Conference on Applied Statistics in Agriculture 2009 - 21st Annual Conference Proceedings

STATISTICAL ISSUES IN NEXT-GENERATION SEQUENCING

Paul L. Auer

R. W. Doerge

STATISTICAL ISSUES IN NEXT-GENERATION SEQUENCING

Paul L. Auer and R.W. Doerge
Department of Statistics, Purdue University, West Lafayette, IN 47907-2066

Abstract

High throughput deep-sequencing or next-generation sequencing has emerged as an exciting new tool in a great number of applications (e.g., variant discovery, profiling of histone modifications, identifying transcription factor binding sites, resequencing, and transcriptome characterization). Even though this technology has generated unprecedented amounts of data in the scientific community few studies have looked carefully at its inherent variability. Recent studies of mRNA expression levels found little appreciable technical variation in Illumina's Solexa sequencing platform (a next-generation sequencing device). Although these results are encouraging, they are limited to a specific platform and application, and have been made without any attention to experimental design. This paper provides an overview of some key issues in data management and experimental design related to Illumina's Solexa Genome Analyzer technology.

Keywords: next-generation sequencing, RNA-Seq, experimental design

1. Introduction

Over the last two years there has been an increasing need for statistical assistance (i.e., consulting projects) in dealing with next-generation sequencing (NGS) data. We do not expect this trend to ease. Given that applications of NGS, in the Statistics, Genomics, and Bioinformatics literature grew by a factor of ten [1] from 2007 to 2008, it is not surprising that a commonly expressed opinion holds that NGS will replace microarrays within the next few years [2]. If NGS is indeed the future of science, then it is incumbent upon statisticians who regularly consult with biologists to familiarize themselves with NGS technology, the questions that scientists are asking, and data that arise.

In order to understand NGS and its applications, it is imperative to gain an appreciation of the history and goals of DNA sequencing. Every cell in every living organism contains instructions for its function and development via its genetic code (called its genome). A genome is made up of a sequence of four nucleic acids, adenine, guanine, cytosine, and thymine (i.e., A, G, C, and T, respectively). In the 1970's new biochemical techniques, known as "traditional" or "Sanger" sequencing, were developed [3, 4] to rapidly identify the DNA code from a sample of an organism's genome (i.e., the sequence of A's, T's, C's, and G's that comprise the genome). Sanger sequencing enjoyed a near monopoly in the biological community until just after the completion of the Human Genome Project (HGP) earlier this decade, and it continues to be the most reliable technique for DNA sequencing. However, because of time and cost constraints, newer NGS technologies have entered the market. Currently, there are three commercially

available NGS technologies, the Genome Sequencer FLX system (GS FLX) produced by 454 Sequencing, Illumina's Solexa Genome Analyzer, and Applied Biosystem's SOLiD platform. In contrast to The Human Genome Project, which was a multi-million dollar, decade long collaborative effort that sequenced the human genome using Sanger sequencing [5], sequencing of the Neandertal genome (similar in size, structure and complexity to the human genome) took less than three years at a fraction of the cost [6] using NGS technologies in a single laboratory. In fact, right now (June 2009) third generation ("next-next" or "next^2") sequencing technology is on the horizon and is focused on the $1000 human genome (i.e., sequencing individual human genomes in a few days for less than $1000 US). The impact of sequencing individuals is revolutionizing personalized medicine [7].

Historically, DNA sequencing has been used in a variety of applications to answer diverse biological questions. NGS has followed this path, having been successfully employed in experiments mapping epigenetic modifications [8, 9], characterizing transcriptomes [10], and assessing differential expression [11]. One popular application is called RNA-Sequencing (RNA-Seq), which uses NGS technologies to characterize and quantify the collection of transcripts in a cell.

2. Overview of RNA-Seq using the Illumina Genome Analyzer

One of the most important cellular functions of DNA is the production of proteins, the primary determinants of biological form and function [12]. A protein consists of a chain of one or more amino acids, which in turn are encoded in codons, or triplets of nucleotides in a DNA sequence [12]. DNA is transcribed into ribonucleic acid (RNA) which is then translated into protein. This information transfer from DNA to protein is known as the "Central Dogma of Molecular Biology" (Figure 1) [13].

Both RNA and DNA are nucleic acids. RNA is typically single-stranded, has ribose sugar in its nucleotides (rather than deoxyribose), contains the nucleotide uracil (U) instead of thymine, and unlike DNA has the ability to catalyze biological reactions [12]. There are two general classes of RNAs, those that encode proteins (called messenger RNA, mRNA) and those that are functional as RNA. Interestingly, RNA can be isolated and measured to infer both the expression of genetic material (i.e., genes) into protein (mRNA), as well as the function of cellular processes.

