Statistical Analysis Of Next Generation Sequencing Data

#next generation sequencing data analysis #NGS statistical methods #bioinformatics data interpretation #genomic sequencing statistics #differential expression analysis NGS

Explore the critical role of statistical analysis in deciphering Next Generation Sequencing (NGS) data, essential for extracting meaningful biological insights from vast genomic datasets. This field involves applying advanced statistical methods to identify patterns, variations, and significant findings in areas like genomics, transcriptomics, and epigenetics, crucial for research and clinical applications.

Statistical Analysis of Next Generation Sequencing Data

Next Generation Sequencing (NGS) is the latest high throughput technology to revolutionize genomic research. NGS generates massive genomic datasets that play a key role in the big data phenomenon that surrounds us today. To extract signals from high-dimensional NGS data and make valid statistical inferences and predictions, novel data analytic and statistical techniques are needed. This book contains 20 chapters written by prominent statisticians working with NGS data. The topics range from basic preprocessing and analysis with NGS data to more complex genomic applications such as copy number variation and isoform expression detection. Research statisticians who want to learn about this growing and exciting area will find this book useful. In addition, many chapters from this book could be included in graduate-level classes in statistical bioinformatics for training future biostatisticians who will be expected to deal with genomic data in basic biomedical research, genomic clinical trials and personalized medicine. About the editors: Somnath Datta is Professor and Vice Chair of Bioinformatics and Biostatistics at the University of Louisville. He is Fellow of the American Statistical Association, Fellow of the Institute of Mathematical Statistics and Elected Member of the International Statistical Institute. He has contributed to numerous research areas in Statistics, Biostatistics and Bioinformatics. Dan Nettleton is Professor and Laurence H. Baker Endowed Chair of Biological Statistics in the Department of Statistics at Iowa State University. He is Fellow of the American Statistical Association and has published research on a variety of topics in statistics, biology and bioinformatics.

Next Generation Sequencing and Data Analysis

This textbook provides step-by-step protocols and detailed explanations for RNA Sequencing, ChIP-Sequencing and Epigenetic Sequencing applications. The reader learns how to perform Next Generation Sequencing data analysis, how to interpret and visualize the data, and acquires knowledge on the statistical background of the used software tools. Written for biomedical scientists and medical students, this textbook enables the end user to perform and comprehend various Next Generation

Sequencing applications and their analytics without prior understanding in bioinformatics or computer sciences.

Bioinformatics

This book contains the latest material in the subject, covering next generation sequencing (NGS) applications and meeting the requirements of a complete semester course. This book digs deep into analysis, providing both concept and practice to satisfy the exact need of researchers seeking to understand and use NGS data reprocessing, genome assembly, variant discovery, gene profiling, epigenetics, and metagenomics. The book does not introduce the analysis pipelines in a black box, but with detailed analysis steps to provide readers with the scientific and technical backgrounds required to enable them to conduct analysis with confidence and understanding. The book is primarily designed as a companion for researchers and graduate students using sequencing data analysis but will also serve as a textbook for teachers and students in biology and bioscience.

Next-Generation Sequencing Data Analysis

A Practical Guide to the Highly Dynamic Area of Massively Parallel SequencingThe development of genome and transcriptome sequencing technologies has led to a paradigm shift in life science research and disease diagnosis and prevention. Scientists are now able to see how human diseases and phenotypic changes are connected to DNA mutation, polymorphi

