NGS is projected to play an increasingly important role in covering novel genomic features with base-pair resolution, biomarkers. Because of its high-throughput capability in dissequencing (NGS) is a promising technology to identify omic profiles,7 alternative splicing patterns,8 and miRNA regulatory networks9 and (2) ChIP-seq, which has enabled researchers to examine the epigenetic mechanisms of CVD by profiling the genome-wide pattern of protein-binding regions (eg, transcription factors and enhancers) or histone modifications.10,11

Although NGS provides a great opportunity for discovering potential CVD biomarkers, a few NGS bioinformatics challenges remain: first, the overwhelming NGS data volume requires huge data storage and computational resources. Second, at each data analysis step, multiple bioinformatics tools are publicly available, which makes it challenging to assemble sensible NGS bioinformatics pipelines. Third, 1 single NGS experiment processed by various NGS bioinformatics pipelines can result in a wide variety of information with different biological and clinical interpretation and validation. The rest of this article is organized as follows: In the NGS Bioinformatics for CVD section, we will review the challenges and progress of bioinformatics in NGS data analysis with a focus on RNA-seq and ChIP-seq. In the Case Studies: RNA-seq and ChIP-seq Bioinformatics for CVD section, we will present 2 CVD case studies to illustrate the applications of NGS bioinformatics. Finally, in the Opportunities of NGS for CVD section, we will summarize future opportunities for CVD research using NGS.

NGS Bioinformatics for CVD

Illumina, Life Technologies/Ion Torrent, and Roche/454 are examples of commercially available NGS platforms. Among these, Illumina is the most prevalent platform that can produce millions of relatively short, fixed-length sequence reads in a single experiment. In this section, we will describe the NGS bioinformatics pipelines tailored to handle Illumina data. As shown in Figure 2, we will review bioinformatics methods such as sequence mapping, expression quantification, expression normalization, and DEG detection for RNA-seq data and sequence mapping and peak calling for ChIP-seq data.
Sequence Mapping

The first step of bioinformatics pipelines for both RNA-seq and ChIP-seq is sequence mapping. It determines the genomic or transcriptomic origin of sequence reads (or reads in short). Sequence mapping using brute-force strategy requires large central processing unit and memory resources, where mapping millions of reads to the 3 billion base pairs of the human genome is extremely time-consuming. Thus, the research of sequence mapping largely focuses on improving computational efficiency while maintaining high mapping accuracy.

Table 1 lists some mapping tools with their mapping strategies and key features. Depending on biological applications and computational resources, mapping algorithms can provide 3 types of alignments: (1) ungapped alignment (e.g., Bowtie)\(^\text{12}\) allows only mismatches between query reads and the reference genome to keep the computational cost low. However, for some applications (e.g., mapping of RNA-seq data to the human genome), ungapped alignment may fail to align a large number of reads. (2) Gapped alignment (e.g., BFAST\(^\text{13}\), Bowtie2\(^\text{14}\), BW A\(^\text{15}\), Novoalign, SHRiMP2\(^\text{16}\), SOAPaligner\(^\text{17}\), and SSAHA2\(^\text{18}\)) allows mismatches, insertions, and deletions. Most gapped alignment tools implement the Smith–Waterman\(^\text{19}\) or Needleman–Wunsch\(^\text{20}\) algorithms. (3) Spliced alignment (e.g., GSNAP,\(^\text{21}\) TopHat,\(^\text{22}\) MapSplice,\(^\text{23}\) OSA,\(^\text{24}\) and SOAPsplice\(^\text{25}\)) allows the long extension of gaps within the query reads. Biologically, such long gaps may represent intronic regions or interchromosomal splitting. Algorithmically, spliced alignment may be achieved by segmenting query reads into smaller sequences (e.g., 25 base pairs), mapping these smaller sequences, and then assembling mapped results for each read into a consensus result. Spliced alignment algorithms are often computationally more expensive than unspliced algorithms. However, spliced mapping is necessary for applications that focus on identifying novel splice junctions using RNA-seq. Unspliced mapping, including gapped and ungapped, is sufficient for ChIP-seq data analysis.

