Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 1, Issue. 1, December 2012, pg.15 – 20

RESEARCH ARTICLE

Comparative Study of Microarray and Next Generation Sequencing Technologies

Charles Edeki [1]

[1]Ph.D. School of Science and Technology, Department of Information Technology,

American Military University

111 W. Congress Street, Charles Town, WV 25414

[1]charles.edeki@mycampus.apus.edu

Abstract— Huge efforts are being made to develop algorithms and procedures for DNA sequences. The purpose of this study is to expand understanding of how biologists, computational biologists, bioinformaticians, medical practitioners and scientists would benefit from next-generation sequencing and microarray technology in analyzing DNA and protein dataset. Microarrays techniques usage in analyzing biological dataset (gene expression) has grown exponentially for the past two decades. Recently, next generation sequencing technologies are revolutionizing the DNA/RNA sequencing tasks. These highly efficient parallel sequencing methods make it possible to generate billions of bases of sequence per day in a biological laboratory. These methods allow individual human genomes to be sequenced in an instant or one to two days. In this paper, a comparative study of next generation sequencing technology and microarray technology would be presented and the performance of the two techniques would be discussed.

I. INTRODUCTION

Gene expression has been studied since 1961 with the discovery of RNA. The interest in gene expression increased in the 1980s (Ermolaeva et al, 1998). The larger scale studies introduced the development microarray technologies in mid 1990s and were first used in 2000 to study genetic variation (Rockman & Kruglyak, 2006). Northern blot technique has been used in the study of gene expression for a long time. But advances in molecular biology have led to the development of new techniques that are more sensitive and efficient (Roth, 2002). There have also been studies of gene expression and Knock out genes. The study was interested in gene expression changes in knockout mice with the purpose of coming up with the role of mutated protein and identifying particular genes under its control and retrieving candidate genes that may explain the altered phenotype in the mutant animal (Valor & Grant, 2007). Also the other methods that have been used are: subtractive hybridization, differential display, serial analysis of gene expression, and microarray hybridization. All these methods in general has helped researchers to identify gene expression and characteristics of diseases and most importantly the DNA sequence information is not is not required to construct and use DNA microarray.

II. HISTORY OF SEQUENCING TECHNOLOGIES

New sequencing technologies have been introduced because DNA sequencing has been tedious and time consuming in the past. The isolation of DNA, sample preparation, sequence production and analysis had to be done manually. Currently better methods for DNA isolation has been introduced making the work more efficient. Analysis software's have also made it possible to speed up the analysis step of DNA sequences therefore making it possible to do more volume in less time. The very first methods introduced were labor intensive since they had to be done manually and relied on terminal-end labeling, DNase digestion, and 2-D electrophoresis on cellulose acetate and DEAE cellulose paper. Efficiency then increased with the development of Sanger Method in 1977 and the bacteriophage λ in 1982. Sanger Method is accomplished by the use in-vitro

DNA synthesis using terminators. To perform sequencing, double stranded DNA must be converted to single stranded DNA. This is accomplished by denaturing it with NAOH. It requires a primer, DNA polymerase, a template, a mixture of nucleotides, and detection system. Incorporation of dideoxynucleotides into growing strand terminates synthesis and synthesized strand sizes are determined by using gel or capillary electrophoresis. Automated sequencing then followed in 1995 reducing the cost per base pair tremendously. In 1998 ABI 3700 genetic analyzer was introduced allowing 900,000bp per day reducing the cost to 0.1 per bp. Several improvements have been made since then making run times shorter with better quality sequences (Chan, 2005).

