DrugBank, 503–4, 506, 515, 523–4
druggable gene, 500
druggable genome, 502
website, 504, 506

ECR comparative genomics browser, 92
BACE1 gene, 92, 93
EHPLUS software, 172, 226, 229
embedded genes, 96–7
EMBOSS package, CpGPlot program, 207, 213, 283, 285
Encyclopaedia of DNA Elements (ENCODE) project, 37, 138, 206, 351
epigenomic data, 206
ncRNA data, 351, 352
Ensembl database, 8, 13, 70, 72–3, 117
cross-species comparisons, 92
database management system used by, 473
genome browser, 71, 72–3, 87, 176, 213, 283
FOXP2 gene, 73, 74
promoter region annotation, 257
HapMap data available, 39
Entrez Gene database, 71, 78, 87, 88, 99, 283
ePCR program, 65, 71

EphB6 gene, 489
epigenetic analysis, Mendelian disorders, 160, 162
epigentic, 12–13, 204
and cancer, 204, 438
integrated into genetic analysis, 205
epigenomic data, 206–7
epigenomics, 204
epistasis, 456–7
Eponine program, 283, 287
ESEfinder tool, 257, 260, 283, 297
ethnic groups, in HapMap, 37, 521
ethnicity, and pharmacogenetics, 521–2
Eukaryotic Promoter Database (EPD), 256, 283, 285
European Bioinformatics Institute (EBI) ArrayExpress repository, 377
database(s), 68, 71
EvoFold RNA secondary structure prediction method, 351
evolution
drug absorption, distribution, metabolism and excretion

false discovery rate (FDR), 392, 395, 451
Family Based Association Tests (FBAT), 176
INDEX

FASTLINK software, 218
FastSNP tool, 270, 275–6
FCGR3B gene, 158, 159
FGFR1 gene, 436
fibroblast growth factors, 314, 315
FirstEF program, 283, 288
FLT3 gene, 423
FlyBase database, 69, 71
Food and Drug Administration (FDA, USA), on pharmacogenetics, 497
formaldehyde-assisted isolation of regulatory elements (FAIRE) procedure, 207
four-box model approach to genome scan prioritization, 476, 477
fourfold degenerate sites, 127
FOXP2 gene, 68
FOXP3 gene, 254
FPC program, 60
fragile X disease, 152
Frataxin, 9
frequentist analysis, 458
Friedreich’s ataxia, 9
FTE text editor, 472
functional polymorphisms
cis-acting, 282
in genes and gene regulatory sequences, 254
identifying, 211–12
trans-acting, 282
G-protein-coupled receptors (GPCRs), 87, 271, 502, 510
database, 506, 510, 513, 515
as our receptors, 503
G2D (Genes to Diseases) tool, 210–11
gain-of-function (GOF) effects, 499, 500
GBrowse genome browser, 71, 80
Gecko microarray database package, 380–1
compared with other tools, 378
GenBank database
BAC clone sequence data, 60, 64
Homo sapiens CAGH44 mRNA, 68
gene(s)
aliases and synonyms, 99
anatomy, 255, 258–65
content, 106–7
definition, 106
embedded genes, 96–7
locating
by gene names and symbols, 99
by genome coordinates, 100
primary accession numbers, 98
by raw sequence data, 97–8
secondary accession numbers, 98
names and symbols, 99
overlapping genes, 96
splicing, 258
synonyms and aliases, 99
gene analysis, reasons for learning how to, 86–8
gene expression, 281
as filter for output of genome scans, 479–81
heritability, 391–3
sources of variation, 391
gene expression data
analysis software, 434–5
in cancer, 432–5
databases, 71, 75, 377, 427, 432
gene expression microarrays, 11, 295
earliest use, 390–1
Gene Expression Omnibus (GEO) data repository, 71, 75, 377, 427, 432
gene finding and analysis, tools, 71–2, 87
gene–gene interactions, 456–7
gene locus
defining from linkage and association data, 189–91
definition, 188–9
gene models, complexities, 95–7
Gene Ontology (GO) project, 69, 71
classification terms, 211, 313
in DAVID annotation tool, 483, 484–5, 486
gene–phenotype association
detection of, 153–7, 175–6
proof required, 180
gene prediction, 65
ab initio 65, 66, 71, 91–2
Acembly program, 71, 77
GENEWISE algorithm, 66, 70, 71, 87
GENSCAN program, 66, 70, 72, 87
reasons for learning how to, 86–8
use of sequence similarity, 65–6, 92–4
gene product(s)
evidence cascade for, 88–94, 95, 101
heterogeneity, 96
gene regulation, factors affecting, 14
gene regulatory networks (GRNs), 281
gene regulatory sequences
analysis, 281–309
resources, 283
GENEHUNTER software, 218, 226
GENEHUNTER-PLUS software, 219
GeneNetwork.org, 14, 390, 401–2
   WebQTL tool, 390, 401
   correlation results, 404
   interval mapping, 405
   query interface, 402
   trait data page(s), 401, 403
Generic Genome Browser, HapMap data available via, 40
Generic Model Organism Database (GMOD) project, microarray data module, 382
Genetic Association Database (GAD), 500, 506
   genetic association network, 407
   genetic association scans, 102
   see also genome scan analysis
   genetic correlation analysis, 397–400
   impact of study design, 398
   reference populations, 398–400
   genetic data
   managing and manipulating, 17–31
   see also data . . .
