Advances in Experimental Medicine and Biology 939

Bairong Shen
Haixu Tang
Xiaoqian Jiang *Editors*

Translational Biomedical Informatics

A Precision Medicine Perspective

Springer

Advances in Experimental Medicine and Biology

Volume 939

More information about this series at http://www.springer.com/series/5584

Bairong Shen • Haixu Tang • Xiaoqian Jiang
Editors

Translational Biomedical Informatics

A Precision Medicine Perspective

Springer

Editors
Bairong Shen
Center for Systems Biology
Soochow University
Jiangsu, China

Haixu Tang
School of Informatics and Computing
Indiana University
Bloomington, IA, USA

Xiaoqian Jiang
Department of Biomedical Informatics
University of California San Diego
La Jolla, CA, USA

ISSN 0065-2598 ISSN 2214-8019 (electronic)
Advances in Experimental Medicine and Biology
ISBN 978-981-10-1502-1 ISBN 978-981-10-1503-8 (eBook)
DOI 10.1007/978-981-10-1503-8

Library of Congress Control Number: 2016957788

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Science+Business Media Singapore Pte Ltd.

Contents

Chapter 1
NGS for Sequence Variants

Shaolei Teng

Abstract Recent technological advances in next-generation sequencing (NGS) provide unprecedented power to sequence personal genomes, characterize genomic landscapes, and detect a large number of sequence variants. The discovery of disease-causing variants in patients' genomes has dramatically changed our perspective on precision medicine. This chapter provides an overview of sequence variant detection and analysis in NGS study. We outline the general methods for identifying different types of sequence variants from NGS data. We summarize the common approaches for analyzing and visualizing casual variants associated with complex diseases on precision medicine informatics.

Keywords Sequence variants • Next-generation sequencing • Sequence alignment • Variant calling • Association testing • Visualization • Precision medicine informatics

1.1 Introduction

Over the last decade, next-generation sequencing (NGS) has dramatically changed the precision medicine field by characterizing patients' genomic landscapes and identifying the casual variants associated with human diseases. The Sanger-based sequencing [48] ("first-generation sequencing") was used to sequence the first human reference genome for the Human Genome Project [3], which took 13 years to finish the draft genome at a total cost of $3 billion. NGS technologies make the sequencing at remarkable price and unprecedented speed by carrying out hundreds of millions of sequencing reactions at once [52, 57]. With the revolutionary technology, we can sequence thousands of genomes in just 1 month, address the biological questions at a large scale, identify the genetic risk factors for human diseases, and provide a more precise way to health care [24]. In particular, NGS can be used to detect a large number of sequence variants in the patients' genomes and identify the casual variants associated with human diseases, which has dramatically

S. Teng (✉)
Department of Biology, Howard University, Washington, DC 20059, USA
e-mail: shaolei.teng@howard.edu

B. Shen et al. (eds.), *Translational Biomedical Informatics*, Advances in Experimental Medicine and Biology 939, DOI 10.1007/978-981-10-1503-8_1

changed our perspective on genetic variants, human diseases, and precision medicine.

Discovery of casual sequence variants associated with certain traits or diseases has become a fundamental aim of genetics and biomedical research. The sequence variants can be classified to single nucleotide variants (SNVs), small insertions and deletions (INDELs), and large structural variants (SVs) based on their sequences in length. SNVs, the most common type of sequence variants, are single DNA base-pair differences in individuals. INDELs are defined as small DNA polymorphisms including both insertions and deletions ranging from 1 to 50 bp in length. SVs are large genomic alterations (>50 bp) including unbalanced variants (deletions, insertions, or duplications) and balanced changes (translocations and inversions). Copy number variants (CNVs), a large category of unbalanced SVs, are DNA alterations that result in the abnormal number of copies of particular DNA segments. Somatic mutations are tumor-specific variants in cancer–normal sample pairs. The different types of sequence variants play important roles in the development of human complex diseases. For example, the SNVs associated with major depression were found in the genes encoding serotonin transporter, serotonin receptor, catechol-o-methyltransferase, tryptophan hydroxylase, and tyrosine hydroxylase [29]. These sequence variants can influence the neurotransmitter functions in multiple ways including changing gene expression level, altering substrate binding affinity, or affecting transport kinetics [19]. A balanced t(1;11) (q42.1;q14.3) translocation in disrupted in schizophrenia 1 (*DISC1*) gene was discovered in a large Scottish family highly burdened for severe mental illnesses, and the family members with the translocation showed a reduced P300 event-related potential associated with schizophrenia [9]. Identifying the casual variants and their clinical effects provides important insight to understand the roles of sequence variants in the causation of human diseases.

Discovery of disease-causing variants from a large number of sequence polymorphisms detected from NGS data is a major challenge in precision medicine. Bioinformatics and statistical methods have been developed for detecting sequence variants and identifying disease-related casual variants. The schematic diagram of NGS variant analysis on precision medicine informatics is shown in Fig. 1.1. The DNA samples are extracted from patients (or normal individuals) and sequenced on NGS platforms. The billions of short sequence reads are produced by the sequencers, and sequence information is stored in FASTQ format files. From here, NGS variant analysis falls into two major frameworks. The first framework is the variant detection. The high-quality sequence reads passed quality control (QC) filters are aligned to a reference genome, and the sequence alignment data is deposited in SAM/BAM format files. Several variant detection tools are used to call small variants including SNVs and INDELs. The somatic mutation callers are applied to tumor–normal patient samples. Multiple SV callers are developed to detect large structural variants. The variants called from these tools can be stored in Variant Call Format (VCF) files or BED format files. The next framework is the variant analysis. The annotation tools are used to predict the functional effects of coding and regulatory variants. The association analysis can identify the common

