Subject Index

a
ABBA-BABA method 241. See also D-statistic
ABC. See approximate Bayesian computation (ABC)
absolute risk model 818–821
software for 820–821
absolute value penalty 961
acceptance rate 371
activators 700
adaptive molecular evolution 369–371
along lineages
likelihood calculation under models of variable \( \omega \) ratios 380–381
in primate lysozyme 381–382
amino acid sites under positive selection
likelihood ratio test under models of variable \( \omega \) ratios 384–386
methods that test one site at time 386
positive selection in HIV-1 vif genes 386–388
computer software 391
likelihood calculation on phylogeny 379–380
limitations of current methods 390–391
Markov model of codon substitution 371–372
non-synonymous and synonymous rates 370
non-synonymous/synonymous rate ratio (\( \omega \) ratio) 370
reconstructed ancestral sequences based methods, comparison with 382–384
synonymous and non-synonymous substitution rates estimation
Bayesian estimation 377
heuristic estimation methods 372–373
maximum likelihood estimation 374–377
numerical example 377–379
testing positive selection
branch-site test 388–389
clade models and other variants 389–390
adaptive peaks 133
additive genetic covariances 423
additive genetic value 422
additive genetic variance 423, 426
ADMIIXTURE 254, 998
admixture 247, 275
admixture events, identifying/dating 262–263
DNA segments inherited from different sources 263–265
measuring decay of linkage disequilibrium 265–267
admixture model 480
African Genome Variation Project 776
age-related macular degeneration (AMD) 598
Akaike information criterion (AIC) 37, 185, 710
ALDER 265–267
alignment benchmark method 333
allele frequencies 2–3, 8, 12–14, 20–21, 32, 34–35, 400, 426
changes in 397, 426–427
genetic models for 535–539
temporal changes in 466–470
allele frequency spectrum (AFS) 260–261
allele specific expression 846
allelic dropouts 471
allelic heterogeneity 576
Allen Brain Atlas 715
Subject Index

All of Us Research Program 816
allogamous species 504
allopolyploids 506
allosteric effects 330
alpha designs 519
ALPHAPHASE software 94
altruistic participation 557
American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP) 766
amino acid replacements 350–351
empirically derived models of 348–351
AMMI (additive main effects, multiplicative interaction) 513
amplicon sequence variant (ASV) 981
analogous folds 334
ancestral alleles 535
ancestral lineage 161
ancestral recombination graph (ARG) 160–163, 412, 581
ancestral selection graph 164
ancestral sequence reconstruction 380
ancestry tract inference methods 263–265
ancient DNA (aDNA) 295–296
challenges in working with, and processing aDNA data 296
contamination 297–299
genetic studies and understanding of human past 310
archaic genomes and admixture with modern humans 310–311
demographic changes during late Neolithic and Bronze Age 312–313
Neolithic transition 311–312
handling sequence data from 300
additional filtering 300–301
downstream analysis, effects of limited data on 301–302
mapping and identification of endogenous DNA 300
preprocessing of NGS data 300
opportunities of 302–303
allele frequency trajectories 308–310
continuity 306–307
demographic inference based on ancient genomes 308
migration and admixture over time 307–308
population differentiation in time and space 303–305
sequence degradation 296–297
animal breeding 502
animal model 430–432
annotator software 682
anomalous gene trees 225
anomaly zone 225
ANOVA model 856
approximate Bayesian computation (ABC) 31–37, 413, 470, 1007
ABC-GLM 35–36
ABC-MCMC 33–34
ABC-REG 35
ABC-SMC 34
improved ABC sampling techniques 33–35
insufficient summary statistics 37
post-sampling adjustments 35–36
approximate techniques
composite likelihood 18–19
Monte Carlo sampling 16–18
using summary statistics 16
archaic hominin admixture 275–276
methods for testing 277
D-statistic 279–282
F_{ST}-statistics 282–283
genetic drift and allele frequency divergence 277
three-population test 277–279
area under the precision–recall curve (AUC) 919
ARGweaver method 415
backSPIN 749
bacterial population genomics 997–998
gene content analysis 1013–1014
Subject Index

- genetic population structure background 998
Bernoulli distribution 866
best linear unbiased estimator (BLUE) 503
best linear unbiased prediction (BLUP) 504, 802
   single step 804–805
beta distribution 20, 182, 536
beta function 21
beta prior probability distribution 182–183
big data approaches 551
BIMBAM 611
binary disease phenotypes 599
binary model 543–545
binomial probability distribution 179
Bioconductor R package 752
bioethics 552
BioGRID 781
biological processes, modeling approaches for 702–704
biosynthetic gene clusters (BGC) 990
birth–death models 1003
birth–death process of cladogenesis 195
bisulfite-sequencing (BS-seq) data 936–939.
   See also DNA methylation
   beta-binomial-based methods 936–938
   direct detection 938–939
bivariate trait vector 707
BLASTClust 333
BLIMP (Best Linear Imputation) 104
blocking, defined 913
blocks of genome 138–140
BLSMM 802
BLUEPRINT project 681
Bohring–Opitz syndrome 770
BOLT-LMM 616
Bonferroni correction 601, 956, 981, 1012
Boolean network modeling 703
BoolTraineR (BTR) 751
bootstrap method 192
BOXSET method 334
branch-site test, of positive selection 388–389
BratNextGen 1000
Bravo TOPMed 771
Bray–Curtis distance 986
Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) model 831
breeder’s equation 425, 426
multivariate 425, 444
in plant breeding 502–503
breeding success 433
breeding systems, of plants 504–505, 506
breeding values 422
Brownian motion 120, 121, 208, 1009
Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm 10
bulk RNA-seq methods 747
Bulmer’s iteration 443
burden tests 619
burn-in 27

c
CADD 682
calibration node 206
Cannings model 118–119
canonical correlation analysis (CCA) 753
capillary electrophoresis (CE) 543, 950
capture–mark–recapture (CMR) approach 464, 470
genetic CMR 471
capture-sequence data 939–940. See also DNA methylation
CardioClassifier 776
cardiocvascular disease (CVD) 817
case–control studies 558
   IBD-based 589–590
   casual inference test (CIT) 710
Catalog of Somatic Mutations in Cancer (COSMIC) 770
CAT approach 354
categorical phenotypes 599
CCMpred method 327
CD-hit 333
CellRanger 741
central limit theorem 426
CFTR2 781
Challis–Schmidler OU model of structural evolution 360
Chapman–Kolmogorov equations 23
Chapman–Kolmogorov theorem 375, 376
chemical master equation (CME) 889
China Kadoorie Biobank 268
chip-based microarray technology 600
ChIP-seq 702
ChromHMM 684, 685
CHROMOPAINTER 73, 256, 266
chromosome painting 41–42, 43, 46
   methods 73
chromosomes 574, 998
Subject Index