RNA-Seq experiments begin by isolating RNA from cells. Each RNA strand can be hundreds to thousands of bases long and is fragmented at random positions and copied into complementary DNA (cDNA). In preparation for sequencing, adapters are attached to the ends of the cDNA fragments. Fragments meeting a certain size specification (e.g., 200-300 bases long) are retained for amplification using Polymerase Chain Reaction (PCR). After amplification the cDNA sample is sequenced using any one of a number of NGS technologies. A more detailed overview of this process can be found in [14, 10].

The Illumina Genome Analyzer is a sequencing technology that consists of a flow-cell (a glass slide) containing eight vertical lanes, each of which is capable of sequencing independent

genomic samples. The Illumina technology sequences cDNA fragments one base at a time until it reaches the 36th base (as of June 2009, Illumina read lengths have increased from 36 bases and are approaching 100 bases). In this way, the first 36 bases of millions of template molecules are sequenced in parallel on a single flow-cell lane. The raw sequencing data from a single lane contain sequencing reads of fixed length with quality scores for each base. The quality scores reflect the confidence with which the Illumina machine assigned a base call to a sequence position. Taken alone, the raw data are somewhat meaningless because the genomic location of each sequence is not known. In order to connect the reads to the genome it is necessary to map them to their location in the genome. Mapping raw sequencing reads constitutes a major computational challenge that currently dominates a sizable portion of the Bioinformatics literature.

3. RNA-Seq Data Processing and Normalization

A "reference genome" represents the current state of knowledge regarding a particular species' genome. In a sense, a reference genome is a continually evolving entity that accumulates information as more individuals from the same species are sequenced. It is, more or less, the consensus genome of all published sequences of individuals in a given species. The reference genome provides the context in which to interpret sequencing reads from an RNA-Seq experiment. To do so, the raw sequencing data are "aligned" to the reference genome by parsing the entire genome for regions that match the sequencing reads. Sequencing reads that match multiple genomic regions are rendered ambiguous. Occasionally, a sequencing read will fail to match any region in the reference genome implying either a mistake in the reference or inaccurate base calls in the sequencing read. Given the size of the sequencing libraries (tens of millions of reads) and the size of reference genomes (tens of millions to billions of bases) the computational challenges involved in alignment seem almost insurmountable. Fortunately the Bioinformatics literature is rich with fast and accurate alignment tools such as ELAND (Illumina product), MAQ [15], and SOAP [16] among others. However, the success of any alignment algorithm is entirely dependent on the available knowledge of a particular species' genome (i.e., its reference genome). For instance, the *Arabidopsis Thaliana* (*AT*) genome is relatively small and very well characterized. The reference sequence is "complete" in some sense, whereas the *Triticum* (wheat) genome is relatively large, highly repetitive, and poorly characterized. It stands to reason that sequencing reads taken from an *AT* sample will align to the *AT* reference at a much higher rate than reads taken from a *Triticum* sample and aligned to the *Triticum* reference. Alignment affects all downstream analyses, so it is important to note that the inferences from an analysis are dependent upon and limited by the available knowledge of an organisms' genome.

3.1 Aligned Sequencing Data

Table 1 illustrates the first few lines (of about 5 million lines total) from an ELAND aligned data file. The first field in the first row shows the sequence of the first 36 bases of a random fragment from the genomic sample. The code "R1" indicates that this sequence mapped to several different locations on the reference genome. The next fields indicate that the sequence mapped nowhere perfectly (i.e., 0), at 32 different places with a one base discrepancy and at 255 different

places with a two base discrepancy. The next row shows the sequence of the first 36 bases of a different random fragment from the same genomic sample. The code "U0" indicates that the sequence mapped perfectly to exactly one location on the reference genome. That location is 90,577,824 bases from the 3' end of chromosome 14. The next line is a third sequencing read from the same genomic sample. The code "NM" indicates that this sequence matches nowhere on the reference genome using a two base discrepancy tolerance. The final line (Table 1) provides a read with the code "U1," which indicates that this sequence matches exactly one spot on the reference genome with a one base discrepancy. That location is 45,758,959 bases from the 3' end of chromosome 1. The discrepancy occurs at base position 32 and appears as a "G" in the reference genome.

Once the results from an alignment program are in hand, gene expression is quantified relative to the annotation of the reference genome. Typically, reference genomes are annotated with known functional elements (e.g., genes and promoter regions). This annotation, especially for a gene, occupies a specific region in the reference genome. Therefore it is possible, for each gene, to count the number of times a sequencing read from the alignment file falls within that gene's annotated region. Using this counting approach gene expression is quantified for every gene in the reference genome.