III. WHY NEW EXPRESSION TECHNOLOGIES WERE NEEDED?

DNA microarrays are important tool to use while investigating differential gene expression for thousands of genes simultaneously. Currently, the fluorescent dyes Cy3 and Cy5 are most commonly for preparation of labeled cDNA for microarray hybridizations. The raw data are an image files that has to be into gene expression format. The process requires manipulation of data due to differences in variations caused by the differences in the physical and chemical dye characteristics. To accomplish this problem the levels of transcripts are calculated from fluorescence ratios to normalize the fluorescence signals. Since the goal of most microarray applications is to identify differences in transcript levels calculated from fluorescence ratios, it is necessary to normalize fluorescence signals to compensate for systematic variations (Bilban et al). Even though microarray technology is an important tool for researchers to study to study gene expression, there also has been some limitation like equipment cost and equipment limitation which can be the number of probes of an array by the resolution of the scanner used. False positive data can also make it difficult for researchers to publish their findings In addition to that, the older technologies to measure mRNA expression like Northern Blot are labor intensive and can only detect up to dozens of genes (Burges, 2001). With improvement in microarray technology a large amount of data can be generated with current designs producing up to 64,000 different oligonucleotides and with time it will even get better. This is because researchers will have a better understanding on how different genes are expressed and be able to make improvements based on their findings (Macmillan, 2002).

IV. MEASURING EXPRESSION WITH MICROARRAYS

One of the most exciting developments in gene expression studies in the past two decades has been the emergence of DNA microarrays. Previous gene expression technologies were limited to only one or a few genes at a time. Microarrays, on the other hand, can assess several thousand genes at a time under a certain condition. They are also relatively simple and flexible compared to past technologies. While most of the experiments focus mainly on differential gene expression between tissue types such as cancer and health tissue, the technology has been expanded into other fields of biological study. Because of these apparent benefits, many researchers have claimed that microarray technology may be the end-all solution to gene expression studies.

A microarray consists of an array of specific gene sequences printed to a solid support made of glass or plastic. Single-stranded DNA (ssDNA) is deposited in a grid-like fashion across the support. Each spot with ssDNA on the support is called a probe. Probes contain the genes that the researcher is wanting to measure. The array is washed with the ssDNA target sequence. The target sequence is usually generated from a biological sample that the researcher is trying to study through reverse transcription of mRNA. The process produces a complimentary DNA strand (cDNA) that compliments a particular sequence on the array. They are labeled with a fluorescent dye, such as Cy3 or Cy5 and washed over the array. Complimentary sequences hybridize to one another on the array. The excess target solution is washed off the array and the array is scanned by a fluorescent scanner. The intensities for each probe on the array are measured and recorded (Draghici, 2001).

There are two main types of microarrays used today: cDNA microarrays and Affymetrix oligonucleotide microarrays. A cDNA microarray is probed using pins dipped in cDNA solution synthesized through PCR. An Affymetrix array is synthesized directly on the chip using the photolithographic method. Unlike cDNA microarrays the DNA in the probes are made a base at a time. Therefore, the researcher knows the exact sequence in each probe. In cDNA microarrays, the researcher may not know the exact sequences in each probe. While both types have their pros and cons both are useful depending on the type of research being done (Draghici, 2001).

V. APPLICATIONS OF MICROARRAYS

Today, the microarray technology is used widely in several areas of biological study. While microarrays only measure one thing, the expression levels of certain genes, it is very flexible in what it can study. Many of the

major diseases today have unique gene profiles that can be viewed using microarrays. One of the earliest, and widest, uses of microarrays is the profiling of cancerous tissue. Profiling diseased tissue allows researchers to have a better understanding tumors and what markers certain tumors may have. For example, one study by Sorlie et al. found gene expression profiles that could be used as prognostic markers for overall and relapse-free survival in breast cancer tissue (Sorlie et al, 2001). Identifying gene profiles for tumors allow researchers to better group and classify certain tumors. This allows development of better drug and treatment procedures because certain classes of tumors may be more treatable with certain drugs than others. While cancer studies receive the most attention they are also useful in the study of complex biochemical pathways under certain conditions by allowing researchers to study how gene expression is affected by signaling events. In recent years, research studying microbial or virus infected tissue has been the most successful. Most notably, they have been used to decipher the underlying mechanisms and interactions between host and microbe. For example, one study showed how certain bacteria in humans can effect gene expression in the intestines. It was found that it can affect certain important functions within the intestine (Shoemaker & Linsley, 2002). Other experiments have focused on the efficacy of antimicrobial drugs and how they affect gene expression in the microbes.