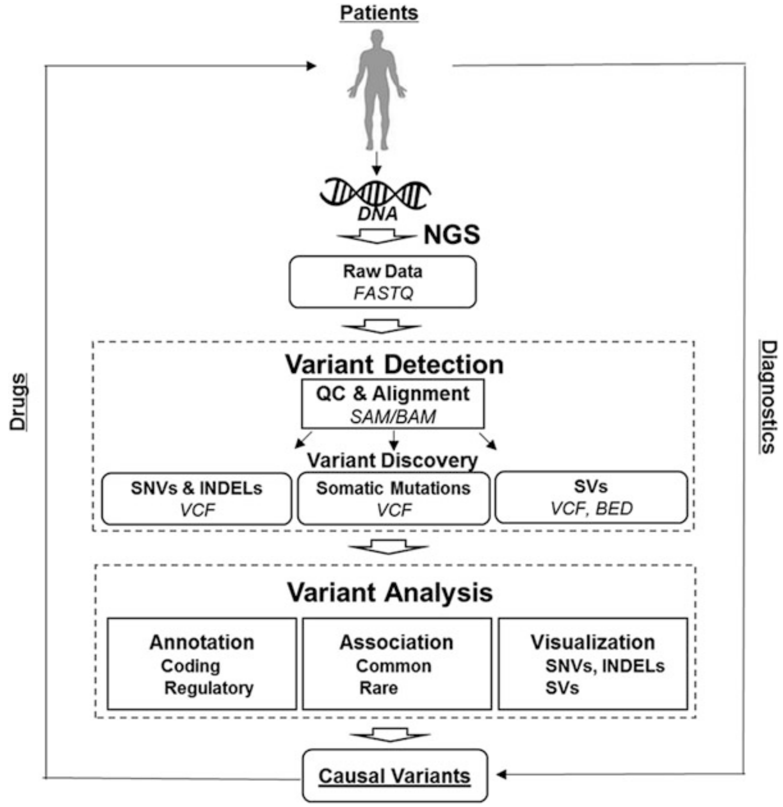

Fig. 1.1 A flowchart of NGS variant analysis in precision medicine

and rare variants associated with certain diseases or traits. Visualization tools are used to view the small and large candidate sequence variants. By combining numerous analyzing tools, the causal variants can be identified and connected with clinical information for precision medicine research. On the one hand, disease-related causal variants provide the genetic biomarkers for diagnostics of complex diseases. On the other hand, the candidate variants offer the targets for developing more precise treatments and drugs for patients. In the following sections, we will review the bioinformatics approaches and provide a guide for detecting and analyzing the sequence variants from NGS data.

1.2 Variant Detection

Variant detection consists of quality control (QC), sequence alignment, and variant calling. The raw data contains a large number of short reads generated by NGS sequencers. Preprocessing and post-processing QC are carried out to remove the potential artifacts and bias from data. The high-quality reads are mapped to positions on a reference genome. The variant calling is performed by comparing the aligned reads with known reference sequences to find which segments are different with the reference genomes. Multiple variant callers have been developed to detect different types of genetic variants including SNVs, INDELs, somatic mutations, and SVs. This section provides an overview on QC and alignment methods, SNV and INDEL callers, somatic mutation tools, and SV detection approaches.

1.2.1 QC and Alignment

The standard outputs of most NGS platforms are files in FASTQ format. The FASTQ files include raw sequence reads together with their Phred-scaled base quality scores. Several tools have been developed to perform preprocessing QC based on FASTQ files (Table 1.1). FastQC [7] provides a comprehensive QC report

Table 1.1 Variant quality control (QC) and alignment tools

Tool	Description	URL	Reference
Preprocessing QC			
FastQC	Tool can provide statistical QC summary report	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/	[7]
Sickle	QC tool can trim low-quality bases	https://github.com/najoshi/sickle	[21]
Trimmomatic	QC tool can remove adaptor and bases	http://www.usadellab.org/cms/?page=trimmomatic	[10]
Hash table alignment			
MAQ	Hashing read aligner that allows two mismatches	http://maq.sourceforge.net/	[32]
SeqMap	Hashing read aligner that allows five mismatches	http://www-personal.umich.edu/~jianghui/seqmap/	[20]
SOAP	Hashing reference aligner	http://soap.genomics.org.cn/	[33]
Suffix tree alignment using Burrows–Wheeler transformation (BWT)			
BWA	BWT aligner using a backward search	http://bio-bwa.sourceforge.net/bwa.shtml	[30]
Bowtie	BWT aligner using a backtracking system	http://bowtie-bio.sourceforge.net	[26]
SOAP2	BWT and hash aligner	http://soap.genomics.org.cn/	[35]
Post-processing QC			
SAMtools	Tool can convert, sort, and index SAM/BAM files	http://samtools.sourceforge.net/	[34]
BamTools	Tool can manage BAM files and filter reads	https://github.com/pezmaster31/bamtools	[8]
Picard	QC tool can remove PCR duplicates	http://broadinstitute.github.io/picard/	

with tables and plots for basic statistics, quality score distribution, read length distribution, sequence duplication levels, and GC content distribution. With the FastQC summary, other QC tools, such as Sickle [21] and Trimmomatic [10], can be used to filter low-quality reads, remove undesired adaptors, and trim incorrectly called bases at the ends of the reads.

The sequence alignment is an essential step for identifying the sequence variants in patients' genomes. Since any errors in alignment will be carried through to the downstream analysis, each of the high-quality sequence reads must be accurately aligned to a reference genome. With the rapid development of NGS technologies, a wide variety of alignment tools have been developed to align the short sequence reads with high efficiency and accuracy (Table 1.1). Most of NGS aligners build indices for reads or references to quickly search potential alignment positions of reads in the reference sequences. Based on the property of the index, these tools can be briefly classified into two groups: hash table approaches or suffix tree approaches [31].

Hash Table Approaches These use a hash-based index to scan either read set or reference genome for rapid searching. Each position of reference is cut into equal-sized fragments and stored into a hash table. The species cut from the read with the same scheme are used as the keys to search the table. The approaches use a seed-and-extend paradigm to identify the matching positions in the reference for the reads. Here, we describe three common hash table tools: MAQ [32], SeqMap [20], and SOAP [33]. MAQ [32] can rapidly align a large number of short reads to the reference sequence and accurately detect small sequence variants including SNVs and INDELs. For sequence alignment, MAQ indexes and hashes the short reads before scanning reference sequence, which allows two mismatches in the first 28 bp of each read. It then searches ungapped match and extends the partial match when a seed match is identified. MAQ utilizes a Phred-scaled mapping quality score to evaluate the reliability of alignments, and the score can measure the probability that a true positive read is not the one found by the mapping algorithm. SeqMap [20] also applied an index filtering algorithm to create index tables for sequence reads. It allows up to five mismatches including substitutions and insertions/deletions. Instead of the construction of hash tables for reads that used in MAQ and SeqMap, SOAP [33] loads the reference genome into memory and constructs index tables for all references sequences. It utilizes a seed strategy for both ungapped and gapped alignments of either single read or paired-end reads.

Suffix Tree Approaches These use Burrows–Wheeler transformation (BWT) [11] to store all suffixes of a string. The reference genome can be converted to a transformed memory-efficient sequence using BWT. Reads are aligned base by base against the transformed reference sequence. The strategy can reduce memory footprint and increase mapping speed. Examples of BWT-based tools include BWA [30], Bowtie [26], and SOAP2 [35]. Burrows–Wheeler Alignment (BWA) tool [30] is the most commonly used NGS aligner. It uses backward search with BWT for exact matching and constructs inexact alignments supported by the exact matches. Bowtie [26] utilizes a novel backtracking system to account mismatches and allows up to two mismatches in the first 28 bp of sequence read. BWA and Bowtie compare the query reads and store the reference to short substrings. The tools compute all combinations of possible mismatches to align the entire reads to reference exactly. SOAP2 [35], an updated version of SOAP, uses BWT to index the reference genome in memory and constructs a hash table to search the location of a read in the reference index. The suffix tree methods run faster than hash table approaches due to the memory efficiency of BWT sequence. The indices of the entire human genome generated by BWT approaches are usually less than 2 GB, whereas the hash table approaches require more than 50 GB.