STATISTICAL ANALYSIS OF HUMAN

This dissertation, "Statistical Analysis of Human Gastrointestinal Microbiota Using Next Generation Sequencing Data" by Youwen, Qin, ƒ Ëa, obtained from The University of Hong Kong (Pokfulam, Hong Kong) and is being sold pursuant to Creative Commons: Attribution 3.0 Hong Kong License. The content of this dissertation has not been altered in any way. We have altered the formatting in order to facilitate the ease of printing and reading of the dissertation. All rights not granted by the above license are retained by the author. Abstract: The human gastrointestinal tract is the niche of both commensal and pathogenic microbes which play an important role in human health. This thesis includes two independent studies relevant to analyzing next-generation sequencing data on the human gastrointestinal microbiota. The first study conducted a comparative analysis on 16S rRNA gene sequencing data obtained from gastritis and gastric cancer patients in the Hong Kong (HK) and Korean cohorts. Neisseriaceae and Lachnospiraceae were the important families in segregating gastritis and cancer samples in the HK dataset while it was Streptococcaceae in the Korean dataset. Proteobacteria, Firmicutes, Bacteroidetes, Actinobacteria and Fusobacteria were the major phyla in the two cohorts, where they made up >= 99% of the total relative abundance. However, when narrowed down to the family level, the two datasets only shared 5 major families among the 15 and 13 major families in the HK and Korean datasets, respectively. Hierarchical clustering showed that samples were segregated into two major clusters according to the relative abundance of Helicobacteria pylori (H. pylori) in the two datasets. Moreover, the cross-prediction results for gastritis versus cancer between two datasets yielded up to 3 times larger error rates compared to the prediction results within the training set. Taken together, the differences between the HK and Korean cohorts in the gastric microbiota outweighed the similarities. The second study developed a computational workflow to improve the draft genomes assembled from shotgun metagenomic sequencing data. The publicly available sequencing data of 396 human stool samples were downloaded for this purpose. Firstly, 3.9 million genes assembled from 396 samples were clustered into 7,381 co-abundance gene groups (CAGs) according to their pairwise correlations. The CAGs (741 CAGs) with more than 700 genes were defined as metagenomic species (MGSs), while the others (6,640 CAGs) were defined as metagenomic units (MGUs). In order to recover the relevant MGSs of the MGUs, the metagenomic deconvolution framework which decomposes the community-level gene content into taxon-specific gene profile was applied. Overall, 377 MGUs were assigned to 354 relevant MGSs, achieving a 9.57% mean improvement in the gene count of MGSs. Most of these MGSs were annotated to phylum Firmicutes. Specifically, the augmented results of 9 MGSs annotated to genus Faecalibacterium by their relative MGUs achieved average improvement of 21.08% and 17.84% in sensitivity and specificity. Importantly, MGUs included essential genes that were missed in MGSs, such as ribosomal genes, metabolism and transport system genes. Hence, the implementation of metagenomic deconvolution after binning improves the draft genomes of metagenomic species. Subjects: Gastrointestinal system - Microbiology Nucleotide sequence

Statistical Analysis of Microbiome Data with R

This unique book addresses the statistical modelling and analysis of microbiome data using cutting-edge R software. It includes real-world data from the authors' research and from the public domain, and discusses the implementation of R for data analysis step by step. The data and R computer programs are publicly available, allowing readers to replicate the model development and data analysis presented in each chapter, so that these new methods can be readily applied in their own research. The book also discusses recent developments in statistical modelling and data analysis in microbiome research, as well as the latest advances in next-generation sequencing and big data in methodological development and applications. This timely book will greatly benefit all readers involved in microbiome, ecology and microarray data analyses, as well as other fields of research.

Statistical Analysis of Human Gastrointestinal Microbiota Using Next Generation Sequencing Data

The new genetic revolution is fuelled by Deep Sequencing (or Next Generation Sequencing) apparatuses which, in essence, read billions of nucleotides per reaction. Effectively, when carefully planned, any experimental question which can be translated into reading nucleic acids can be applied.In Deep Sequencing Data Analysis, expert researchers in the field detail methods which are now commonly used to study the multi-facet deep sequencing data field. These included techniques for compressing of data generated, Chromatin Immunoprecipitation (ChIP-seq), and various approaches for the identification of sequence variants. Written in the highly successful Methods in Molecular Biology series format, chapters include introductions to their respective topics, lists of necessary materials and reagents, step-by-step, readily reproducible protocols, and key tips on troubleshooting and avoiding known pitfalls. Authoritative and practical, Deep Sequencing Data Analysis seeks to aid scientists in the further understanding of key data analysis procedures for deep sequencing data interpretation.

Deep Sequencing Data Analysis

Next generation sequencing (NGS) has surpassed the traditional Sanger sequencing method to become the main choice for large-scale, genome-wide sequencing studies with ultra-high-throughput production and a huge reduction in costs. The NGS technologies have had enormous impact on the studies of structural and functional genomics in all the life sciences. In this book, Next Generation Sequencing Advances, Applications and Challenges, the sixteen chapters written by experts cover various aspects of NGS including genomics, transcriptomics and methylomics, the sequencing platforms, and the bioinformatics challenges in processing and analysing huge amounts of sequencing data. Following an overview of the evolution of NGS in the brave new world of omics, the book examines the advances and challenges of NGS applications in basic and applied research on microorganisms, agricultural plants and humans. This book is of value to all who are interested in DNA sequencing and bioinformatics across all fields of the life sciences.

Next Generation Sequencing

Recent improvements in the efficiency, quality, and cost of genome-wide sequencing have prompted biologists and biomedical researchers to move away from microarray-based technology to ultra high-throughput, massively parallel genomic sequencing (Next Generation Sequencing, NGS) technology. In Next Generation Microarray Bioinformatics: Methods and Protocols, expert researchers in the field provide techniques to bring together current computational and statistical methods to analyze and interpreting both microarray and NGS data. These methods and techniques include resources for microarray bioinformatics, microarray data analysis, microarray bioinformatics in systems biology, next generation sequencing data analysis, and emerging applications of microarray and next generation sequencing. Written in the highly successful Methods in Molecular BiologyTM series format, the chapters include the kind of detailed description and implementation advice that is crucial for getting optimal results in the laboratory. Authoritative and practical, Next Generation Microarray Bioinformatics: Methods and Protocols seeks to aid scientists in the further study of this crucially important research into the human DNA.