   Genetic Data Analysis (GDA) tool, 230
   genetic markers, combining data using Perl, 23–6
   Genetic Power Calculator website, 471
   genetic reference populations, 398–400
   recombinant inbred mouse strains, 398–9
   and relations of gene expression to complex phenotypes, 399–400
   standard inbred mouse strains, 399
   genetic study design and analysis, 8–10, 145–246
   genetic traits, role of bioinformatics in understanding, 4–5
   genetic variation
   diseases caused by, 249
   functional analysis, integrated tools for, 274–6
   sources, 209
   genetic/genomic data interface, analysis at, 10–12, 369–528
   genetical genomics, 390
   genetics research, role of bioinformatics, 4–12
   GENEWISE algorithm, 66, 70, 71, 87
   genome
   annotation, 59–60, 64–80
   conserved non-genic sequences (CNGs), 133
   coordinates, location using, 100
   copy number
   analysis and visualization, 429
   break-point analysis, 428–9
   data analysis tools, 428–31
   data resources, 419, 427–8
   measurement technologies, 425–7
   functional sequences, 110, 133
   localizing markers in, 192–3, 194
   portal inspection, 100, 101
   pre-assembled, 117
   repetitive elements/sequences in, 107–8, 123
   segmental duplications, 109, 115–16
   sequence mutation, 108–9
   ultra-conserved regions, 133–5
   ‘unknown unknown’ elements, 14
   variations in, 108–9
   visualization tools, 71, 87, 213
   see also human genome
genome-scale datasets, practicalities of analysis, 471–3
   genome scan, prioritization approaches, 476, 477, 488
   genome scan analysis, 449–59
   bioinformatics, 469–89
   filtering and annotating output, 479–89
   maximization of inclusivity, 477–9
   genome-scanning technologies, 459–69
   genome-wide association analysis, 11, 447
   Bayesian analysis used, 457–9
   follow-up strategies, 474–6
   Parkinson’s disease case study, 470–1
   reasons for use, 448–9
   replication studies, 455–6, 475–6
   SNP genotyping panels used, 460–8
   genome-wide mapping of single-gene disorders, 148–52
   microsatellite mapping approaches, 148, 150
   SNP-mapping approaches, 150–2
   genome-wide SNP genotyping panels
   Affymetrix design, 460–1, 461, 466, 467
   evaluation of, 463–6
   case study, 466–8
   Illumina design, 461, 463, 466, 467
   Genomes OnLine database, 59, 71
   genomic control, 167–8
   genomic databases, 8
   genomic prediction, 91–2
genomic sequence
annotation, 59–60, 64–80
future developments, 78–80
nucleotide level, 64–7
preliminary annotation, 59
process level, 68–70
protein level, 67–8
assembly
CG method, 60–1, 61–2
IHGSC method, 60, 61
NCBI method, 62
characterization
at locus under investigation, 168–9
tools, 71, 72–3, 75–8, 87, 176
defining and obtaining, 116–18
extracting and annotating across locus, 194–5
'finished', 63, 113
vertebrates, 102, 113–15
visualization of alignments, 114, 125–6
genomic sequence alignments, 118–25
anonymous/unannotated sequences, 128
distribution of control and test sequences, 129–31
global alignments, 121–2
local alignments, 120–1
multiple-sequence alignments, 122–4
tools, 119
evaluation of success/accuracy, 124
visualizing, 114, 125–6
whole-genome, 124–5
genomics, comparative, 8, 92–4, 105–44
Genomics Unified Schema (GUS), 381
genotyping, improvements in technology, 3, 4
GENSCAN gene prediction program, 66, 70,
72, 87
germline polymorphism, and somatic
mutation, 415
glutamate, 317, 319, 328–9
glutamic acid, 317
glutamine, 319, 320, 329
glutathione-S-transferase (GST) enzymes, 517
glycine, 317, 320, 332–3
GNF gene-expression data, 203
GOLD program, 42, 230
GOLDsurfer program, 42, 230
Google, 7
Google Scholar, 99, 202
GPCRDB, 506, 510, 513, 515
grapefruit juice, effect on drug efficacy, 509
haemoglobin genes, mutations in, 311, 339