clade models 389–390
ClinGen 779, 784
clinical trait QTL (cQTL) 708
ClinVar 781, 782
ClonalFrame 1005
CLUMPAK 999
CLUSTAL-Omega program 333
ClusterFinder 990
clustering algorithms 251–252
admixture LD model 254–255
admixture model 253–254
on allele frequency patterns 252–253
fineSTRUCTURE model 255–258
interpreting genetic clusters 258–259
STRUCTURE model 252–253
Clustering through Imputation and Dimensionality Reduction (CIDR) 746
cluster of orthologous gene sequences (COG) 1010
cluster score 602
coalescence trees 150, 156
coalescent 145–146, 171
approximation 148–151
topology and branch lengths 148–149
and ‘classical’ population genetics 169
data analysis 145–146
diploidy and segregation 157–159
hermaphrodites 157–159
males and females 159
fundamental insights 146–148
generalizing 151–155
population structure on different time-scales 153–155
robustness and scaling 151–152
variable population size 152–153
geographic structure and strong-migration limit 155–157
structured coalescent 155–156
Kingman 117–118
multispecies (see multispecies coalescent)
neutral mutations 168–169
and phylogenetics 169–171
recombination 159–164
ancestral recombination graph 160–163
properties and effects of 163–164
selection 164–168
background selection 168
balancing selection 165–166
selective sweeps 166–167
sequentially Markov 139
structured 137
coalescent-based likelihood methods 469
coalescent history 221–224
coevolutionary independent sites (CIS) 236
coevolutionary modelling 62–67
genealogical history in 63
LD in populations with geographical subdivision/admixture 67
LD in recombining regions 65–66
LD patterns in absence of recombination 63–65
sequence of correlated trees in 63
coalescent theory 17, 1003
coalescent unit 220, 239
Cochran–Armitage trend test 607
codon-based models 355–356
codon model 203–204
coefficient of coancestry 423
coefficient of gene differentiation 476
coherence 718
collider bias 670
COLOC 713
combined probability of exclusion (CPE) 542
combined probability of inclusion (CPI) 542
common disease common variant (CDCV) hypothesis 598
CommonMind 680
community consultation 567
complete-data likelihood 11
complete-data log-likelihood 12, 15, 88
complex diseases 598
composite likelihood methods 18–19, 70–71, 412–413
computational analysis 739
computational annotations 685
conditional probability 179
conditional random fields (CRFs) 287–289
conditional structured coalescent 164
confidence intervals, construction of 192
confidentiality, in research 564–565
confounders 404
confounding factors 606–607
conjugate priors 21, 183
consent, in research 559–563
conservation genetics 457, 491–492
aim of 457
census size, estimation of 470–472
conservation genetics (Continued)
capture–mark–recapture (CMR) approach 470–471
methods based on pedigree reconstruction 474–475
multilocus genotype mismatch method 472
pairwise relatedness analysis 472–473
pairwise relationships approach 473–474
deintrogression strategies 487–489
effective population size, estimation of 458–470
from heterozygosity excess approach 459–460
from linkage disequilibrium-based methods 460–463
present-time 460–462
in recent past 462–463
from methods based on relatedness 463–466
temporal approach 466–470
genetic species delimitation 489–491
genetic structure, inferring 475
genetic differentiation, measurement of 475–477
inferring levels of recent gene flow 483–486
landscape genetics 486–487
population assignment 477–479
population clustering and ancestry proportions 479–483
overview 457–458
conservative migration 157
ConTest 333
contingency-table test 57
continuous model 546
continuous-time Markov models 199. See also substitution models
covarion model 205–206
Cox–Ingersoll–Ross process 208
Cox proportional hazards model 820
CpG dinucleotide 933
CpG islands (CGI) 933
Cramér–von Mises test 751
credible intervals 1003
CRISPR 777–778
CRISPR-Cas9 551
Critical Assessment of techniques for protein Structure Prediction (CASP) 330–332
crossovers 575
cross-sectional study 433
curse of dimensionality 23
dDarwin’s theory of natural selection 115
data compression 74–75
data mining approaches 618
Dayhoff and Eck model 348–349
Decipher 781
Deciphering Developmental Disorders Study 776
deep coalescence 170
deep learning method 336–338
convolutional neural networks 337–338
DeepSea 684
deintrogression strategies 487–489
delimitation 489
DeltaSVM 684
demographic histories 259, 262
demographic inference 17–18, 19
de novo mutations 766
DESeq 747, 751
detailed balance equations 25, 372
differential expression (DE) 742
differentially methylated cytosines (DMC) 936
differentially methylated regions (DMR) 936
direct detection of 938–939
diffusion approximation 61–62
diffusion coefficient 120
diffusion process 120
direct coupling analysis (DCA) 326–327, 1012
directed acyclic graph (DAG) 3
directional selection 398–399, 441, 442
directional negative selection 398–399
directional positive selection 398
directional selection differential 424–425, 434, 435, 439
directional selection gradient 439
Dirichlet distribution 39
Dirichlet–Laplace prior 862
Dirichlet process mixtures (DPM) 853
DiscovEHR 771
discrete-time Markov chains (DTMCs) 23
discriminant analysis of principal components (DAPC) 1001
disease risk models
  absolute risk model 818–821
  background 815–818
  breast cancer 829–832
  challenges 832–833
  clinical utility 828–829
  future directions 832
  model validation 827–828
  polygenic risk score 821–826
  PRSs and epidemiologic factors 826–827
disomic inheritance 506
dispersion tests 619
disruptive selection 441, 442
distance-based method 178, 479
DIVAN 685
divergence time estimation 206–211
DNA 573–574
  ancient DNA 275 (see also ancient DNA (aDNA))
  descent of 575–576 (see also genetic mapping)
  inheritance of 575
  library 296
  markers 531–532
  polymorphisms 801
  sequencing 547
DNA conservation metrics 681
DNA methylation 701, 933–934
differential
  bisulfite-sequencing data 936–939
  capture-sequence data 939–940
  HumanMethylation array data 940–941
  measuring 934–936
  double reduction 506–507
doublet model 202–203
downstream analysis 301–302
drift coefficient 120
  drift load 126
dropClust 749
dropEst 741
Drop-seq 737
Drosophila 433, 617
D-statistic 279–282, 284
duoHMM approach 93
dynamical Bayesian networks (DBN) 900
dysbiosis 978
e
  EAGLE2 method 96
  EAGLE v1 method 94
  East London Genes and Health 776
eCaviar 713
ecovalence 512
edgeR 747
EEMS software 487
effective migration rates 487
effective population size 152, 458
  defined 458
  estimation of 458–470
efficiency, defined 718
EIGEN 684, 685
EIGENSTRAT (PCA software) 250
EigenTHREADER 334
elastic net penalty 961
EM algorithm. See expectation-maximization (EM) algorithm
emission probabilities 90
EMMAX 616
empirical Bayes analysis 194
  for latent variable problems 30–31
  empirical research 552
Encyclopedia of DNA Elements (ENCODE) 680
entropy measures, for Y-STR markers 540
epigentic annotation data 681. See also genome-wide association studies (GWAS)
epigeneitcs 933
epigenome-wide association studies (EWAS) 941
epistasis 617
equilibrium additive genetic variance 429
An Essay toward Solving a Problem in the Doctrine of Chances (Thomas Bayes) 20
ethics 552
  bioethics 552
  defined 552
Subject Index

ethics (Continued)
and governance, case studies on 554–555
confidentiality and security 564–565
consent 559–563
100,000 Genomes Project 556
incentives to participate 562–563
leaving the study 562
recruitment of participants 558–559
returning individual genetic research results 563–564
scientific and clinical value of research 556–558
UK Biobank 555–556
voluntariness 560–561
of population genetic research 553
benefit maximisation models 554
risk control models 553–554
stewardship and social issues 565–566
benefit sharing 566–567
community involvement and public engagement 567
race, ethnicity and genetics 567–568
EVCouplings 326
EvolBoosting 414
evolutionary quantitative genetics 421
additive genetic variances and covariances 423
Fisher’s genetic decomposition 422
fitness 432
episodes of selection 433–434
individual 432–433
fitness functions and characterization of selection 441
gradients and local geometry of fitness surfaces 442–443
individual and mean fitness functions 441–442
inference of parameters 430–432
inference of selection gradients 447
flexible inference of fitness functions 449–450
normality and selection gradients 450–451
ordinary least squares analysis 448–449
infinitesimal model 426
allele frequency changes under 426–427
changes in variances 427–429
equilibrium additive genetic variance 429
linearity of parent–offspring regressions under 426
multivariate selection 444
effects of genetic correlations 444–446
multivariate breeder’s equation 444
selection gradients and traits affecting fitness 446–447
opportunity for selection 437–438
parent–offspring regressions and response to selection 423–424
genetic and phenotypic covariance matrices 425
multiple-trait parent–offspring regressions 425
multivariate breeder’s equation 425
selection differentials and breeder’s equation 424–425
single-trait parent–offspring regressions 424
Robertson–Price identity 435
selection coefficients 438–439
measures of selection on mean 439
measures of selection on variance 439–441
theorems of selection 434
description of 435–436
empirical operationalization 436–437
Ewens sampling formula 64–65
Ewens–Watterson homozygosity test 65
ExAC 771
Exomiser 781
expectation-maximization (EM) algorithm 11–12, 262
application of 12–14
with numerical optimization 14–16
expected Fisher information 21
expected lod score 587
explicit assumptions, on probability distributions 3–4
explicit incorporation of prior information, in Bayesian analysis 181
expression matrix, analysis of 747–753
clustering 748–749
combining data sets 752–753
differential expression and marker genes 750–751
dimensionality reduction and visualization 747
Subject Index

- feature selection 747–748
- network inference 751–752
- pseudotime 749–750
- expression quantitative trait loci (eQTL) 681, 698, 844
- for gene expression traits 701
- modeling for 705
- clinical trait linkage mapping 708–714
- heritability of expression traits 705–706
- joint 706–708
- single-trait 706
- extended haplotype homozygosity (EHH) 73, 410
- extended selection gradient 447
- External RNA Control Consortium 739

f
- factor loadings 962
- false alleles 471
- false discovery rate (FDR) 601, 706, 773, 855, 956
- family-based association studies 611–612
- fastPHASE 91
- fastsimcoal2 17, 19
- FDR. See false discovery rate (FDR)
- Felsenstein equation 227
- Felsenstein’s pruning algorithm 178, 348, 352–353, 357
- Felsenstein’s substitution model 15
- fineSTRUCTURE 73, 257, 998, 1000
- Firth’s correction 1011
- Fisher information matrix 8–9
- Fisher’s genetic decomposition 422
- fitness components 433
- frequency-dependent 441
- function 441
- individual 432–433
- fitness landscapes 441
- fixed-effects likelihood (FEL) model 386
- fixed-sites models 384
- Fluidigm C1 chip 736
- fluorescent activated cell sorting (FACS) 737
- fMRI data 715
- Fokker–Planck equation 123
- forensic genetics 531–532
- behavior of likelihood ratio 546–547
- mixtures 542
- combined probabilities of inclusion and exclusion 542
- likelihood ratios 542–546
- principles of interpretation 532–534
- profile probabilities 534–535
- genetic models for allele frequencies 535–539
- multi-locus dependencies 538–539
- population structure 535–537
- relatedness 537–538
- Y-STR profiles 539–542
- single nucleotide polymorphism, sequence and omic data 547–548
- forward-backward algorithm 42–43
- fossilized birth–death process 211
- four-gamete test 164
- four-population test. See D-statistic
- four-state general time-reversible (GTR) models 349
- fragmentation of DNA strands 296–297
- Framingham Risk Score 817
- FRAPPE 254
- free induction decay (FID) 951
- frequentist statistics 1. See also statistical inference, model-based
- F-statistics 856, 998
- F$_4$-statistics 282–283
- functional gene 574
- fundamental theorem of selection 435
- FUN-LDA 685
- funnel plots 644

g
- Gail model 817
- gamma rate variation model 205
- Gaussian graphical model (GGM) 967
- Gaussian kernel 745
- Gaussian latent variable model 744
- Gaussian process (GP) 879
- bulk time series expression data 883–884
- differential equation models 888–889
- identifying differential expression 884–885
- replicates and clusters 887–888
- two time course experiments 885–887
- covariance function 880–882
- inference 882–883
Subject Index