3.2 Data Reduction

Recall that the lanes on the Solexa sequencing platform are independent and that independent genomic samples are loaded into the different lanes. Each lane produces a file of raw sequencing reads and each of these files is aligned to a reference genome, independently, as just described. In a typical RNA-Seq experiment, the files from the alignment occupy approximately 1GB of disk-space per sample (or per lane) and can become unwieldy since most experiments have at least six independent samples (i.e., using at least 6GB). While the alignment file may require 1GB per sample, the file summarizing the per gene expression quantification requires 10MB or less of disk-space making it easy to work with on a laptop or PC with standard statistical software (e.g., R or SAS). The process of translating the aligned sequence reads (per gene) to a manageable data file is easily accomplished with the following UNIX command line:

```
awk '{print($4);}' alignmentfile | sort -n | uniq -c | awk '{print($2,$1);}'  > table.txt
```

This code assumes that the "alignmentfile" is in the format of Table 2. The output (Table 3) from the code, "table.txt," appears as a single column summarizing the gene expression counts into a gene expression matrix. Notice that Table 3 adheres to the standard format for a typical gene expression analysis from a microarray experiment. Of course, by reducing the alignment files into a gene expression matrix an enormous amount of information is discarded or ignored (e.g., allele specific expression, alternative splicing, unknown transcription events, and exon level expression). Specifically, reads mapping to multiple locations are removed and may reduce the data file up to 40% [10, 11]. Fortunately, RNA-Seq experiments are focused solely on testing differential expression, therefore only the gene expression matrix is required for the statistical analysis.

3.3 Normalizing RNA-Seq Data

Similar to microarray based gene expression experiments, the "parameter being measured is many steps removed from the parameter being inferred" [17]. Recall that in a typical RNA-Seq experiment, cells are isolated, RNA is harvested, randomly fragmented, copied into cDNA, amplified, loaded into a sequencing device which in turn amplifies the sample again, and the sequencing device then uses laser excitation along with fluorescently labeled nucleotides to decode the sequence. The sequence is then analyzed with an alignment program which effectively labels the sequencing reads with annotation from a reference genome. From this point a frequency table summarizing the annotated sequencing reads represents a measure of gene expression for any gene present in the reference database. Clearly, with so many steps involved, experimental errors and computational assumptions accumulate, all of which affect the accuracy of the gene expression quantification. For RNA-Seq experiments these distortions occur on a per-sample (or lane) basis making it necessary to rely on normalization methods to make samples comparable.

If we let y_{gi} denote the gene expression quantification for the g^{th} gene in the i^{th} sample then, as reasoned by Sebastiani et al. [17], the observed gene expression y_{gi} masks the true expression level $\tilde{y}_{gi}$ had all samples been conducted under the exact same experimental conditions. Thus,

$$y_{gi} = f(\tilde{y}_{gi}), \tag{1}$$

and normalization consists of estimating $f(\cdot)$ for the purpose of recovering

$$\tilde{y}_{gi} = f^{-1}(y_{gi}). \tag{2}$$

Currently, there are two standard normalization techniques in the RNA-Seq literature, quantile normalization [18] and "Reads Per Kilobase of exon model per Million" (RPKM) [10] mapped reads. The quantile normalization method [19] gained popularity for the analysis of microarray data because it is computationally fast, it is easy to understand, and it is effective. The goal of quantile normalization is to make the distribution of gene expression measurements the same across samples by substituting the respective quantile means, for each of G genes, for the original data. By forcing the tails of the distributions to be the same across samples, gene expression values on the endpoints of the range are made identical across samples. As such, quantile normalization can be problematic when working in the tails of the distribution [19] if the data do not warrant this sort of adjustment. In fact, with respect to RNA-Seq data, it is not wise to use quantile normalization because these data enjoy a characteristic called "dynamic range." Specifically, there is no background noise in RNA-Seq data, so genes with very low expression values (0-10) and genes with very high expression values (1,000 or more) provide reliable data that can all be used to test differential expression. If the tails of the distributions were forced to be the same across samples the gene expression values on the endpoints of the range would be identical across samples, thus robbing RNA-Seq data of one of its most advantageous features, sensitivity at the endpoints of the data range (i.e., "dynamic range").

RPKM was introduced specifically for normalizing RNA-Seq data. It adjusts/divides each cell in the gene expression matrix (Table 3) by both the corresponding column total and gene size which allows for inter-gene and inter-sample comparisons. "Sequencing depth" represents the redundancy with which a single fragment is sequenced and is often different both within and between samples. Dividing each cell (Table 3) by the column total corrects for differential sequencing depth between samples. Because the isolated RNA is randomly sheared (early in the process), it is expected that longer strands of RNA will produce a greater number of random fragments than smaller strands of RNA. Therefore, in order to make accurate inter-gene comparisons, the RPKM technique divides each cell by the corresponding gene size. Although this last step is necessary for inter-gene comparisons, the column total is a poor substitute for an accurate per fragment estimate of depth. To date there is no consensus on how to estimate fragment level depth, and it is still not clear that counting redundant fragments adds accuracy to the measure of gene expression, since there is very small probability that a random shearing mechanism would cut two strands of RNA in the identical position.

