Journal of Proteomics & Bioinformatics - Open Access

Review Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000138
JPB/Vol.3 Issue 6

Statistical Analyses of Next Generation Sequence Data: A Partial Overview

Susmita Datta[1]*, Somnath Datta[1], Seongho Kim[1], Sutirtha Chakraborty[1] and Ryan S. Gill[2]

[1]*Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA*
[2]*Department of Mathematics, University of Louisville, Louisville, KY 40202, USA*

Abstract

Next generation sequencing has revolutionized the status of biological research. For a long time, the gold standard of DNA sequencing was considered to be the Sanger method. However, in 2005, commercial launching of next generation sequencing has made it possible to generate massively parallel and high resolution DNA sequence data. Its usefulness in various genomic applications such as genome-wide detection of SNPs, DNA methylation profiling, mRNA expression profiling, whole-genome re-sequencing and so on are now well recognized. There are several platforms for generating next generation sequencing (NGS) data which we briefly discuss in this mini overview. With new technologies come new challenges for the data analysts. This mini review attempts to present a collection of selected topics in the current development of statistical methods dealing with these novel data types. We believe that knowing the advances and bottlenecks of this technology will help the researchers to benchmark the analytical tools dealing with these data and will pave the path for its proper application into clinical diagnostics.

Keywords: DNA; Sequencing; Deep sequencing; High throughput; Sequence reads; RNA; ChIP-seq; Intensities

Introduction

Next generation sequencing (Shendure and Ji, 2008), also known as deep sequencing, is a transformative technology for today's biomedical research. The growing importance of next generation sequencing for the clear understanding of various biological systems has indirectly triggered a competition among several companies, each trying to come up with a sequencing platform which can produce high quality longer read sequences with greater throughput and reduced cost. We begin this mini review by discussing various technologies for next-generation sequencing.

Roche 454

NGS using Roche 454 technology became commercially available in 2005 (Margulies et al., 2005). This technology uses bead-based emulsion polymerase chain reaction (em- PCR) to amplify copies of templates of DNA molecule (Dressman et al., 2003). The amplified beads are sequenced in parallel by pyrosequencing (Marsh, 2007). In pyrosequencing, four different nucleotides are flowed in a sequential manner through a solid surface containing wells into which single beads can fit. This process goes on for cycles and the signal intensity per flowing nucleotide is recorded for each bead over time and is analyzed to generate good quality sequence.

This platform is lot more high throughput than any capillary based sequencing. In the Titanium version of the Roche 454 platform, the output has several hundred mega bases of 400-500 base reads per run. Hence, it is more cost effective than Sanger's chain termination method (Sanger et al., 1977) which was the old standard of DNA sequencing. The 454 technology does not suffer from the G-C rich content and does not skip the unclonable segments as the process does not rely on cloning. However, it is to be noted that 454 technology suffers from detecting subsequences of repetitive DNA sequences or homopolymers in a DNA sequence. As pyrosequencing depends on intensities of light, the light emitted for detecting TAAAA or AAAAA could be very similar. Also, while 454 sequencing is cheaper and faster per base, each run is quite expensive (over $8000), and so it is not suitable for sequencing targeted fragments from small numbers of DNA samples.

SOLiD by applied biosystem

In this technology, similar to 454, DNA fragments are amplified by em-PCR onto beads (Dressman et al., 2003). The difference between the SOLiD and the 454 platform is that the SOLiD beads are much smaller than 454 beads (1 μm vs. 28 μm). This results in much denser packing of beads into the same area in SoLiD (100 million beads per sequencing run). This platform can produce approximately 20 Gb of short-read sequence data per run (25-50 bases) and so is preferable to resequencing for de novo assembly. SOLiD uses a unique ligation-mediated sequencing strategy which is less prone to the errors involved with pyrosequencing method in 454 platform. In the SOLiD system, each data point represents two adjacent bases, and each base is interrogated twice. Hence it can discriminate between sequencing errors and true polymorphisms. The drawback of this platform is that the data is collected in color space and it provides information about two adjacent bases but not the definitive identification. So, they have to be decoded in order to be mapped to a reference genome and conventional alignment tools can't be used for the mapping process. Direct conversion from color to sequence data also is prone to error such as reads that contain sequencing errors can not be converted accurately. Error rate of this platform is significantly higher than traditional Sanger sequencing.

Illumina/solexa genome analyzer

This platform was first introduced by Solexa in 2006 and later on re-branded as Illumina Genome Analyzer (GA). This technology does not depend on the em-PCR to amplify the template DNA strands like

***Corresponding author:** Susmita Datta, Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA, Tel: 1-5028520081; E-mail: susmita.datta@louisville.edu

Received April 23, 2010; **Accepted** June 03, 2010; **Published** June 03, 2010

Citation: Datta S, Datta S, Kim S, Chakraborty S, Ryan SJ (2010) Statistical Analyses of Next Generation Sequence Data: A Partial Overview. J Proteomics Bioinform 3: 183-190. doi:10.4172/jpb.1000138

the other two platforms mentioned before. Instead, adapter ligated template molecules flow into the flow cell (hollow glass slide). Template DNA hybridizes to the primers on the flow cell surface and gets copied onto the flow cell as an extension to the hybridization primer. This results in a generation of reverse complimentary copy of the template. This newly synthesized strand serves as templates for isothermal amplification reaction and results in clusters of amplified strands. Due to the terminator nucleotides, each DNA strand within a cluster incorporates the same nucleotide within a cycle. The clusters are imaged and the next round of nucleotide incorporation begins after removing the imaged blocked groups and the flurophores of the newly incorporated nucleotides. Analysis of the images generates a separate sequence for each cluster.

An Illumina Genome Analyzer is currently capable of producing a sequence up to 10-Gb per 76-cycle paired-end run. However, beyond this length the frequency of the substitution errors are high due to cluster phasing and de-phasing.

