

Index

A

- Aligned pattern clustering, 132, 172
 - class A scavenger receptors, 176–178
 - pattern and data spaces of, 173
 - WeMine system (*see* WeMine aligned pattern clustering system)
- Aligned residue associations
 - definition of, 175
 - discovered, 177
 - pattern and data spaces of, 173
- Alport syndrome, 12
- Array-based methods, 39
- Assembled contigs
 - correctly and incorrectly, 122–124
 - k*-mers, 125

B

- Balanced translocation, 39
- Batch-effect correction, scRNA-seq, 26
- Bayesian inference
 - complex models with many parameters, 68, 69
 - differential gene expression, 79–85
 - gene expression modeling, 77–79
 - likelihood distribution, 66
 - posterior distribution, 68
 - prior distribution, 67, 68
 - single gene expression, 69–72
 - Stan case study, 79–85
 - whole transcriptome expression, 72–77
- Benchmark. *See* simulated benchmark transcriptome datasets

- Bernoulli distribution, 70, 73
- Beta distribution, 70
- Binary fingerprints, 42
- Binomial distribution, 70–71
- Biomarker
 - definition of, 54
 - toxicological studies, 54
- Biomarker discovery
 - classification, 54
 - feature selection, 55
 - machine learning techniques, 54–55
- Biomarker discovery evaluation methods
 - comparative study of, 56–57
 - “gold standard” gene sets, 55
 - prediction accuracy, 56
 - stability, 55–56
- Biomedical Entity Search Tool, 4
- Biosequence pattern analysis. *See* WeMine aligned pattern clustering system

C

- Categorical distribution, 73
- Cell cluster annotation, scRNA-seq, 28–29
- Cell quality control, scRNA-seq, 25
- Chromosomal translocation, 41
- Class A scavenger receptors APC, 176–178
- Class association, 138, 146–147
- Cluster analysis, scRNA-seq, 28–29
- Computational analysis, scRNA-seq
 - alignment of raw data, 25
 - batch-effect correction, 26
 - cell quality control, 25
 - data analysis platforms, 23–24
 - data correction, 26

- data interpretability, 27–28
- data processing stages, 24
- dimensionality reduction, 27–28
- expression recovery, 27
- external data integration, 26
- feature selection, 27
- flow chart summary of, 24
- generation and normalization, count matrix, 25–26
- visualizations, 28
- Conjugate distribution, 73
- Co-occurrence based algorithms, 2
- Co-occurring pattern association, 136, 144–146
- Copy number variations, 39

D

- Data correction, scRNA-seq, 26
- Data interpretability, 27–28
- Data perturbation, 57, 58, 59
- Differential expression testing, 30
- Dimensionality reduction, 27–28
- Dirichlet distribution, 73–74
- Discordant reads, 41
- Disease maps, scRNA-seq, 30–31
- Downstream analysis, scRNA-seq
 - cell cluster annotation, 28–29
 - cluster analysis, 28–29
 - differential expression testing, 30
 - disease maps, 30–31
 - gene expression dynamics, 29–30
 - gene regulatory networks, 30–31
 - pseudotime, 29–30
 - trajectory analysis and inference, cellular origins, 29
- Droplet-based techniques, scRNA-seq, 21

E

- Ensemble approach, 113, 125
 - combination study of, 58, 59
 - data perturbation, 57, 58, 59
 - function perturbation, 57, 58, 59
- Error Correction Evaluation Toolkit, 90
 - accuracy of, 94–96
 - input data, 93–94
 - limitations, 90–91
 - process of, 91–93
 - Illumina read error-correction tools, 96–99, 101, 102
 - ONT read error-correction tools, 97, 104
 - PacBio error-correction tools, 97, 101
 - software availability, 104
 - TGS read error-correction tools, 98
 - Error-free reads, 90
- Expression recovery, 27
- External data integration, 26

F

- Feature selection, 27, 55
- Fluorescent activated cell sorting, 21
 - biomarkers, 11
 - cause of CKD, 7
 - text-mining algorithms, 13
- Function perturbation, 57, 58, 59

G

- Gamma distribution, 72
- Gene expression dynamics, 29–30
- Gene expression modeling
 - biological variability, 77–78
 - negative binomial distribution, 77–78
 - normalization factors, 78–79

Gene expression shrinkage, hierarchical
model of, 76–77
Gene ranking, 3–4
Gene regulatory networks, 30–31
Genie algorithm, 4
Genome-guided approach, 111,
120–121, 122
Gold standard gene sets, 55

H

High-throughput chromosome
conformation capture (Hi-C)-
based methods, 46–48
cross-links, 46
ligation, 46, 48
process diagram, 46
transposable elements, 47
Hydatidiform moles, 4–7
complete and partial, 5
integration of genes, 6
pathophysiology of, 4–5

I

Illumina read error-correction tools,
96–99, 101, 102
Integrated structural variant calling
pipeline, 44
Interactive web interface, 8, 11–14
Intra-protein interaction, 138

K

k-mers, 90
in assembled contigs, 125
counting method, 75–76
de novo assemblies, 121–123, 125

L

Likelihood of protein-protein interaction
prediction, 146
Log likelihood, 84

M

Machine learning techniques, 54–55
Microwell-based techniques,
scRNA-seq, 21
Multinomial distribution, 73–74
Multiple sequence alignment, 152, 174

N

Natural language processing (NLP)
algorithms, 2
Negative binomial distribution, 77–78
Next generation sequencing (NGS)
method, 40–44
binary fingerprints, 42
discordant reads, 41
Normalization
count matrix, 25–26
gene expression modeling, 78–79

O

Omics techniques, 66
ONT read error-correction tools, 97, 104
Over-dispersion parameter, 82