Gaussian process (GP) (Continued)
modelling single-cell data
  dimensionality reduction and pseudotime inference 891–892
modelling branching dynamics 892–893
modelling single-cell trajectory data 889–891
Gaussian process latent variable model (GPLVM) 891
Gaussian random field 487
Gauss–Markov process 890
GEMMA 616
GenABEL 606
genealogical distortion 167
gene-based analyses 619
GENEBPM 619, 620
GENECAP 472
GeneCards 782
GENECLUSTER 619–620
gene content analysis 1013–1014
gene editing 509
gene expression 850
gene-level association analysis,
  imputation-based 688–690
gene-level read counts 850–851
GeneMatcher 784
Gene Ontology (GO) 716
general genotype model 608
general linear model (GLM) 743
general time-reversible (GTR) model 200–201, 1002
gene regulatory networks
  Bayesian network reconstruction process 721–722
  building from the bottom up or top down 714–715
  integrating genetic data 720–721
  predictive Bayesian networks 722
  predictive gene networks 714
  predictive network models 718–720
  reconstruct coexpression networks 715–718
genetics 551, 552. See also ethics
genetic analysis 573
 genetic architecture of traits, in plants 510–511
 genetic covariance and partitioning heritability 685–688
gene differentiation among populations
  401. See also natural selection
gene diversity 51
gene drift 121, 148, 277, 398
  strength of 458
gene evidence, interpretation of 532–533
genetic interference 575
genetic linkage 576
genetic map distance 575
genetic mapping 510, 573–574
  association mapping 576
  associations between markers and traits,
    IBD-based detection of 583–589
    goal of 574
    IBD-based case–control studies 589–590
    IBD-based linkage likelihoods
      for major gene models 585–587
      for random-effects models 587–589
    IBD-based mapping 576–577
  inference of IBD from genetic marker data 577
  IBD at a locus 577–579
  inferring local IBD from marker data 581–583
  modeling probabilities of patterns of IBD 580–581
  probabilities of patterns of IBD 579–580
  meiosis and descent of DNA 574–576
  patterns of IBD in affected relatives 590–592
  SNP marker data for 574
  trait data probabilities
    for major gene models 583–584
    for random-effects models 584–585
genetic recombination 157
genetic research 553
genetics 573
Genetics Home Reference 779
genetic species delimitation 489–491
genetic variation 458
  background 651–652
  causal inference with genetic data 665–667
  Mendelian randomization 652–655, 667–670
  monogenic Mendelian randomization analyses 655–656
  polygenic Mendelian randomization analyses 656–664
<table>
<thead>
<tr>
<th>Subject Index</th>
<th>1119</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene tree</td>
<td>219</td>
</tr>
<tr>
<td>likelihood based on</td>
<td>229–230</td>
</tr>
<tr>
<td>probabilities</td>
<td>221–227</td>
</tr>
<tr>
<td>and species tree</td>
<td>170, 219–221</td>
</tr>
<tr>
<td>topology probabilities</td>
<td>221–225</td>
</tr>
<tr>
<td>gene tree topologies, likelihood based on</td>
<td>229</td>
</tr>
<tr>
<td>genic variance</td>
<td>426, 427</td>
</tr>
<tr>
<td>GenoCanyon</td>
<td>684, 685</td>
</tr>
<tr>
<td>genome</td>
<td>51</td>
</tr>
<tr>
<td>blocks of</td>
<td>138–140</td>
</tr>
<tr>
<td>transmission of</td>
<td>138</td>
</tr>
<tr>
<td>genome browsers and annotator software</td>
<td>682</td>
</tr>
<tr>
<td>genome-scale metabolic models (GEM)</td>
<td>990</td>
</tr>
<tr>
<td>1000 Genomes Project</td>
<td>74, 87, 95, 103, 104, 609, 613, 771</td>
</tr>
<tr>
<td>100,000 Genomes Project</td>
<td>556</td>
</tr>
<tr>
<td>in bacteria</td>
<td></td>
</tr>
<tr>
<td>background</td>
<td>1008–1009</td>
</tr>
<tr>
<td>phylogenetic methods</td>
<td>1009–1011</td>
</tr>
<tr>
<td>regression-based methods</td>
<td>1011–1012</td>
</tr>
<tr>
<td>of binary disease outcomes</td>
<td>604, 606, 615</td>
</tr>
<tr>
<td>design concepts</td>
<td>599</td>
</tr>
<tr>
<td>genome-wide significance and correction for multiple testing</td>
<td>601–602</td>
</tr>
<tr>
<td>GWAS genotyping technology design</td>
<td>600–601</td>
</tr>
<tr>
<td>phenotype definition</td>
<td>599</td>
</tr>
<tr>
<td>replication</td>
<td>602</td>
</tr>
<tr>
<td>sample size considerations</td>
<td>601</td>
</tr>
<tr>
<td>structure of common genetic variation</td>
<td>599–600</td>
</tr>
<tr>
<td>functional annotation data in DNA conservation</td>
<td>681</td>
</tr>
<tr>
<td>epigenetic annotation data</td>
<td>681</td>
</tr>
<tr>
<td>transcriptomic annotation data</td>
<td>680–681</td>
</tr>
<tr>
<td>genetic structure in</td>
<td>611–612</td>
</tr>
<tr>
<td>identification of related individuals</td>
<td>612–613</td>
</tr>
<tr>
<td>identify ethnic outliers and account for population stratification</td>
<td>613–614</td>
</tr>
<tr>
<td>mixed modelling approaches</td>
<td>614–615</td>
</tr>
<tr>
<td>software</td>
<td>615–616</td>
</tr>
<tr>
<td>meta-analysis</td>
<td>640–647</td>
</tr>
<tr>
<td>methods to integrate functional annotations in future directions</td>
<td>690</td>
</tr>
<tr>
<td>gene-level association analysis</td>
<td>688–690</td>
</tr>
<tr>
<td>partitioning heritability and genetic covariance</td>
<td>685–688</td>
</tr>
<tr>
<td>multiple SNP association analysis</td>
<td>616</td>
</tr>
<tr>
<td>gene-based analyses</td>
<td>619</td>
</tr>
<tr>
<td>haplotype-based analyses</td>
<td>616–617</td>
</tr>
<tr>
<td>SNP–SNP interaction analyses</td>
<td>617–618</td>
</tr>
<tr>
<td>software</td>
<td>619–620</td>
</tr>
<tr>
<td>personalised medicine, delivery of</td>
<td>620</td>
</tr>
<tr>
<td>quality control</td>
<td>602–603</td>
</tr>
<tr>
<td>sample quality control procedures</td>
<td>604–605</td>
</tr>
<tr>
<td>SNP quality control procedures</td>
<td>603–604</td>
</tr>
<tr>
<td>software for</td>
<td>606</td>
</tr>
<tr>
<td>replication</td>
<td>632–635</td>
</tr>
<tr>
<td>single SNP association analysis</td>
<td>606</td>
</tr>
<tr>
<td>accounting for confounding factors</td>
<td>606–607</td>
</tr>
<tr>
<td>Bayesian methods</td>
<td>611</td>
</tr>
<tr>
<td>coding of SNP genotypes</td>
<td>607–609</td>
</tr>
<tr>
<td>generalised linear modelling framework</td>
<td>606</td>
</tr>
<tr>
<td>imputed genotypes</td>
<td>609</td>
</tr>
<tr>
<td>interactions with non-genetic risk factors</td>
<td>609–610</td>
</tr>
<tr>
<td>software for</td>
<td>611</td>
</tr>
<tr>
<td>visualisation of results of</td>
<td>609</td>
</tr>
<tr>
<td>synthesise annotation data</td>
<td></td>
</tr>
<tr>
<td>computational annotations</td>
<td>685</td>
</tr>
<tr>
<td>genome browsers and annotator software</td>
<td>682</td>
</tr>
<tr>
<td>supervised learning methods</td>
<td>682–684</td>
</tr>
<tr>
<td>unsupervised learning methods</td>
<td>684</td>
</tr>
<tr>
<td>use of genotype imputation in</td>
<td>98–99</td>
</tr>
<tr>
<td>winner's curse</td>
<td>635–640</td>
</tr>
<tr>
<td>genome-wide epistasis analysis</td>
<td>997, 1012–1013</td>
</tr>
<tr>
<td>genomic best linear unbiased prediction (GBLUP) method</td>
<td>489, 516</td>
</tr>
<tr>
<td>genomic estimates of breeding values (GEBVs)</td>
<td>515, 517, 518</td>
</tr>
<tr>
<td>genomic medicine and variant interpretation</td>
<td></td>
</tr>
<tr>
<td>current challenges</td>
<td>761–765</td>
</tr>
<tr>
<td>effect of variant</td>
<td>765–771</td>
</tr>
<tr>
<td>functional assays of genetic variation</td>
<td>777–779</td>
</tr>
<tr>
<td>future challenges</td>
<td>783–786</td>
</tr>
<tr>
<td>holistic variant interpretation</td>
<td>782–783</td>
</tr>
</tbody>
</table>
Subject Index