Deep Sequencing Data Analysis: Challenges and Solutions

Population genomics is a recently emerged discipline, which aims at understanding how evolutionary processes influence genetic variation across genomes. Today, in the era of cheaper next-generation sequencing, it is no longer as daunting to obtain whole genome data for any species of interest and

population genomics is now conceivable in a wide range of fields, from medicine and pharmacology to ecology and evolutionary biology. However, because of the lack of reference genome and of enough a priori data on the polymorphism, population genomics analyses of populations will still involve higher constraints for researchers working on non-model organisms, as regards the choice of the genotyping/sequencing technique or that of the analysis methods. Therefore, Data Production and Analysis in Population Genomics purposely puts emphasis on protocols and methods that are applicable to species where genomic resources are still scarce. It is divided into three convenient sections, each one tackling one of the main challenges facing scientists setting up a population genomics study. The first section helps devising a sampling and/or experimental design suitable to address the biological question of interest. The second section addresses how to implement the best genotyping or sequencing method to obtain the required data given the time and cost constraints as well as the other genetic resources already available, Finally, the last section is about making the most of the (generally huge) dataset produced by using appropriate analysis methods in order to reach a biologically relevant conclusion. Written in the successful Methods in Molecular BiologyTM series format, chapters include introductions to their respective topics, lists of the necessary materials and reagents, step-by-step, readily reproducible protocols, advice on methodology and implementation, and notes on troubleshooting and avoiding known pitfalls. Authoritative and easily accessible, Data Production and Analysis in Population Genomics serves a wide readership by providing guidelines to help choose and implement the best experimental or analytical strategy for a given purpose.

Next Generation Microarray Bioinformatics

Nucleic acid sequencing techniques have enabled researchers to determine the exact order of base pairs - and by extension, the information present - in the genome of living organisms. Consequently, our understanding of this information and its link to genetic expression at molecular and cellular levels has lead to rapid advances in biology, genetics, biotechnology and medicine. Next-Generation Sequencing and Sequence Data Analysis is a brief primer on DNA sequencing techniques and methods used to analyze sequence data. Readers will learn about recent concepts and methods in genomics such as sequence library preparation, cluster generation for PCR technologies, PED sequencing, genome assembly, exome sequencing, transcriptomics and more. This book serves as a textbook for students undertaking courses in bioinformatics and laboratory methods in applied biology. General readers interested in learning about DNA sequencing techniques may also benefit from the simple format of information presented in the book.

Data Production and Analysis in Population Genomics

The new genetic revolution is fuelled by Deep Sequencing (or Next Generation Sequencing) apparatuses which, in essence, read billions of nucleotides per reaction. Effectively, when carefully planned, any experimental question which can be translated into reading nucleic acids can be applied.In Deep Sequencing Data Analysis, expert researchers in the field detail methods which are now commonly used to study the multi-facet deep sequencing data field. These included techniques for compressing of data generated, Chromatin Immunoprecipitation (ChIP-seq), and various approaches for the identification of sequence variants. Written in the highly successful Methods in Molecular Biology series format, chapters include introductions to their respective topics, lists of necessary materials and reagents, step-by-step, readily reproducible protocols, and key tips on troubleshooting and avoiding known pitfalls. Authoritative and practical, Deep Sequencing Data Analysis seeks to aid scientists in the further understanding of key data analysis procedures for deep sequencing data interpretation.

Next-Generation Sequencing and Sequence Data Analysis

Advances in the field of genomics over the last years have led to diverse types of measurements with vast amounts of next generation sequencing (NGS) data. As genomic measurements become faster and cheaper with higher-throughput technologies, larger quantities of data are being generated, and awaiting novel insights and discoveries. Multiple NIH-funded large consortia projects such as the Encyclopedia of DNA Elements, the Roadmap Epigenomics Mapping Consortium, and the Genotype-Tissue Expression Project have generated a myriad of data types of DNA accessibility (DNAse-seq and ATAC-seq), protein-DNA interaction (ChIP-seq, and ChIP-exo/nexus), RNA transcription (RNA-seq), and DNA methylation (Methyl-seq) among many other assays. In this thesis we introduce two methods. First, we present ChIPexoQual, an R/Bioconductor quality control pipeline that enables exploration and analysis of ChIP-exo and related experiments. ChIPexoQual evaluates a number of key issues