About mapping accuracy, it may be affected by the reporting strategy of mapping tools. Uniquely mapped reads provide more definite information than multi-mapped reads. If a query read is mapped to multiple genomic loci because of insufficient read length, the ambiguous mapping happens and a mapping tool may randomly report 1 optimal mapping.
Table 1. Summary of RNA Sequencing and Chromatin Immunoprecipitation Sequencing Sequence-Mapping Tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Mapping Strategy and Usage</th>
<th>Algorithmic Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFAST</td>
<td>Hash table, Smith–Waterman local alignment</td>
<td></td>
</tr>
<tr>
<td>Bowtie</td>
<td>Burrows–Wheeler transform and FM-index</td>
<td></td>
</tr>
<tr>
<td>Bowtie2</td>
<td>Burrows–Wheeler transform, FM-index–assisted seed alignment, dynamic programming</td>
<td></td>
</tr>
<tr>
<td>BWA</td>
<td>Unspliced mapping to transcriptome or genome</td>
<td>Burrows–Wheeler transform</td>
</tr>
<tr>
<td>Novoalign</td>
<td>Commercial software, algorithm unpublished</td>
<td></td>
</tr>
<tr>
<td>SHRiMP2</td>
<td>Multiple spaced-seed indexing, Smith–Waterman local alignment</td>
<td></td>
</tr>
<tr>
<td>SOAPaligner</td>
<td>Bidirectional Burrows–Wheeler transform</td>
<td></td>
</tr>
<tr>
<td>SSAHA2</td>
<td>Hash table</td>
<td></td>
</tr>
<tr>
<td>GSNAP</td>
<td>Spliced mapping to genome and unspliced mapping to transcriptome or genome</td>
<td>Minimal sampling strategy, oligomer chaining for approximate alignment, sandwich dynamic programming</td>
</tr>
<tr>
<td>MapSplice</td>
<td>Uses Bowtie for alignment, segmented mapping</td>
<td></td>
</tr>
<tr>
<td>OSA</td>
<td>Two-stage transcriptome and genome alignment, segmented mapping</td>
<td></td>
</tr>
<tr>
<td>SOAPsplice</td>
<td>Spliced mapping to genome</td>
<td>Burrows–Wheeler transform, segmented mapping</td>
</tr>
<tr>
<td>TopHat</td>
<td>Uses Bowtie or Bowtie2 for alignment, segmented mapping</td>
<td></td>
</tr>
</tbody>
</table>

BFAST indicates BLAT-like fast accurate search tool; BWA, Burrows–Wheeler aligner; GSNAP, genomic short-read nucleotide alignment program; FM, full-text index in minute space; OSA, Omicsoft sequence aligner; SHRiMP2, short read mapping package, version 2; SOAPaligner, short oligonucleotide analysis package aligner; SOAPsplice, short oligonucleotide analysis package for splice junction detection; and SSAHA2, sequence search and alignment by hashing algorithm, version 2.

randomly selected from all optimal mappings or report all optimal mappings. However, multi-mapped reads may benefit the downstream quantification algorithms in model training and expression estimation.

To improve the computational efficiency of sequence mapping, auxiliary data structures can be used to reduce the similarity search space such as to index either the reference genome or the query reads using hash tables (BFAST, GSNAP, SHRiMP2, and SSAHA2 are representatives) or to index the reference genome using the Burrows–Wheeler transform with suffix/prefix arrays (Bowtie, Bowtie2, BWA, and SOAPaligner are representatives).26

Expression Quantification

The second step of the RNA-seq bioinformatics pipeline is expression quantification of genes, transcripts, or other functional small RNAs. Because a read may map to multiple genomic loci, the accuracy of gene or transcript expression estimation depends on the ability of the quantification algorithm to resolve the ambiguities from the sequence-mapping step. In addition, a gene may have multiple alternatively spliced isoforms sharing a common set of exons, where a read mapped to the shared exons may belong to any one of the isoforms. Currently, the handling of these ambiguities involves building a probabilistic framework and then estimating gene/transcript expression using either the expectation-maximization algorithm or Bayesian inference.27–29

Quantification algorithms can be categorized into 3 groups: count-based, linear model-based, and Poisson model-based.30 Table 2 lists common RNA-seq quantification tools, categorized in terms of the model, the estimation algorithm, and quantifiable targets. Count-based quantifiers (eg, ERANGE,31 HTSeq,32 NEUMA,33 and ALEXA-Seq)34 assign each read to its mapped location with a probability of one. Each quantifier implements a proprietary filtering criterion, and the expression profile is the accumulated read count on each targeted gene or transcript. Linear model-based quantifiers (eg, rQuant35 and IsoInfers)36 assume that read counts are normally distributed, and least squares can be applied to infer expression estimates. Poisson model-based quantifiers (eg, RSEM,27 Cufflinks,28 MISO,29 and IsoEM)37 probabilistically assign multi-mapped reads based on the assumption that reads from genomic loci follow the Poisson distribution. Because count-based quantifiers do not rely on a predefined model, they usually have lower computational complexity than the other model-based quantifiers. However, the expression estimates of count-based quantifiers might deviate from the truth because of the naïve way in which multi-mapped reads are handled.

Expression Normalization

The third step of the RNA-seq bioinformatics pipeline is normalization. Because of variations introduced in sequencing and bioinformatics processes, inter-sample comparison of RNA-seq expression estimates can only be done after normalization. Most normalization methods for RNA-seq are based on scaling, in which the gene or transcript expression of any biological sample is normalized by multiplying or dividing by a fixed scaling factor. Therefore, the fundamental challenge for RNA-seq expression normalization is to estimate a set of robust scaling factors for samples in the data set. Table 3 lists commonly used RNA-seq normalization methods.