Helicos true single-molecule-sequencing (tSMS) technology

This sequencing platform has been launched in 2008, and is considered as the next-next generation or 3rd generation sequencing. This platform is based on the technology that was published in 2003 (Braslavsky et al., 2003). This sequencing also deals with millions of templates parallely; however, tSMS differs from the existing next generation sequencing described earlier in that it does not amplify the template molecules. So it is free from any errors due to the amplification process. Also, the library preparation is simple and rapid. In this technology, single molecules are the substrates for the sequencing reaction. Fluorescent nucleotides are added singly. The flow cell is visualized to identify strands that hybridized the particular nucleotide with the help of the fluorescence of the nucleotide. The incorporation of the nucleotides to the strand depends on the compatibility of the template strand with the order of the nucleotide addition. Hence the length of the sequences is variable but on average they are of lengths of 25-30 bases.

This technology is free from phasing errors. However, sensitivity seems to be an issue with this platform (Harris et al., 2008). The true error rate can be reduced from 2 to 3% to below 1% by repeating the reading of the same strand twice. However, it increases the running time (Harris et al., 2008).

Other than these major platforms, some noteworthy emerging platforms are mentioned here. The SMRT technology by Pacific Biosciences (Eid et al., 2009) has recently showed promising early results using single-molecule real-time DNA sequencing. Dover Systems' Polonator was announced in early 2008 by George Church at MIT and arose from collaboration between George Church's laboratory and Danaher Corporation. Although this system uses bead based em-PCR (Dressman et al., 2003) and sequencing by ligation (Shendure et al., 2005), this is very high throughput and can generate data up to 3 Gb per day. However, read lengths are short and so it is difficult to use them for vertebrate-size genome. An appealing feature of this technology is that it is open source and the users can buy all the reagents from any supplier. Other NGS platforms include: BASE (single-molecule sequencing technology) by Oxford Nanopore Technologies/Illumina, one by Intelligent Bio-systems using proprietary sequencing-bysynthesis technology (Ju et al., 2006) and single-molecule sequencing technology based on fluorescence resonance energy transfer (FRET) by VisiGen Biotechnologies etc. Sequencing technologies are also under development by Affymetrix, Reveo, Base4innovation, Genome Corp, and Complete Genomics, among others. Detailed technical reviews of various NGS platform appeared in Mardis (2008) and Metzker (2010).

A brief comparison of the three most popular NGS platforms

Roche 454 gives longer reads (500-700 bp) than both Illumina and SOLiD but it suffers from low accuracy in the long homopolymeric regions.

The price for sequencing each nucleotide is several times reduced in the Illumina technology compared to the Roche 454 pyrosequencing platform. In terms of the total analysis time and sequencing throughput, Illumina and SOLiD platforms are close to each other (flow cell construction being costlier in Illumina and sequencing in SOLiD).

From the perspective of practical applications, both Illumina and SOLiD platforms have their respective cutting edges. In high-throughput resequencing of large genomes SOLiD is more accurate than Illumina whereas for RNA sequence analysis (RNA-seq), Illumina is more suitable.

Base calling techniques

Most of the work in this important research area has taken place primarily for the llumina (Solexa) platform. Its base calling can suffer from three dominant noise factors (Erlich et al., 2008) as follows. In sequencing-by-synthesis, each single-stranded nucleotide fragment is amplified around the initial attachment in the flow cell, resulting in a cluster of about 1,000 identical copies of each fragment. Each

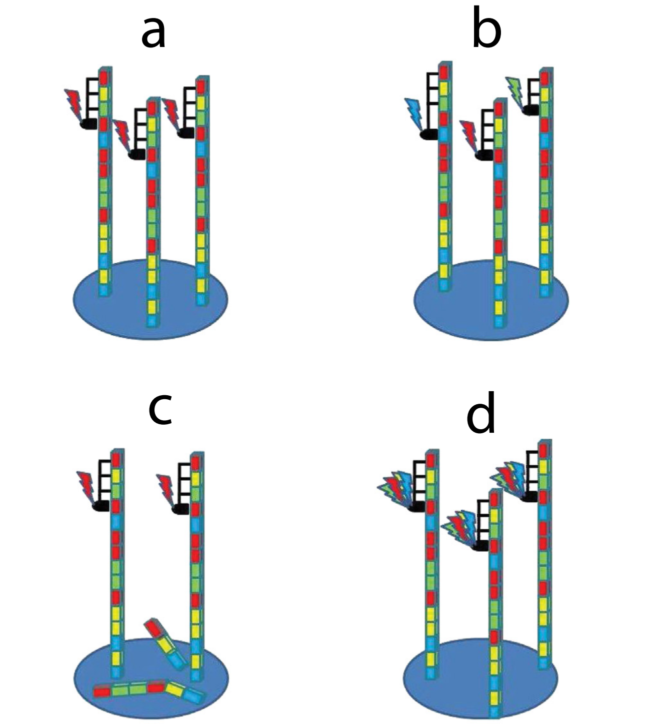

Figure 1: a: The Illumina noise factors. Here the identical DNA templates are depicted as the colored boxes and the black ovals denote DNA polymerases. With no noise, the current signal (red lightning) is strong enough to infer the DNA sequences directly. **b:** When the phasing noise is introduced, it causes lagging (green lightning) or leading (blue lightning). **c:** The loss of material can cause reducing the signal intensity (fading noise). **d:** The fluorophore cross-talk can cause misinterpretation of the current signal.

Review Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000138
JPB/Vol.3 Issue 6

terminal nucleotide in all the clusters is then excited by lasers and its signal is detected by charged coupled device (CCD) images of fluorescence emission. Ideally, the current position for synthesizing will be the same within a cluster, generating a strong signal (Figure 1a). However, the unstable chemistry causes stochastic failures in reading the next nucleotide, introducing phasing (lagging; no new base synthesized) and prephasing (leading; two bases synthesized) noises (Figure 1b). The second noise factor is due to loss of copies of fragments so that the signal intensity is reduced, which is called the fading noise (Figure 1c). The third noise factor is known as the fluorophore cross-talk, causing misinterpretation of the signal (Figure 1d).