Haplopter tool, 522
haplotypes, 170–5, 225
construction of, 172
reconstruction of, 225–9
statistical analysis, 170, 175–6, 225–9
HaploView, 40, 42, 197, 213, 231
LD plots, 41, 233
worked example, 231–4
HapMap, 3, 7, 13, 35–58, 172
accessing HapMap data, 38–42
application in association studies, 42–53,
449
databases, 39
defining locus, 195–7
downloading HapMap data, 38–40
bulk download, 39–40
via Generic Genome Browser, 40
via HapMart, 40
ENCODE regions, 460, 464, 466
etnic groups covered, 37, 521
and fine-mapping experiments, 53
future developments, 54
genome browser, 40, 41, 195, 196, 213
as data-mining/analysis tool, 196–7
genotyping of SNPs, 37–8
historical background, 35–6
linkage disequilibrium data, 40–2, 459,
460
quantitative trait analysis, 11
reasoning behind, 36, 173
SNP ascertainment strategy, 450, 466
subject populations, 36–7
criteria used to assign membership, 37
tag SNPs, 176, 178
viewing HapMap LD data, 40–2
website, 38, 39, 213
HapMart tool, 40, 52, 197, 198, 213
HAPPY software, 236
Hardy–Weinberg equilibrium, 225
testing for, 176, 454
heritability
estimation in microarray analysis, 392–3
gene expression, 391–3
heterozygosity, loss of, in cancer, 431–2
HEXB gene, 267
HGVBase, 500
Highwire search engine, 202
Hirschprung’s disease, 135, 136
histidine, 319, 320, 325
INDEX

histograms of conservation, genomic sequence alignments visualized by, 114, 126
HLA-B*5701 520, 521
HMMER software, 68, 72, 317
homologues, 110
homozygosity mapping, 151–2
Human ABC-Transporter Database, 515, 518
Human Epigenome Project (HEP), 13, 207
Human Gene Mutation Database (HGMD), 150, 419, 424
human genome annotation
  Ensembl browser, 70, 71, 72–3, 74
  future developments, 78–80
  nucleotide level, 64–7
  preliminary annotation, 59
  process level, 68–70
  protein level, 67–8
  UCSC browser, 71, 73, 75–6
browsers, 8
'finished' sequence, 63, 113
length/size, 106
locating known genes, 97–100
number of protein-coding genes, 94, 107, 258, 347
and other vertebrate genomes, 4, 102, 113–15
selective constraint in, 132–3
sequence, 186–7
  characterization, 3–4
  'finished' sequence, 63, 113
  gene location using, 97–8
sequence assembly
  CG method, 60–1, 61–2
  IHGSC method, 60, 61
  NCBI method, 62
'unknown unknown' elements, 347
Human Genome Organization (HUGO), gene symbols, 99
Human Membrane Transporter Database (HMTD), 515, 518
Human Structural Variation Database, 158
example of use, 149
huntingtin, 9
Huntington's disease, 9, 152
hypersensitivity reaction, 520
IARC, TP53 Mutation Database, 419, 424–5
IBD1 locus, 192
IBD5 locus
  in silico characterization of, 191–209
  building biological rationale around candidate genes, 210–13
  defining locus in HapMap, 195–7
  definition of known and novel genes across locus, 197–201
  evaluating epigenomic and epigenetic effects, 204–9
  evaluating structural variation across locus, 209
  extracting and annotating genomic sequence across locus, 194–5
  gene expression analysis, 203–4
  localizing markers in genome, 192–4
  identical by descent (IBD) alleles, 219, 223
  identical by state (IBS) alleles, 219
  IdSelect program, 44
  Illumina, 240
  SNP arrays/panels, 150, 426, 461, 463
  compared with Affymetrix chips, 461, 466, 467, 469
  immune-mediated adverse drug reactions, 519–21
in silico analysis, gene regulatory polymorphism, 281–309
in silico predictions, 102
  prioritization for further investigation, 276
in silico science, bioinformatics perceived as, 12
indel-based measures of selective constraint, 131
inflammatory/irritable bowel diseases, 192, 202
biological rationale for various genes, 209–11
see also Crohn's disease; IBD1; IBD5 locus
Inpharmatica, list of druggable targets, 503
insertions and deletions (indels), 131
see also indel...