including strand imbalance, library complexity, and signal enrichment of data. Assessment of these features are facilitated through diagnostic plots and summary statistics computed over regions of the genome with varying levels of coverage. We evaluated our QC pipeline with both large collections of public ChIP-exo/nexus data and multiple, new ChIP-exo datasets from Escherichia coli. ChIPexoQual analysis of these datasets resulted in guidelines for using these QC metrics across a wide range of sequencing depths and provided further insights for modeling ChIP-exo data. Next, we introduce FM-HighLD, a fine-mapping method for variants in high linkage disequilibrium that, by using ChIP-seq based annotation data, is capable of calculating SNP probabilities for being causal for single or multiple traits by using annotation data. We provide a simple pipeline to calculate the allelic skew, a measure of allele-specific transcriptional activity calculated with in-vivo sequencing data. We evaluated FM-HighLD computationally by data-driven simulations, and to the best of our knowledge used for the first time Massively Parallel Reproducible Assays (MPRA) as gold-standards with which to evaluate causal predictions of eQTL associations in lymphoblastoid cell lines of variants in high LD. These simulations resulted in a number of key findings regarding FM-HighLDs' performance. First, FM-HighLD outperforms state-of-the-art fine-mapping methods when the causal variants are randomly selected from the variants deemed as regulatory by Tewhey et al. Second, FM-HighLD exhibits higher recall rate when fewer causal variants are considered in a polygenic model. Finally, when the causal variants are selected in LD clusters composed by SNPs in high LD, FM-HighLD outperforms PAINTOR, and is comparable to CAVIAR. Finally, in this thesis, we showcase two R/Bioconductor packages to facilitate the analysis of next generation sequencing data, and explore RNA-seq differential expression results. First, ChIPUtils is a package that calculates typical ChIP-seq quality control metrics, and generates a few visualizations to explore the data. Second, Segvis is a package for computing the genomic coverage of high throughout sequencing data along genomic segments, and allows the user multiple visualizations across different samples and peak sets. In other words, this package is capable of generating multiple visualization summarizing the binding patterns across individual or multiple peaks of different samples aligned to the genome.

Deep Sequencing Data Analysis

Big Data in Omics and Imaging: Association Analysis addresses the recent development of association analysis and machine learning for both population and family genomic data in sequencing era. It is unique in that it presents both hypothesis testing and a data mining approach to holistically dissecting the genetic structure of complex traits and to designing efficient strategies for precision medicine. The general frameworks for association analysis and machine learning, developed in the text, can be applied to genomic, epigenomic and imaging data. FEATURES Bridges the gap between the traditional statistical methods and computational tools for small genetic and epigenetic data analysis and the modern advanced statistical methods for big data Provides tools for high dimensional data reduction Discusses searching algorithms for model and variable selection including randomization algorithms, Proximal methods and matrix subset selection Provides real-world examples and case studies Will have an accompanying website with R code The book is designed for graduate students and researchers in genomics, bioinformatics, and data science. It represents the paradigm shift of genetic studies of complex diseases– from shallow to deep genomic analysis, from low-dimensional to high dimensional, multivariate to functional data analysis with next-generation sequencing (NGS) data, and from homogeneous populations to heterogeneous population and pedigree data analysis. Topics covered are: advanced matrix theory, convex optimization algorithms, generalized low rank models, functional data analysis techniques, deep learning principle and machine learning methods for modern association, interaction, pathway and network analysis of rare and common variants, biomarker identification, disease risk and drug response prediction.

Statistical Methods for ChIP-exo/nexus Quality Control, and Fine-mapping of Multi-trait SNPs in High LD by Using Next-generation Sequencing Data

This book is unique in covering a wide range of design and analysis issues in genetic studies of rare variants, taking advantage of collaboration of the editors with many experts in the field through large-scale international consortia including the UK10K Project, GO-T2D and T2D-GENES. Chapters provide details of state-of-the-art methodology for rare variant detection and calling, imputation and analysis in samples of unrelated individuals and families. The book also covers analytical issues associated with the study of rare variants, such as the impact of fine-scale population structure, and with combining information on rare variants across studies in a meta-analysis framework. Genetic association studies have in the last few years substantially enhanced our understanding of factors

underlying traits of high medical importance, such as body mass index, lipid levels, blood pressure and many others. There is growing empirical evidence that low-frequency and rare variants play an important role in complex human phenotypes. This book covers multiple aspects of study design, analysis and interpretation for complex trait studies focusing on rare sequence variation. In many areas of genomic research, including complex trait association studies, technology is in danger of outstripping our capacity to analyse and interpret the vast amounts of data generated. The field of statistical genetics in the whole-genome sequencing era is still in its infancy, but powerful methods to analyse the aggregation of low-frequency and rare variants are now starting to emerge. The chapter Functional Annotation of Rare Genetic Variants is available open access under a Creative Commons Attribution 4.0 International License via link.springer.com.

Big Data in Omics and Imaging

Good,No Highlights,No Markup,all pages are intact, Slight Shelfwear,may have the corners slightly dented, may have slight color changes/slightly damaged spine.