Illumina developed the built-in base-caller *Bustard* to transform observed intensities into sequences. *Bustard* consists of three steps and each step deals with the three main noise factors separately. It first handles the fluorophore cross-talk by transforming intensities to concentrations. To do this, it defines the cross-talk matrix and removes the overlapping fluorophore effect from intensities by taking the inverse crosstalk matrix. Next renormalization of concentrations is performed by dividing by the average concentration to eliminate the fading noise. The third step involves fitting a Markov model to eliminate the phasing noise resulting in the estimated sequences.

Rougemont et al. (2008) used probabilistic modeling and model-based clustering to identify and code ambiguous bases and to arrive at decisions to remove uncertain bases towards the ends of the reads. *Alta-Cyclic* was developed by Erlich et al. (2008) based on support vector machine (SVM), requiring a control lane containing a sample with a known reference genome for supervised learning. Another attempt to improve the Illumina basecaller led to *Swift* by Whiteford et al. (2009). They devoted it to the image analysis.

One of the primary challenges in base calling is the dependency among cycles. Bustard, including *Alta-Cyclic*, assumes that all the cycles are performed independently. Recently, several cycle-dependent base-callers have been introduced. *Ibis* (Improved base identification system) was developed based on the SVM by Kircher et al. (2009). They used the multiclass-SVM to provide for a cycle-dependent model differently from *Alta- Cyclic* in which univariate SVM was used (Erlich et al., 2008). Bravo and Irizarry (2009) came up with their own modeling to quantify the read/base-cycle effects. Recently, Kao et al. (2009) developed *BayesCall* based on a stochastic Bayesian modeling. A somewhat complex dynamic modeling strategy is used in *BayesCall* which is schematically described in Figure 2, where L refers to the total number of cycles (length of fragments) in a run, $S_k = (S_{1,k}, S_{2,k}, \ldots S_{L,k})$ represents the complementary DNA sequence with length L in cluster k, $I_{t,k} = (I^A_{t,k}, I^C_{t,k}, I^G_{t,k}, I^T_{t,k})' \in R^{4\times 1}$ denotes the observed fluorescence intensities of the *A, C, G, T* channels at cycle t in cluster k, and $\Lambda_{t,k}$ denotes the active template concentration in cluster k at the t-th cycle. One of the novelties of *BayesCall* is the capability to use cycle-dependent parameters in its modeling, adding greater flexibility. To avoid over-fitting, the read length is divided into non-overlapping windows and it is assumed that the parameters remain constant within each window. In general, three types of algorithms are used to estimate the parameters in *BayesCall*, namely, MCEM (Wei and Tanner, 1990), ECM (Meng and Rubin, 1993) and simulated annealing (Kirkpatrick et al., 1983). Finally, a quality score for a call is calculated based on its estimated posterior probability. For further details we refer the readers to the original paper by Kao et al. (2009).

For the Roche (454 Life Sciences) platform, there exist two base callers that are the built-in 454 base caller and *Pyrobayes* (Quinlan et al., 2008). The Applied Biosystems (SOLiD) uses a different style to detect the signal by the two base color code and there currently is only its own built-in base-caller.

Data quality and reproducibility

Several papers have examined the reliability and reproducibility of data from next generation sequencing platforms. While some studies have found next generation sequencing data to be superior to competing methods, others have found systematic problems with the reads obtained in next generation sequencing. Most of these studies used data obtained from the Illumina platform.

Marioni et al. (2008) observed that next generation sequencing data from Illumina are highly reproducible and very reliable, and overall they found it to be superior to the data produced by microarray technology. They used Illumina to sequence each sample on seven lanes across two plates. The gene counts were highly correlated across lanes (Spearman correlation average = 0.96).

To test for a lane effect by comparing each pair of lanes, Marioni et al. (2008) tested the null hypothesis that gene counts in one lane represent a random sample from the reads in both lanes for each mapped gene. Let, for a sample t, x_{tk} denote the observed number of counts in lane k and let C_k denote the number of reads in lane k for $k = a, b$. For a clear understanding of this test for a lane effect, it is helpful to let X_{tk} denote the random variable representing the number of counts in lane a in an experiment with $x_{ta}+x_{tb}$ total counts, C_a reads from lane a, and C_b reads from lane b. Now, they tested the null hypothesis that the gene counts in one lane represent a random sample from the reads in both lanes for each mapped gene; symbolically, this is a test of the null hypothesis $H_0 : \pi = P_0(x_{ta})$ versus the alternative $H_A : \pi \neq P_0(x_{ta})$ where $\pi = P(X_{ta} = x_{ta})$ and

$$P_0(x) = \frac{\binom{C_a}{x}\binom{C_b}{x_{ta}+x_{tb}-x}}{\binom{C_a+C_b}{x_{ta}+x_{tb}}}$$

is the probability that $X_{ta}=x$ if the null hypothesis is true (in which case X_{ta} follows a hypergeometric distribution). They used this test to compute the P-values for each gene and plotted the quantiles of the uniform distribution against the observed quantiles of the P-values for each gene, and they found that less than 0.5% of the genes had small P-values when the pair of lanes had the same concentration of samples. However, a larger proportion of genes indicated a lane

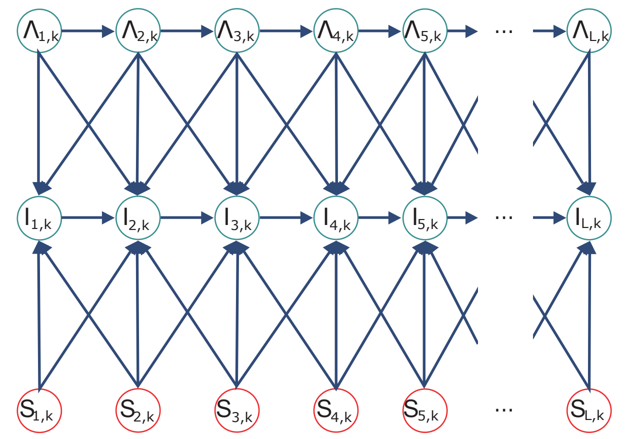

Figure 2: The graphical modeling for *BayesCall* when the window size is 1.

effect when the pair of lanes had different concentrations.