Institute for Genomic Research, Multiple Experiment Viewer, 435
insulin gene, 9, 180
internal ribosome entry site (IRES) elements, 263, 264
International Human Genome Sequencing Consortium (IHGSC)
  genomic sequence assembly method, 60, 61
  compared with other assembly methods, 62–3
International Protein Index (IPI), 87, 91, 506
INDEX

Internet, 7, 99
InterPro database, 68, 71, 87, 270
interspersed repetitive elements (IREs), 107–8, 127–8
intracellular proteins, 312
- amino-acid substitutions, 269, 534
- introduction to bioinformatics, 1–31
- intronic splicing enhancers and silencers (ISE/ISS), 255, 259
- examples, 254
- introns, 255
- inverse genetics, 266–7
- example of use, 269
- isoleucine, 319, 321–2, 511
- IUPAC codes
  - amino acids, 530
  - nucleotides, 529

Jagged1 protein, evaluation of mutation in, 273–4
JASPAR database, 283, 290
JSNP database, HapMap data available, 39

Kcnj9 gene
- association network, 407
- QTL mapping analysis, 403–5, 408
KCNJ11 gene, 251
KEGG pathway tool, 478
- and DAVID annotation tool, 483, 486, 487
Kimweb (kinases database), 506, 513, 515
knowledge management, 5–7
- 'known' gene product, meaning of term, 86
- known gene, locating in human genome, 97–100
Kozak consensus sequence, 262, 263
KRAS gene, 424

laboratory information-management systems (LIMS), 7
laboratory notebook discipline, in data analysis, 18–19
lactase gene, inter-ethnic differences, 522
lactate dehydrogenase, 338
Lagan alignment tool, 119, 121, 122, 124
language-acquisition disorders, 68
leucine, 319, 322
linear discriminant analysis (LDA), 287
linkage analysis, 165–80, 217–23
- compared with association analysis, 9, 10
- genetic loci defined using, 190
nonparametric approach, 219–20
- MERLIN used, 219, 221–3
parametric approach, 218–19
- Perl used, 22
- preliminary, 176–8
- study population, 166–8
linkage disequilibrium (LD), 169–70, 229–34
- characterization using HapMap, 7, 35
- genome scan analysis inclusivity maximized using LD data, 477–9
- HapMap data, 40–2, 459, 460
- maximum-likelihood methods, 230
- measures, 229
- moment method, 230
- SNP selection using, 43–53
- software, 42, 230–1
- squared correlation coefficient, 229–30
Lipinski’s 'rule-of-five', 504
literature search, 6–7
locked nucleic acid (LNA) probes, 355
locus definition/identification, 187–9
locus refinement, 178–9
LOD (log of odds) score, linkage region defined by, 190, 218, 222–3
Longhorn Array Database (LAD), 384
loss of heterozygosity (LOH) in cancer, 431–2
- data analysis tools, 431–2
- loss-of-function (LOF) effects, 499, 500
- LS-SNP tool, 270, 274
- lung cancers, 424
- lysine, 319, 320, 327–8
McDonald–Kreitman test, 131–2
McNemar’s test, 224
McPromoter program, 283, 287
malaria, protection against, 153
malate dehydrogenase, 338
Map Manager QTX software, 236
- data input, 237
- simple interval mapping, 238–9
- single marker association testing, 238, 239
- worked example, 236–9
mapping
- candidate gene, 152
- quantitative trait locus (QTL), 235–9
markers
- genomic sequence, combining data, 23–6
- multi-allelic, 224–5
- simple tandem repeat, 165, 174–5
| MatchMiner, 99 | microdeletion syndromes, 153 |
| MAVID alignment tool, 119, 123 | micro-RNAs (miRNAs), 4, 10, 14, 133 |
| MECP2 gene, 160 | databases, 345, 351, 354, 363 |
| Mega2 data-handling program, 21, 31 | expression studies, 354–6 |
| melanomas, 416 | first discovered, 343, 349 |
| genetic alterations, 435–6 | number in Man, 14, 343 |
| MEME program, 283, 291 | role in cancer, 356–8 |
| Mendelian disorders, 8–9 | role in regulation of multiple genes, 346, 348, 349 |
| mutations, 259 | see also miRNA genes |
| identifying, 8–9, 147–64 | microsatellites, 8, 148, 174 |
| Mendelian traits, epigenetic effects, 160, 162 | data format, 26 |
| MERLIN software, 150, 218–19 | mapping of single-gene disorders using, 148, 150 |
| data input, 220–1 | as markers, 192 |
| haplotype reconstruction, 226 | see also simple tandem repeat markers |
| NPL analysis, 219, 221–3 | Microsoft Excel (spreadsheet software), 20 |
| graphical output, 222 | limitations, 472 |
| worked example, 220–3 | Minimal Information About a Microarray Experiment (MIAME) standard, 376 |
| MEROPS resource, 505, 506, 507, 513, 515 | MiRanda tool, 345, 353, 354 |