Several naïve methods such as RPM/FPM, median normalization, and upper-quartile normalization are mathematically similar. RPM/FPM adjusts expression estimates of each sample by the total number of mapped reads/fragments in the sample. Median and upper-quartile normalizations use the median and upper-quartile read/fragment counts, respectively, of each sample as the substitute for the total mapped reads/fragments. With the Illumina sequencing protocol, longer genes or transcripts tend to produce a larger number of sequence fragments. Thus, some methods such as RPKM/FPKM11 and TPM2 further adjust expression estimates by gene or transcript length, which in turn...
enables both inter- and intra-sample comparisons. However, there exist limitations for the aforementioned normalization methods. For the RPKM/FPKM, gene or transcript length cannot be precisely defined. In addition, some methods such as RPM/FPM and RPKM/FPKM are sensitive to data sets that have a small number of extreme DEGs. Therefore, several methods such as TMM\textsuperscript{50} and RLE\textsuperscript{40} assume that most genes are not differentially expressed and use robust estimates of library size as the scaling factors. Dillies et al\textsuperscript{41} systematically evaluated a few normalization methods and recommended TMM or RLE as the most robust method for most RNA-seq data.

### DEG Detection

The fourth key step in the identification of potential CVD biomarkers from RNA-seq data is the detection of DEGs between 2 groups of samples. RNA-seq expression estimates from the 2 groups are first fit to a statistical distribution, followed by statistical hypothesis testing to determine whether the statistical distributions between the 2 groups are significantly different for a targeted gene. Soneson et al\textsuperscript{42} and Rapaport et al\textsuperscript{43} have conducted comprehensive quantitative evaluations for DEG detection methods. Thus, this section will mainly focus on qualitative categorization.

DEG detection methods can be nonparametric or parametric. Nonparametric methods such as SAMseq\textsuperscript{44} and NOISeq\textsuperscript{45} use resampling and counting techniques to avoid making assumptions about the underlying distribution of RNA-seq expression estimates. Data permutation is a common technique for estimating false discovery rates for nonparametric methods.

Parametric methods use Poisson-based models to fit RNA-seq read count data.\textsuperscript{46} The Poisson distribution has the variance equal to the mean. However, overdispersion (ie, when the variance is significantly greater than the mean) often occurs in RNA-seq data. Therefore, the negative binomial distribution, which is a 2-parameter extension of the Poisson distribution, introduces an additional parameter to capture the high variability.\textsuperscript{47} Selecting an appropriate statistical model is the key for a parametric DEG detection method. For example, DESeq\textsuperscript{47} applies the Poisson distribution to model RNA-seq read count data; edgeR,\textsuperscript{48} baySeq,\textsuperscript{49} and DESeq\textsuperscript{40} use the negative binomial model to capture the overdispersion; Myrna\textsuperscript{50} models the data as either the Gaussian distribution or the negative binomial model to capture both overdispersion and uncertainty in the fragment count of a transcript. After constructing the statistical model, most parametric methods assess the significance level of each gene using either \( P \) values computed from the likelihood ratio test or the Fisher’s exact test or posterior probabilities estimated from the empirical Bayes method, whereas Cuffdiff\textsuperscript{25} assumes that RNA-seq data is normally

<table>
<thead>
<tr>
<th>Normalization Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>Scaling by median of all counts</td>
</tr>
<tr>
<td>Quantile</td>
<td>Matching distributions of counts</td>
</tr>
<tr>
<td>RLE</td>
<td>Scaling by median ratio to median library</td>
</tr>
<tr>
<td>RPKM/FPKM</td>
<td>Scaling by library size and gene/transcript length</td>
</tr>
<tr>
<td>RPM/FPFM</td>
<td>Scaling by library size</td>
</tr>
<tr>
<td>TMM</td>
<td>Scaling by estimate of relative RNA production</td>
</tr>
<tr>
<td>TPM</td>
<td>Scaling by mean length of expressed genes/transcripts</td>
</tr>
</tbody>
</table>

#### Upper quartile

<table>
<thead>
<tr>
<th>Normalization Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper quartile</td>
<td>Scaling by upper quartile of all counts</td>
</tr>
</tbody>
</table>

**ALEXA-Seq** indicates alternative expression analysis by sequencing; EM, expectation maximization; ERANGE, enhanced read analysis of gene expression; HTSeq, analyzing high-throughput sequencing data with Python; IsoEM, isoform quantification by expectation maximization; IsoInfer, inference of isoforms from short sequence reads; MISO, mixture of isoforms; NEUMA, normalization by expected uniquely mappable area; RNA-seq, RNA sequencing; rQuant, transcript quantification with RNA-seq data; and RSEM, RNA-seq by expectation maximization.