Marioni et al. (2008) also suggested a global test for lane effects by comparing all L lanes. For each sample i, they assume that the number of reads mapped to gene j for lane k follows independent Poisson distributions with mean $c_{ik}\lambda_{ijk}$ where c_{ik} is the total rate that lane k produces reads at and λ_{ijk} is the rate of reads to gene j in lane k relative to other genes. To test the null hypothesis H_0: $\lambda_{ij1} = \ldots = \lambda_{ijL}$ (i.e., λ_{ijk} are equal for each lane $k = 1,\ldots, L$) versus the alternativeH_A: λ_{ijk} are not equal for all k, they used a goodness-of-fit statistic which follows a chi-square distribution with L -1 degrees of freedom when the null hypothesis is true. After plotting the Chi-squared quantiles against the observed quantiles for each gene j, they found that only about 0.5% of the genes had extra Poisson variation when lanes sequenced the same sample at the same concentrations.

Marioni et al. (2008) also used the Poisson model to identify differentially expressed genes. Specifically, for each gene j, they tested the null hypothesis that the rate of reads λ_{ijk} are the same for all i and k versus the alternative that the liver and kidney sample have different read rates λ_{ijk}^{A} and λ_{ijk}^{B} . Here, they used the likelihood ratio test statistic which follows a X_1^2-distribution under the null hypothesis. Using this method, 11493 genes were found to be differentially expressed in the liver-versus-kidney samples. This list of differentially expressed genes obtained with the Illumina data was compared with the results based on Affymetrix U133 Plus 2 arrays where an empirical Bayes approach was used to identify differentially expressed genes. Of 8113 differentially expressed genes found by the array, 81% were also found to be differentially expressed using Illumina. Finally, quantitative PCR (qPCR) was used to examine discrepancies, and overall, the qPCR results agreed more with Illumina than with the arrays.

Fu et al. (2009) arrived at a similar conclusion by comparing the relative accuracy of transcriptome sequencing (RNA-seq) and microarrays with protein expression data from adult human cerebellum using 2D-LC MS/MS. They found that the next generation sequencing provided more accurate estimation of absolute transcript levels.

Wall et al. (2009) used simulation models to compare next generation sequencing with traditional capillary-based sequencing and concluded that next generation sequencing offers great benefit in terms of coverage over capillary-based sequencing. However, they suggest combining sequencing methodologies such as FLX and Solexa to achieve optimal performance at a modest cost.

On the other hand, a number of authors have reported problems and systematic biases with the sequence reads obtained in next generation sequencing. Dohm et al. (2008) considered two Solexa read data sets and found that error rates were greater at the end of reads (0.3% at the beginning compared with 3.8% at the end) and wrong base calls are often preceded by base G. Also, base substitution errors were significantly disproportionate with A to C substitution error being 10 times more frequent than the C to G substitution. Similar artifacts were observed by Bravo and Irizarry (2009) who considered data from the control lane of an Illumina ChIP-seq experiment and reported A to T miscall to be the most common error in their calibration study. They also reported that the error rates vary with the position on the read and questioned the utility of the quality scores supplied by the manufacturers with a base call. These and other systematic biases may lead to wrong statistical conclusions. Finally, Oshlack and Wakefield (2009) considered three data sets including sequencing data from Illumina and SOLiD and demonstrated for each data set that when gene expression is calculated using aggregated tag counts for each gene in RNA-seq technology the ability to call differentially expressed genes (or ranking) between samples is strongly associated with the length of the transcript.

Statistical tools for using sequence reads

There are a number of notable papers in the area of transcriptome analysis using NGS technology: Nagalakshmi et al. (2008) in yeast; Cloonan et al. (2008); Morin et al. (2008); Marioni et al. (2008) in human; Mortazavi et al. (2008) in mouse; Vera et al. (2008) in butterfly, and so on. A next generation sequencing technology obtains millions of short reads from the transcript population of interest and by mapping these reads to the genome, RNA-Seq produces digital (counts) rather than analog signals and offers the chance to detect novel transcripts. Obviously, there are several protocols for transcript quantification for NGS data.

Mapping software such as MAQ by Li et al. (2008) are useful in assembling short sequence reads to match a reference genome. MAQ uses a Bayesian calculation to produce a phred-scaled probability (−10 times the common logarithm of the probability) that an individual alignment is mapped incorrectly. It also includes the capability to use mate-pair information for paired-end read alignment in diploid samples. MAQ can assimilate the mapping quality and raw sequence base quality scores and uses a Bayesian analysis to make a final genotype call. More recent mapping tools such as Bowtie (Langmead et al., 2009) and BWA (Li and Durbin, 2009) utilize computational advantages of string matching theory via the Burrows-Wheeler transform to provide much faster algorithms for short read alignment against a large reference genome with a small memory footprint. Bowtie extends the Burrows-Wheeler transform for alignment using backtracking to allow mismatches as well as double indexing to avoid excessive backtracking. Although it is a greedy algorithm which will not necessarily find the best match if it is inexact, Bowtie gives users options to improve its accuracy in return for computational costs. The BWA algorithm uses an efficient backward search with the Burrows-Wheeler transform and allows for inexact matching and gapped alignment. Li and Durbin (2009) describe the computational performance and accuracy of MAQ, Bowtie, and BWA on some simulated and real data examples.