| metabolic disorders, 508 | miRNA database, 345, 351, 354, 363 |
| methionine, 319, 323 | miRNA genes |
| MiFOLD structure prediction tool, 264, 345, 350 | paralogues, 360, 361 |
| microarray(s), 371–87 | prediction, 350–1 |
| analysis | machine-learning approach, 350 |
| complementary approaches, 377 | species conservation approach, 350–1 |
| experimental process, 375 | variants, as disease alleles, 359–61 |
| false discovery rate (FDR), 392, 395 | miRNA targets |
| heritability estimated in, 392–3 | genetic variation, as disease alleles, 361–2 |
| data | prediction, 353–4 |
| analysis, 376 | miRNAMap, 345, 363 |
| annotation of, 375 | MirScan algorithm, 345, 350 |
| Bayesian approaches, 400 | MiRseeker algorithm, 350 |
| filtering and selecting, 376 | Mitelman Database, 357, 419, 427 |
| QTL mapping of, 394–7 | lung tumour data, 430, 431 |
| research database packages, 377–85 | mLagan alignment tool, 119, 123 |
| sharing and publication of, 376, 377 | MLINK output, use by Perl, 29, 30 |
| storage of, 375 | MODBASE protein database, 270, 271 |
| transformations, 375, 393 | monogenic disorders see Mendelian disorders |
| visualization of, 376 | Mouse Genome Database, 69, 71, 236 |
| gene expression microarrays, 11, 295, 390–1 | mRNA |
| as genomics platform(s), 11, 371–87 | alternative splicing, 258, 296 |
| miRNA expression studies, 354–5 | databases, 89–91 |
| splicing, 299–300 | processing and translation, regulatory control, 265 |
| technologies, 372–3 | regulatory, 266 |
| Affymetrix technology, 372 | secondary structure, 264 |
| principles, 373–6 |
mRNA (cont.)
splicing
  mutations affecting, 259
  prediction of splice sites, 296
transcript polymorphisms, 263
analysis, 261–2
transcripts
  anatomy, 262
  delineating, 5′ and, 3′ ends, 95–6
  initiation of translation, 262–4
  regulatory elements, 262, 266
multidrug resistance proteins (MRPs), 518
multigenetic diseases see complex genetic
diseases
Multimapper software, 236
MultiPipMaker alignment visualization tool,
114, 119
multiple QTL mapping, 235, 396
‘multiple rare variant’ hypothesis, 449, 469
multispecies conservedseq (MCS) elements,
348
MultiZ alignment tool, 119, 125
mutation matrices, amino acids classified by,
316–17, 533–5
mutations
  cancer, 420–5
  databases, 421–5
  correlated, 335
  disease, 152
  DNA, 109, 110
  Mendelian disorders, 8–9, 152–60, 259
  potentially deleterious, 276
  protein structures affected by, 334–9
  role in determining phenotypes, 281
MVP viewer, 207, 213, 438
example, 208
MySQL database management system, 473
  use in microarray database packages, 378,
  379, 384
NCBI genomic sequence assembly method,
62
  compared with other assembly methods,
62–3
NCBI Map Viewer browser, 71, 77–8, 87, 176,
213
FOXP2 gene, 77, 78
NCI-60 tumour-derived cell lines, 417
Needleman–Wunsch algorithm, 122
neuroblastomas, 416, 425, 439
nitrous oxide bioactivity, inter-ethnic
differences, 521
N-myc gene, 416
NOD2 gene, 9, 166, 174, 192
non-coding RNAs (ncRNAs), 10, 133, 266,
343–67
annotation, 72, 344
  classification, 344–5
  computational analysis, 349–56
  identification of novel small ncRNAs, 351,
  352
  prediction, 66–7
  role in cancer, 356–8
  role in complex disease, 359–62
  variation
    assessing impact, 362–3
    role in disease, 356–62
  see also micro-RNA (miRNA)
nonparametric linkage (NPL) analysis,
219–20
MERLIN used, 219, 221–3
  graphical output, 222
non-synonymous coding polymorphisms,
  functional analysis, 268–74
non-synonymous single-nucleotide
  polymorphisms (nsSNPs)
  functional annotation, 274
  prediction tools, 270
NOS1AP gene, 452
novel genes
  across IBD5 locus, 199–201
  analysing, 101–2
  evidence for, 200–1
  meaning of term, 86
  novel regulatory elements, 267
  identification tools, 257, 268
  nuclear hormone receptor (NHRs), 504
  NucleaRDB, 506, 513, 515
  nucleotide level annotation, 64–7
  nucleotides, IUPAC codes, 529

National Cancer Institute (NCI)
  Cancer Genome Atlas, 418
  Center for Bioinformatics, 418, 419
  see also NCI-60
National Centre for Biotechnology
  Information see NCBI
National Human Genome Research Institute
  (NHGRI), Cancer Genome Atlas, 418
NCBI database(s), 71, 75, 377
  genomic sequence assemblies, 71, 117
polymorphisms (cont.)