### Table 2. Summary of RNA Sequencing Expression Quantification Tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Mathematical Model</th>
<th>Estimation</th>
<th>Gene/Isiform</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALEXA-Seq</td>
<td></td>
<td>Average coverage of mapped reads</td>
<td>Yes/yes</td>
</tr>
<tr>
<td>ERANGE</td>
<td></td>
<td>Accumulated counts, read assigns</td>
<td>Yes/no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>proportionally to expression level</td>
<td></td>
</tr>
<tr>
<td>HTSeq</td>
<td>Count-based model</td>
<td>Accumulated counts, read assigns</td>
<td>Yes/no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with probability 1</td>
<td></td>
</tr>
<tr>
<td>NEUMA</td>
<td></td>
<td>Accumulated counts of informative</td>
<td>Yes/yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>reads</td>
<td></td>
</tr>
<tr>
<td>IsoInfer</td>
<td></td>
<td>Maximum likelihood estimation from</td>
<td>Yes/yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>convex quadratic programming</td>
<td></td>
</tr>
<tr>
<td>rQuant</td>
<td>Linear model</td>
<td>Minimize read coverage deviation</td>
<td>Yes/yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with quadratic programming</td>
<td></td>
</tr>
<tr>
<td>Cufflinks</td>
<td></td>
<td>Maximize likelihood with the maximum</td>
<td>Yes/yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a posteriori estimates using</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bayesian inference</td>
<td></td>
</tr>
<tr>
<td>IsoEM</td>
<td>Poisson model</td>
<td>Maximum likelihood estimation with</td>
<td>Yes/yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EM algorithm</td>
<td></td>
</tr>
<tr>
<td>MISO</td>
<td></td>
<td>Posterior mean estimates using</td>
<td>Yes/yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bayesian inference</td>
<td></td>
</tr>
<tr>
<td>RSEM</td>
<td></td>
<td>Maximum likelihood estimation with</td>
<td>Yes/yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EM algorithm</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3. Summary of RNA Sequencing Expression Normalization Methods

<table>
<thead>
<tr>
<th>Normalization Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>Scaling by median of all counts</td>
</tr>
<tr>
<td>Quantile</td>
<td>Matching distributions of counts</td>
</tr>
<tr>
<td>RLE</td>
<td>Scaling by median ratio to median library</td>
</tr>
<tr>
<td>RPKM/FPKM</td>
<td>Scaling by library size and gene/transcript length</td>
</tr>
<tr>
<td>RPM/FPFM</td>
<td>Scaling by library size</td>
</tr>
<tr>
<td>TMM</td>
<td>Scaling by estimate of relative RNA production</td>
</tr>
<tr>
<td>TPM</td>
<td>Scaling by mean length of expressed genes/transcripts</td>
</tr>
<tr>
<td>Upper quartile</td>
<td>Scaling by upper quartile of all counts</td>
</tr>
</tbody>
</table>

**ALEXA-Seq** indicates alternative expression analysis by sequencing; EM, expectation maximization; ERANGE, enhanced read analysis of gene expression; HTSeq, analyzing high-throughput sequencing data with Python; IsoEM, isoform quantification by expectation maximization; IsoInfer, inference of isoforms from short sequence reads; MISO, mixture of isoforms; NEUMA, normalization by expected uniquely mappable area; RNA-seq, RNA sequencing; rQuant, transcript quantification with RNA-seq data; and RSEM, RNA-seq by expectation maximization.
distributed after a particular transformation and uses the \(t\)-test to determine the statistical significance of each DEG.

**Peak Calling**

For the ChIP-seq bioinformatics pipeline, the second step is peak calling. Sequence reads generated from ChIP-seq mostly originate from DNA sequences around targeted protein-binding regions. After mapping reads to the reference genome and identifying uniquely mapped reads, genomic loci that accumulate a large number of reads (ie, peaks) indicate putative protein-binding regions. Peak-calling tools distinguish true peaks from background noise by (1) generating a signal profile along each chromosome, (2) defining a background noise model, (3) identifying candidate peak locations, and (4) assessing the significance of each candidate peak.\(^{52}\) Peak-calling tools in earlier time quantify fold enrichment between samples of interest and expected background and then apply the Poisson model to assess the significance of the enriched regions.\(^{52}\) Recently developed peak-calling tools use the strand-dependent bimodality information and adopt a more realistic background model to capture local variations.\(^{53}\)

**Case Studies: RNA-Seq and ChIP-Seq Bioinformatics for CVD**

In this section, we will demonstrate the use of NGS bioinformatics for cardiovascular research by applying RNA-seq and ChIP-seq pipelines to publicly available CVD RNA-seq and ChIP-seq data sets downloaded from the National Center for Biotechnology Information Sequence Read Archive repository (Figure 2). In each case study, we will illustrate the NGS bioinformatics solution for CVD research, evaluate the performance of the critical step that identifies CVD biomarkers, and discuss the remaining bioinformatics challenges.

**Cardiovascular Transcriptomics Using RNA-Seq**

**Data Set**

The RNA-seq data set (Sequence Read Archive accession: SRP009662) was acquired to investigate the effects of \(Ezh2\) deletion on postnatal cardiac development, homeostasis, and gene expression.\(^{54}\) The authors reported that the loss of \(Ezh2\) gene in cardiac precursors would lead to cardiac hypertrophy and fibrosis. This data set contains wild-type and \(Ezh2\)-deficient adult mouse right ventricle samples, each with 2 biological replicates. Each sample was sequenced with the Illumina HiSeq 2000 platform and contains \(\approx 30\) million \(2 \times 50\) base pair read pairs.