Assessing Rare Variation in Complex Traits

Introduces readers to core algorithmic techniques for next-generation sequencing (NGS) data analysis and discusses a wide range of computational techniques and applications This book provides an in-depth survey of some of the recent developments in NGS and discusses mathematical and computational challenges in various application areas of NGS technologies. The 18 chapters featured in this book have been authored by bioinformatics experts and represent the latest work in leading labs actively contributing to the fast-growing field of NGS. The book is divided into four parts: Part I focuses on computing and experimental infrastructure for NGS analysis, including chapters on cloud computing, modular pipelines for metabolic pathway reconstruction, pooling strategies for massive viral sequencing, and high-fidelity sequencing protocols. Part II concentrates on analysis of DNA sequencing data, covering the classic scaffolding problem, detection of genomic variants, including insertions and deletions, and analysis of DNA methylation sequencing data. Part III is devoted to analysis of RNA-seq data. This part discusses algorithms and compares software tools for transcriptome assembly along with methods for detection of alternative splicing and tools for transcriptome quantification and differential expression analysis. Part IV explores computational tools for NGS applications in microbiomics, including a discussion on error correction of NGS reads from viral populations, methods for viral quasispecies reconstruction, and a survey of state-of-the-art methods and future trends in microbiome analysis. Computational Methods for Next Generation Sequencing Data Analysis: Reviews computational techniques such as new combinatorial optimization methods, data structures, high performance computing, machine learning, and inference algorithms Discusses the mathematical and computational challenges in NGS technologies Covers NGS error correction, de novo genome transcriptome assembly, variant detection from NGS reads, and more This text is a reference for biomedical professionals interested in expanding their knowledge of computational techniques for NGS data analysis. The book is also useful for graduate and post-graduate students in bioinformatics.

Statistical Analysis of DNA Sequence Data

Microbiome research has focused on microorganisms that live within the human body and their effects on health. During the last few years, the quantification of microbiome composition in different environments has been facilitated by the advent of high throughput sequencing technologies. The statistical challenges include computational difficulties due to the high volume of data; normalization and quantification of metabolic abundances, relative taxa and bacterial genes; high-dimensionality; multivariate analysis; the inherently compositional nature of the data; and the proper utilization of complementary phylogenetic information. This has resulted in an explosion of statistical approaches aimed at tackling the unique opportunities and challenges presented by microbiome data. This book provides a comprehensive overview of the state of the art in statistical and informatics technologies for microbiome research. In addition to reviewing demonstrably successful cutting-edge methods, particular emphasis is placed on examples in R that rely on available statistical packages for microbiome data. With its wide-ranging approach, the book benefits not only trained statisticians in academia and industry involved in microbiome research, but also other scientists working in microbiomics and in related fields.

Computational Methods for Next Generation Sequencing Data Analysis

This book presents a guide to building computational gene finders, and describes the state of the art in computational gene finding methods, with a focus on comparative approaches. Fully updated and expanded, this new edition examines next-generation sequencing (NGS) technology. The book also discusses conditional random fields, enhancing the broad coverage of topics spanning probability theory, statistics, information theory, optimization theory and numerical analysis. Features: introduces the fundamental terms and concepts in the field; discusses algorithms for single-species gene finding, and approaches to pairwise and multiple sequence alignments, then describes how the strengths in both areas can be combined to improve the accuracy of gene finding; explores the gene features most commonly captured by a computational gene model, and explains the basics of parameter training; illustrates how to implement a comparative gene finder; examines NGS techniques and how to build a genome annotation pipeline.

Assembly and Analysis of Next-generation Sequencing Data

Probabilistic models are becoming increasingly important in analysing the huge amount of data being produced by large-scale DNA-sequencing efforts such as the Human Genome Project. For example, hidden Markov models are used for analysing biological sequences, linguistic-grammar-based probabilistic models for identifying RNA secondary structure, and probabilistic evolutionary models for inferring phylogenies of sequences from different organisms. This book gives a unified, up-to-date and self-contained account, with a Bayesian slant, of such methods, and more generally to probabilistic methods of sequence analysis. Written by an interdisciplinary team of authors, it aims to be accessible to molecular biologists, computer scientists, and mathematicians with no formal knowledge of the other fields, and at the same time present the state-of-the-art in this new and highly important field.

Statistical Analysis of Microbiome Data

This textbook describes recent advances in genomics and bioinformatics and provides numerous examples of genome data analysis that illustrate its relevance to real world problems and will improve the reader’s bioinformatics skills. Basic data preprocessing with normalization and filtering, primary pattern analysis, and machine learning algorithms using R and Python are demonstrated for gene-expression microarrays, genotyping microarrays, next-generation sequencing data, epigenomic data, and biological network and semantic analyses. In addition, detailed attention is devoted to integrative genomic data analysis, including multivariate data projection, gene-metabolic pathway mapping, automated biomolecular annotation, text mining of factual and literature databases, and integrated management of biomolecular databases. The textbook is primarily intended for life scientists, medical scientists, statisticians, data processing researchers, engineers, and other beginners in bioinformatics who are experiencing difficulty in approaching the field. However, it will also serve as a simple guideline for experts unfamiliar with the new, developing subfield of genomic analysis within bioinformatics.