  regulatory, 11
  analysis of, 281–309
  evaluating functional importance, 300–2
see also copy number polymorphisms (CNPs); microsatellites; single nucleotide polymorphisms (SNPs)
PolyPhen tool, 270, 274
population genetics
  and comparative genomics, 131–2
  growth in topic, 241
population stratification, 167
position shift loci (PSLs), 270
PPAR\(\gamma\)/PPARG gene, association with diabetes, 9, 251, 448, 452
Prader–Willi syndrome (PWS), 157
pre-assembled genomes, 117
'predicted', meaning of term, 86
presenilins, 507
Princeton University Microarray Database, 433
process level annotation, 68–70
processed pseudogenes, 107–8, 117, 128–9
Promoter tool, 256, 257
promoter regions, 255
  anatomy, 256–8, 284
  characteristics/definition, 201, 284–5
  databases and tools, 256, 283, 285–6
  prediction, 286–8
  tools, 283, 286, 287, 288
  UCSC browser information, 211
promoters, CpG-related vs non-CpG-related, 287–8
prostate cancer, 166, 175, 300, 437–8
proteases
  databases listing, 505, 506, 515
  as drug targets, 505, 507
protein(s)
  3D structure analysis tools, 270
  amino-acid substitutions, 269, 533–5
  cellular location, 269, 312–13
  Clusters of Orthologous Groups (COGs), 67, 71
  databases, 87, 91, 99, 271, 340, 345, 504
  disulphide bonds in, 273–4, 312
  duplication, 313
  environments, 269, 312–13
  evolution, 313–14
  extracellular, 312
  amino-acid substitutions, 269, 534
  function, 314
  effect of mutations, 334–9
  intracellular, 312
  amino-acid substitutions, 269, 534
  orthologues, 67, 313
  paralogues, 67, 313
  post-translational modification, 316
  as product(s) of genome, 106–7
  secondary structure, prediction, 270
  speciation, 313
  structure, 270–2, 313, 334
  effect of SNPs, 336
  paired amino acids in, 337
  tertiary structure, analysis, 270–2
  transmembrane, amino-acid substitutions, 269, 535
  protein-coding genes, 94, 106–7
  number in human genome, 94, 107, 258, 347
  regulation by micro-RNAs, 348
see also non-coding RNAs (ncRNAs)
Protein Data Bank (PDB), 270, 271
protein kinases, 324, 333
protein level annotation, 67–8
protein quantity locus (PQL), 390
protein sequences, experimentally determined, 88–9
proteome analysis, 79, 87
Proteome Browser, 76
pseudogenes, 96, 128–9
  processed, 107–8, 117, 128–9
  and regulatory mRNA, 266
PTEN gene, 437–8
public databases, accessing, 7
PubMed resource, 303, 499–500, 504
pupaSNP tool, 270, 275, 283
QTDFT software, 223, 225
QTL Cartographer software, 235, 236
QTPHASE program, 227, 228
quantitative trait loci (QTLs)
cis QTLs, 394–5
cis/trans test, 395
effects, 390
trans QTLs, 394, 395
quantitative trait locus (QTL) mapping
  composite interval mapping, 39, 235
  distributional assumptions, 393
  in experimental crosses, 235–9
  microarray data, 394–7
  multiple-QTL mapping, 235, 396
INDEX

multi-trait mapping, 396–7
in multistage study, 452, 453
single-locus models, 394–5

R programming language, 473
RAD microarray database package, 381–2
compared with other tools, 378
RankVISTA tool, 130
recessive disease, mutation, 152
redundancy, miRNA, 359, 360
RefSeq database, 71, 77, 87, 89, 99, 504
RefSeqNP database, 99
regulatory elements, 106, 136–7, 257–8
cross-species conservation, 287
identification and analysis, 266–7
regulatory polymorphisms, 11
analysis of, 281–309
evaluating functional importance, 300–2
regulatory regions
physical properties, 287
predicting, 282, 284–8
relational database management systems
(RDBMS), 473
see also MySQL; ORACLE
RepeatMasker software, 61, 65, 72, 123
RESCEU-ESE tool, 257, 260, 283, 298
RET gene, 135, 433
retinitis pigmentosa, X-linked, 136
Rett syndrome, 160, 162
reuse code, 19–20
Rfam database, 72, 345, 346, 347
RLMM software, 432
RNA abundance, 371, 373
RNA Abundance Database (RAD), 378, 381–2
RNA-primed array-based Klenow assay
(RAKE), 355
RNA processing, 295–6
RNA structure prediction tools, 363
RNAfold structure prediction tool, 345, 363
RNAHybrid tool, 345, 353, 354
RPGR gene, 136
Saccharomyces Genome Database, 69, 71