The sequence alignments are stored in SAM/BAM files [34]. Sequence Alignment/Map (SAM) file contains the read alignment data, and BAM file is the binary version of SAM file. SAMtools [34] can be used to convert SAM/BAM format and sort, index, and merge the alignment files. BamTools [8] can manage the BAM files and filter properly reads with high mapping quality. Picard can be used to remove PCR duplicates caused by the sorting from merged alignment files. These post-processing QC tools generate clean aligned sequencing files suitable for further variant detection.

1.2.2 SNV and INDEL Discovery

After mapping the short reads to a reference sequence, the variants can be discovered by comparing the sample genome to the reference genome. Many variant callers have been developed to detect small variants including SNVs and INDELs (Table 1.2). These computational tools use either heuristic or probabilistic approaches. Since probabilistic approaches can estimate sequencing error and monitor the accuracy of calling, they are more generally used for variant calling [40]. We introduce three probabilistic callers MAQ [32], SAMtools [34], and GATK [38] below.

MAQ [32] is the first widely used tool for variant calling in NGS data. It uses a Bayesian statistical model to generate consensus genotype sequence from the alignments. MAQ compares the consensus sequence to the reference genome to identify potential SNVs and filtered them using some predefined rules. SAMtools [34] uses a revised MAQ model to measure statistical uncertainty of called genotypes and applies given likelihood for each possible genotype. It uses a subset of commands, called BCFtools, to call SNVs and INDELs. The true small variants can be filtered by base alignment quality scores computed from the depth of coverage, numbers of reads in alternate and reference alleles, average quality scores, and mapping quality of reads.

Genome Analysis Toolkit (GATK) [38] is the most frequently used toolkit for small variant calling. It provides a structured Java programming MapReduce framework for NGS analysis. The GATK package includes coverage analyzer, local realigner, quality score recalibrator, and variant caller. It inputs SAM/BAM

Table 1.2 Variant discovery tools

Tool	Description	URL	Reference
SNV and INDEL discovery			
MAQ	Tool can detect small variants using a Bayesian statistical model	http://maq.sourceforge.net/	[32]
SAMtools	Tool can detect genotypes and small variants using a revised MAQ model	http://samtools.sourceforge.net/	[34]
GATK	Package including coverage analyzer, local realigner, quality score recalibrator, and variant caller	https://www.broadinstitute.org/gatk/	[38]
Somatic mutation discovery			
VarScan2	Caller can detect somatic mutations using Fisher's exact test	http://varscan.sourceforge.net/	[23]
Strelka	Caller can detect somatic mutations using Bayesian probability model	https://sites.google.com/site/strelkasomaticvariantcaller/	[49]
SomaticSniper	Caller can compute Phred-scaled scores to detect somatic mutations using Bayesian probability model	http://gmt.genome.wustl.edu/packages/somatic-sniper/	[27]
JointSNVMix	Caller can detect somatic mutations using two Bayesian probability-based models	http://compbio.bccrc.ca/software/jointsnvmix/	[47]
Structural variant discovery			
CNVnator	Read-depth caller can detect deletions and duplications	http://sv.gersteinlab.org/cnvnator/	[2]
BreakDancer	Read-pair caller can detect insertion, deletions, inversions, and translocations	http://breakdancer.sourceforge.net/	[12]
Pindel	Split-read caller can detect large deletions and medium insertions	https://github.com/genome/pindel	[61]
CONTRA	Read-depth caller can detect CNVs from exome sequencing data	https://sourceforge.net/projects/contra-cnv/	[36]
XHMM	Read-depth caller can detect CNVs using hidden Markov model from exome sequencing data	https://atgu.mgh.harvard.edu/xhmm/	[18]

files from initial read mapping. Then, the tool carries out a local INDEL realignment and computes base quality scores for recalibration. A program, called UnifiedGenotyper, is used to identify all potential SNVs and INDELs. GATK applies machine learning approaches to filter true variants from machine artifacts in NGS technologies [56]. GATK recently developed a HaplotypeCaller [16] program which performs a local de novo assembly of aligned reads and calls SNVs and INDELs simultaneously. It provides a greater quality for INDELs calling than UnifiedGenotyper program [42]. In addition, HaplotypeCaller can handle the non-diploid samples and work well for the region including different types of sequence variants close to each other. The outputs of the most variant callers are

Variant Call Format (VCF) files. The VCF is used for storing sequence variants such as SNVs, INDELs, as well as SVs. The genetic variant information in VCF files includes variant positions, unique identifiers, reference and alternate alleles, quality scores, filters, annotations, and genotypes.

1.2.3 Somatic Mutation Discovery

Discovery of somatic mutations associated with oncogenesis is essential for identifying appropriate treatments for cancer patients. Several callers have been developed for detecting the somatic mutations that present in tumor cells but not in normal tissue (Table 1.2). VarScan2 [23] screens the genotypes that are above certain coverage and quality thresholds from the cancer and normal samples, respectively. The variant calls with minimum variant frequency of all reads greater than 20 % are classified as either heterozygous calls or homozygous calls. For each position with genotypes that do not match in tumor and normal, VarScan2 uses a one-tailed Fisher's exact test to check significant difference of allele frequency across samples. The somatic mutations are called if the normal samples are homozygous reference or heterozygous as loss of heterozygosity, but the calls in tumor samples do not match.

Several tools based on Bayesian probability model have been developed for the discovery of somatic mutations in matched cancer–normal pairs. Strelka [49] carries out a realignment around INDELs in the tumor and normal sequence alignment files like GATK. It uses a Bayesian probability approach to model the normal sample allele frequencies as diploid genotypes and tumor sample allele frequencies as a mixture of the normal sample with somatic variation. The approaches also apply some priors for strand bias, mapping qualities, somatic mutation rates, and estimated heterozygosity rates of the normal sample. SomaticSniper [27] uses a Bayesian probability model to compute the probability of all possible combined genotypes for the cancer–normal pair samples. The likelihood is given by the observed as well as prior information from the rates of population mutation, sequencing error, and somatic mutation. Each variant call in tumor samples is assigned a Phred-scaled score indicating the probability that the cancer and normal genotypes are different. JointSNVMix [47] utilizes a different Bayesian method with a mixed binomial model to call each variant in the tumor and normal samples. It analyzes the allelic count in paired cancer–normal samples using two probabilistic graphical models: JointSNVMix1 that assumes the base calls and read numbers and follows a perfect binomial distribution and JointSNVMix2 that weighs priors for base call and mapping quality.

1.2.4 Structural Variant Discovery

Structural variants (SVs) are widespread in human genomes and play important roles in the development of human diseases. As the growing number of SVs has been demonstrated to have clinical relevance, SV discovery is critical in precision medicine and cancer genomics. NGS technologies have revolutionized SV studies. Compared to traditional hybridization-based approaches such as array CGH and SNP microarrays, sequencing-based bioinformatics methods can detect multiple types of SVs at a wide size range [5]. Most of these approaches distinguish SVs based on two read mapping signatures including depth of coverage and paired-end mapping [39]. The first type of approaches searches the regions with abnormal read counts; the second type of tools investigates the configurations of the paired-end mappings [60]. In this section, we describe the computational approaches (Table 1.2) based on the two signatures below.