While most of the attention has been paid to gene expression profiling they are often used in other biological areas as well. In recent years, they have been extended into protein interaction profiling, toxicology, evolutionary biology, disease characterization, physiology and stress responses, and many others (Coppee, 2008). For example, microarrays have been very useful in the understanding of the signaling pathways in plants used for plant defense and how certain environmental conditions change these pathways (Reymond, 2000). They have also been useful in parasitology. For example, microarrays were used to study the gene profiles of *I. multifillis*, a fresh water fish parasite. The study successfully identified differentially expressed genes between the three major lifecycles of the parasite (Abernathy et al, 2011).

VI. LIMITATIONS OF MICROARRAYS

While microarrays are widely used and have showed great promise over the years, they do have their limitations. One of the major limitations is the amount of noise produced in the experiment. Sometimes, they are so noisy that researchers can do two experiments with the same samples and same materials and obtain different results. There are several reasons for the high level of noise, including non-specific hybridization and differences in temperature, labeling, humidity, the amount of the target, the equipment, etc. In fact, every step and procedure in the experiment introduces some level of noise. One major consequence of this varying amount of noise is that two different experiments may not be easily compared. Each experiment may use different equipment and procedures, and therefore, have a varying amount of noise. In order to combat this, microarray experiments must be normalized to account for this noise difference. However, this is not always an easy task. There is not a general consensus on how exactly normalization should be done. Normalization procedures tend to be different from experiment to experiment because each one was administered in a different way (Draghici, 2001).

Another limitation is the overall cost of the experiment. While every experiment is different many experiments require several arrays to be administered with several different samples. This could potentially increase the cost of the experiment substantially. Also, microarray experiments yield a large amount of complex data that can only be normalized and analyzed through computer programs. Therefore, experiments may require a large amount of technical skill to do.

VII. NEXT GENERATION SEQUENCING

Next generation sequencing (NGS) represents a new breed of sequencing technologies that offer unprecedented high-throughput and low cost sequencing platforms. The technologies are characterized by an ability to process millions of sequences reads in parallel, higher yields and low cost per read. The development of technology was triggered by a need for a higher sequencing throughput as highlighted by the whole human genome sequencing project. It was also motivated by the limitations with existing capillary-based sequencing methods.

NGS was not relatively used until after 2005. Before, the best technology was shotgun sequencing called Sanger Methodology. The most common NGS platforms fall in two categories:

- Second- generation sequencing technologies such as Genome Sequencer from Roch 454 life sciences, the Solexa Genomer Analyzer from Illumina, SOLiD Sequencer from Applied Biosystems, Commercialized Polonator, Heliscope from Helicos; and
- Third - generation sequencing technologies such as Single Moledule Real Time by Pacific Biosciences and a new technology by Visigen Biotechnologies Inc.

These platforms share characteristics such as highly parallel operations, higher yield, simpler operations, much lower cost per read, and to their disadvantage much shorter reads. They differ in characteristics such as read length, number of DNA molecules sequenced in parallel, raw accuracy, sequencing chemistry and a few other attributes shown in table below (Miller et al, 2010).

The following table summarizes the differences and similarities between the common NGS platforms

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330[a]	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome de novo assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA$_x$	Frag, MP/ solid-phase	RT	75 or 100	4[a], 9[b]	18[a], 35[b]	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7[a], 14[b]	30[a], 50[b]	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G007	MP only/ emPCR	Non cleavable probe SBL	26	5[b]	12[b]	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RT	32*	8[a]	37[a]	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

*Average read-lengths. [a]Fragment run. [b]Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

Table 1 Differences and Similarities between the common NGS platforms

NGS has transformed the way we study biological problems and made a number of sequencing projects and applications feasible. These include whole-genome sequencing, targeted re-sequencing, discovery of transcription factor binding sites, noncoding RNA expression profiling, mutation discovery, defining DNA-protein interactions, and enabling metagenomics (Mardis, 2009). NGS is also being used to confirm and extend associations between single nucleotide polymorphisms. The technologies are extremely powerful for studying genetic variation and have proven useful in finding gaps between phenotypes and genotypes in a population of study (Pettersson et al, 2009). Another significant application of NGS is in gene expression analysis where the technology is poised to challenge microarrays as the primary tool for gene expression studies or in some cases as a complementary technology (Matsumura et al, 2010).