**Bioinformatics Pipeline and Performance Evaluation Metrics**

The bioinformatics pipeline for identifying DEGs using RNA-seq data includes sequence mapping, expression quantification, expression normalization, and DEG detection (Figure 2). We use the same sequence mapper, TopHat,\(^{22}\) and expression quantifier, Cufflinks,\(^{25}\) with 8 different DEG detection tools to construct 8 pipelines. Each DEG detection tool implements an expression normalization method that optimizes its DEG detection performance. TopHat maps the 4 RNA-seq samples to the GRCm38/mm10 mouse genome\(^{45}\) with the guidance of the RefSeq genome annotation.\(^{45}\) Cufflinks quantifies gene/transcript expression in terms of raw read counts. The 8 DEG detection tools (Figure 2, left table) include both nonparametric (eg, SAMseq\(^{44}\) and NOISeq\(^{45}\)) and parametric methods (eg, baySeq,\(^{49}\) Cuffdiff\(^{2}\),\(^{51}\) DESeq2,\(^{40}\) dispersion shrinkage for sequencing (DSS)),\(^{57}\) edgeR,\(^{40}\) and Limma+Voom\(^{48}\) that represent a wide variety of prevalent and novel algorithms. The significant DEGs are identified with adjusted \(P\) values \(<0.05\). To evaluate the performance of the 8 pipelines, we have designed 5 metrics:

1. The authors of the original paper\(^{44}\) performed quantitative reverse transcription polymerase chain reaction validation for 16 cardiac hypertrophy- or fibrosis-related genes. Among 16 genes, only 12 were reported to be significantly differentially expressed. To assess the power of RNA-seq, we use the concordance of these 12 DEGs between quantitative reverse transcription polymerase chain reaction and 8 RNA-seq pipelines as the first metric. For genes detected by \(<4\) of the 8 pipelines, we further investigated those expression patterns to understand the cause.
2. To assess the biological relevance of DEGs, we use the ToppFun web-based tool in the ToppGene Suite\(^{30}\) to annotate the functions of DEGs in terms of significant gene ontology (GO) terms and pathways with adjusted \(P\) values \(<0.05\). The 114 GO terms and 4 pathways associated with the 16 cardiac hypertrophy- or fibrosis-related genes are defined as the ground-truth functional annotation (Figure 3B). We use the concordance between the pipeline-specific annotations and the ground truth as the second metric.
3. To assess the reproducibility among various DEG detection tools, we compute the number of overlapping DEGs among the 8 tools as the third metric.
4. To assess the expression profiles of DEGs, as defined in Equation 1, we use the distribution of the ratio of the dominant read count (ie, the largest read count of a DEG across all samples) to the total read count for any DEG \(g\) as the fourth metric.

\[
R_{\text{dominance, } g} = \frac{\max(A_{g, 1}, A_{g, 2}, B_{g, 1}, B_{g, 2})}{\sum(A_{g, 1}, A_{g, 2}, B_{g, 1}, B_{g, 2})} \tag{1}
\]

where \(A_{g, 1}\) and \(B_{g, 2}\) are normalized read counts after adjusting the sequencing depth effect for samples \(A_{g, 1}, A_{g, 2}, B_{g, 1}\), and \(B_{g, 2}\) for any DEG \(g\). \([A_{g, 1}, A_{g, 2}]\) and \([B_{g, 1}, B_{g, 2}]\) are biological replicates for the wild-type and \(Ezh2\)-deficient samples, respectively. Given Equation 1, the range of \(R_{\text{dominance, } g}\) is from 25% (ie, \(A_{g, 1}=A_{g, 2}=B_{g, 1}=B_{g, 2}\)) to 100% (ie, \(\max=\sum\)) with a few possible scenarios: (1) if a gene is not significantly differentially expressed and the variability between replicates is small, the normalized read counts \(A_{g, 1}\) and \(B_{g, 2}\) will only differ slightly from one another, and the \(R_{\text{dominance, } g}\) will be close to 25%; (2) if a gene is highly differentially expressed and the variability between replicates is small, \(R_{\text{dominance, } g}\) will be \(\approx 50\%\); and (3) if the variability between replicates is large, \(R_{\text{dominance, } g}\) can be significantly \(>50\%\) (eg, \(R_{\text{dominance, } g}=80\%\) if \([A_{g, 1}=A_{g, 2}=B_{g, 1}=B_{g, 2}]=[120, 30, 0, 0]\)).
5. To assess the capability of each tool for detecting highly expressed and low-expressed DEGs, we calculate the mean read count of each DEG \(g\) from the normalized
Figure 3. Biological interpretation and quantitative assessment of various differentially expressed gene (DEG) detection tools using RNA sequencing data. **A**, Checkmarks indicated concordant DEGs between quantitative reverse transcription polymerase chain reaction (qRT-PCR) validation conducted by the original study and the results from our 8 DEG detection tools. Genes marked in red were identified by <4 of the 8 tools. **B**, The top 20 significant gene ontology (GO) terms and all 4 significant pathways of the ground-truth functional annotation, which was established by annotating the 16 qRT-PCR validated genes. **C**, The number of concordant GO terms and pathways between ground-truth and pipeline-specific annotations. **D**, The number of DEGs supported by 1, 2, or all 8 tools for each DEG detection pipeline. **E**, The distribution of the ratio of dominant read counts to total read counts of all DEGs for each DEG detection pipeline. **F**, The distribution of the mean read counts of all DEGs for each DEG detection pipeline. baySeq indicates empirical Bayesian analysis of patterns of differential expression in count data; DESeq2, differential gene expression analysis based on the negative binomial distribution, version 2; DSS, dispersion shrinkage for sequencing data; edgeR, empirical analysis of digital gene expression data in R; Limma+Voom, linear models for microarray data followed by variance modelling at the observational level; and SAMseq, significance analysis of sequencing data.
read counts (ie, $A_{1,g}$, $A_{2,g}$, $B_{1,g}$, and $B_{2,g}$) and use its distribution as the fifth metric.