Depth of Coverage The approaches assume that read mapping follows a Poisson distribution and the divergence from this distribution indicates the SV signatures. The duplication has more reads mapping to region, and deletions show significantly reduced coverage. CNVnator [2] can detect the deletions and duplications using a statistical analysis of read mapping density for single-end and paired-end reads. It captures the read-depth signatures by dividing sequencing regions into equal-sized bins and computing the counts of reads in each bin. The partitioning of the signatures is based on a mean-shift approach with additional filters such as GC-bias correction. The statistical significance test is used to identify the regions with abnormal signals for detecting possible deletions or duplications. The read-depth approaches can predict the absolute copy numbers of genomic segments. However, they cannot detect the balanced SVs such as translocations and inversions.

Paired-End Mapping The approaches can be classified into two types of strategy: read pair and split read. Read-pair methods analyze the span and orientation of paired-end reads and identify the read pairs that are mapped with discordant separation distances or orientation. Read-pair approaches can detect all classes of SVs. BreakDancer [12] can detect read pairs with mapping span and orientation that are inconsistent with the control. It has two models: BreakDancerMax can identify five types of SVs including insertion, deletions, inversions, and intrachromosomal and interchromosomal translocations, while BreakDancerMini is used to detect INDELs. Split-read approaches are used to search split-read signatures to identify the breakpoints of SVs. The deletions and duplications can be identified from the continuous stretch of gaps in the sequence reads or references, respectively. Split-read methods are suitable for long reads, but some algorithms can use short reads to identify the breakpoints of large SVs. For example, Pindel [61] uses a pattern growth algorithm to find large deletions and medium insertions from short paired-end reads. The algorithm can align the gapped short sequences to reference sequences with local alignment, which can reduce memory and increase speed for searching potential split reads.

Structure variant discovery from targeted or whole-exome sequencing data is very challenging due to the noncontiguous reads in exons. The targeted sequencing results in some biases in sample collection, targeted genomic hybridization, and GC content. Multiple tools have been developed to overcome these biases. CONTRA [36] is a read-depth tool for CNV discovery. It uses BAM/SAM alignments as inputs and builds an average baseline across multiple samples as the control. CONTRA then computes the base-level log-ratios with corrections for imbalanced library size bias and GC content bias. It calculates two-tailed P-values to detect CNVs. XHMM [18] applies principal component analysis to normalize read depth in targets. It uses hidden Markov model (HMM) to detect CNVs across multiple samples (>50 samples). In addition to VCF files, Browser Extensible Data (BED) format files can be used to store and display large structural variants for further analysis.

1.3 Variant Analysis

Causal variant discovery is the key step in precision medicine informatics. Identifying the disease-related variants promises to dramatically expand current aspects of biomedical research in disease diagnostics and drug design. Multiple bioinformatics tools have been developed to distinguish the causal variants associated with human diseases from the massive number of nonfunctional variants detected by NGS variant callers. Annotation methods determine the possible functional impact of all identified variants. Association analyses connect the variants with complex diseases or clinical traits. Visualization tools provide the graphic views of identified causal variants. The disease-related casual variants can be identified by combining these approaches and stored in public variant databases such as ClinVar [25] and HGMD [54]. The Human Variome Project (http://www.humanvariomeproject.org/) has curated the gene-/disease-specific databases to collect the sequence variants and genes associated with diseases. In this section, we summarize the variant analysis approaches for identifying the most promising causal variants underlying human diseases.

1.3.1 Variant Annotation

Variant annotation can be used to determine the effects of sequence variants on genes and proteins and filter the functional important variants from a background of neutral polymorphisms. Coding mutations, such as nonsynonymous SNVs, could change amino acid sequences and affect protein structures and functions. They are more likely to be involved in the development of diseases. Regulatory variants located in noncoding regions could modulate the gene expressions and work as the causative modifiers of human diseases. Here, we describe the common computational tools for predicting the effects of coding mutations and regulatory variants. We also introduce the generally used annotation toolkits to access the prediction results generated from these tools.

Damaging Nonsynonymous Mutation Prediction With the advent of NGS technologies, particularly of exome sequencing, there is a significant need to interpret the coding variants. A number of tools have been developed to distinguish deleterious mutations from a large number of harmless nonsynonymous polymorphisms. Sorting Intolerant From Tolerant (SIFT) [41] is a commonly used method for predicting the effects of coding mutations on protein function. The algorithm assumes that important protein sites should be conserved throughout evolution and mutations located in these sites could alter protein functions. SIFT searches the target sequence in protein database and constructs the sequence alignments using closely related sequences. It computes the degree of conservation of protein residues to distinguish the deleterious and neutral coding mutations. Polymorphism Phenotyping v2 (PolyPhen2) [4] is another popular tool for predicting deleterious missense mutations. The PolyPhen2 prediction is based on sequence annotations, structural attributes, and comparative evolutionary considerations. PolyPhen2 uses an iterative greedy algorithm to extract sequence-based and structure-based features. Then, it constructs the supervised machine learning classifiers to predict missense variants as benign, possibly damaging, or probably damaging mutations. PolyPhen2 uses two data sets (HumDiv and HumVar) for training. HumDiv data set collects all damaging mutations associated with human Mendelian diseases from UniProtKB and non-damaging mutations between the proteins and their closely related mammalian homologs. HumDiv model can be used to analyze rare variants mildly deleterious at functionally important regions such as the regions involved in complex phenotypes or identified from genome-wide association studies (GWAS). HumVar data set uses all disease-causing mutations from UniProtKB as positive data and the common sequence variants not involved in disease as negative instances. HumVar model can be used to identify the damaging mutations with significant effects for Mendelian disease research. Other common in silico programs include likelihood ratio test (LRT) [13], which identifies the damaging mutations that disrupt significantly conserved amino acid positions within the human proteome, and MutationTaster [51] which evaluates the deleterious sequence variants using a naive Bayesian model constructed from features including splice-site alterations, mRNA changes, loss of protein, and evolutionary conservation.

Regulatory Variant Effect Prediction The majority of disease-related variant hits identified from GWAS fall in noncoding DNA region, which indicate the regulatory variants located in noncoding regions are critical in human disease. Regulatory variants play important roles in gene expression and protein modification. Several bioinformatics tools have been developed for predicting the functional effects of regulatory variants. Genome-wide annotation of variants (GWAVA) [45] uses a random forest algorithm to construct three classifiers to distinguish the functional sequence variants in regulatory regions from a background of neutral variants. The classifiers integrate genomic features such as evolutionary conservation and GC content and range of epigenomic annotations from the Encyclopedia of DNA

Elements (ENCODE) project [15]. Combined Annotation Dependent Depletion (CADD) [22] is a score that can be used to prioritize the functional variants including coding variants and regulatory variants. CADD tool constructs support vector machine classifiers to integrate various genomic and epigenomic annotations into a single measure (C score) for each sequence variant. Recently, deep learning algorithm has been applied for interpretation of regulatory variants. DeepSEA [62] is a deep learning-based tool for predicting the effects of noncoding variant and prioritizing regulatory variants. The software uses deep learning algorithms to learn regulatory sequence code from large-scale chromatin-profiling data and predict the effects of noncoding variants on chromatin accessibility such as DNase I sensitivities, transcription factor binding, and histone marks at regulatory elements.