The software package F-seq, developed by Boyle et al. (2008), employs kernel smoothing in converting high-throughput sequencing reads into continuous signals along a chromosome whose output can be displayed directly in the UCSC Genome Browser. This type of data summary will be useful to identify specific sequence features, such as transcription factor binding sites (ChIP-seq) or regions of open chromatin (DNase-seq). F-seq provides a more statistically rigorous tool to researchers who would otherwise use histograms to calculate regions of highly dense sequence reads. Also, Zhang et al. (2008a) developed MACS (Model-based Analysis of ChIP-seq) that utilizes Poisson modeling and to capture local biases in the genome resulting in more robust predictions of binding sites.

Jiang and Wong (2009) used statistical inference of isoform expression using high throughput RNA sequencing (RNA-Seq) data by Poisson modeling and solving a convex optimization problem. The measure RPKM (reads per kilobase of the transcript per million mapped reads to the transcriptome) was originally introduced by Mortazavi et al. (2008). Normalizing the counts of reads mapped to a gene (or to exons belonging to gene) against the transcript length and the sequencing depth, this RPKM measure can compare the expression measures across different genes and different experiments. However, reads that are mapped to a gene are frequently shared by multiple isoforms. Consequently, Jiang and Wong (2009) developed the following statistical model for this isoform expression estimation problem. Let G be the set of genes and F be the set of isoforms for all possible isoforms for all genes. Let l_f be the length and let k_f be the number of copies of the transcripts in the form of an isoform $f \in F$. Assuming every read is independently and uniformly sampled from all possible nucleotides in the sample, the probability that a read comes from isoform f is $k_f\, l_f / L$, where the total length of the transcripts in the sample is $L = \Sigma_{f \in F} k_f l_f$. If w denotes the total number of mapped reads, then the number of reads coming from a region of length l in f can be modeled by a Binomial random variable with w trials and probability of success $k_f\, l / L$. Furthermore, if w is large and p is small, the law of rare events allows this distribution to be approximated by a Poisson distribution with mean $\lambda = k_f\, lw / L$. Now, assume that there are m exons with respective lengths $l_1,\ldots,l_m$ and n isoforms with respective expressions $\theta_1,\ldots,\theta_n$. The set of observations falling into a region can be modeled by a Poisson random variable with mean $\lambda = l_g w \Sigma_{f=1}^{n} c_{fg}\theta_f$ where c_{fg} is an indicator variable that equals 1 if isoform f contains exon g and equals 0 otherwise. The counts for exon-exon junctions can be modeled by a Poisson random variable with mean $\lambda = lw\Sigma_{f=1}^{n} c_{fg}c_{fh}\theta_f$. In the multiple isoform case, numerical methods (e.g., hill climbing) must be used to obtain the maximum likelihood estimate of the θ's; fortunately, the joint log-likelihood is concave, so any local maximum is also guaranteed to be a global maximum. Standard numerical calculations based on the Fisher information matrix can be problematic when some of the isoforms have low expressions, so in these cases, a Bayesian alternative using importance sampling is proposed for making statistical inferences.

In a recent article, Bullard et al. (2010) explore the effects of different systematic sources of variability in measuring the differential expression of genes using three platforms (mRNA-seq data from Illumina sequencing, microarray and quantitative real time PCR assay data), all of which are based on the context of the Microarray quality control project (MAQC). In addition, it is also shown that using an auto-calibration instead of Illumina's standard way of reserving one flowcell lane for the control can help improving the mapping quality of the reads thereby ensuring a much more cost-effective and efficient experimental design. Normalization strategies are suggested to get rid of these biases.

Other notable contributions leading to broad data analytic tools include Johnson et al. (2007, mapping techniques); Fejes et al. (2008, enrichment analysis); Ji et al. (2008, ChIP-seq data); Sharon et al. (2008, protein-DNA interactions); Zhang et al. (2008b, ChIP-seq data); Rozowsky et al. (2009, ChIP-seq data); Langmead et al. (2009, alignment tool); Xie and Tammi (2009, DNA copy number variation). For a comprehensive review of methods for ChIP-seq and RNA-seq data, see Pepke et al. (2009).

R and bioconductor packages

Already, a number of R (http://www.r-project.org/) and Bioconductor (http://www.bioconductor.org/) packages/tools for analyzing NGS data have been developed. The *rtracklayer* (Lawrence et al., 2009) package provides an interface between R and genome browsers. This package includes functions that import/export, track data and control/query external genome browser sessions/views. The *chipseq* (Kharchenko et al., 2008) package provides useful tools for design and analysis of ChIP-seq experiments and detection of protein-binding positions with high accuracy. These tools include functions that improve tag alignment and correct for background signals. The *Biostrings* 2 (Pages, 2009) package allows users to manipulate big strings easily and quickly by introducing new implementations and new interfaces into *Biostrings 1*. The *ShortRead* package (Morgan et al., 2009) provides useful tools for analyzing highthroughput data produced by Solexa, Roche 454, and other sequencing technologies. These tools include input and output, quality assessment, and downstream analysis functions. The *IRanges package* (Pages et al., 2009) includes functions for representation, manipulation, and analysis of large sequences and subsequences of data as well as tools for attaching information to subsequences and segments. The *BSgenome* package (Pages, 2009) provides infrastructure for accessing, analyzing, creating, or modifying data packages containing full genome sequences of a given organism. The *biomaRt* package (Durinck et al., 2006) allows users to connect to and search *BioMart* databases and integrates them with software in Bioconductor. This package includes functions that annotate identifiers with genetic information and allow retrieval of data on genome sequences and single nucleotide polymorphisms. The *ChIPpeakAnno* package (Zhu et al., 2009) provides users with facilitation tools for the batch annotation of the peaks identified from either ChIP-chip or ChIP-seq experiments. These tools include functions that find the nearest gene, exon, miRNA or transcription factor binding sites as well as identify Gene Ontology (GO) terms followed by GO enrichment test. The *TileQC* package (Dolan and Denver, 2009) can be used with Solexa output; it identifies bias and error in data by flow cell tiles through graphical means. The PICS package (Zhang et al., 2010; http://www.bioconductor.org/packages/2.6/bioc/html/PICS.html) can identify enriched regions by extracting information from ChIP-Seq aligned-read data via a Bayesian hierarchical t-mixture model. The rGADEM package (Droit et al., http://bioconductor.org/packages/2.6/bioc/html/rGADEM.html) provides users with an efficient de novo motif discovery tool for large-scale genomic sequence data. Several of these packages work in consort as shown in Figure 4.