Genomic medicine and variant interpretation
(Continued)
human and model phenotype resources 779–782
large human reference cohorts 771–776
Genomic rearrangements, in plants 509–510
Genomics England 565, 566
genomic sequencing 566
GenoSkyline 685
genotype 347
genotype calling 5, 31
from microarrays 94
from sequencing 94
genotyped markers 87
Genotype–environment interaction, in crops 511–514
Genotype frequencies 5, 38, 39–40
Genotype imputation 97–98
accuracy, factors affecting 104–105
ancestry 105–106
genotyping microarray 106
imputation methods 107
reference panel size and SNP allele frequency 105
basic idea of 98
future directions 109
haplotype imputation 99–100
methods
Beagle 101–102
imputation servers 103
IMPUTE 100–101
MaCH/minimac 101
positional Burrows–Wheeler transform 102
SNP tagging approaches 102–103
quality control for imputed data 107–109
summary statistic imputation 104
and testing for association 103–104
use of, in GWASs 98–99
Genotype likelihood (GL) 94
genotype likelihood models 4
Genotype–phenotype relationship 347
Genotype-Tissue Expression (GTEx) project 680
genotyping errors 471
genotyping microarray 106
GGE analysis 513

Gibbs sampling 999
Gibbs variable selection (GVS) 858
Gillespie algorithm 889
Glimmer 990
Global Alliance for Genomics & Health (GA4GH) 784
Global Lipids Genetics Consortium 655
GLOBETROTTER 73, 263, 266–267
globular protein, structure of 328
GMMAT 616
gnomAD 771
GNOVA 688
GODAMBE information matrix 19
GPflow package 892
GQT toolkit 74
Grand Prix package 892
Granger causality 752
GREML algorithm 686
GREMLIN method 327–328, 330
GREMLIN server 334
group penalty 961
GTEx 781
Gubbins 1005
GWASs. See genome-wide association studies (GWASs)
GWAVA 682

h
Haldane genetic map 575
HAPGEN program 96
HapHedge 96
HAPI-UR (Haplotype Inference for Unrelated Samples) 93
haplotype imputation 99–100
haplotype 51, 87, 616
analyses 616–617
clustering 617, 619
defined 87
heterozygosity 54
use of 87
haplotype-based methods, to detect selection 410–412
haplotype estimation 87–88
challenges in 87–88
hidden Markov models for 89–93
Beagle 91
fastPHASE 91
HAPI-UR 93
IMPUTE and MaCH 90–91
Subject Index

PHASE and Li and Stephens model 89–90
SHAPEIT approach 91–92
measuring phasing performance 96–97
from reference panel 95–96
in related samples 93–94
simple haplotype frequency model 88–89
using sequencing data 94–95
Haplotype Reference Consortium (HRC) project 94
haplotype reference panels 98
haplotyping. See haplotype estimation
HapMap Project 102, 105
HapMap SNPs 687
HapMix 73
hard sweep 403
Hardy–Weinberg equilibrium (HWE) 3, 6, 459, 464, 534, 603
deviation from 604
Hardy–Weinberg genotype frequency 459
Hastings ratio 25, 26, 28
heritability 424
in plant breeding 503–504
hermaphrodites 157–159
heterochaly 354
heterotic pool 518
heterozygosity 14–16
heterozygosity excess, effective population size estimation from 459–460
heuristic algorithms 740
heuristic estimation methods 372–373
HGMD 781
Baum–Welch algorithm 44–46
Bayesian inference of hidden states 42–43
for phasing 89–93
Viterbi algorithm 43–44
hierarchical Bayesian analysis 194
hierarchical models 2
hierBAPS 998, 1000, 1001
higher-level abstraction 336
highest posterior density (HPD) 22
high-throughput sequencing (HTS) 94, 978
hitchhiking effect 401–403
HKA test 406
HKY mutation model 371
HLI 771
HMM forward backward algorithm 101
HMMs. See hidden Markov models (HMMs)
homozygosity mapping 591
Hotelling’s $T^2$ test 939
Hudson’s composite likelihood approach 71
Human Gene Mutation Database (HGMD) 682
human genome 51
Human Genome Epidemiology Network (HuGENet) 640
Human Metabolome Database 952
Human Methylation array data 940–941.
See also DNA methylation
Human MethylationEPIC array 935
Human Phenotype Ontology (HPO) 782
human versus experimental models 704–705
hybrid breeding 518
Hybrid-Coal (software) 225
hybridization and gene flow 240–241
hybridization parameter 240
hypergeometric model 137
hypermutator strains 1006
hypothesis-free approach 598
identical by descent (IBD) 94, 574–575
detection of associations between markers and traits 583–589
mapping based on 576–577 (see also genetic mapping)
segments 466
single-locus 61
two-locus 61
identity by state (IBS) metric 612–613
identity matrix 707
iHS test 73
Illumina HiSeq 935
improper priors 21
imputation servers 103
IMPUTE 100–101
IMPUTE v1 100
IMPUTE v2 90–91, 100–101
IMPUTE v4 101
inbreeding 28–29
inbreeding effective size 459
incomplete lineage sorting (ILS) 279
incomplete selective sweeps, identification of 73–74
incomplete sweep 408
Subject Index

independence assumptions 2–3
independent-rate models, for rate variation 209
index of total selection 437
individual fitness 432–433
individual fitness function 441–442
InDrop 737
inference methods and algorithms 4. See also statistical inference, model-based
inference of genotype frequencies 5
infinite-alleles model 169
infinitely many alleles model 118
infinitely many sites model 118
infinitesimal model 137–138, 426
allele frequency changes under 426–427
changes in variances 427–429
equilibrium additive genetic variance 429
linearity of parent–offspring regressions under 426
infinite-sites model 169
Infinium Human-Methylation450K 935
inflammatory bowel disease (IBD) 722
information measures 108
inheritance probability 240
instrument strength independent of direct effect (InSIDE) 662
integrated haplotype score (iHS) 410–411
intermediate quantitative traits 599
International Breast Cancer Intervention Study (IBIS) 830
International HapMap Consortium 609, 613
International HapMap Project 599–600
International Maize and Wheat Improvement Center (CIMMYT) 503
introgressed likelihoods 16–18
introgressed sequences identification 283–284
advantages and disadvantages of 289
hidden Markov and conditional random field models 287–289
\(S^*\)-statistic 284–287
invariance property, ML estimator 6
IQ-TREE 1002

\(j\)
jackHMMER 335
Jacobian matrix 10
Jeffreys’ prior 21–22
Jensen’s inequality 8

JLIM 713
jModelTest 1002
Jones group 327
Jones–Taylor–Thornton (JTT) model, of amino acid replacement 350

\(k\)
Kalman filter 890, 953
KEGG 715
kernel-based regression, in microbiome studies 986–987
kernel function 880
Kimura’s stepping stone model 127–130
kinetic models 703
Kingman coalescent 117–118
kinship matrix 707, 1011
\(k\)-mers 1009
\(k\)-nearest-neighbours batch effect test (\(k\)BET) 744
Kolmogorov equations 123–124
backward equation 123
forward equation 123–124
Kolmogorov–Smirnov test 751
Kruskal–Wallis test 751
Kullback–Leibler information 589