General Variant Annotation Multiple annotation toolkits have been developed to determine the impacts of sequence variants on genes and proteins and access their functional effects from above predictors. ANNOVAR [58] is a command-line Perl software for annotating SNVs and INDELs based on genes, regions, or filters. In gene-based annotation, it can annotate whether sequence variants affect protein amino acid sequences (nonsense, missense, splice site, etc.). In region-based annotation, it can identify the variants located in ENCODE-annotated regions such as transcribed regions, enhancer regions, DNase I hypersensitivity sites, transcription factor binding site, and transcription factor ChIP-Seq data. In filter-based annotation, ANNOVAR can extract the information (allele frequency and identifier) of a sequence variant in public databases such as dbSNP [53], ClinVar [25], 1000 Genomes Project [1], and Exome Variant Server (http://evs.gs.washington.edu/EVS/). In addition, it can be used to access the annotations from damaging mutation predictors (SIFT, PolyPhen2, LRT, MutationTaster, etc.) for nonsynonymous mutations and CADD for regulatory variants. SnpEff [14] is another popular annotation package to estimate the functional effects of SNVs, INDELs, and multiple nucleotide polymorphisms. Based on the functional impacts of the sequence variants, SnpEff classifies the variants to four classes: high, moderate, low, and modifier. It also provides the annotations for regulatory variants. SnpEff provides a summary HTML page to display overall statistics for sequences and variants (Table 1.3).

1.3.2 Variant Association Testing

Understanding how genetic variants contribute to diseases is the key challenge in precision medicine. There are two hypotheses for interpreting the genetic contribution of sequence variants in complex diseases such as cancers and mental disorders [50]. The "common disease–common variant" hypothesis states that a few common variants, usually defined as the allele frequency greater than 1 % in the population, make the major contributions for the genetic variance in complex disease susceptibility. In contrast, the "common disease–rare variant" hypothesis argues that multiple risk variants, each of which has low frequency (e.g., allele frequency less than 1 %) in the population, are the major contributors to the genetic

Table 1.3 Variant annotation tools

Tool	Description	URL	Reference
Damaging nonsynonymous mutation prediction			
SIFT	Tool can predict deleterious and neutral mutations based on sequence homology	http://sift.jcvi.org/	[41]
PolyPhen2	Tool can predict probably damaging, possibly damaging, and benign mutations based on sequence and structure features	http://genetics.bwh.harvard.edu/pph2/	[4]
LRT	Tool can predict deleterious, neutral, or unknown mutations using likelihood ratio test	http://www.genetics.wustl.edu/jflab/lrt_query.html	[13]
MutationTaster	Tool can predict disease-causing and polymorphism mutations using naive Bayesian model	http://www.mutationtaster.org/	[51]
Regulatory variant effect prediction			
GWAVA	Tool can predict the regulatory variant effects using random forest algorithm	https://www.sanger.ac.uk/sanger/StatGen_Gwava	[45]
CADD	Tool can predict the effects of coding and noncoding variants using support vector machine algorithm	http://cadd.gs.washington.edu/	[22]
DeepSEA	Tool can predict the regulatory variant effects using deep learning algorithm	http://deepsea.princeton.edu/	[62]
General variant annotation			
ANNOVAR	Perl annotation toolkit based on genes, regions, and filters	http://annovar.openbioinformatics.org/	[58]
SnpEff	Java annotation package based on genes	http://snpeff.sourceforge.net/	[14]

susceptibility to complex diseases. NGS technologies can detect the full spectrum of sequence variants including the rare variants that are difficult to be captured by traditional genotyping arrays. Here, we describe the generally used case–control association approaches for common and rare variants.

Case–Control Data QC The first step in any case-control association analysis is the data quality control [6]. The samples and variants with poor quality should be removed to reduce the numbers of false-positive and false-negative associations. The samples with outlying heterozygosity rates, high missing data rates, and discordant sex information have poor quality and should be removed firstly. In addition, the related samples or samples from divergent ancestry should not be used for case-control analysis. If the variants showed a high rate of missing genotypes, departure from Hardy–Weinberg equilibrium, or a different missing genotype rate between cases and controls, these variants should be excluded from case-control analysis.

Common-Variant Association Analysis The genome-wide association study (GWAS) is a generally used approach to identify the common variants associated with complex diseases and traits. The common methods used in GWAS are carried out based on a single-variant level. The variants are tested individually, and multiple testing correction should be used to control the family-wise error rate (FWER). PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/) is the most commonly used software package for GWAS analysis in large-scale studies [43]. It provides numerous useful tools for genetic data management, data quality control, and association tests. Multiple association tests implemented in PLINK can be used to identify the common variants associated with diseases based on their minor allele frequencies between cases and controls. Fisher's exact test can be used for case-control traits in small-sized samples; permutation methods should be applied to control for FWER. Linear regression test can be utilized for complex quantitative traits, and permutation approaches should be performed to generate empirical P-values to avoid issues with the test statistic distribution caused by the combination of variants and traits that deviate greatly from normality. Another popular association testing tool is PLINK/SEQ (https://atgu.mgh.harvard.edu/plinkseq/). The toolset performs Fisher's exact test for single-variant association, on the contrary, based on the alternate allele frequencies of variants in cases and controls.

Rare-Variant Association Analysis GWAS research has identified many common variants strongly implicated in complex diseases. However, most of the common variants have modest effects on the disease risk and much of the genetic contribution to complex diseases remains unexplained [37]. Recent sequencing studies revealed the rare genetic variants have large effects on the risk for complex diseases such as schizophrenia [44]. The rare-variant association tests are usually carried out on a gene, or gene set level due to single-variant analysis is underpowered for rare variants unless the sample sizes are very large. The general rare-variant burden test collapses the rare variants across all samples into a single variable and compares the cumulative effects in cases with controls within a gene to evaluate the significance of the difference. The sequence kernel association test (SKAT, https://cran.r-project.org/web/packages/SKAT/) is particularly designed for the rare-variant analysis from NGS data [59]. It uses a kernel machine regression approach to aggregate the associations between variants in a gene region and a continuous or dichotomous trait. SKAT-O [28] test applies a unified test to search the optimal linear combination of the general burden test and SKAT test to maintain the power in both scenarios. In addition, the SKAT package provides "SKAT_CommonRare" function to evaluate the combined effects of rare and common variants. The permutation method can be used in rare-variant association analysis to control FWER.