**Results and Discussion**

We have evaluated the performance of the 8 DEG detection tools by 5 metrics (Figure 3A–3F). Using the 12 quantitative reverse transcription polymerase chain reaction validated DEGs, Figure 3A showed that nonparametric methods, such as NOISeq and SAMseq, and parametric methods, such as Cuffdiff2 and edgeR, were able to identify at least half of these 12 DEGs. In contrast, parametric methods, such as baySeq, DESeq2, DSS, and Limma+Voom, were able to identify only 1 or 2 of these 12 DEGs. Six genes were difficult to detect by RNA-seq-based methods (marked in red). The expression pattern of these difficult genes showed that they have either smaller fold changes (eg, Tgfb3) or higher between-replicate variability (eg, Actn3).

Figure 3B listed the top 20 GO terms and all 4 pathways from the ground-truth functional annotations, most of which were linked to the mechanisms of muscle contraction and heart development. Figure 3C summarized the concordance between the pipeline-specific annotations and the ground truth in terms of the top 20 GO terms (ranking by $P$ values), all GO terms, and all pathways. DEGs detected by baySeq, DSS, and Limma+Voom were associated with zero GO terms and only a few pathways, which suggested that these tools detected DEGs with very diverse functions. DEGs detected by edgeR and Cuffdiff2 had more high-ranking functional annotations concordant with the ground-truth annotation. In contrast, DEGs detected by DESeq2, SAMseq, and NOISeq were linked to many GO terms and pathways that were biologically irrelevant to the original study, with no concordance appeared in the top 20 GO terms.

Figure 3D showed the number of DEGs supported by 1, 2, or all 8 tools for each DEG detection method. baySeq, DSS, and Limma+Voom identified a fewer number of DEGs (ie, 32, 23, and 12, respectively) that were highly reproducible among various tools (ie, each DEG was supported by $\geq2$ other tools). Tools with more detected DEGs, such as SAMseq, NOIseq, and DESeq2, tended to have more pipeline-specific or unique DEGs. However, as discussed earlier, a higher number of DEGs did not necessarily lead to more biologically relevant results. Figure 3E demonstrated the distribution of $R_{dominance,g}$ for DEGs. Most DEGs had $R_{dominance,g}$ in the range of 25% to 60% following the scenarios (4-1) and (4-2) we have discussed in the Bioinformatics Pipeline and Performance Evaluation Metrics section. Such observation indicated that most DEGs detected by RNA-seq pipelines did not have huge variability between biological replicates. DEGs with larger between-replicate variability resulted in $R_{dominance,g} > 60%$. Such high variability can be the nature of biological replicates or biases introduced in the sequencing or bioinformatics processes. The nonparametric NOISeq method had the highest percentage of DEGs with $R_{dominance,g} > 60%$ because it identified many genes with low read counts (eg, $[A_{1,g}, A_{2,g}, B_{1,g}, B_{2,g}]=[1,1,0,0]$). Thus, a small deviation in the read counts may have caused a huge variation in $R_{dominance,g}$. For parametric methods, higher $R_{dominance,g}$ indicated that the read counts may not follow a negative binomial distribution. edgeR and Cuffdiff2 had a higher chance of detecting this type of genes as DEGs. Figure 3F showed the distribution of the mean read counts of DEGs. NOISeq had a bimodal distribution because of its tendency to identify some DEGs with low read counts. The other 7 tools shared a similar range of the mean read counts of DEGs, with baySeq slightly skewed to the left (ie, lower mean read counts).

In summary, the original paper used RNA-seq to study the effect of Ezh2 deletion on gene expression profiles. It identified a set of DEGs relevant to cardiac tissue development and remodeling. Our study examined the functions of DEGs detected by the 8 RNA-seq pipelines, and edgeR and Cufflinks yielded the most functionally relevant DEGs. The nonparametric methods such as NOISeq and SAMseq identified many more DEGs than other tools, yet a large proportion of these DEGs may have been less reliable (eg, DEGs with low read counts) and irrelevant to the biology of the original study.

**Remaining Bioinformatics Challenges**

RNA-seq technology provides an opportunity to comprehensively study the transcriptome. While fixing sequence mapping and expression quantification steps and focusing on evaluating only DEG detection methods, we found that different tools generated different DEG sets. Therefore, translating the computational findings into real clinical applications requires integrative biological interpretation and large-scale experimental validation. Capturing the full dynamics of the pipeline and forming a guideline of pipeline selection require a factorial experiment for studying the effect of each module in the pipeline. Finally, RNA-seq technology can be unreliable for low-expressing genes, but currently no standardized methods are capable of handling them properly. Thus, distinguishing true signals from noise for low-expressing genes remains a challenge.