P

PacBio error-correction tools, 97, 101
Pairwise alignment methods, 152
Pattern class association, 138, 146–147
Pattern directed align pattern clustering, 141

- Pattern discovery, 134, 135
- Pattern discovery and disentanglement
- applications of, 172
 - area under curve, 180
 - association discovery on class A
 - scavenger receptors APC, 176–178
 - binding site prediction, 178–180
 - definition of, 174
 - methodology of, 174–176
- Pattern extension method, 142
- Pattern gaps and mutations, 141–142
- Pattern refinement, 136, 140, 141–142
- Pattern summarization, 134, 135, 139–141
- Percentage similarity, 95
- Plant transcriptome assembly.
- See* transcriptome assembly
- Poisson distribution
- single gene expression, 71–72
 - whole transcriptome expression, 73–74
- Poisson limit theorem, 71
- Posterior distribution, 68
- Prior distribution, 67, 68
- binomial distribution, 70–71
 - differentially expressed genes, 69
 - Poisson distribution, 71–72
 - single gene expression, 69–72
- Profile hidden Markov models (Profile HMMs)
- description of, 152
 - integrated approach for viral research, 163–164, 165
 - MinionDB, 164, 166
 - multigene elements, cellular organisms, 163, 164
 - rational design of, 157–160
 - roadmap for rational design of, 155–157
 - screening sequencing datasets, 160–162
 - targeted sequence reconstruction, 162
 - viral (*see* viral profile hidden Markov models)
- Protein-protein interacting pair prediction, 178–180
- Protein-protein interaction prediction, 137, 146
- Pseudotime, 29–30
- p*-value, 54
- ## R
- Reference-based transcriptome assembly
- performance metrics, 114–115
- Reference-free transcriptome assembly
- performance metrics, 114
- Reprojected statistical residual vector space, 174, 176
- Residue-residue contact (R2R-C), 175
- Residue-to-residue interactions (R2R-I)
- prediction, 178–180
- RNA velocity, 29
- RNA-Seq based methods, 44–46
- RNA-sequencing (RNA-seq)
- definition of, 110
 - simulation examples, 117, 120
 - simulation methods, 117
 - single-cell (*see* single-cell RNA-sequencing (scRNA-seq))
- ## S
- Screening sequencing datasets, 160–162
- Segmental duplication, 39
- Simulated benchmark transcriptome datasets
- assembly performance metrics, 115–116

- RNA-seq simulation examples, 117, 120
- RNA-seq simulation methods, 117
- Single cell transcriptome, 7–8
 - of human adult kidney, 11
- Single gene expression, probabilistic models
 - binomial distribution, 70–71
 - Poisson distribution, 71–72
- Single nucleotide variant, 38
- Single-cell RNA-sequencing (scRNA-seq)
 - computational analysis (*see* computational analysis, scRNA-seq)
 - for data analysis, 21–22
 - downstream analysis (*see* downstream analysis, scRNA-seq)
 - droplet-based techniques, 21
 - flow chart guide, 23
 - full-length based protocols, 21–22
 - microwell-based techniques, 21
 - selection of right approach, 22
 - tag-based protocols, 21–22
- Statistical hypothesis test, 54
- Statistical residual vector space, 174, 176
- Structural variation
 - definition of, 38
 - gene fusions, 39
 - Hi-C based methods, 46–48
 - next generation sequencing method, 40–44
 - RNA-Seq based methods, 44–46
 - types of, 38
- Student's *t*-test, 54
- T**
- Targeted sequence reconstruction, 162
- Text mining gene selection
 - bioinformatics process diagram, 9
 - bioinformatics tools, 3–4
 - co-occurrence based algorithms, 2
 - focal segmental glomerulosclerosis, 7–8
 - gene ranking, 3–4
 - hydatidiform moles, 4–7
 - integration of genes, 6
 - interactive web interface, 8, 11–14
 - natural language processing algorithms, 2
 - online result visualization, 8, 11–14
 - single cell transcriptome, 7–8
 - visualization of results, 2–3
- TGS read error-correction tools, 91, 98
- Third-generation sequencing, 90
- Transcription factor binding sites, 133
- Transcriptome, 110
- Transcriptome assembly
 - de novo* approach, 112–113
 - definition of, 110
 - ensemble approach, 113
 - genome-guided approach, 111
- Transcriptome assembly performance
 - comparison of methods, 118–119, 126
 - reference-based metrics, 114–115
 - reference-free metrics, 114
 - simulated benchmark data, 115–116
- Translocation
 - balanced, 39
 - chromosomal, 41
 - Hi-C-based methods, 46–48
 - phenomenon of, 38
- Transposable elements, 47
- U**
- Unique molecular identifiers, 22
- V**
- Viral discovery, 151, 157, 162, 167
- Viral metagenomics, 167

Viral profile hidden Markov models
databases for, 152–155
integrated approach, 163–164, 165
MinionDB, 164, 166

W

WeMine aligned pattern clustering
system
biological applications of, 142–148
class association, 138, 146–147
comparative study of, 143–144
co-occurring pattern association, 136,
144–146
description of, 132
overview of, 133–134, 139
pattern discovery, 134, 135
pattern extension method, 142
pattern gaps and mutations, 141–142

pattern refinement, 136, 140, 141–142
pattern summarization, 134, 135,
139–141
protein-protein interaction prediction,
137, 146
proteomic application, 141
Whole genome sequencing, 39
Whole transcriptome expression
binomial to multinomial distribution,
73–74
categorical distribution, 73
Dirichlet distribution, 73–74
gene expression shrinkage, hierarchical
model, 76–77
k-mer counting method, 75–76
multi-mapping reads, 74–75
Poisson distribution, 73–74
Wilms' tumor protein, 11

Doi: <https://doi.org/10.36255/exonpublications.bioinformatics.2021.ind>