4. Statistical Design and Analysis

In any experimental design the experimental unit constitutes the fundamental quantity for analysis. In an RNA-Seq experiment, independent genomic samples are loaded in different lanes of the flow-cell, thus lane can be considered the experimental unit. Consider a situation where RNA-Seq data have been collected from two treatment groups A and B for the purpose of testing differential expression. Suppose that each treatment group contains four independent biological replicates. Then a reasonable experimental design would randomly assign each of the 8 samples to a lane (experimental unit) on the flow-cell (Table 4). Randomizing and replicating across lanes provides the best protection against systematic lane effects. Of course with such a simple design, the statistical analysis is straightforward and uses the model

$$f(Y_{ijk}) = \mu + T_i + G_j + (TG)_{ij} + \varepsilon_{ijk}$$
$$\varepsilon_{ijk} \sim N(0, \sigma^2), \quad (3)$$

where Y_{ijk} is the normalized gene expression measure from the k^{th} biological replicate of the j^{th} gene from the i^{th} treatment group, i=1,2, j=1,...g, and k=1,…,4. T is the treatment effect, G is the gene effect, and TG is the treatment by gene interaction. The function f is generally a variance stabilizing transformation and ε_{ijk} is the random unexplained variation. One can either assume constant variance across genes, or a per gene variance, the former is rarely true and the latter lacks statistical power. There are methods that find a compromise between these two approaches [20]. Nevertheless, the null hypothesis testing differential expression of gene j between treatment groups [21] is

$$H_{0j} : T_1 + (TG)_{1j} = T_2 + (TG)_{2j}. \quad (4)$$

4.1 Experimental Design and Reproducibility of RNA-Seq Results

Results from any analysis are only scientifically valuable if they are reproducible. Evaluating the reproducibility of RNA-Seq results entails a comprehensive study of the technical variation

in the experimental process that produced the data. There has been some preliminary work on technical variation in RNA-Seq data [11] as well as comparisons of results between microarrays and NGS [22]. To date, virtually no attention has been paid to the design of RNA-Seq experiments. This is disappointing when one considers the significant contributions that proper statistical design lent to microarray experiments [23, 24]. The cost constraints associated with RNA-Seq almost certainly play a role, especially considering that a single NGS run using a single lane of an Illumina sequencer costs approximately $1,500 (after the machine has been purchased). Nevertheless, in anticipation of cheaper sequencing runs and to illustrate how such an experiment should be designed, we continue with our previous example.

Suppose that for each of the 8 biological samples, each sample is replicated 4 times, giving a total of 32 samples (8 biological replicates, each with 4 technical replicates). When deciding on an experimental design, the decision is often based on optimizing an objective function of the design space. D-optimality [25] is often used as a criterion, where the D-optimal design utilizes the design matrix X and maximizes the determinant of the X'X matrix. For the example under consideration, Table 5 illustrates a D-optimal design. Experimental designs utilizing technical replicates can be thought of as repeated measures designs where the order of the repeated measure is inconsequential, thus making it a split-plot design. When dealing with a split-plot design there are two different experimental units, the whole-plot experimental unit (the biological sample) and the sub-plot experimental unit (the technical replicate). Reorganizing Table 5 illustrates (Figure 2) the advantages of this D-optimal design. Specifically, the flow-cell itself forms the whole-plot block allowing one to account for the variation between flow-cells (or between sequencing runs). Using this D-optimal design (Table 5) and testing per-gene differential expression can be accomplished with the following model

$$\begin{aligned} f(Y_{ijk}) &= \mu + T_i + R_j + \delta_{ij} + L_k + (TL)_{ik} + \varepsilon_{ijk} \\ \delta_{ij} &\sim N(0, \sigma_\delta^2) \\ \varepsilon_{ijk} &\sim N(0, \sigma_\varepsilon^2), \end{aligned} \tag{5}$$

where Y_{ijk} is the observed gene expression in the kth technical replicate of sequencing run j, in treatment group i, i=1,2, j=1,...,4, and k=1,...,4. T is the whole plot factor (treatment effect), R is the whole plot block (sequencing run or flow-cell), δ is the whole plot error, L (lane) is the sub-plot factor (technical replicate), and ε is the sub-plot error. Table 6 shows the ANOVA table for this design (Table 5) with appropriate degrees of freedom (df) and expected mean squares (EMS). Differential expression can be tested using

$$\begin{aligned} H_0 &: T_1 = T_2 \\ \frac{MS(T)}{MS(\delta)} &\sim F_{1,3}. \end{aligned} \tag{6}$$

Lane effect (i.e., technical variation) can be tested using

$$\begin{aligned} H_0 &: L_1 = L_2 = L_3 = L_4 \\ \frac{MS(L)}{MS(\varepsilon)} &\sim F_{3,18}. \end{aligned} \tag{7}$$

Although this is a hypothetical example that simultaneously studies differential expression and technical reproducibility, it illustrates some of the statistical issues involved.