1.3.3 Variant Visualization

Visualizing the individual genomes and causal variants based on the existing knowledge provides critical supports for biomedical research. Various standalone visualization tools have been developed for interactive exploration of NGS data from public resources and researchers' own studies. Integrative Genomics Viewer

(IGV) is a high-performance tool that provides a rapid visualization for large genomic data sets [46]. IGV tool (https://www.broadinstitute.org/igv/) can load sequence alignment BAM files, annotation data, and reference genomes from local computers or remote sites. IGV includes tools for data tiling and file format sorting and indexing [55]. It provides both stand-alone GUI desktop version and command-line scripts for generating different image snapshots. IGV is capable of displaying various sequence variants including SNVs and INDELs. As shown in Fig. 1.2, IGV provides the views of chromosome ideogram, genomic coordinates, coverage plot, sequence reads, and gene annotation tracks. The base mismatches compared to reference are highlighted with color bars in coverage plot or color bases in read tracks. Figure 1.2a shows an intronic SNV (rs3812384) in Src-like-adaptor

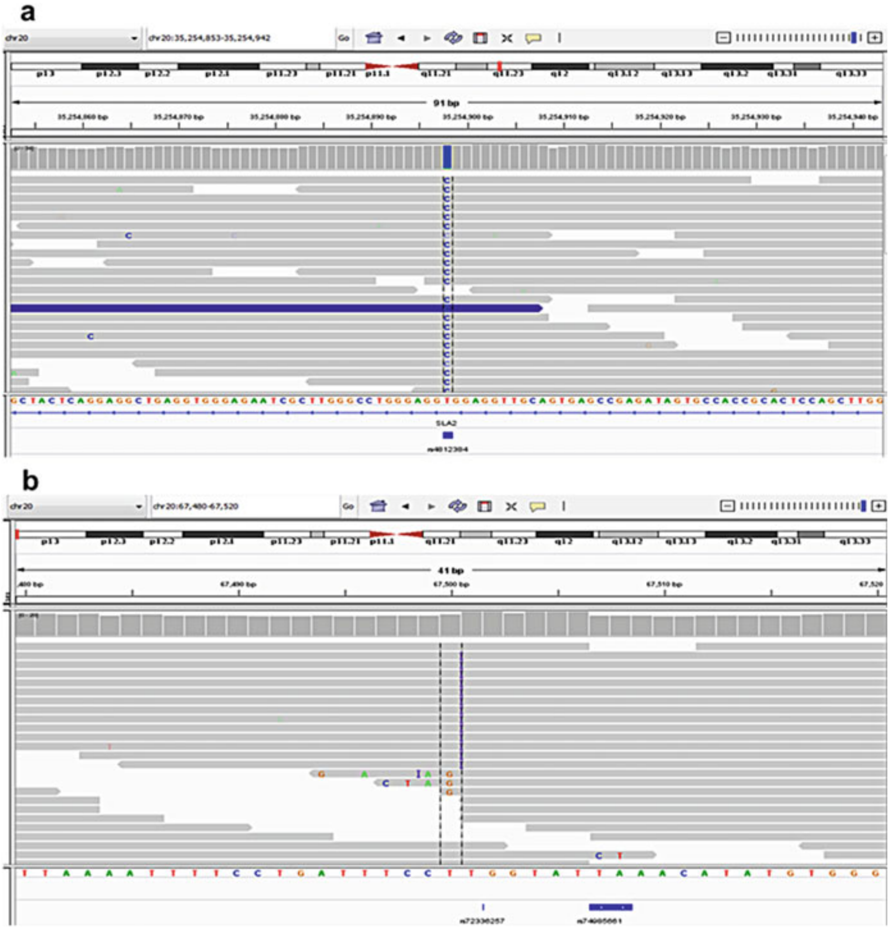

Fig. 1.2 IGV views of small sequence variants. Snapshots of IGV showing (**a**) an intronic SNV (rs3812384) in *SLA2* gene and (**b**) an insertion (rs72336257) in *DEFB125* gene

2 (*SLA2*) gene, and Fig. 1.2b displays an insertion (rs72336257) in defensin beta 125 (*DEFB125*) gene.

Several visualization tools have been developed for displaying the large SVs from NGS data. Sequence Annotation, Visualization, and ANalysis Tool (Savant) is a viewer for analyzing and visualizing sequence reads and variants [17]. Savant browser (http://genomesavant.com) uses a modular docking framework to show each module in a separate window. It provides the track module, bookmark module, and table view to analyze the NGS data. In particular, Savant provides multiple visualization modes to view and compare SVs in different samples (Fig. 1.3). For example, a 150 kb duplication presenting in case but not in control shows a higher coverage in zinc finger and AT-hook domain containing (*ZFAT*) gene in the

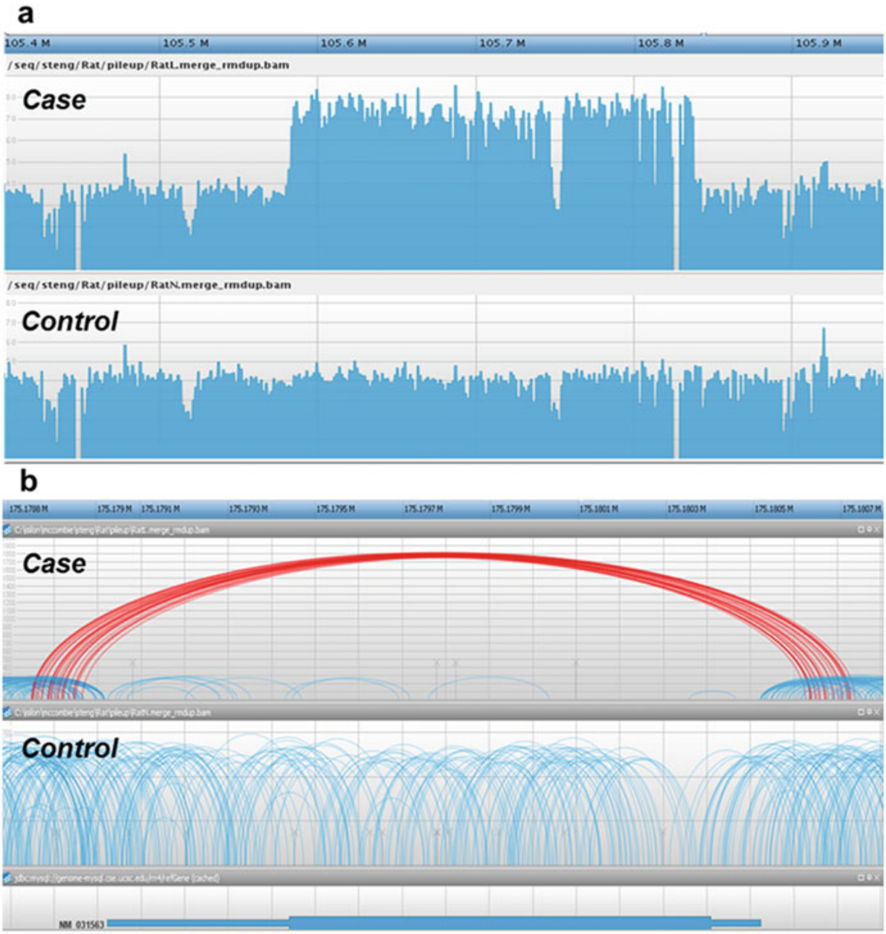

Fig. 1.3 Savant views of large SVs. Plots of Savant displaying the large SVs that present in case but not in control including (**a**) a 150 kb duplication in *ZFAT* gene and (**b**) a 1.5 kb deletion in *YBX1* gene

coverage mode (Fig. 1.3a). The Matepair (arc) mode displays the relative distance between paired-end reads. The red and taller arcs indicate a larger distance between the two reads of a pair, suggesting a 1.5 kb deletion in Y box binding protein 1 (*YBX1*) gene is only carried by the case sample (Fig. 1.3b).