salt bridges, amino acids in, 326, 327, 328
schizophrenia, 205
segmental duplications, 109, 115–16
sequence mutation, 108–9
sequence similarity, 109
gene prediction using, 65–6, 92–4
sequence-tagged sites (STSs), 97
serial analysis of gene expression (SAGE)
technique, 373
miRNA characterization by, 355
serine, 316, 320, 330
'shotgun sequencing', 60
Shuffle-Lagan alignment tool, 119, 122
sib transmission disequilibrium test (S-TDT), 168, 224
application, 475
sickle cell anaemia, 311
sickle cell gene, 153
SIFT tool, 270, 274
simple interval mapping (SIM), 235
Map Manager QTX example, 238–9, 240
simple tandem repeat markers, 165
monogenic trait linkage analysis using,
174–5
see also microsatellites
SimWalk2 software, 223, 226, 230
single-gene disorders
genome-wide mapping of, 148–52
microsatellite mapping approaches, 148, 150
SNP- mapping approaches, 150–2
identifying mutations in, 8–9, 147–64
mutation in, 152–60
detection of comparative genomic
hybridization, 157–60
detection by sequencing, 153–5
detection of uniparental disomy,
157–60
other detection approaches, 155–7
see also Mendelian disorders
single-marker association testing, Map
Manager QTX example, 238, 239
single-marker linkage analysis, automation
using Perl, 28–30
single-nucleotide polymorphisms (SNPs)
association analyses on marker-by-marker
basis, 47–8
chips/arrays, 150, 151, 158, 426, 432
in conserved non-coding regions, 135
databases, 7, 8, 71, 72, 130
data not represented, 150
ethnic group specific, 522
density, 250
functional analysis, 211
genotyping of, 3, 11, 37–8
IBD5 risk haplotype, 212
iterative testing, 179
single-nucleotide polymorphisms (cont.)
mapping of monogenic disorders using, 150–2
non-synonymous
functional annotation, 274
prediction tools, 270
novel, 179
protein structure affected by, 336
selection of, 37–8
use of linkage disequilibrium to inform, 43–53
tag SNP selection, 44–7
dependency on haplotype block structure, 51–2
dependency on physical distance, 52
limitations of tagging methodology, 48–9
performance of HapMap-derived tags in other populations, 49–50
processing burden, 52
relevance of statistical methods used to test for association, 47–8
tools, 46, 173, 174
site-directed mutagenesis, 337
'six degrees of separation' concept, 5–6
SKY/M-FISH and CGH Database, 419, 427
SLC22A1 gene, epigenomic data, 207, 208
SLC22A4 gene, 204, 210
epigeneic features across, 206
SLC22A5 gene, 204, 210
'sliding windows' (alignment) approach, 129–30
SLITRK1 gene, 361–2
SMD package, 382–4
compared with other tools, 378
SMN2 gene, 297
SNAI2 locus, 415
SNPs see single-nucleotide polymorphisms
software
documentation for, 19–20
modularity, 19
SOLAR software, 219
somatic mutation
data resources, 150
and germ-line polymorphism, 415
*Sonic Hedgehog (SHH)* gene, 111, 136–7
spinal muscular atrophy (SMA), 297
splice sites, 255, 259
prediction, 260, 296
tools, 260, 283
splicing enhancers and silencers, identification of, 297–9
splicing enhancers and silencers, 180, 255, 259
eamples, 254
*in silico* derivation, 298–9
splicing microarrays, 299–300
splicing regulation, 258–9
predicting regulatory elements, 295–300
Spotfire software, 472
spreadsheet programs, 20
SR-protein-binding sites, 297
SSAHA program, 65, 72, 87
Stanford Microarray Database (SMD), 378, 382–4, 427, 433
*see also* Longhorn Array Database (LAD); SMD package
statistical analysis
general comments, 179–80
haplotypes, 170, 175–6
stochastic context-free grammar, 345
Structural Classification of Proteins (SCOP)
database, 345
Structured Query Language (SQL), 473
study population, 166–8
advantage of using, 168
Stxbp1 gene, 408
SwissProt database, 87, 91, 99, 270, 271, 504
SymAtlas tool, 478, 480, 481, 482, 515, 524–5
syonyms, checking, 99
synteny, 187
system genetic analysis, 400–2
*see also* GeneNetwork.org
system genetics, 390, 400
SYT11 gene, 479, 481
tag SNP selection see single-nucleotide polymorphisms, tag SNP selection
Tagger software, 40, 46, 52, 53, 173, 174, 178, 197, 213
target family databases, 506