\(l\)
labeled histories 195
Lactase gene 54
Lagrange multiplier 5, 6
Lagrangian function 6
Lander–Green algorithm 93
landscape genetics 486–487
Langevin equation 890
Lassosum 825
latent variable models 16
LC-MS. See liquid chromatography – mass spectrometry (LC-MS)
LD. See linkage disequilibrium (LD)
LD score regression 686–687
Leiden Open Variation Database 3.0 781
Lewontin’s paradox 136
lifetime fitness 432–433
lifetime reproductive success 432
likelihood 374
calculation of 374–375
likelihood-based partition method 475
likelihood computation, for multi-allelic markers 468
likelihood function 4, 181
likelihood methods, for estimating population recombination rate 70
likelihood of hypothesis 179
likelihood principle 4
likelihood ratios 533–534, 542–543
behavior of 546–547
binary model 543–545
continuous model 546
semi-continuous model 545–546
likelihood ratio statistic 37–38
likelihood ratio test (LRT) 37, 58, 184, 374
likelihood tuning parameter 257
limma-voom 751
lineage association 1009
lineage sorting 170
linear mixed model (LMM) 1011
linear noise approximation (LNA) 890
linear regression methods 959–960. See also metabolomics, statistical methods in
linkage analysis 93
linkage disequilibrium (LD) 51–53, 131, 163, 248, 423, 535, 576, 600, 681, 800, 824, 1000
and additive genetic variance 427
admixture LD 248, 254–255
background LD 248–249, 255
coalescent modelling 62–67
complete LD 600
data analysis 69–75
decay of 265–267
effective population size estimation from 460–463
extensions of two-locus LD measures 60
haplotype patterns with different levels of 52
historical sketch of mathematical treatments of 60–62
hitchhiking and 401
LD coefficients 52
matrix methods and diffusion approximations 61–62
measuring 53–60
relating genealogical history to 67–69
score regression 666–667
single-number summaries of 54–56
spatial distribution of 56–60
spatial structure of 59
and two-locus identity by descent (IBD) 61
linkage equilibrium 52, 130
liquid chromatography – mass spectrometry (LC-MS) 952–954
L₁ + L₂ penalty 961
local clocks 209
local geometry of fitness surfaces 442–443
locations 574
LocusZoo 609
lod score 586
LOFTEE 770
logic models 703
log-likelihood function 5, 6, 375
longitudinal study 433
long-range phasing (LRP) 94
loss function 22
Louvain algorithm 749
lowest common ancestor (LCA) 981
lymphoblastoid cell lines 704
m
MACAU 937
machine learning 335, 413–414, 804
deep learning method 336–338
MetaPSICOV method 336
PconsC 335–336
phylogeny constraints 339–340
sequence pairing 338–339
MaCH (Markov Chain Haplotyping) method 90–91, 101
Madsen–Browning scheme 619
MAF (minor allele count) 604
major histocompatibility complex (MHC) 370
Manhattan plot 609, 610
map_align algorithm 334
marginal likelihood 20, 181
computing 185
marker genes 751
marker imputation, in plant breeding 519
Markov Affinity-based Graph Imputation of Cells (MAGIC) 745
Markov chain law of large numbers 197
Markov chain Monte Carlo (MCMC) 23–29, 185, 844, 903, 999
algorithm 89
for approximating posterior probability of phylogenies 195–198
convergence and mixing 26–28
<table>
<thead>
<tr>
<th>Subject Index</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Markov chain Monte Carlo (MCMC)</strong></td>
</tr>
<tr>
<td>(Continued)</td>
</tr>
<tr>
<td>Markov chains 23–25</td>
</tr>
<tr>
<td>Metropolis–Hastings algorithm 25–26, 28–29</td>
</tr>
<tr>
<td>Markov chains 23–25, 1000</td>
</tr>
<tr>
<td>aperiodic 24</td>
</tr>
<tr>
<td>discrete-time 23</td>
</tr>
<tr>
<td>irreducible 24</td>
</tr>
<tr>
<td>reducible 24</td>
</tr>
<tr>
<td>reversibility 24</td>
</tr>
<tr>
<td>Markov clustering algorithm 999</td>
</tr>
<tr>
<td>Markov equivalent 719</td>
</tr>
<tr>
<td>Markovian dependence 928</td>
</tr>
<tr>
<td>Markov model of codon substitution 371–372</td>
</tr>
<tr>
<td>Markov process 371</td>
</tr>
<tr>
<td>Markov property 23</td>
</tr>
<tr>
<td>Mash algorithm 1001</td>
</tr>
<tr>
<td>mass spectrometry (MS) 950</td>
</tr>
<tr>
<td>match probability 533, 538</td>
</tr>
<tr>
<td>mate selection 517–518</td>
</tr>
<tr>
<td>mathematical models in population genetics 115–116</td>
</tr>
<tr>
<td>multi-locus models 130–131</td>
</tr>
<tr>
<td>linkage disequilibrium 134–140</td>
</tr>
<tr>
<td>linkage equilibrium 131–134</td>
</tr>
<tr>
<td>single-locus models 116–117</td>
</tr>
<tr>
<td>diffusion approximations 120–126</td>
</tr>
<tr>
<td>of panmictic populations 116</td>
</tr>
<tr>
<td>random drift and Kingman coalescent 117–120</td>
</tr>
<tr>
<td>spatially structured populations 126–130</td>
</tr>
<tr>
<td>mating systems 459</td>
</tr>
<tr>
<td>of flowering plants 504</td>
</tr>
<tr>
<td>of hermaphroditic populations 504</td>
</tr>
<tr>
<td>MaxEntScan 767</td>
</tr>
<tr>
<td>maximum a posteriori (MAP) estimate 22</td>
</tr>
<tr>
<td>maximum a posteriori (MAP) tree 234</td>
</tr>
<tr>
<td>maximum clade credibility (MCC) tree 234</td>
</tr>
<tr>
<td>maximum composite likelihood estimator (MCLE) 19</td>
</tr>
<tr>
<td>maximum likelihood (ML) 4, 179–180 estimation 374–377</td>
</tr>
<tr>
<td>mechanics of 192–193</td>
</tr>
<tr>
<td>maximum likelihood estimator (ML estimator) 4–6, 468</td>
</tr>
<tr>
<td>asymptotically unbiased 7–8</td>
</tr>
<tr>
<td>properties of 6–8</td>
</tr>
<tr>
<td>quantifying confidence with 8–9</td>
</tr>
<tr>
<td>maximum likelihood inference 4–19, 60</td>
</tr>
<tr>
<td>maximum posterior probability (MAP) estimate 181</td>
</tr>
<tr>
<td>McDonald–Kreitman (MK) test 406</td>
</tr>
<tr>
<td>MCMC. See Markov chain Monte Carlo (MCMC)</td>
</tr>
<tr>
<td>mean-field substitution model 359</td>
</tr>
<tr>
<td>mean fitness 132</td>
</tr>
<tr>
<td>mean fitness landscape 441</td>
</tr>
<tr>
<td>medical ethics 552</td>
</tr>
<tr>
<td>meiosis 573, 574–575</td>
</tr>
<tr>
<td>Mendelian randomization (MR) 621, 652–655, 712</td>
</tr>
<tr>
<td>monogenic 655–656</td>
</tr>
<tr>
<td>polygenic 656–664</td>
</tr>
<tr>
<td>interactions and subsetting 663–664</td>
</tr>
<tr>
<td>median estimation methods 660</td>
</tr>
<tr>
<td>modal estimation methods 660</td>
</tr>
<tr>
<td>MR-Egger method 661–663</td>
</tr>
<tr>
<td>multivariable methods 663</td>
</tr>
<tr>
<td>practical advice 664</td>
</tr>
<tr>
<td>regularization methods 660–661</td>
</tr>
<tr>
<td>robust methods 661</td>
</tr>
<tr>
<td>Mendel’s first law 574</td>
</tr>
<tr>
<td>mental capacity 559</td>
</tr>
<tr>
<td>31-mers 1009</td>
</tr>
<tr>
<td>messenger RNA (mRNA) 700, 843</td>
</tr>
<tr>
<td>MetaboAnalyst 3.0 957</td>
</tr>
<tr>
<td>metabolome-wide significance level (MWSL) 956</td>
</tr>
<tr>
<td>metabolomics, statistical methods in 949–950</td>
</tr>
<tr>
<td>metabolite identification and pathway analysis</td>
</tr>
<tr>
<td>pathway and metabolite set analysis 971–972</td>
</tr>
<tr>
<td>statistical correlation spectroscopy 969–970</td>
</tr>
<tr>
<td>multivariate methods and chemometrics techniques 958–959</td>
</tr>
<tr>
<td>linear regression methods 959–960</td>
</tr>
<tr>
<td>shrinkage methods 960–961</td>
</tr>
<tr>
<td>network analysis 966–969</td>
</tr>
<tr>
<td>orthogonal projection methods 961</td>
</tr>
<tr>
<td>onto latent structures 965–966</td>
</tr>
<tr>
<td>partial least squares 964–965</td>
</tr>
<tr>
<td>principal components analysis 962–964</td>
</tr>
</tbody>
</table>
### Subject Index