Comparative Gene Finding

The past three decades have witnessed an explosion of what is now referred to as high-dimensional `omics' data. Bioinformatics Methods: From Omics to Next Generation Sequencing describes the statistical methods and analytic frameworks that are best equipped to interpret these complex data and how they apply to health-related research. Covering the technologies that generate data, subtleties of various data types, and statistical underpinnings of methods, this book identifies a suite of potential analytic tools, and highlights commonalities among statistical methods that have been developed. An ideal reference for biostatisticians and data analysts that work in collaboration with scientists and clinical investigators looking to ensure rigorous application of available methodologies. Key Features: Survey of a variety of omics data types and their unique features Summary of statistical underpinnings for widely used omics data analysis methods Description of software resources for performing omics data analyses

Biological Sequence Analysis

This book presents the statistical aspects of designing, analyzing and interpreting the results of genome-wide association scans (GWAS studies) for genetic causes of disease using unrelated subjects. Particular detail is given to the practical aspects of employing the bioinformatics and data handling methods necessary to prepare data for statistical analysis. The goal in writing this book is to give statisticians, epidemiologists, and students in these fields the tools to design a powerful genome-wide study based on current technology. The other part of this is showing readers how to

conduct analysis of the created study. Design and Analysis of Genome-Wide Association Studies provides a compendium of well-established statistical methods based upon single SNP associations. It also provides an introduction to more advanced statistical methods and issues. Knowing that technology, for instance large scale SNP arrays, is quickly changing, this text has significant lessons for future use with sequencing data. Emphasis on statistical concepts that apply to the problem of finding disease associations irrespective of the technology ensures its future applications. The author includes current bioinformatics tools while outlining the tools that will be required for use with extensive databases from future large scale sequencing projects. The author includes current bioinformatics tools while outlining additional issues and needs arising from the extensive databases from future large scale sequencing projects.

Genome Data Analysis

This book addresses complex problems associated with crop improvement programs, using a wide range of programming solutions, for genomics data handling and sustainable agriculture. It describes important concepts in genomics data analysis and sequence-based mapping approaches along with references. The book contains 16 chapters on recent developments in several methods of genomic data analysis for crop improvements and sustainable agriculture, all authored by eminent researchers who are experts in their fields. These chapters focus on applications of a wide range of key bioinformatics topics, including assembly, annotation, and visualization of next-generation sequencing (NGS) data; expression profiles of coding and noncoding RNA; statistical and quantitative genetics; trait-based association analysis, quantitative trait loci (QTL) mapping, and artificial intelligence in genomic studies. Real examples and case studies in the book will come in handy when applying the techniques. The relative scarcity of reference materials covering bioinformatics applications as compared with the readily available books also enhances the utility of this book. The targeted readers of the book are scientists, researchers, and bioinformaticians from genomics and advanced breeding in different areas. The book will appeal to the applied researchers engaged in crop improvements and sustainable agriculture by using bioinformatics tools, students, research project leaders, and practitioners from the various marginal disciplines and interdisciplinary research.

Bioinformatics Methods

This book is comprised of presentations delivered at the 5th Workshop on Biostatistics and Bioinformatics held in Atlanta on May 5-7, 2017. Featuring twenty-two selected papers from the workshop, this book showcases the most current advances in the field, presenting new methods, theories, and case applications at the frontiers of biostatistics, bioinformatics, and interdisciplinary areas. Biostatistics and bioinformatics have been playing a key role in statistics and other scientific research fields in recent years. The goal of the 5th Workshop on Biostatistics and Bioinformatics was to stimulate research, foster interaction among researchers in field, and offer opportunities for learning and facilitating research collaborations in the era of big data. The resulting volume offers timely insights for researchers, students, and industry practitioners.

Design, Analysis, and Interpretation of Genome-Wide Association Scans

Uncertainty quantification (UQ) is a mainstream research topic in applied mathematics and statistics. To identify UQ problems, diverse modern techniques for large and complex data analyses have been developed in applied mathematics, computer science, and statistics. This Special Issue of Mathematics (ISSN 2227-7390) includes diverse modern data analysis methods such as skew-reflected-Gompertz information quantifiers with application to sea surface temperature records, the performance of variable selection and classification via a rank-based classifier, two-stage classification with SIS using a new filter ranking method in high throughput data, an estimation of sensitive attribute applying geometric distribution under probability proportional to size sampling, combination of ensembles of regularized regression models with resampling-based lasso feature selection in high dimensional data, robust linear trend test for low-coverage next-generation sequence data controlling for covariates, and comparing groups of decision-making units in efficiency based on semiparametric regression.