**Cardiovascular Epigenomics Using ChIP-Seq**

**Data Set**

The ChIP-seq data set (Sequence Read Archive accession: SRP008658) investigated the genome-wide map of human heart enhancers with a pan-specific antibody that targets 2 closely related transcriptional coactivator proteins, p300 and CBP (cAMP response element-binding protein–binding protein). This data set contains tissue samples from 1 fetal and 1 adult human heart. Each sample was sequenced with Illumina Genome Analyzer and contains around 27 million 36 base pair single-ended reads.

**Bioinformatics Pipeline and Performance Evaluation Metrics**

The bioinformatics pipeline for identifying genome-wide protein-binding regions using ChIP-seq includes sequence mapping and peak calling (Figure 2). We use the same sequence mapper, Bowtie, and 6 different peak-calling tools to construct totally 6 pipelines. Bowtie12 maps sequence reads (or sequence tags in ChIP-seq) to the GRCh37/hg19 human genome and reports only uniquely mapped tags. The 6 peak-calling tools (Figure 2, right table), including SISSRs, MACS, FindPeaks, SWEMBL, SICER, and F-Seq, represent a wide variety of algorithms for determining statistically significant peaks. We run these tools using their default or recommended parameters with a $P$ value threshold of $10^{-4}$. The identified peaks are putative protein-binding regions for p300 and CBP proteins.

To assess the performance of the 6 peak-calling tools, we have designed 5 metrics: (1) to visualize sequence-mapping and peak-calling information using the Integrative Genomics Viewer, (2) to count the total number of peaks called by each tool, (3) to investigate the distribution of $N_{p}$, the normalized tags per peak,
as defined in Equation 2, (4) to compute the average length of peaks called by each tool, as defined in Equation 3:

\[ N_i = \frac{\text{(number of tags)}_i}{\text{(peak length /100)}}, \]

\[ L_{\text{average}} = \frac{\sum_{i=1}^{N} \text{length(peak)}_i}{N}, \]

where \( N \) is the total number of peaks, and (5) to biologically validate the peaks by investigating the percentage of peaks that contain \( \geq 1 \) p300 motif using FIMO (find individual motif occurrences)\(^{69}\) with a \( P \) value threshold of \( 10^{-4} \). The input information for FIMO includes DNA sequences corresponding to these peaks and the position-specific scoring matrix for the p300 motif retrieved from the SwissRegulon Portal.\(^{70}\)

**Results and Discussion**

We have investigated the performance of the 6 peak-calling tools by 5 metrics (Figure 4A–4E). Visualizing by the Integrative Genomics Viewer, Figure 4A showed the peak regions called by the 6 tools with corresponding coverage information from the Bowtie alignment in the upstream region of the INPP5A gene (inositol polyphosphate-5-phosphatase, 40 kDa). The selective nature of SISSRs and MACS resulted in sparse and short peaks. In contrast, FindPeaks and SWEMBL tended to call long peaks with lengths >10 kbp. Figure 4B demonstrated the number of peaks called by the 6 tools. SICER called the largest number of peaks, followed by FindPeaks and F-Seq. SICER failed to form longer peaks by merging nearby peaks, resulting in a relatively higher number of peaks. FindPeaks called 2 separate peaks although 2 protein-binding regions were in close proximity; thus, FindPeaks also tended to call more peaks than the other tools. Figure 4C depicted the distribution of the number of tags per peak normalized by the peak length. Larger numbers indicated that the detected peaks were supported by more evidence. SISSRs, MACS, SWEMBL, and F-Seq exhibited a moderate to high number of tags per peak. In contrast, FindPeaks and SICER detected some peaks with a low number of tags per peak. These peaks may not have been reliable because of limited evidence. Figure 4D showed the average length of peaks called by the 6 tools. Among them, SWEMBL had the longest average length, which may not have been a reasonable length for protein binding. The selective nature of SISSRs and MACS resulted in sparse and short peaks. In contrast, FindPeaks and SWEMBL tended to call long peaks with lengths >10 kbp. Figure 4B demonstrated the number of peaks called by the 6 tools. SICER called the largest number of peaks, followed by FindPeaks and F-Seq. SICER failed to form longer peaks by merging nearby peaks, resulting in a relatively higher number of peaks. FindPeaks called 2 separate peaks although 2 protein-binding regions were in close proximity; thus, FindPeaks also tended to call more peaks than the other tools. Figure 4C depicted the distribution of the number of tags per peak normalized by the peak length. Larger numbers indicated that the detected peaks were supported by more evidence. SISSRs, MACS, SWEMBL, and F-Seq exhibited a moderate to high number of tags per peak. In contrast, FindPeaks and SICER detected some peaks with a low number of tags per peak. These peaks may not have been reliable because of limited evidence. Figure 4D showed the average length of peaks called by the 6 tools. Among them, SWEMBL had the longest average length, which may not have been a reasonable length for protein binding.