<table>
<thead>
<tr>
<th>Term</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>preprocessing and deconvolution</td>
<td>952–954</td>
</tr>
<tr>
<td>liquid chromatography – mass spectrometry</td>
<td>952–954</td>
</tr>
<tr>
<td>nuclear magnetic resonance spectroscopy</td>
<td>950–952</td>
</tr>
<tr>
<td>univariate methods</td>
<td>954–956</td>
</tr>
<tr>
<td>metabolome-wide significance levels</td>
<td>956–957</td>
</tr>
<tr>
<td>sample size and power</td>
<td>957–958</td>
</tr>
<tr>
<td>MetaNeighbour</td>
<td>752</td>
</tr>
<tr>
<td>MetaPSICOV method</td>
<td>327</td>
</tr>
<tr>
<td>methicillin-resistant <em>Staphylococcus aureus</em> (MRSA)</td>
<td>1004</td>
</tr>
<tr>
<td>method of Lagrange multipliers</td>
<td>5</td>
</tr>
<tr>
<td>MethylKit</td>
<td>937</td>
</tr>
<tr>
<td>Metropolis–Hastings algorithm</td>
<td>25–26, 195–196</td>
</tr>
<tr>
<td>in Bayesian inference</td>
<td>28–29</td>
</tr>
<tr>
<td>Metropolis–Hastings step</td>
<td>94, 859, 904</td>
</tr>
<tr>
<td>microarray data</td>
<td>851–856</td>
</tr>
<tr>
<td>multi-class data</td>
<td>855–856</td>
</tr>
<tr>
<td>microbial diversity index</td>
<td>983</td>
</tr>
<tr>
<td>microbiome</td>
<td></td>
</tr>
<tr>
<td>covariate</td>
<td>985–987</td>
</tr>
<tr>
<td>future prospects of</td>
<td>991</td>
</tr>
<tr>
<td>in human health and disease</td>
<td>977–978</td>
</tr>
<tr>
<td>integrative analysis</td>
<td>989–991</td>
</tr>
<tr>
<td>mediator</td>
<td>987–989</td>
</tr>
<tr>
<td>methods for analysis</td>
<td>983–985</td>
</tr>
<tr>
<td>16S rRNA and shotgun metagenomic sequencing data</td>
<td>980–983</td>
</tr>
<tr>
<td>microsatellites</td>
<td>477, 597</td>
</tr>
<tr>
<td>midparent–offspring regression</td>
<td>424</td>
</tr>
<tr>
<td>migration</td>
<td></td>
</tr>
<tr>
<td>conservative</td>
<td>157</td>
</tr>
<tr>
<td>rate estimation</td>
<td>485</td>
</tr>
<tr>
<td>strong</td>
<td>156–157</td>
</tr>
<tr>
<td>minimum mean squared error (MMSE) estimate</td>
<td>22</td>
</tr>
<tr>
<td>minimum spanning tree (MST)</td>
<td>750</td>
</tr>
<tr>
<td>minor allele frequency (MAF)</td>
<td>803</td>
</tr>
<tr>
<td>missense tolerance ratio (MTR)</td>
<td>771</td>
</tr>
<tr>
<td>mitochondrial contamination estimation approach</td>
<td>298</td>
</tr>
<tr>
<td>mixed stock analysis</td>
<td>480</td>
</tr>
<tr>
<td>mixture models (MIX)</td>
<td>900</td>
</tr>
<tr>
<td>mixture over marker (MOM)</td>
<td>866</td>
</tr>
<tr>
<td>mnnCorrect</td>
<td>752</td>
</tr>
<tr>
<td>mode of selection on trait</td>
<td>441</td>
</tr>
<tr>
<td>molecular clock hypothesis</td>
<td>206</td>
</tr>
<tr>
<td>molecular phylogenetics</td>
<td>170</td>
</tr>
<tr>
<td>moment estimators</td>
<td>466–467</td>
</tr>
<tr>
<td>moment methods, for estimating population recombination rate</td>
<td>69–70</td>
</tr>
<tr>
<td>Monocle</td>
<td>750, 751</td>
</tr>
<tr>
<td>Monte Carlo sampling, for intractable likelihoods</td>
<td>16–18</td>
</tr>
<tr>
<td>Monte Carlo simulations</td>
<td>1010</td>
</tr>
<tr>
<td>morality</td>
<td>552. See also ethics</td>
</tr>
<tr>
<td>Moran model</td>
<td>120</td>
</tr>
<tr>
<td>morgan</td>
<td>575</td>
</tr>
<tr>
<td>most recent common ancestor (MRCA)</td>
<td>146, 577</td>
</tr>
<tr>
<td>Mouse Genome Informatics (MGI)</td>
<td>781</td>
</tr>
<tr>
<td>MP-EST (Maximum Pseudo-likelihood for Estimating Species Trees)</td>
<td>232, 233</td>
</tr>
<tr>
<td>MR. See Mendelian randomization (MR)</td>
<td></td>
</tr>
<tr>
<td>MrBayes program</td>
<td>202</td>
</tr>
<tr>
<td>MULSEL method</td>
<td>333</td>
</tr>
<tr>
<td>MultiBLUP</td>
<td>802</td>
</tr>
<tr>
<td>multi-dimensional diffusion</td>
<td>124</td>
</tr>
<tr>
<td>multi-dimensional scaling (MDS)</td>
<td>251, 613</td>
</tr>
<tr>
<td>multilocus genotype mismatch method</td>
<td>472</td>
</tr>
<tr>
<td>multi-marker multi-trait analysis</td>
<td>865–868</td>
</tr>
<tr>
<td>multinomial distribution</td>
<td>6</td>
</tr>
<tr>
<td>multiple hits</td>
<td>373</td>
</tr>
<tr>
<td>multiple loci, population genetics of</td>
<td>130–131</td>
</tr>
<tr>
<td>linkage disequilibrium</td>
<td>134–140</td>
</tr>
<tr>
<td>applications</td>
<td>136–137</td>
</tr>
<tr>
<td>approximations</td>
<td>137–138</td>
</tr>
<tr>
<td>blocks of genome</td>
<td>138–140</td>
</tr>
<tr>
<td>genotype frequencies, representing</td>
<td>134–136</td>
</tr>
<tr>
<td>linkage equilibrium</td>
<td>131–134</td>
</tr>
<tr>
<td>random drift</td>
<td>133–134</td>
</tr>
<tr>
<td>selection gradients</td>
<td>131–133</td>
</tr>
<tr>
<td>multiplexed assays for variant effect (MAVE)</td>
<td>778</td>
</tr>
<tr>
<td>multispecies coalescent</td>
<td>219–221</td>
</tr>
<tr>
<td>estimation of parameters at population and species levels</td>
<td>239</td>
</tr>
<tr>
<td>hybridization and gene flow</td>
<td>240–241</td>
</tr>
<tr>
<td>speciation times and population sizes</td>
<td>239–240</td>
</tr>
<tr>
<td>species delimitation</td>
<td>241</td>
</tr>
<tr>
<td>future prospects</td>
<td>242</td>
</tr>
</tbody>
</table>
Subject Index

multispecies coalescent (Continued)
genetic trees and species tree 219–221
probability distributions under 221
genetic tree probabilities 221–227
model assumptions and violations 230
site pattern probabilities 227–229
species tree likelihoods 229–230
species tree inference under 231
Bayesian full-data methods 234–235
empirical examples 237–238
multilocus versus SNP data 236–237
site pattern-based methods 235–236
summary statistics methods 231–234
multi-tissue HESS (MT-HESS) model 866
multi-locus parent–offspring regression 425
multivariate beta function 39
directional selection 400–403
directional selection 398–399
neutral theory 369
neutral Wright–Fisher model 116, 117, 146
Newton’s algorithm 9, 10
Newton’s method 9–11
next generation sequencing (NGS) 296, 325, 997
NGS. See next generation sequencing (NGS)
non-homogeneous dynamic Bayesian networks (NH-DBN)
NHLBI Exome Sequencing Project 771
NMR Suite 951
NNSPLICE 767
non-admixture model 480
ncRNA 701
non-homogeneous dynamic Bayesian networks
(application 900–901
methodology
Bayesian linear regression 902–905
Bayesian piecewise linear regression 905–908
computational complexity 920
coupled regression coefficients 908–914
dynamic Bayesian network modelling 918–920
dynamic Bayesian networks 901–902
time-varying network structures 916–918
nonparametric differential expression for single cells (NODES) 742
nonparametric statistical approach 3
Subject Index

nonparametric techniques 55
non-random mating 460
non-ribosomal peptides (NRPS) 990
normal distribution 9, 26, 27
normality
  of phenotype 451
  of residuals 450
  and selection gradients 450–451
NormExpression 743
nuclear magnetic resonance (NMR)
spectroscopy 328, 950–952
nucleotide substitution model 227
null hypothesis 717

O
odds ratio 608
ODE. See ordinary differential equation (ODE)
OD-seq analysis tool 333
OLS. See ordinary least squares (OLS)
omic data 547
OMIM 779
omni-genic model 679
one-step transition matrix 23
Online Mendelian Inheritance in Animals (OMIA) 781
On the Origin of Species (Darwin) 177
operational taxonomic units (OTU) 980
opportunity for selection 437–438
optimal contribution theory 518
oracle penalty 961
ordinary differential equation (ODE) 703, 752, 888
ordinary least squares (OLS) 959, 960
Ornstein–Uhlenbeck (OU) process 126, 208, 360, 880
Orphanet 779
orthogonal signal correction (OSC) 965
outcrossing species 505
output feature map 337
overlapping generations, models of 120
overlapping mixture of Gaussian processes (OMGP) 892
Oxford Phasing Server 96

P
pain in the torus 127, 128
pairwise relatedness analysis 472–473
pairwise sequentially Markovian coalescent (PSMC) model 261–262
paleobiotechnology 383
pan-genome 998
PANINI 1014
parametric modeling 3
parental haplotypes 94
parent-independent mutation 133
parent–offspring phenotypic covariance 423
parent–offspring regressions 423–424
breeder’s equation 425
directional selection differential 424–425
genetic and phenotypic covariance matrices 425
multiple trait 425
response to selection 424–425
single-trait 424
Parisi–Echave technique 357
parsimony method 178
partial least squares (PLS) 804, 961, 964–965
regression 703
PCA. See principal components analysis (PCA)
peak of polymorphism 166
Pearson correlation 707
Pearson correlation coefficient 57
pedigree-based linkage analysis 585–587
pedigree-derived relationship matrix 430
peer review 557
penetrance probabilities 583
percent splicing index (PSI) 746
permutation procedures 601–602
person of interest (POI) 532, 533, 547
PHASE software 90, 619
phasing. See haplotype estimation
PhastCons 681
Phen-Gen 685
phenotype 347
definition 599
phenotype prediction and DNA variants 799–800
accuracy of prediction 806–808
additive genetic values 802–806
Bayesian genomic selection models 808–809
DNA polymorphisms 801
genetic variation affecting phenotype 800–801
genomic prediction 809–810
phenotypic covariance matrix 425
phenotypic value of offspring 422
phenotypic variance 503
Subject Index