Genomics Data Analysis for Crop Improvement

Scientists strive to develop clear rules for naming and grouping living organisms. But taxonomy, the scientific study of biological classification and evolution, is often highly debated. Members of a species, the fundamental unit of taxonomy and evolution, share a common evolutionary history and a common

evolutionary path to the future. Yet, it can be difficult to determine whether the evolutionary history or future of a population is sufficiently distinct to designate it as a unique species. A species is not a fixed entity â€" the relationship among the members of the same species is only a snapshot of a moment in time. Different populations of the same species can be in different stages in the process of species formation or dissolution. In some cases hybridization and introgression can create enormous challenges in interpreting data on genetic distinctions between groups. Hybridization is far more common in the evolutionary history of many species than previously recognized. As a result, the precise taxonomic status of an organism may be highly debated. This is the current case with the Mexican gray wolf (Canis lupus baileyi) and the red wolf (Canis rufus), and this report assesses the taxonomic status for each.

New Frontiers of Biostatistics and Bioinformatics

Next Generation Sequencing (NGS) Technology in DNA Analysis explains and summarizes next generation sequencing (NGS) technological applications in the field of forensic DNA analysis. The book covers the transition from capillary electrophoresis (CE)-based technique to NGS platforms and the fundamentals of NGS technologies, applications, and advances. Sections provide an overview of NGS technology and forensic science, including information on processing biological samples for forensic analysis, sequence analysis, and data analysis software as well as the analysis of NGS data. The book explores the valuable applications of NGS-based forensic DNA analysis and covers the validations and interpretation guidelines of NGS workflows. With chapter contributions from an international array of experts and the inclusion of practical case studies, this book is a useful reference for academicians and researchers in genetics, biotechnology, bioinformatics, biology, and medicine as well as forensic DNA scientists and practitioners who aim to learn, use, apply, and validate NGS-based technologies. Describes the fundamentals of NGS and its advances for forensic applications Explains the transition from CE-based technique to NGS technology Includes case studies related to NGS and DNA fingerprinting Explores the future use and applications of NGS technologies

Uncertainty Quantification Techniques in Statistics

Provides a global view of the recent advances in the biological sciences and the adaption of the pathogen to the host plants revealed using NGS. Molecular Omic’s is now a major driving force to learn the adaption genetics and a great challenge to the scientific community, which can be resolved through the application of the NGS technologies. The availability of complete genome sequences, the respective model species for dicot and monocot plant groups, presents a global opportunity to delineate the identification, function and the expression of the genes, to develop new tools for the identification of the new genes and pathway identification. Genome-wide research tools, resources and approaches such as data mining for structural similarities, gene expression profiling at the DNA and RNA level with rapid increase in available genome sequencing efforts, expressed sequence tags (ESTs), RNA-seq, gene expression profiling, induced deletion mutants and insertional mutants, and gene expression knock-down (gene silencing) studies with RNAi and microRNAs have become integral parts of plant molecular omic’s. Molecular diversity and mutational approaches present the first line of approach to unravel the genetic and molecular basis for several traits, QTL related to disease resistance, which includes host approaches to combat the pathogens and to understand the adaptation of the pathogen to the plant host. Using NGS technologies, understanding of adaptation genetics towards stress tolerance has been correlated to the epigenetics. Naturally occurring allelic variations, genome shuffling and variations induced by chemical or radiation mutagenesis are also being used in functional genomics to elucidate the pathway for the pathogen and stress tolerance and is widely illustrated in demonstrating the identification of the genes responsible for tolerance in plants, bacterial and fungal species.

Evaluating the Taxonomic Status of the Mexican Gray Wolf and the Red Wolf

This book addresses the difficulties experienced by wet lab researchers with the statistical analysis of molecular biology related data. The authors explain how to use R and Bioconductor for the analysis of experimental data in the field of molecular biology. The content is based upon two university courses for bioinformatics and experimental biology students (Biological Data Analysis with R and High-throughput Data Analysis with R). The material is divided into chapters based upon the experimental methods used in the laboratories. Key features include: • Broad appeal--the authors target their material to researchers in several levels, ensuring that the basics are always covered. • First book to explain how to use R and Bioconductor for the analysis of several types of experimental data in the field of molecular biology. •

Focuses on R and Bioconductor, which are widely used for data analysis. One great benefit of R and Bioconductor is that there is a vast user community and very active discussion in place, in addition to the practice of sharing codes. Further, R is the platform for implementing new analysis approaches, therefore novel methods are available early for R users.