Besides, there are several packages/tools for visualizing NGS data. The *HilbertVis* package (Anders, 2009) provides several functions for visualizing long vectors of integer data by means of Hilbert curves. The *GenomeGraphs* packages (Durinck et al., 2009) allows users to plot different data types such as array CGH, gene expression, sequencing and other data, together in one plot using the same genome coordinate system. These tools include functions to convert the Eland and Q-score data contained within the Solexa text files to a more flexible database form.

Some selected applications and statistical analyses

In earlier sections, we have mentioned several papers (and packages) developing statistical tools for use with NGS data, many of which are broad based while others are specific to certain types of applications. In this section, we selectively review a number of additional papers applying the next generation sequencing technology in a multitude of biological investigations along with brief descriptions of the statistical techniques used; each of these papers employ interesting novel statistical methods for downstream analyses of NGS data for solving the problem at hand. For a general review of applications of NGS technology, see the article by Fox et al. (2009).

In a recent article, Choi et al. (2009) used NG ChIP-seq data together with array hybridization data towards enhancing the detection of transcription factor binding sites. There are a number of

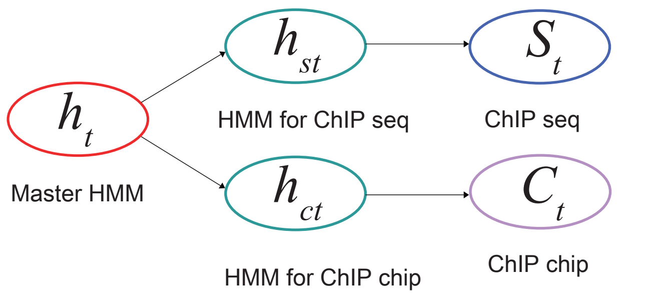

Figure 3: Joint modeling of ChIP-seq and ChIP-chip data.

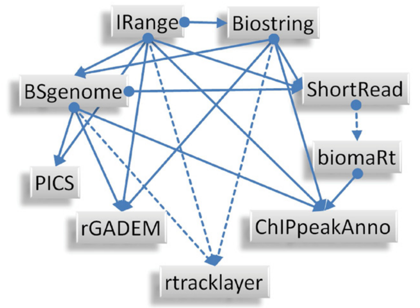

Figure 4: The dependency among the released R/Bioconductor packages. The solid lines represent the direct dependency and the dotted lines the indirect dependency.

reasons for combining these two platforms. ChIP-seq offers genome-wide coverage in a single base pair resolution at low cost; however, with ChIPseq, different mapping strategies may identify mutually exclusive peak regions as candidate binding sites and massively parallel sequencing may not work well for all DNA fragments uniformly. Other mapping methods not relying on direct sequencing, e.g. ChIP-chip, can be a valuable source to complement the weakness of the sequencing technology. See Schones and Zhao (2008) for an excellent review of various technologies and their combination for studying chromatin modifications genome-wide. This rather interesting analysis by Choi et al. (2009) uses a hierarchical hidden Markov model to combine individual hidden Markov models used with each data types. Regular hidden Markov models (HMMs) have been a standard tool in modeling ChIP-chip data (Humburg et al., 2008). The main difficulties in combining data from these two sources arise from the distinct nature of these two data types. The peaks identified by ChIP-seq are expected to form regions that are much sharper than those in ChIP-chip due to its superior resolution, whereas ChIP-chip tends to report broader regions with moderate significance including potential false positives. The signals from the two data sources have to be appropriately weighted in order to keep the overall false positive rates low and obtain good sensitivity in the joint analysis. This is done through a mostly Bayesian strategy. Individual HMMs $\{h_{st}\}$ and $\{h_{ct}\}$ are fit to both ChIP-seq data $\{S_t\}$ and ChIP-chip data $\{C_t\}$ which, in turn, are controlled by a master or hierarchical HMM $\{h_t\}$ consisting of either ChIP enriched or background states (Figure 3). The states in the individual HMMs were generated from a multinomial distribution given the emissions of the master HMM. HMM in ChIP-chip followed the uniform and normal distributions, respectively, for the ChIP enriched and the background states. The counts in ChIP-seq data in the two states were modeled by a generalized Poisson and a zero-inflated Poisson (to reflect the empty reads), respectively. Posterior probabilities of the master states are computed and a state is declared to be ChIP enriched if this probability exceeds a given threshold, say, 90%.