Phevor 685
philosophical ethics 552
phylogenetic inference 170
phylogenetic networks 240
*Phylogenetic Systematics* (Hennig) 177
phylogenetic trees 1001
calibrating 209–211
phylogeny
constraints 339–340
likelihood calculation on 379–380
phylogeny estimation, by likelihood-based methods 177–179, 212
applications of likelihood-based methods 199–211
divergence time estimation 206–211
expanding the model around groups of sites 202–204
rate variation across sites 204–206
substitution models 199–202
Bayesian inference 180–184
mechanics of 192–198
calculating likelihood for phylogenetic model 186
calculating probability of character history 187–188
character matrices and alignments 186
continuous-time Markov model 188–189
marginalizing over character histories 189–191
phylogenetic model 186–187
choosing among models 184–185
cladists vs. pheneticists 177–178
maximum likelihood method 179–180
mechanics of 192–193
statistical phylogenetics 178
PhyloP 681
phyloscanner 1006
piecewise constant population sizes 261
PINES 685
plant breeding 501–502
breeding systems 504–505, 506
experimental design and analysis 519–521
genetic architecture of traits 510–511
genomic rearrangements 509–510
genomic selection
and cross prediction 517
genomic prediction of hybrid performance and heterosis 518
genotype–environment interaction 514–515
marker imputation 519
mate selection 517–518
and phenotyping cost 517
quantitative trait loci and major genes 516–517
sequential selection 518
heritability and breeder’s equation in 502–504
polyploidy in plants and genetic consequences 505–509
response to environment and plasticity 511–514
plasmids 998
pleiotropy 423
pleiotropy effects test (PET) 716
PLINK 606, 619, 824
PLINKv2 606, 611, 615
plmDCA method 326, 327
PLS. See partial least squares (PLS)
PMEN2 lineage 1014
Poisson-beta distribution 747
Poisson distribution 739
Poissonization 168
Poisson model 850
Pólya–gamma latent variables 860
polygenic effect 585
polygenic risk score (PRS) 803, 817
polygenic selection 399, 403
polyketides (PKS) 990
polymerase chain reaction (PCR) 737, 765, 934
polyploidy in plants 505–509
Pool-seq 5, 10–11
population assignment analysis 477–479, 484
population-based association studies 597–598
population branch statistic (PBS) 409
population, definition of 251–252
population genetics 53, 116
multi-locus models 130–131
linkage disequilibrium 134–140
linkage equilibrium 131–134
single-locus models 116–117
diffusion approximations 120–126
of panmictic populations 116
Subject Index

random drift and Kingman coalescent 117–120
spatially structured populations 126–130
population mean fitness 441–442
population recombination rate 55. See also recombination rate, estimation of
population-scale sequencing projects 74
population size changes and split times, inferring 259–260
allele frequency spectrum approach 260–261
whole-genome sequencing approaches 261–262
population splits 247
population-structure parameter 535
positional Burrows–Wheeler transform (pBWT) 74, 92, 102
posterior distribution 20
posterior mean estimate 22
posterior probability 184
of tree 195
post-mortem degradation (PMD) 298–299
potential scale reduction factors (PSRF) 919
Potts models 1012
Prdm9 75
precision medicine 556
PredictABEL 820, 821
PrediXcan 689, 690
PRIMUS 615
principal components analysis (PCA) 249–251, 613, 703, 747, 892, 961–964
Principles of Numerical Taxonomy (Sokal and Sneath) 177
prior distribution 4, 20
choice of 21–22
probabilistic model of protein evolution 347–348, 360–361
codon-based models 355–356
dependence among positions 357–359
heterogeneity of replacement rates among sites 351
models of amino acid replacement
Dayhoff and Eck model 348–349
descendants of Dayhoff model 350–351
models with physicochemical basis 355
protein structural environments 351–353
stochastic models of structural evolution 359–360
variation of preferred residues among sites 353–355
probabilistic quotient normalization (PQN) 954
probability distributions 3–4
emission probabilities 41–42
initial state distribution 41
under multispecies coalescent 221
gene tree probabilities 221–227
model assumptions and violations 230
site pattern probabilities 227–229
species tree likelihoods 229–230
transition probabilities 41
Procrustean design 519
product of approximate conditionals (PAC) scheme 71
professional ethics 552
profile drift 332
proportion of variance explained (PVE) 863
protein
disordered 329
globular 328
protein–protein interactions 329–330
protein evolution, probabilistic model of 347–348, 360–361
codon-based models 355–356
dependence among positions 357–359
heterogeneity of replacement rates among sites 351
models of amino acid replacement
Dayhoff and Eck model 348–349
descendants of Dayhoff model 350–351
models with physicochemical basis 355
protein structural environments 351–353
stochastic models of structural evolution 359–360
variation of preferred residues among sites 353–355
protein–protein interaction (PPI) 721
protein structural environments 351–353
protein-truncating variants (PTV) 766
pseudocounted quantiles (pQ) 742
pseudo-haploid model 95
pseudo-likelihood method (PLM) 233, 326
pseudo likelihoods 18–19
pseudotime 749–750
PSI-BLAST 335
PSICOV method 327
Subject Index

public engagement 567
p-value 547, 716

q
QCTOOL 606
Q-function 11–13
QIIME 981
quadratic convergence 10
quadratic function 22
quadratic penalty 960–961
quadratic selection differential 442
quadratic selection gradient 441
quality control (QC) 953–954
GWAS 602–603
automated procedures 603
sample quality control procedures 604–605
SNP quality control procedures 603–604
software for 606
quantile–quantile plot, GWAS 612
quantitative genetics 422. See also

evolutionary quantitative genetics
quantitative trait loci (QTLs) 502, 509, 511, 514, 516–518, 844
multiple-response models 864–868
single-response models 863–864
quantitative traits 136
quasi likelihoods 18–19
quasi-Newton methods 10

r
RADMeth 937
random drift 133–134
random regression model 804
random-sites models 384
RapidNJ 1001
RaptorX-contact 337
rare alleles 467
rates of substitution 400
rate variation over protein sites 351
RAXML 1002
real-time polymerase chain reaction (RT-PCR) 922
recent common ancestry, use of 72–73
recombination 159–160, 575
ancestral recombination graph 160–163
breakpoints 161–163
clocks 73
properties and effects of 163–164
recombination rate, estimation of 69
approximating coalescent 71–72
approximating likelihood 70–71
likelihood methods 70
moment methods 69–70
RECON project 1007
recruitment of participants 558–559
reduced representation bisulfite sequencing (RRBS) 935
reference genome 846
reference panels 105
reference transcriptome 846
rejection algorithm 31–32
relaxed-clock models 207–208
autocorrelated models 208–209
grouping branch rates 209
independent branch-rate models 209
relaxed molecular clocks 207–209
remove unwanted variation (RUV) 744
replicates and clusters, hierarchical models of 887–888
replication 632–635
forms of 632–634
heterogeneity 635
motivation 632
significance thresholds 634–635
two-stage GWAS design 634
repressor proteins 700
reproduction ‘events’ 130
residual blocks 337
RettBASE 781
reverse graph embedding 750
reversible jump Markov chain Monte Carlo 71, 202, 490
rhodopsin parsed contact maps 335
ribosomally encoded and posttranslationally modified peptides (RiPP) 990
ridge regression 961
risk control models, of genetics 553–554
risk function 22
R language 1000
RNA-sequencing (RNA-seq) 697, 735
data 846–847
RNA structure prediction 328–329
Roadmap Epigenomics Project 680, 681
Robertson–Price identity 435, 437
Robinson model, of protein-coding DNA sequence evolution 358
ROLLOFF 265–267
*Saccharomyces cerevisiae* 617
sample-centric approach (genotypes storage) 74
Sanger Imputation Server 102
SCODE 752
scone package scores 742
score function 5
SCOTTI 1007
SCUBA 750
secondary theorem of selection 435
seemingly unrelated regression (SUR) model 868
SEER software 1011
segment parsing 334–335
segregation load 126
Segway 684, 685
selection coefficients 398, 438–441
selection differentials 424–425, 434
selection, fundamental and secondary theorems of 435–436
selection gradients 443, 446–447
correlational 448
inference of 447–451
normality and 450–451
quadratic 448
selection intensity 437, 439
selection on mean 439
selection on trait, detecting 439
selection on variance 439–441
selective sweeps 166–167, 401
semi-continuous model 545–546
sense codons 371
separation of time-scales 134, 157
sequence alignment 332
alignment benchmarking and improvement 333
family membership validation 332–333
sequence covariance analysis, in biological polymers 325–326, 340–341
applications
allostery and dynamics 330
CASP 330–332
globular protein fold prediction 328
protein disordered regions 329
protein–protein interactions 329–330
RNA structure prediction 328–329
transmembrane protein prediction 328
CCMpred method 327
comparison to known structures 333–334
DCA method 326–327
GREMLIN method 327–328
machine learning 335
deep learning method 336–338
MetaPSICOV method 336
PconsC 335–336
phylogeny constraints 339–340
sequence pairing 338–339
plmDCA method 327
PSICOV method 327
segment parsing 334–335
sequence alignment 332
alignment benchmarking and improvement 333
family membership validation 332–333
sequence data 138
sequence pairing 338–339
direct contact iteration 339
phylogenetic similarity 339
sequence path 358
sequencing reads 95
sequential kernel association test (SKAT) 619
sequentially Markov coalescent (SMC) 65, 139, 415
sequential Monte Carlo (SMC) 34
sequential selection 518
SeqWell 737
Seurat 751
Seurat algorithm 749
Seurat R package 753
sex systems 504
sexual selection 433
SHAPEIT approach 91–92, 96, 619
SHAPEIT v1 92, 93
SHAPEIT v2 92, 93, 95, 96
SHAPEIT v3 92
SHAPEIT v4 92
Sherlock 713
SHiC 414
shifting balance model of evolution 133
shortcut methods. See summary statistics methods
shrinkage methods 960–961. See also
metabolomics, statistical methods in
shuttle breeding 503
sibship frequency approach 464–465
SINCERA 749
Subject Index