1.4 Conclusions

NGS has significantly benefited the discovery of disease-related sequence variants, which greatly facilitated the improvement of diagnosis and treatment methods in precision medicine. Computational approaches have been developed to detect different types of sequence variants (SNVs, INDELs, somatic mutations, and structural variants) from NGS data. Bioinformatics methods have been applied to annotate, filter, and visualize the casual variants associated with complex diseases. There are limit standards regarding best practices in NGS variant detection and analysis. To meet the challenges in precision medicine, some international scientific organizations, such as Human Variome Project, are developing standardized workflows for analyzing sequence variants implicated in human diseases.

References

1. Abecasis GR, Altshuler D, Auton A, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73. doi:10.1038/nature09534.
2. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21:974–84. doi:10.1101/gr.114876.110.
3. Adekoya E, Ait-Zahra M, Allen N, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
4. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9. doi:10.1038/nmeth0410-248.
5. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12:363–76. doi:10.1038/nrg2958.
6. Anderson CA, Pettersson FH, Clarke GM, et al. Data quality control in genetic case–control association studies. Nat Protoc. 2010;5:1564–73. doi:10.1038/nprot.2010.116.
7. Andrews S. FastQC: a quality control tool for high throughput sequence data. babraham Bioinforma 1. 2010. doi: citeulike-article-id:11583827.
8. Barnett DW, Garrison EK, Quinlan AR, et al. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011;27:1691–2. doi:10.1093/bioinformatics/btr174.
9. Blackwood DH, Fordyce A, Walker MT, et al. Schizophrenia and affective disorders-cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: clinical and P300 findings in a family. Am J Hum Genet. 2001;69:428–33.
10. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20. doi:10.1093/bioinformatics/btu170.
11. Burrows M, Wheeler D. A block-sorting lossless data compression algorithm. Algorithm, Data Compression 18. 1994. doi: 10.1.1.37.6774.

12. Chen K, Wallis JW, McLellan MD, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6:677–81. doi:10.1038/nmeth.1363.
13. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19:1553–61. doi:10.1101/gr.092619.109.
14. Cingolani P, Platts A, le Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6:80–92. doi:10.4161/fly.19695.
15. Consortium TEP. The ENCODE (ENCyclopedia Of DNA Elements) project. Science. 2004;306:636–40. doi:10.1126/science.1105136.
16. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8. doi:10.1038/ng.806.
17. Fiume M, Williams V, Brook A, Brudno M. Savant: genome browser for high-throughput sequencing data. Bioinformatics. 2010;26:1938–44. doi:10.1093/bioinformatics/btq332.
18. Fromer M, Moran JL, Chambert K, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet. 2012;91:597–607. doi:10.1016/j.ajhg.2012.08.005.
19. Hahn MK, Blakely RD. Monoamine transporter gene structure and polymorphisms in relation to psychiatric and other complex disorders. Pharmacogenomics J. 2002;2:217–35. doi:10.1038/sj.tpj.6500106.
20. Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics. 2008;24:2395–6. doi:10.1093/bioinformatics/btn429.
21. Joshi N, Fass J. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. 2011. Available at https://github.com/najoshi/sickle2011.
22. Kircher M, Witten DM, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5. doi:10.1038/ng.2892.
23. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568–76. doi:10.1101/gr.129684.111.
24. Koboldt DC, Steinberg KM, Larson DE, et al. The next-generation sequencing revolution and its impact on genomics. Cell. 2013;155:27–38.
25. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42:D980–5. doi:10.1093/nar/gkt1113.
26. Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25. doi:10.1186/gb-2009-10-3-r25.
27. Larson DE, Harris CC, Chen K, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012;28:311–7. doi:10.1093/bioinformatics/btr665.
28. Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case–control whole-exome sequencing studies. Am J Hum Genet. 2012;91:224–37. doi:10.1016/j.ajhg.2012.06.007.
29. Levinson DF. The genetics of depression: a review. Biol Psychiatry. 2006;60:84–92. doi:10.1016/j.biopsych.2005.08.024.
30. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60. doi:10.1093/bioinformatics/btp324.
31. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. 2010;11:473–83. doi:10.1093/bib/bbq015.
32. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18:1851–8. doi:10.1101/gr.078212.108.
33. Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics. 2008;24:713–4. doi:10.1093/bioinformatics/btn025.

34. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9. doi:10.1093/bioinformatics/btp352.
35. Li R, Yu C, Li Y, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25:1966–7. doi:10.1093/bioinformatics/btp336.
36. Li J, Lupat R, Amarasinghe KC, et al. CONTRA: copy number analysis for targeted resequencing. Bioinformatics. 2012;28:1307–13. doi:10.1093/bioinformatics/bts146.
37. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53. doi:10.1038/nature08494.
38. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303. doi:10.1101/gr.107524.110.
39. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods. 2009;6:S13–20. doi:10.1038/nmeth.1374.
40. Mielczarek M, Szyda J. Review of alignment and SNP calling algorithms for next-generation sequencing data. J Appl Genet. 2015;57:1–9. doi:10.1007/s13353-015-0292-7.
41. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31:3812–4.
42. Pirooznia M, Kramer M, Parla J, et al. Validation and assessment of variant calling pipelines for next-generation sequencing. Hum Genomics. 2014;8:14. doi:10.1186/1479-7364-8-14.
43. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75. doi:10.1086/519795.
44. Purcell SM, Moran JL, Fromer M, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014;506:185–90. doi:10.1038/nature12975.
45. Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11:294–6. doi:10.1038/nmeth.2832.
46. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6. doi:10.1038/nbt.1754.
47. Roth A, Ding J, Morin R, et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics. 2012;28:907–13. doi:10.1093/bioinformatics/bts053.
48. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74:5463–7.
49. Saunders CT, Wong WSW, Swamy S, et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28:1811–7. doi:10.1093/bioinformatics/bts271.
50. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev. 2009;19:212–9. doi:10.1016/j.gde.2009.04.010.
51. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7:575–6. doi:10.1038/nmeth0810-575.
52. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26:1135–45.
53. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308–11.
54. Stenson PD, Mort M, Ball EV, et al. The human gene mutation database: 2008 update. Genome Med. 2009;1:13. doi:10.1186/gm13.
55. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–92. doi:10.1093/bib/bbs017.
56. Van der Auwera G a., Carneiro MO, Hartl C, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr Protoc Bioinfor. 2013;43:11.10.1–33. doi:10.1002/0471250953.bi1110s43.

57. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30:418–26.
58. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38, e164. doi:10.1093/nar/gkq603.
59. Wu MC, Lee S, Cai T, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89:82–93. doi:10.1016/j.ajhg.2011.05.029.
60. Xi R, Kim T-M, Park PJ. Detecting structural variations in the human genome using next generation sequencing. Brief Funct Genomics. 2010;9:405–15. doi:10.1093/bfgp/elq025.
61. Ye K, Schulz MH, Long Q, et al. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25:2865–71. doi:10.1093/bioinformatics/btp394.
62. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12:931–4. doi:10.1038/nmeth.3547.

Chapter 2
RNA Bioinformatics for Precision Medicine

Jiajia Chen and Bairong Shen

Abstract The high-throughput transcriptomic data generated by deep sequencing technologies urgently require bioinformatics methods for proper data visualization, analysis, storage, and interpretation. The involvement of noncoding RNAs in human diseases highlights their potential as biomarkers and therapeutic targets to facilitate the precision medicine. In this chapter, we give a brief overview of the bioinformatics tools to analyze different aspects of RNAs, in particular ncRNAs. We first describe the emerging bioinformatics methods for RNA identification, structure modeling, functional annotation, and network inference. This is followed by an introduction of potential usefulness of ncRNAs as diagnostic, prognostic biomarkers and therapeutic strategies.

Keywords RNA • Precision medicine • Biomarkers • Bioinformatics • Cancer

2.1 Introduction

RNAs are polymeric molecules that carry genetic information and are implicated in protein synthesis. Recently, it is discovered that only a minor fraction of human genomes encode for proteins [10, 15], and the remaining large fraction of the transcripts are known as noncoding RNAs (ncRNAs).

ncRNAs could be broadly grouped into distinct classes based on the lengths. Some classes have been known for long, e.g., ribosomal RNAs, transport RNAs, and small nucleolar RNAs (snoRNAs). Several novel ncRNA classes have been discovered, e.g., microRNAs (miRNAs), small interfering RNAs (siRNAs), PIWI-interacting RNAs (piRNAs), hairpin RNAs (hpRNAs), and long noncoding RNAs (lncRNAs) [31, 36, 51, 56]. The repertoire of ncRNAs continues to expand.

J. Chen
School of Chemistry, Biology and Material Engineering, Suzhou University of Science and Technology, No1. Kerui road, Suzhou, Jiangsu 215011, China

B. Shen (✉)
Center for Systems Biology, Soochow University, No1. Shizi Street, 206, Suzhou, Jiangsu 215006, China
e-mail: bairong.shen@suda.edu.cn

B. Shen et al. (eds.), *Translational Biomedical Informatics*, Advances in Experimental Medicine and Biology 939, DOI 10.1007/978-981-10-1503-8_2

Each of the ncRNA class has its unique biogenesis routes, three-dimensional structure, and modes of action. Therefore, the functional spectrum of ncRNAs is richer than expected. RNAs function as critical players at the epigenetic, transcriptional, and posttranscriptional level [40]. They regulate gene expression through diverse mechanisms, e.g., by mediating imprinting [57], alternative splicing [61], and modification of other small noncoding RNAs.

The huge amount of biological data generated by high-throughput sequencing technologies have opened new possibilities for RNA research field. On the other hand, these data in turn give rise to an urgent demand for proper visualization, analysis, storage, and interpretation of the data. It's imperative to find efficient bioinformatics methods to utilize these rich data source for a better understanding of the RNA world.

Clinical transcriptomics will have great impact on the therapeutic strategy of human disorders. The involvement of the RNA in human diseases has widely been investigated for microRNAs. MicroRNAs, however, are just the tip of the iceberg. The heterogeneous family of long noncoding RNAs (lncRNAs) may also participate in the progression of human diseases.

In the year of 2015, US government invested $215 million to launch the Precision Medicine Initiative (PMI). Precision medicine (PM) is a customized healthcare model. This model takes into account the individual variability and tailors the medical treatment to the individual patient. Precision medicine is dependent on molecular diagnostics or other molecular and cellular analyses to select appropriate therapies based on the context of a patient's genetic content [45].

Bioinformatics analyses integrating high-throughput microarray, next-generation sequencing (NGS), and genotyping data in disease cases and matched healthy controls consistently reveal changes in gene expression of both protein-coding and regulatory noncoding RNAs. The strong correlations between deregulated lncRNAs and disease development and prognosis highlight the potential of ncRNA as biomarkers and therapeutic targets to facilitate the precision medicine.

In this chapter, we give a brief overview of the bioinformatics tool to analyze several different aspects of RNAs, in particular ncRNAs. We first describe the emerging bioinformatics methods for RNA identification, structure modeling, functional annotation, and network inference. This is followed by an introduction of potential usefulness of ncRNAs as diagnostic, prognostic biomarkers and therapeutic strategies.

2.2 RNA Detection

RNA-seq is now a rich source to discover new ncRNA transcripts. It is also an interesting topic to further validate any novel transcripts discovered.

2.2.1 Detection of Small ncRNAs

Most of the currently available methods or algorithms that investigate NGS-generated transcriptomic data aim at detecting, predicting, and quantifying small RNAs, in particular microRNAs.

miRDeep [19] is the first stand-alone tool that identifies both novel and known microRNAs from large-scale sRNA-seq data. miRDeep evaluates the possibility of a hairpin structure by using Bayesian probability controls during the processes of microRNA biogenesis. miRDeep successfully identifies new miRNA candidates by searching for the characteristic read profile covering the mature miRNA and its complement miRNA*. In miRDeep2, the well-conserved structure of ncRNA families is used as a supplementary step to confirm a novel or known ncRNA [49]. miRDeep* [1] employs a miRNA precursor prediction algorithm to minimize the putative range of the precursor loci. It outperforms both versions of miRDeep by reducing false negatives.

In addition to the microRNA-specific tools, other approaches or pipelines generally concentrate on the whole family of small RNAs.

The web service DARIO is a comprehensive approach designed to predict and analyze different types of small RNAs generated from any next-generation RNA sequencing experiments [17]. The web server CPSS [67] makes further improvement on DARIO. CPSS is able to analyze small RNA-seq data that originate from single or paired samples. CPSS also predicts target genes for interested miRNAs and performs functional annotation of the predicted target genes. Different from CPSS, ncPRO-seq [7] is an integrated pipeline that identifies regions significantly enriched with short reads which do not belong to any known ncRNA families, thus allowing the identification of novel ncRNA- or siRNA-producing regions.

2.2.2 Detection of lncRNAs

Compared with messenger RNAs or small ncRNAs, detection of long ncRNAs from assembled transcripts is more complicated, because more deep sequencing reads are needed for lncRNAs to achieve sufficient coverage.

The main problem in the lncRNA detection is to reconstruct the transcripts from the short sequencing reads. In order to meet the ever-increasing need for effective NGS data mining, a variety of bioinformatics tools have been developed for NGS-based RNA transcriptome investigation. Below we will give a brief introduction on the currently available tools.

Sun et al. [59] developed a computational pipeline lncRScan to predict novel lncRNA from transcriptome sequencing data. lncRScan uses expression level as a filter to preclude artifacts or mRNAs from the initially assembled candidate transcripts. iSeeRNA [60] trained a support vector machine-based classifier using conservation, open reading frame, and sequence information as features. iSeeRNA