VIII. NGS AND GENE EXPRESSION

NGS technologies provide the ability to study hundreds or even thousands of gene loci simultaneously and so facilitate performance of applications that require many gigabases of sequences to be examined at the same time. NGS technologies produce a digital record of the numerical frequency of a sequence in a sample which is tends to be a better approximation of actual transcript content in a sample. Gene expression studies with NGS involve sequencing-based transcriptome analysis which in many ways is superior to microarrays because the sequencing-based method is digital, highly accurate, and easy-to-perform (Willenbrock et al, 2009).

Some platforms like SOLiD System offer hypothesis-neutral analysis differential gene expression on a genome-wide scale and so are the platform of choice for discovery including novel transcript discovery and splice variant discovery because they offer scalability, sensitivity, high accuracy and a broad dynamic range" (Willenbrock et al, 2009). The use of NGS in gene expression analysis has catalyzed development of techniques

like Digital Gene Expression TAG (DGE-TAB), DeepSAGE and RNA-Seq which has demonstrated increased sensitivity, specificity and accuracy compared to microarray platforms (Matsumura et al, 2010).

IX. NGS STRENGTHS AND LIMITATIONS

The strength of NGS platforms lies in their ability to lay millions of DNA fragments on a single chip and then simultaneously process the fragments. This increases throughput and permits a larger dynamic range of input data. NGS technologies have greater sensitivity, specificity and accuracy than microarrays.

Unlike microarrays which are considered a 'closed system' because they can only account for sequences that are targeted by probes on the array, NGS technologies represent an 'open' system suitable for cataloguing gene diversity (including discovery of novel gene diversity), without a priori sequence information". NGS permits ultra-deep sequencing with massive DNA sequence reads from amplified PCR products. This allows insight into diversity of data (Ron et al, 2010). Other advantages of NGS technologies include lower background, better sensitivity and ability to provide quantitative measurements (University Health Network Microarray Centre, 2009).

Most NGS technologies suffer from short read-lengths & lower sequence quality. With the exception of 454, all existing platforms produce read lengths of 50-100bp. Shorter read lengths tend to complicate interpretation of results in some instances. Shorter read lengths also translate to short tag reads (21 bp) in DGE-TAG technique which sometimes leaves tag-to-gene annotation more difficult. For RNA-Seq, it means there is a requirement for large amount of sequence reads to fully cover the dynamic range and to provide a truly quantitative gene expression profiling. So in such cases, a reliable protocol of tag-based gene expression profiling based on sequencing of longer tag fragments is highly desirable (Matsumura et al, 2010).

Another limitation to NGS is that the number of samples processed in a single run is determined by the physical partitioning or sample-specific barcoding approach utilized (Ron et al, 2010). Each instrument run may take days to weeks. Also, complex sample preparation is involved in NGS systems. Most NGS technologies are highly dependent on PCR based amplification which makes them prone to bias introduced during the amplification step (Willenbrock, 2009).

NGS has limited bioinformatics to handle the large amounts of data efficiently (University Health Network Microarray Centre, 2009). The analysis of the resulting dataset requires sophisticated computer systems, bioinformatics tools and a significant time contribution (Ron et al, 2010). Although the overall costs of NGS systems has been falling, there are substantial initial costs associated with the technology such as massive information infrastructure for data storage, data transfer, data analysis, and training of personnel (University Health Network Microarray Centre, 2009).

X. CONCLUSION

Results of this comparative study were cleared that both microarray and next generation sequencing technologies have their benefits and limitations. The major limitation in microarray technology is the high noise produced during the experiment, this leads an exact experiment with the same attributes or datasets can yield different result. The overall cost of microarray experiment is high comparing to next generation sequencing technology.