5. Discussion

NGS technology continues to provide statisticians with yet another unique opportunity to contribute to science. While this paper focuses on one specific application of NGS (RNA-Seq) using one specific sequencing device (Illumina's Solexa sequencer) much of what is presented and discussed generalizes to other applications and technologies.

NGS technologies were initially developed to make resequencing projects faster and less expensive. Resequencing has led to the continual updating of reference genomes (across the taxonomy of life) by focusing on the accuracy of sequencing reads and their alignment to the reference (unlike RNA-Seq which focuses on the abundance of particular DNA fragments). NGS applications in epigenomics using the Chromatin Immunoprecipitation Sequencing (ChIP-Seq) technique to investigate epigenetic events (e.g., histone modifications and DNA methylation) are also quite common [8, 9]. Although both ChIP-Seq and RNA-Seq rely on NGS to quantify DNA fragments, these data emanate from two entirely different biological processes making the respective normalization methods and statistical analyses distinct. Even though these applications are vastly different, issues that are central to technical variation and reproducibility are entirely relevant to all three (resequencing, RNA-Seq, ChIP-Seq) yet need to be addressed individually.

SOLiD sequencing technology [26] is similar in many respects to Illumina's Solexa technology. Both technologies are flow-cell based, rely on eight lanes per flow-cell, and produce sequencing reads of similar length (30-60 bases) that give rise to about 1GB of data per lane [18]. Although the experimental protocols and details of the sequencing reactions are quite different between the two platforms, from an analysis perspective both Solexa and SOLiD data are similar in alignment, normalization, and analysis. A third technology based on a completely different biochemical approach, is the 454 GS FLX sequencing platform [26]. It stands alone in many respects. The 454 sequencing reaction uses a technique called "pyro-sequencing," and the device itself does not have lanes or flow-cells per-se. Furthermore, each sequencing run produces a magnitude less data (100MB) and read lengths are much longer (200 bases or longer) [26]. The corresponding alignment of 454 reads is therefore more trustworthy than those obtained from a Solexa or SOLiD sequencing run. Read length notwithstanding, the Solexa and SOLiD platforms produce much higher average depth of coverage per input fragment, and thus enjoy a decided advantage over 454 data in sensitivity (i.e., dynamic range).

Regardless of the application or choice of NGS platform, there are several realities that both scientists and statisticians must accept. First, although the cost of sequencing has dropped dramatically over the past decade, sequencing is still quite expensive ($1,500 US for a single lane on a Solexa machine) and replication is often considered an unaffordable luxury. R.A. Fisher [27] offers articulate advice that has withstood the test of time. Namely, without appropriate replication "perhaps these should not be called experiments at all, but be added

merely to the body of experience on which, for lack of anything better, we may have to base our opinions". Second, even though NGS is replacing microarrays as the preferred platform in high-throughput biology, there is no consensus as to the magnitude of technical variation in NGS devices. Mostly because the expense of doing these experiments, in biologists' opinions, outweighs the worth of the information gained. Finally, the shear magnitude of the raw data provided from NGS platforms requires considerable computational and bioinformatic finesse and may overwhelm any unprepared analyst. Specifically, a laptop or personal computer simply cannot provide the computing requirements (memory or RAM) necessary for the bioinformatics and statistical analyses. Moreover, statistical software packages (e.g., SAS, R, and STATA) are ineffective tools for carrying out the necessary bioinformatics. We have found that a 64-bit Linux server with 32GB of RAM and 500GB of disk-space along with a working knowledge of Perl and UNIX is a reasonable place to start.

Some resources are available for statisticians that want to become involved in this new and fast paced world of NGS. Developers of Bioconductor [28], at the R-project [29], are developing infrastructure for dealing with NGS data (e.g., "chipseq" and "shortread" libraries). Furthermore, there is currently an effort to establish a Minimum Information for Sequencing Experiments (MINSEQE) [30] standard (similar to MIAME [31]) which will undoubtedly play a role in determining the format of publicly available sequencing datasets. Lastly, the journal Bioinformatics has set up a repository for published journal articles that deal with NGS [1].

6. Glossary of Sequencing Terms

Base Call - The process by which an image (fluorescence) from a sequencing device is interpreted as one of four nucleotides (A, T, C, or G). This is usually accomplished by image recognition software that is part of the sequencing platform.

Chromatin Immunoprecipitation Sequencing (ChIP-Seq) - An experimental method which uses NGS to sequence, map, and quantify a ChIP product. This is a useful technique for identifying transcription factor binding sites. ChIP-seq is a recent alternative to ChIP-chip which uses microarrays to study epigenetic events.