Figure 4. Qualitative and quantitative assessment of various peak-calling tools using chromatin immunoprecipitation sequencing data. A. Integrated genomics viewer visualized peaks called by the 6 tools in the upstream region of the INPP5A gene using the adult heart sample. The black histogram on the top represented the coverage of Bowtie alignment in the same region. B. The number of peaks detected by each peak-calling pipeline. C. The distribution of the number of tags per peak normalized by the peak length for each peak-calling pipeline. D. The average length of peaks for each peak-calling pipeline. E. The percentage of peaks that contains \( \geq 1 \) p300 motif identified by FIMO with a \( P \) value threshold of \( 10^{-4} \) for each peak-calling pipeline. IGV indicates Integrative Genomics Viewer; MACS, model-based analysis of ChIP-seq; SICER, spatial clustering for identification of ChIP-enriched regions; and SISSRs, site identification from short sequence reads.
DNA-binding sites. SISSRs, MACS, and F-Seq exhibited the average peak length of <400 base pair, which was close to the designed fragment length from the Illumina sequencing protocol. Using FIMO, Figure 4E demonstrated the percentage of peaks that contained the p300 motif. MACS performed the best with 15% to 22% of the peaks containing the motif. SISSRs and F-Seq performed moderately well with their motif discovery rate ranging from 6% to 8%. Although FindPeaks and SICER detected a significantly larger number of peaks than the others, only 2% to 3% of these peaks contained the p300 motif, exposing their relatively high false positive rate. Around 11% to 22% of peaks called by SWEMBL contained the p300 motif. However, despite such high performance, the peaks were not reliable because SWEMBL had extremely long peaks on average, which increased the probability of identifying the motif by chance alone.

In summary, the original study used ChiP-seq with the antibody that recognizes the enhancer-associated coactivator proteins p300 and CBP to annotate candidate heart enhancers that may regulate the expression of heart development-related genes in the human genome. By examining the percentage of peak regions (ie, candidate heart enhancer regions) that contained the p300 motif, MACS achieved the highest motif discovery rate among the 6 tools, which suggested that MACS identified more biologically relevant peaks than the others. In addition, MACS’s peaks had the second highest tag coverage and the reasonable average peak length. In contrast, SICER identified peaks with the lowest motif discovery rate and low tag coverage.

Remaining Bioinformatics Challenges

Similar to the case for RNA-seq, ChiP-seq requires a factorial experiment for studying the effect of either sequence-mapping or peak-calling step. Most peak-calling tools need control samples for building background models essential for conducting statistical tests. These background models can be local or global. The global model is easier to build but lacks the consideration of local biases. Accurately identifying peaks requires an adaptive background signal model that can dynamically change parameters to accommodate local variations and different ChiP-seq experiments.

Opportunities of NGS for CVD

The prevention or treatment of CVDs can benefit from personalized care that tailors clinical decisions on a patient-by-patient basis. Most common CVDs such as atherosclerosis and coronary artery disease have complex phenotypes and are affected by both genetic and environmental factors.

Recently, omic biomarkers (eg, genomic, transcriptomic, and epigenomic) have emerged as a complement to traditional CVD risk factors. Facilitated by NGS bioinformatics, 1 distinctive feature of the NGS technology is its capability to identify a variety of omic biomarkers. As shown in the Case Studies: RNA-seq and ChiP-seq Bioinformatics for CVD section, the benefits of NGS bioinformatics are (1) it can comprehensively detect DEGs or peaks in the entire genome, (2) it can define the boundaries of peaks using ChiP-seq data at base-pair resolution, and (3) it can discover previously unknown DEGs or peaks as candidate omic biomarkers. Currently, most research uses a single omic data modality captured by NGS to study only 1 aspect of biological mechanisms of diseases. However, for complex diseases such as CVDs, candidate biomarkers based on a single omic data modality suffer from the reproducibility issue. To address this issue, we need to research novel NGS bioinformatics methods that not only can integrate multiple omic data modalities (eg, genetic variations, transcriptional regulation, and epigenetic modifications) but also can associate heterogeneous omic information with CVD phenotypes. Associations (eg, correlation and causality) resulted from NGS bioinformatics present a great opportunity for researchers to obtain further insights on disease mechanisms and to improve risk predictions for complex CVDs.

Sources of Funding

This work was supported in part by grants from the National Institutes of Health (National Heart, Lung, and Blood Institute 5U01HL080711, U54CA119338, and 1RC2CA148265), the Georgia Cancer Coalition (Distinguished Cancer Scholar Award to Dr. Wang), the Children’s Healthcare of Atlanta, Microsoft Research, and the Hewlett-Packard Company.

Disclosures

None

References


J Mol Biol

J Mol Biol

Bioinformatics

2013;14:R95.


Cardiovascular Transcriptomics and Epigenomics Using Next-Generation Sequencing: Challenges, Progress, and Opportunities

Circ Cardiovasc Genet. 2014;7:701-710
doi: 10.1161/CIRCGENETICS.113.000129
Circulation: Cardiovascular Genetics is published by the American Heart Association, 7272 Greenville Avenue,
Dallas, TX 75231
Copyright © 2014 American Heart Association, Inc. All rights reserved.
Print ISSN: 1942-325X. Online ISSN: 1942-3268

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circgenetics.ahajournals.org/content/7/5/701

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation: Cardiovascular Genetics can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation: Cardiovascular Genetics is online at:
http://circgenetics.ahajournals.org//subscriptions/