proteases example, 505, 507
target identification (of drugs), 498–501
role of bioinformatics, 499
target validation and tractability, 501
targetome, 501–5
TargetScan tool, 345, 353, 354
TBLASTN, 102
therapeutic response, effect of target polymorphism, 509–10
Therapeutic Target Database (TTD), 503, 506
therapeutics development, use of genetic studies, 11
threaded blockset aligner (TBA) tool, 119, 123
threonine, 316, 320, 330–1, 511
TIGR Gene Index, 87, 89
TM4 microarray database package, 384–5
compared with other tools, 378
TNFRSF11A gene, 155
tourette’s syndrome, 361–2
transcript abundance, 371
as complex phenotype, 390–3
transcription factor binding sites (TFBSs)
consensus-based motifs, 289
de novo motif finding, 290–3
conserved motifs, 292–3
discriminative motifs, 291–2
most-over-represented motifs, 291
identification, 286
matrix-based motifs, 289–90
modelling, 288–93
position weight matrix (PWM) representation, 289–90
predicting novel, 293–4
transcription factors (TFs), number in humans, 290
transcription unit, gene as, 106, 255
transcriptional regulatory region analysis, 94
transcriptional start site (TSS), 255, 256–7, 284
database, 256, 283, 285
TRANSFAC database, 283, 290
MATCH program, 283, 294
transmembrane proteins, amino-acid substitutions, 269, 535
transmission disequilibrium test(s) (TDT), 168, 175–6, 224–5
transporters, 517–18
Transterm database, 265
TRES tool, 257, 268
trisomy, 21 157
trypsin (enzyme), 314, 315, 328
tryptophan, 319, 324
tumorigenesis, 414
Tumour Gene Database, 419, 424
tumour modelling, 438–9
tumour protein 53 (TP53), 424
mutation database, 424–5
tumour suppressor genes, 414, 415, 424, 425, 437–8
tumours, miRNA expression in, 358
tylosis with oesophageal cancer (TOC), 155
type 1 errors (false-positives), 11, 167, 187
causes, 454
correcting, 180, 451
in Parkinson’s disease case study, 470
type 2 errors (false-negatives), 167
typhoid fever, protection against, 153
tyrosine, 316, 319, 320, 324
tyrosine kinase genes, 422–3, 422
UCHL1 gene, 481, 482
UCSC Coordinate conversion tool, 187, 195
UCSC database(s), 71
genomic sequence assemblies, 71, 123–4
UCSC GeneSorter, 203, 204, 213
UCSC genome browser, 8, 13, 71, 73, 75–6, 87, 117, 176, 213, 283
BACE1 gene, 89, 90, 100, 101
conserved non-coding sequences, 267
cross-species comparisons, 92
custom tracks, 199
database management system used by, 473
ENCODER project data accessed via, 138, 206, 351, 352
epigentic data presentation, 206–7
FOXP2 gene, 75, 76
GNF gene-expression atlas, 203
HapMap data available, 39
in silico PCR tool, 193
LiftOver tool, 117
localization of markers in genome, 192–3, 194
micro-RNA targets, 354
microsatellite maps, 148, 149
Net alignments, 117, 131
non-coding RNA analysis, 351, 352
TFBS annotation, 257
WSSD duplication tracks, 116, 158, 160, 209
UCSC table browser, 275
ultra-conserved regions, 133–5
ultra-high-density genome-scanning technologies, 459–69
UniGene database, 71, 72, 87
uniparental disomy (UPD), detection of, 157–60
UniProt dataset/tool, 87, 91, 269, 510, 511, 513, 517
‘unknown’ gene product, meaning of term, 86
UNPHASED program suite, 226–7
worked example, 227–9
INDEX

uridine diphosphate-glucuronosyl transferase (UGT) enzymes, 517
UTRdb database, 265
valine, 322–3
vertebrate genomes, 102, 113–15
virtual mRNAs, meaning of term, 86
VISTA alignment visualization tool, 115, 126
VITESSE software, 218
Waardenburg syndrome, 415
WebQTL tool, 390, 401
query interface, 402
see also GeneNetwork.org
Wellcome Trust Case Control Consortium (WTCCC), 478
genome-wide association data, 490, 500
Whole Genome Association (WGA) database, 478, 500, 506
whole-genome association scans, 168–9
see also genome scan analysis
whole-genome sequence alignments, 124–5
whole-genome shotgun sequencing strategy, 60, 61, 113, 115
Williams syndrome, 157
WNT2 gene, 114, 130
WOMBAT, list of druggable targets, 503
yeast (S. cerevisiae) genome, 65
ZRS, 136–7
conservation in vertebrate development, 137

Index compiled by Paul Nash