Flow-cell - A glass slide onto which genomic samples are attached. Both Solexa and SOLiD employ these glass slides with their respective sequencing technologies.

Human Genome Project - A collaborative effort, spanning more than a decade, culminating in a draft sequence of the human genome in 2004.

Illumina Solexa Genome Analyzer - A NGS technology that is popular for RNA-Seq and ChIP-Seq experiments. It produces short sequencing reads (36 bases) from millions of DNA fragments.

Next-Generation Sequencing (NGS) - Sequencing technologies developed since 2000 which sequence DNA in a highly parallel fashion. Read lengths are typically shorter than with Sanger

sequencing, but is much less expensive (per base) and much higher throughput (1,000-10,000 times higher than Sanger sequencing).

Sanger/Traditional Sequencing - The experimental method developed in the 1970's that identifies the DNA code from a sample of an organism's genome. This method can sequence 1,000 bases at a time and is highly accurate. However, it is expensive and low throughput in the sense that it only processes at most 100 fragments in parallel.

Reads Per Kilobase of exon model per Million mapped reads (RPKM) - Introduced in Mortazavi et al. [10], it is the most popular normalization technique for RNA-Seq data. Essentially, it divides each gene count in each sample by the length of the gene and the number of reads that mapped back to the reference sequence in that sample

Reference Genome - The consensus genome of all published sequences in a given species. It is a continually evolving entity that accumulates information as more individuals from the same species are sequenced.

RNA-Sequencing (RNA-Seq) - An experimental method which uses NGS to sequence, map, and quantify a sample of transcripts isolated from a cell. It is a recent alternative to microarrays for studying differential expression.

SOLiD - A NGS technology, produced by Applied Biosystems, similar in many respects to Solexa technology. Read lengths and throughput are similar, but the biochemical techniques are quite different.

454 GS FLX - A NGS technology, produced by Roche. It is somewhat different in technology than that produced by SOLiD and/or Solexa. It uses a biochemical technique called "pyro-sequencing," producing longer reads (500 bases) with lower throughput (100,000 reads).

7. Summary

Statistical consulting projects that involve the design and analysis of NGS experiments are quickly becoming commonplace. Fortunately, because of similarities between NGS data and microarray data, the learning curve for statisticians, analysts, and bioinformaticans has been less steep than with microarrays. However, NGS data ups the ante by increasing the data file size by at least an order of magnitude. Even though the Bioinformatics and Biostatistics communities have risen to the challenge of dealing with the unprecedented amounts of data offered by microarray technology, NGS represents the next step in high-throughput biology and may prove to be more challenging. As such, research pertaining to the experimental design, pre-processing, and analysis of NGS data must keep pace. Here, we have only begun to briefly summarize the current state of NGS technology. In order for NGS to efficiently revolutionize genomics and personalized medicine, statisticians, bioinformaticians, and analysts alike must remain actively involved in this fast paced and rapidly developing field.

8. Acknowledgements

We thank Professors Rob Martienssen and Scott Jackson for their patience in answering many well intended, yet naive questions about biology and sequencing, and for providing us with a wealth of data. We also thank the RWD research group for their suggestions and support, as well as Doug Crabill and My Truong for their computational guidance. This work was funded by a NSF Plant Genome grant to RWD (DBI-0733857).

9. References

[1] Bioinformatics Next-Generation Sequencing Virtual Issue: http://www.oxfordjournals.org/our_journals/bioinformatics/nextgenerationsequencing.html

[2] Shendure, J. The beginning of the end for microarrays? *Nature Methods*, **2008**, *5*, 585-587.

[3] Sanger, F. et al. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, **1977**, *265*, 687-95.

[4] Sanger, F. & Coulson, A. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, **1975**, *94*, 441-448.

[5] International Human Genome Sequencing Consortium, *Nature*, **2001**, *409*, 860-921.

[6] The Neandertal Genome Project: http://www.eva.mpg.de/neandertal/index.html.

[7] Shendure, J.; Mitra, R. D.; Varma, C. & Church, G. M. Advanced sequencing technologies: methods and goals. *Nature Reviews Genetics*, **2004**, *5*, 335-344.

[8] Mikkelsen, T. S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **2007**, *448*, 553-560.

[9] Valouev, A. et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods*, **2008**, *5*, 829-834.

[10] Mortazavi, A.; Williams, B. A.; McCue, K.; Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **2008**, *5*, 621-628.

[11] Marioni, J. C.; Mason, C. E.; Mane, S. M.; Stephens, M. & Gilad, Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, **2008**, *18*, 1509-1517.