A similar combination of data types was used by Zang et al. (2009), who looked for spatial clusters of signals, for identification of ChIP enriched signals for histone modification profiles. Chu et al. (2009) applied whole genome sequencing to diagnose the fetal genetic disease using cell-free DNA from maternal plasma samples in the first trimester of pregnancy. Cokus et al. (2008) used NG sequencing to identify novel components of the Arabidopsis for methylation. In a rather potentially high impact application, Quon and Morris (2009) developed a statistical method to identify the primary origin of a cancer sample via next generation sequencing. This utilizes a detail profile of tissues of each primary origin and not a data based classifier. Friedländer et al. (2008) used deep sequencing technology to identify small RNAs (miRNAs). They were able to identify and experimentally validate four novel miRNAs for the worm *Caenorhabditis elegans* and altogether over two hundred potential miRNAs using data from *C. elegans*, dog and human those were previously unknown. They computed a test statistic (i.e., a score) based on the compatibility of the position and frequency of sequenced RNA with the secondary structure of the miRNA precursor. The false positive rates (or sizes) of their test were estimated using a permutation algorithm. Meng et al. (2008) studied targeted gene inactivation in zebra fish using engineered zinc-finger nucleases (ZFNs). They demonstrated that co-injection of mRNAs encoding these ZFNs (that were engineered to recognize certain sequence) into one-cell-stage zebra fish embryos led to mutagenic lesions at the target site that were transmitted through the germ line with high frequency. They showed this by comparing the Solexa sequence data from target sites versus off-target sites at each ZFN dose; Fisher's exact test was employed to test whether these two groups had different insertion/deletion rates in the sequence. Dalevi et al. (2008) considered the problem of matching individual short reads sampled from the collective genome of a microbial community to protein families. They found that assignments based on proxy-genes, where full-length protein sequences with high similarity to the translated sequences are identified, were typically more accurate than direct assignment. However, proxy-gene assignments may lead to redundancy, so hierarchical clustering was used to significantly reduce the size of the dataset while still maintaining the quality of the functional information obtained from the analysis. Use of NGS for discovering structural variation is reviewed in Medvedev et al. (2009). Very recently, Goya et al. (2010) developed novel statistical methods of predicting single nucleotide variant from NGS data using mixtures of binomial distributions to model allelic counts. Their methodologies were developed specifically to work with cancer data where earlier simpler methods (e.g., Li et al., 2008) did not work adequately.

Concluding Remarks

With new technology come new challenges for the data analysts and next generation sequencing is no exception. There seems to be a general perception that given the high quality of NGS data, replication is hardly necessary. While this may be true for technical replicates it cannot be the case with biological replicates in experiments where a conclusion is being reached about certain genetic aspect of a population from a biological sample from that population. The high dimensionality of the data makes direct use of classical statistical techniques difficult if not outright impossible. The success stories thus far seem to come from mostly Bayesian statistical techniques; however, often these are combined with frequentist

Review Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000138
JPB/Vol.3 Issue 6

calculations. In many instances, the entire analysis combines various statistical methods of varied complexities in a mostly ad hoc manner. Although, simulation studies are generally performed to demonstrate the effectiveness of the combined approach, its overall statistical properties are difficult to assess from a theoretical standpoint; in particular, no assessment of optimality of the overall statistical procedure can be assessed this way.

There is also a misconception amongst some practitioners that Bayesian methods are immune from the sample size requirement. While it is true that one can always get a Bayesian answer even with a small number of biological replicates, for good empirical statistical properties such as posterior consistency, a large sample size is necessary; this issue is directly linked with overall robustness with respect to prior misspecification and the overall reliability of the answers from a Bayesian calculation. Next generation sequencing also presents some of the same statistical challenges presented by other high throughput genomic data types, namely, high dimensionality, global error rate control, and correlation amongst counts at different sites.

The challenges described above also present new opportunity for the statisticians for collaborative (interdisciplinary) as well as methodological development in this exciting area of research. In particular, there is a need for development of systematic statistical methods that adhere to fundamental statistical principles while addressing the practical needs of the researchers. There is still scope of employing novel statistical methods as new applications to this technology emerge. Also, methods to get a better handle of the issues mentioned in the previous paragraph are needed sooner than latter. Finally, there needs to be more work towards development of study designs and establishing global statistical standard in these platforms.

Acknowledgement

We acknowledge funding from National Science Foundation (grant DMS-0805559 to Susmita Datta) and National Institute of Health (grant numbers CA133844 and 1P30ES014443 to Susmita Datta). We thank the reviewers for their constructive comments.

References

1. Anders S (2009) Visualisation of genomic data with the Hilbert curve. Bioinformatics 25: 1231-1235.
2. Boyle AP, Guinney J, Crawford GE, Furey TS (2008) F-Seq: A feature density estimator for high-throughput sequence tags. Bioinformatics 24: 2537- 2538.
3. Braslavsky I, Hebert B, Kartalov E, Quake SR (2003) Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci USA 100: 3960-3964.
4. Bravo HC, Irizarry RA (2009) Model-Based Quality Assessment and Base-Calling for Second-Generation Sequencing Data. Biometrics [Epub ahead of print].
5. Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11: 94
6. Choi H, Nesvizhskii A, Ghosh D, Qin ZS (2009) Hierarchical hidden Markov model with application to joint analysis of ChIP-chip and ChIP-seq data. Bioinformatics 25:1715-1721.
7. Chu T, Bunce K, Hogge WA, Peters DG (2009) Statistical model for whole genome sequencing and its application to minimally invasive diagnosis of fetal genetic disease. Bioinformatics 25: 1244-1250.
8. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, et al. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5: 613-619.
9. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, et al. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452: 215-219.
10. Dalevi D, Ivanova NN, Mavromatis K, Hooper SD, Szeto E, et al. (2008) Annotation of metagenome short reads using proxygenes. Bioinformatics 24: i7-i13.
11. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36: e105.
12. Dolan PC, Denver DR (2009) TileQC: A system for tile-based quality control of the Solexa data. BMC Bioinformatics 9: 250.
13. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci USA 100: 8817-8822.
14. Droit A, Gottardo R, Robertson G, and Li L rGADEM: de novo motif discovery. http://bioconductor.org/packages/2.6/bioc/html/rGADEM.html
15. Durinck S, Bullard J, Spellman PT, Dudoit S (2009) Genome graphs: integrated genomic data visualization with R. BMC Bioinformatics 10: 2.
16. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, et al. (2006) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21: 3439-3440.
17. Eid J, Fehr A, Gray J, Luong K, Lyle J, et al. (2009) Real-time DNA sequencing from single polymerase molecules. Science 323: 133-138.
18. Erlich Y, Mitra PP, delaBastide M, McCombie WR, Hannon GJ (2008) Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. Nat Methods 5: 679-682.
19. Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, et al. (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 24: 1729-1730.
20. Fox S, Filichkin S, Mockler TC (2009) Applications of ultra-high-throughput sequencing. Methods Mol Biol 553: 79-108.
21. Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, et al. (2008) Discovering microRNAS from deep sequencing data using miRDeep. Nat Biotechnol 26: 407-415.
22. Fu X, Fu N, Guo S, Yan Z, Xu Y, et al. (2009) Estimating accuracy of RNA-Sequencing and microarray with proteomics. BMC Genomics 10: 161.
23. Goya R, Sun MG, Morin RD, Leung G, Ha G, et al. (2010) SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics 26: 730-736.
24. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, et al. (2008) Single-Molecule DNA Sequencing of a Viral Genome. Science 320: 106-109.
25. Humburg P, Bulger D, Stone G (2008) Parameter estimation for robust HMM analysis of ChIP-chip data. BMC Bioinformatics 9: 343.
26. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, et al. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol 26: 1293-1300.
27. Jiang H, Wong W (2009) Statistical inferences for isoform expression in RNA-Seq. Bioinformatics 25: 1026-1032.
28. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497-1502.
29. Ju J, Kim DH, Bi L, Meng Q, Bai X, et al. (2006) Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc Natl Acad Sci USA 103: 19635-19640.
30. Kao W, Stevens C, Song Y (2009) Bayes Call: A model-based basecalling algorithm for high-throughput short-read sequencing. Genome Res 19: 1884-1895.
31. Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26: 1351-1359.
32. Kircher M, Stenzel U, Kelso J (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol 10: R83.
33. Kirkpatrick S, Gelatt CD Jr, Vecchi MP (1983) Optimization by simulated annealing. Science 220: 671-680.

34. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

35. Lawrence M, Gentleman R, Carey V (2009) rtracklayer: an R package for interfacing with genome browsers. Bioinformatics 25: 1841-1842.

36. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

37. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851-1858.

38. Mardis ER (2008) Next generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9: 387-402.

39. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.

40. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genom Res 18: 1509-1517.

41. Marsh S (2007) Pyrosequencing applications. Methods Mol Biol 373: 15-24.

42. Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 6: S13-S20.

43. Meng X, Noyes MB, Zhu LJ, Lawson ND, Wolfe SA (2008) Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. Nat Biotechnol 26: 695-701.

44. Meng XL, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80: 267-278.

45. Metzker ML (2010) Sequencing technologies — the next generation. Nat Rev Genet 11: 31-46.

46. Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, et al. (2009) Shortread: a Bioconductor package for input, quality assessment and exploration of high throughput sequence data. Bioinformatics 25: 2607-2608.

47. Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, et al. (2008) Application of massively parallel sequencing to micro RNA profiling and discovery in human embryonic stem cells. Genome Res 18: 610-621.

48. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods 5: 621-628.

49. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344-1349.

50. Oshlack A, Wakefield M (2009) Transcript length bias in RNA-sequencing data confounds systems biology. Biol Direct 4: 14.

51. Pages H (2009) BSgenome: Infrastructure for Biostrings-based genome data packages. R package version 1.12.3.

52. Pages H, Aboyou P, Lawrence M (2009) IRanges: Infrastructure for manipulating intervals on sequences. R packages version 1.2.3.

53. Pages H, Aboyoun P, Gentleman R, DebRoy S (2009) String objects representing biological sequences, and matching algorithms. Biostrings available at: http://www.bioconductor.org/packages/bioc/html/Biostrings.html

54. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods 6: S22-S32.

55. Quinlan AR, Stewart DA, Strömberg MP, Marth GT (2008) Pyrobayes: An improved base caller for SNP discovery in pyrosequences. Nat Methods 5: 179-181.

56. Quon G, Morris Q (2009) ISOLATE: A computational strategy for identifying the primary origin of cancers using high throughput sequencing. Bioinformatics 25: 2882-2889.

57. Rougemont J, Amzallag A, Iseli C, Farinelli L, Xenarios I, et al. (2008) Probabilistic base calling of Solexa sequencing data. BMC Bioinformatics 9: 431.

58. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, et al. (2009) Peakseq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 27: 66-75.

59. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463–5467.

60. Schones DE, Zhao K (2008) Genome wide approaches to studying chromatin modifications. Nat Rev Genet 9: 179-191.

61. Sharon E, Lubliner S, Segal E (2008) A feature-based approach to modeling protein-DNA interactions. PLoS Comput Biol 4: e1000154.

62. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nature Biotechnology 26: 1135-1145.

63. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, et al. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309: 1728-1732.

64. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, et al. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol Ecol 17: 1636-1647.

65. Wall PK, Leebens-Mack J, Chanderbali AS, Barakat A, Wolcott E, et al. (2009) Comparison of next generation sequencing technologies for transcriptome characterization. BMC Genomics 10: 347.

66. Wei GCG, Tanner MA (1990) A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. J Am Stat Assoc 85: 699-704.

67. Whiteford N, Skelly T, Curtis C, Ritchie ME, Löhr A, et al. (2009) Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics 25: 2194-2199.

68. Xie C, Tammi MT (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics 10: 80.

69. Zang C, Schones DE, Zeng C, Cui K, Zhao K, et al. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25: 1952-1958.

70. Zhang X, Gottardo R, Droit A (2010) Probabilistic inference of ChIP-seq. PICS available at: http://www.bioconductor.org/packages/2.6/bioc/html/PICS.html

71. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. (2008a) Model-based Analysis of ChIP-Seq (MACS). Genome Biol 9: R137.

72. Zhang ZD, Rozowsky J, Snyder M, Chang J, Gerstein M, et al. (2008b) Modelling ChIP Sequencing In Silico with Applications. PLoS Comput Biol 4: e1000158.

73. Zhu LJ, Pages H, Gazin C, Lawson N, Lin S, et al. (2009) Batch annotation of the peaks identified from either ChIP-seq or ChIP-chip experiments. ChippeakAnno available at: http://www.bioconductor.org/packages/2.5/bioc/html/ChIPpeakAnno.html