Next Generation Sequencing (NGS) Technology in DNA Analysis

A comprehensive introduction to modern applied statistical genetic data analysis, accessible to those without a background in molecular biology or genetics. Human genetic research is now relevant beyond biology, epidemiology, and the medical sciences, with applications in such fields as psychology, psychiatry, statistics, demography, sociology, and economics. With advances in computing power, the availability of data, and new techniques, it is now possible to integrate large-scale molecular genetic information into research across a broad range of topics. This book offers the first comprehensive introduction to modern applied statistical genetic data analysis that covers theory, data preparation, and analysis of molecular genetic data, with hands-on computer exercises. It is accessible to students and researchers in any empirically oriented medical, biological, or social science discipline; a background in molecular biology or genetics is not required. The book first provides foundations for statistical genetic data analysis, including a survey of fundamental concepts, primers on statistics and human evolution, and an introduction to polygenic scores. It then covers the practicalities of working with genetic data, discussing such topics as analytical challenges and data management. Finally, the book presents applications and advanced topics, including polygenic score and gene-environment interaction applications, Mendelian Randomization and instrumental variables, and ethical issues. The software and data used in the book are freely available and can be found on the book's website.

Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches

This eBook is a collection of articles from a Frontiers Research Topic. Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact.

Molecular Data Analysis Using R

Dr. Anton Buzdin (AB) is employed by Omicsway Corp. (USA). AB received grants from Amazon and Microsoft Azure to support cloud computations. Dr. Xinmin Li is director of JCCC Shared Genomics Resource, the University of California, Los Angeles, CA Dr. Ye Wang is Director of Gene testing Department (Core Lab) of Qingdao Central Hospital, the Second Affiliated Hospital of Qingdao University

An Introduction to Statistical Genetic Data Analysis

This book is a printed edition of the Special Issue Transcriptional Regulation: Molecules, Involved Mechanisms and Misregulation that was published in IJMS

Next-Generation Sequencing of Human Antibody Repertoires for Exploring B-cell Landscape, Antibody Discovery and Vaccine Development

The analysis of molecular genetic data has driven the fields of molecular biology, genetics, population genetics, and quantitative genetics for over half a century. Only recently though has technology advanced to the point where molecular genetic data can be acquired cheaply and efficiently for the entire genome of several individuals enabling scientists to conduct genome-wide comparisons between several individuals or several population samples, and ask comprehensive questions regarding the nature of genetic variation in extant populations and the evolutionary forces in the population's history which generated and influenced this variation. Several challenges exist to utilizing these new technologies successfully however and in most cases both experimental optimization of laboratory protocols and the customization or de novo implementation of computational and statistical analysis methods are required to obtain adequate results. Even when the raw physical data acquired by these technologies has been successfully rendered into biologically meaningful molecular genetic data, the analysis

of these large, genome-wide datasets is formidable and again requires advanced and customized methods to ask biologically motivated questions and produce conclusive results which may not have been obtainable without complete genome information. Here, I discuss two main technologies for the acquisition of genome-wide molecular data, next-generation sequencing technologies and fixed-array highly multiplexed SNP genotyping, and discuss the challenges in applying them in plant systems. Additionally, I demonstrate a population genetic analysis for the detection of recent selective sweeps in four subpopu- lations of Oryza sativa (cultivated Asian rice) and one Oryza rufipogon population (wild Asian rice) utilizing the genome-wide molecular data acquired by next-generation sequencing. The development of an improved and accurate statistical method to detect selection in population genomic analysis combined with genome-wide data in each of these subpopulations allowed the extent and location of selective sweeps in Oryza sativa subpopulations and its wild progenitor Oryza rufipogon to be quantified and compared for the first time, revealing that each cultivated subpopulation appears to have a largely unique and independent selective and domestication history, but several advantageous alleles for cultivation of rice that originated and were selected for in one subpopulation have been introduced into other subpopulations by introgression. (Abstract).

Improving the Clinical Effectiveness of Metagenomic Next Generation Sequencing (mNGS) in Infection Disease Diagnosis and Treatment: Linking the NGS Specialists and Clinicians

Heterogeneity, or mixtures, are ubiquitous in genetics. Even for data as simple as mono-genic diseases, populations are a mixture of affected and unaffected individuals. Still, most statistical genetic association analyses, designed to map genes for diseases and other genetic traits, ignore this phenomenon. In this book, we document methods that incorporate heterogeneity into the design and analysis of genetic and genomic association data. Among the key qualities of our developed statistics is that they include mixture parameters as part of the statistic, a unique component for tests of association. A critical feature of this work is the inclusion of at least one heterogeneity parameter when performing statistical power and sample size calculations for tests of genetic association. We anticipate that this book will be useful to researchers who want to estimate heterogeneity in their data, develop or apply genetic association statistics where heterogeneity exists, and accurately evaluate statistical power and sample size for genetic association through the application of robust experimental design.

Next Generation Sequencing Based Diagnostic Approaches in Clinical Oncology

Transcriptional Regulation: Molecules, Involved Mechanisms and Misregulation