[12] Griffiths, A. J. F.; Wessler, S. R.; Lewontin, R. C. & Carroll, S. B. Tenney, S. (2008) *Introduction to Genetic Analysis*, 9th ed. New York: W.H. Freeeman and Company**.**

[13] Crick, F. Central Dogma of Molecular Biology. *Nature,* **1970**, *227*, 561-563.

[14] Morozova, O. & Marra, M. A. Applications of next-generation sequencing technologies in functional genomics. *Genomics,* **2008**, *92*, 255-264.

[15] Li, H.; Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research,* **2008**, *18*, 1851-1858.
[16] Li, R.; Li, Y.; Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics*, **2008**, *24*, 713-714.

[17] Sebastiani, P.; Gussoni, E.; Kohane, I. S. & Ramoni, M. F. Statistical Challenges in Functional Genomics. *Statistical Science,* **2003**, *18*, 33-70.

[18] Cloonan, N. et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods,* **2008**, *5*, 613-619.

[19] Bolstad, B.; Irizarry, R. A.; Astrand, M. & Speed, T. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **2003**, *19*, 185-193.

[20] Smyth, G. K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, **2004**, *3*.

[21] Craig, B. A.; Black, M. A. & Doerge, R. W. Gene Expression Data: The technology and statistical analysis. *Journal of Agricultural, Biological, and Environmental Statistics*, **2003**, *8*, 1-28.

[22] 't Hoen, P. et al. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Research,* **2008**, *36*, e141.

[23] Churchill, G. A. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics*, **2002**, *32*, 490-495.

[24] Kerr, M. K. & Churchill., G. A. Experimental design for gene expression microarrays. *Biostatistics*, **2001**, *2*, 183-2001.

[25] Goos, P. (2002) *The Optimal Design of Blocked and Split-Plot Experiments*. New York: Springer-Verlag.

[26] Mardis, E. R. Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics,* **2008**, *9*, 387-402.

[27] Fisher, R. A. (1951). *The Design of Experiments,* 6th ed. Edinburgh: Oliver and Boyd.

[28] Gentleman, R. C. et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 2004, *5*, R80.

[29] R Development Core Team. R: A Language and Environment for Statistical Computing. 2009. http://www.R-project.org.

[30] Minimum Information about a high-throughput SeQuencing Experiment - MINSEQE *(Draft Proposal).* http://www.mged.org/minseqe/.

[31] Brazma, A. et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics,* 2001, *29*, 365 - 371.

10. Tables and Figures

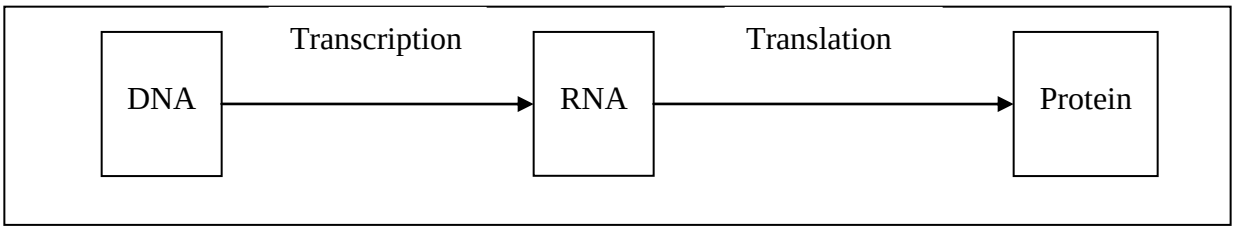

Figure 1: The Central Dogma of Molecular Biology. DNA is transcribed into RNA and then translated to protein.

Table 1: Example of four lines from output of the ELAND alignment tool. The first column shows the sequencing read, the second column displays a code denoting if and how the sequencing read aligns to the reference genome (R1=matches multiple positions in the reference genome, U0=matches one position in the reference genome perfectly, NM=no match in the reference genome, and U1=matches one position in the reference genome with a single base discrepancy). Columns 3-5 show the frequency with which the read maps with no discrepancy, a one base discrepancy, and a two base discrepancy, respectively. Column 6 shows the chromosome and position from its 3' end at which the sequencing read aligns. The last column details the position of the base, in the sequencing read, at which a discrepancy exists and the corresponding base found in the reference.

CAATAAAGAACCTACCAACCAAAAAATGCTCTGGAT	R1	0	32	255		
GATCTGAAGTGAAGAAGATTGAGACACAAAAAAATT	U0	1	0	0	chr14 90577824	
GATCTACTCATTCAGCATCTGCATCTCATCACATCC	NM	0	0	0		
CGAGCAAAGTAATGAACATATCTGTCACCTGATGTA	U1	0	1	0	chr1 45758959	32 G

Table 2: Example of an annotated alignment file. The first column shows the sequencing read, the second column displays a code denoting if and how the sequencing read aligns to the reference genome (R1=matches multiple positions in the reference genome, U0=matches one position in the reference genome perfectly, NM=no match in the reference genome, and U1=matches one position in the reference genome with a single base discrepancy). Column 3 shows the chromosome and position from its 3' end at which the sequencing read aligns. The last column shows the gene that resides on the chromosome and position in column 3.