Single-cell Analysis Via Expression Recovery (SAVER) 745
single-cell consensus clustering (SC3) 749
single-cell Interpretation via Multikernel LeaRning (SIMLR) 749
single-cell latent variable models (sCLVM) method 744
SINgle CELL Regularized Inference using Time-stamped Expression profiles (SINCERITIES) 752
Single Cell rEgulatory Network Inference and Clustering (SCENIC) 749
single-cell RNA-sequencing (scRNA-seq) 735–736
experimental platforms and low-level analysis computational analysis 739
high-throughput methods 737–739
low-throughput methods 736–737
novel statistical challenges estimating transcript levels 739–747
expression matrix analysis 747–753
single-locus models, of panmictic populations 116–117
diffusion approximations 120–126
adding selection and mutation 121–122
diffusion process 120–121
Gaussian fluctuations and drift load 125–126
Kolmogorov equations 123–124
multiple alleles 124–125
Wright–Fisher model and 121
random drift and Kingman coalescent 117–120
adding mutation 118
Cannings model 118–119
Kingman coalescent 117–118
limitations 119–120
Moran model 120
neutral Wright–Fisher model 117
spatially structured populations 126–130
continuous space 126–127
Kimura’s stepping stone model 127–130
spatial Lambda–Fleming–Viot process 130
single-marker multi-trait analysis 865
single nucleotide polymorphisms (SNPs) 53, 87, 98, 104, 301, 547–548, 574, 597, 598, 600, 633, 679, 706, 800, 817, 942
coding of SNP genotypes 607–609 (see also genome-wide association studies (GWASs))
low-quality SNPs 603, 604
quality control procedures 603–604
tag SNPs 600
single nucleotide variants (SNV) 765
SingleSplice 746
singleton density score (SDS) 411
singular value decomposition (SVD) 744, 962
site frequency spectrum (SFS) 400. See also allele frequency spectrum (AFS)
site-likelihoods 348
site pattern probabilities 227–229
site-wise likelihood ratio (SLR) test 386
SKAT-O 619
SLICER 750
SMARTPCA 616
Smartseq2 737
snapclust 1000, 1001
SNAPP 237
SNPs. See single nucleotide polymorphisms (SNPs)
SNP tagging approaches 102–103
SNPTEST 611
SNPttools approach 94
social issues, by genetics research 567–568
soft sweep 403, 404
software
genetic structure in GWAS 615–616
gene tree topology probabilities 225
for GWAS quality control 606
for haplotype reconstruction in GWASs 619–620
for single-variant analysis 611
STRUCTURE 136
software library 74
solidary participation 557
South Asian Genome project 776
spatial Lambda–Fleming–Viot process 130
Spearman correlation network 752
speciation times and population sizes, estimation of 239–240
species delimitation 239, 241
species tree 170, 219–221. See also multispecies coalescent
species tree inference, under multispecies coalescent 231
Bayesian full-data methods 234–235
empirical examples 237–238
multilocus versus SNP data 236–237
site pattern-based methods 235–236
summary statistics methods 231–234
species tree likelihoods, under multispecies coalescent 229–230
gene trees and 229–230
gene tree topologies and 229
multilocus data and 230
speed breeding platform 503
SpeedGene data format 74
splines 449
SpydrPick 1013
squared correlation 105
SQUAREM method 11–12
'S'-statistic 284–287
stabilizing selection 439, 441
differential 440
gradients 448
STARNET 680
stationary distribution 24
stationary frequencies 201
statistical coupling 330
statistical genetics 1
statistical inference, model-based 1
Bayesian inference 20–37
hidden Markov models 40–46
maximum likelihood inference 4–19
model selection 37–40
statistical models and inference 1–4
statistical models 1, 2. See also statistical inference, model-based
independence assumptions 2–3
probability distributions 3–4
statistical phylogenetics 178
statistical total correlation spectroscopy (STOCSY) 969, 970
stepping stone model 127–130
stewardship 565–567
STITCH (Sequencing To Imputation Through Constructing Haplotypes) method 95
stochastic differential equation 120
stochastic models 115
of structural evolution 359–360
stochastic search variable selection (SSVS) 858
stop codons 371
Storey's method 706
strict clock 1003
strong migration 156–157
structural variants (SV) 765
STRUCTURE 136, 480–483, 998
clustering algorithms 252–253
structured coalescent 137, 155–156
structured Wright–Fisher model 154
Structure (v2) model 1000
subjective probabilities, in Bayesian analysis 181
subset quantile normalization (SQN) 940
substitution-based methods, to detect selection 405–406
substitution models 184, 199–202
GTR model 200–201
Jukes and Cantor (1969) model 199–200
Kimura (1980) model 200
reversible-jump MCMC 202
time-reversible 201–202
subtree pruning and regrafting (SPR) 192
succinct tree sequence 75
summary statistics methods 231–234
SuperDCA 1013
supervised learning methods 682–684
support vector machine (SVM) 414
SVDQuartets method 236
sweep
hard 137, 403, 411
incomplete 408
selective 166–167, 401
soft 137, 403, 404, 411
sweep from standing variation (SSV) 403
Swendsen–Wang algorithm 861
switching model 390
symmetrical models 137
symmetric matrix 425
Systematic Long Range Phasing (SLRP) 94
systematic review 557

T
Targeting Induced Local Lesions in Genomes (TILLING) 509
TASC 751
technical replicates 845
temporal method 466–470
TensorFlow package 892
TESLA 921
theory of speed and scale 123
therapeutic misconception 557, 559
therapeutic privilege 563
**Subject Index**

three-population test 277–279
threshold number 472
Time reconstruction in Single-Cell RNAseq ANalysis (TSCAN) 750
time-reversible model 201
time to the most recent common ancestor (TMRCA) 261
tip dating approach 211
TopMed study 105
topology 148–149
total evidence approach 211
total inbreeding coefficient 535
trait SDTS (tSDS) 411
trans-acting elements 701
transcription 700–702
transcription factor binding-associated proteins (TAF) 700
transcription factor binding site (TFBS) 721
transcriptomic annotation data 680–681.
  See also genome-wide association studies (GWASs)
transcripts per million (TPM) 742
trans-eQTL 712, 713
trans-ethnic fine-mapping 622
transgressive segregation 502
transition probabilities 89, 90, 348
transition–transversion rate difference 373, 378–379
TransPhylo 1006
transversions 200
TraP 767
tree
  neighbors 192
topology 192
tree bisection and reconnection (TBR) 192
TreeTime 1004
treeWAS 1011
t-SNE 1014
TWAS 689, 690

**U**

UK Biobank 268, 555–556, 566, 622
UK10K 771
UK National Health Service (NHS) 555
Ultimate MRCA 161
unbiased estimator 7

UniProt 781
unique molecular identifiers (UMI) 739–740
UNPHASED 619
unsupervised learning methods 684
UTMOST 690

**V**

variance effective size 459
variance selection differential 442
variance-standardized selection differential 437
Variant Call Format 74
variant-centric approach (data compression) 74
Variational Bayes (VB) 864
VEGAS 690
Viterbi algorithm 43–44

**W**

WAG model 350
Wald confidence intervals 9
Wald statistic 8
Wald test 606
walk through tree space 160, 163
Wanderlust 750
Waterfall 750
Watterson estimator 90
Wellcome Trust Case Control Consortium (WTCCC) 598, 633, 637
WhatsHap method 95
whole-exome sequencing (WES) 765
whole-genome bisulfite sequencing (WGBS) 935
whole genome sequencing (WGS) 105, 275, 547, 601, 765, 999. See also genome-wide association studies (GWASs)
Wilcoxon test 955
Wilcoxon rank-sum test 751
winner’s curse 635–640
Wishart distribution 868
Wright–Fisher diffusion 121
Wright–Fisher model 25, 116, 117, 121
Wright–Fisher transition matrix 468
Wright–Malécot formula 127, 128
Wright’s fixation index 475
Wright’s F-statistics 535
Wright’s inbreeding coefficient 605
Subject Index

x
X chromosome, genotype data from 605

y
Y-STR database 539
Y-STR profiles 539–542
Y-typing 539

z
Zebrafish Information Network (ZFIN) 781
Zellner–Siow Cauchy 863
zero-inflated factor analysis (ZIFA) 747
zero-inflated negative binomial-based wanted variation extraction (ZINB-WaVE) 747
Z-scores 104, 687