CAATAAAGAACCTACCAACCAAAAAATGCTCTGGAT	R1		
GATCTGAAGTGAAGAAGATTGAGACACAAAAAAATT	U0	chr14 90577824	Gene 1
GATCTACTCATTCAGCATCTGCATCTCATCACATCC	NM		
CGAGCAAAGTAATGAACATATCTGTCACCTGATGTA	U1	chr1 45758959	Gene 99
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.

Table 3: Example of a gene expression matrix. For each gene, the number of sequencing reads mapping to that gene is tabulated per lane along with the total gene size (last column). The RPKM normalization technique divides each cell in the matrix by the corresponding column total (per 10^6), and gene size (per 10^3). For instance, 100 is converted into $\frac{100*10^6*10^3}{1256723*3250}$.

	$Lane_1$	$Lane_j$	Size(kbp)
$Gene_1$	100	2	3,250
.	.	.	.
.	.	.	.
.	.	.	.
$Gene_g$	16	1,257	163
Total	1,256,723	3,561,006	

Table 4: A completely randomized design with treatment groups A and B with (four) biological replicates A_1, …, A_4, B_1, … , B_4, on a single flow-cell. The eight samples are randomly assigned to the eight Lanes of a flow-cell.

Lane 1	Lane 2	Lane 3	Lane 4	Lane 5	Lane 6	Lane 7	Lane 8
A_1	B_4	B_2	A_2	B_3	A_4	B_1	A_3

Table 5: A D-optimal split plot design using four Illumina flow-cells (i.e., sequencing runs). There are two treatment groups A and B with (four) biological replicates A_1, …, A_4, B_1, … , B_4, and four technical replicates per biological replicate (technical replicates are randomly assigned to the eight Lanes in a flow-cell). This design uses the flow-cell as the whole-plot block, the biological replicate as the whole-plot experimental unit, and the technical replicate as the sub-plot experimental unit.

Run	Lane 1	Lane 2	Lane 3	Lane 4	Lane 5	Lane 6	Lane 7	Lane 8
1	A_1	B_1	B_1	A_1	B_1	A_1	B_1	A_1
2	A_2	B_2	A_2	A_2	B_2	B_2	B_2	A_2
3	B_3	A_3	B_3	A_3	A_3	B_3	A_3	B_3
4	A_4	B_4	B_4	B_4	A_4	A_4	A_4	B_4

Run 1	
A_1	B_1
Lane 1	Lane 2
Lane 4	Lane 3
Lane 6	Lane 5
Lane 8	Lane 7

Run 2	
A_2	B_2
Lane 1	Lane 2
Lane 3	Lane 5
Lane 4	Lane 6
Lane 8	Lane 7

Run 3	
A_3	B_3
Lane 2	Lane 1
Lane 4	Lane 3
Lane 5	Lane 6
Lane 7	Lane 8

Run 4	
A_4	B_4
Lane 1	Lane 2
Lane 5	Lane 3
Lane 6	Lane 4
Lane 7	Lane 8

Figure 2: A D-optimal split plot design illustrating whole-plot blocks over four Illumina flow-cells (i.e., sequencing runs). There are two treatment groups (A and B) with (four) biological replicates A_1,...,A_4 and B_1,...,B_4, and four technical replicates per biological replicate (technical replicates are randomly assigned to the eight Lanes in a flow-cell). This design uses the flow-cell as the whole-plot block, the biological replicate as the whole-plot experimental unit, and the technical replicate as the sub-plot experimental unit.

Table 6: The degrees of freedom (df) and Expected Mean Squares (EMS) for a D-optimal split plot design with two whole-plot treatment groups (T_1, T_2), four whole-plot blocks (R_1,..., R_4), whole-plot error δ, four sub-plot treatment groups (L_1,..., L_4), sub-plot treatment by whole-plot treatment interaction (TL), and sub-plot error ε.

	df	**EMS**
Wholeplot		
T_i	1	$\sigma_\varepsilon^2 + 4\sigma_\delta^2 + 16 f(T)$
R_j	3	$\sigma_\varepsilon^2 + 4\sigma_\delta^2 + 8\sigma_R^2$
δ_{ij}	3	$\sigma_\varepsilon^2 + 4\sigma_\delta^2$
Subplot		
L_k	3	$\sigma_\varepsilon^2 + 8 f(L)$
$(TL)_{ik}$	3	$\sigma_\varepsilon^2 + 4 f(TL)$
ε_{ijk}	18	σ_ε^2