The next generation sequencing platforms are advancing quickly with large increase in read-length and reduction in cost comparing to microarray technology. Another benefit of next generation sequencing technology over microarray is the base pair accuracy in the evaluation of transcriptional start and stop sites. The major limitation of next generation sequencing is the capability of short read-length, roughly between 50-100bp, but scientist are working in improving the technology, the read-length is increasing. In summary the microarray and next generation sequence technologies have their benefits and limitation, selection of each technique will depend on the researcher and the experiment.

REFERENCES

[1] Abernathy, J., Xu, D.H., Peatman, E., Kucuktas, H., Klesius, P., Liu, Z. (2011). Gene expression profiling of a fish parasite ichthyophthirius multifiliis: insights into development and senescence-asssociated avirulence. *Comparative Biochemistry and Physiology, 6*(D), 382-392. doi:10.1016/j.cbd.2011.08.003

[2] Bilban, M., Buehler, L. K., Head, S., Desoye, G., Quaranta, V. (2002). Normalizing DNA Microarrays Data. *Mol. Biol, 4,* 57-64

[3] Burges, J. K. (2001). Gene Expression Studies using Microarrays. *Clinical and Experimental Pharmacology and Physiology 28,321-328.*

[4] Chan, E. Y. (2005). Advances in sequencing technology. *Mutation Research 573, 13-40*. doi:10.1016/j.mrfmmm.2005.01.004

[5] Coppee, J.Y. (2008). Do dna microarrays have their future behind them? Microbes and Infection, 10, 1067-1071. doi:10.1016/j.micinf.2008.07.003

[6] Draghici, S. (2001). Data analysis tools for dna microarrays. Boca Raton, FL: Chapman & Hall/CRC.

[7] Ermolaeva, O., Rastogi, M., Pruitt, K. D., Schuler, G. D., Bittner, M. L., Chen, Y., Simon, R., Meltzer, P., Trent, J. M., Boguski, M. S. (1998). Data management and analysis for gene expression. Nature America Inc.

[8] Macmillan magazine Ltd (2002). Microarray technology an array of opportunities. Retrieved from www.nature.com

[9] Mardis, E.R. (2007). The impact of next-generation sequencing technology on genetics. Cell Press, 133-141. doi:10.1016/j.tig.2007.12.007

[10] Matsumura, H., Yoshida, K., Luo,S., Kimura, E., Fujibe, T., Albertyn, Z., Barrero, R.A., et al. (2010). High-Throughput SuperSAGE for digital gene expression analysis of multiple samples using Next Generation Sequencing.
Plos One,5(8), 1-8, Retrieved from http://www.plosone.org/article/info:doi/10.1371/journal.pone.0012010

[11] Metzker, M. (2010). Comparison of next-generation sequencing platforms. Nature Reviews Genetics, 11, 31-46, doi:10.1038/nrg2626

[12] Miller, J.R., Koren, S., Granger, S. (2010). Assembly algorithms for next-generation sequencing data. Genomics, 95, 315-327, doi:10.1016/j.ygeno.2010.03.001.

[13] Pettersson, E., Lunderberg, J., Ahmadian, A. (2009). Generations of sequencing technologies. Genomics. 93, 105-11. doi:10:1016/j.ygeno.2008.10.003

[14] Reymond, P. (2000). Dna microarrays and plant defence. Plant Physiology and Biochemistry, 39, 313-321. doi:10.1016/S0981-9428(00)01235-3

[15] Rockman, M. V., Leonid Kruglyak, L. (2006). Genetics of global gene expression. Retrieved from www.nature.com/reviews/genetics

[16] Roh, S.W., Guy C.J. Abell, G.C.J., Kim, K., Nam, Y., & Bae, J.(2010). Comparing microarrays and next generation sequencing technologies for microbial ecology research. Trends in Biotechnology 28(2010) 291-299, doi:10.1016/j.tibtech.2010.03.001

[17] Roth, C. M. (2002). Quantifying Gene Expression.Mol. Biol.4, 93-100.

[18] Shoemaker, D.D. & Linsley, P.S. (2002). Recent developments in dna microarrays. Current Opinions in Microbiology, 5, 334-337. doi:10.1016/S1369-5274(02)00327-2

[19] University Health Network Microarray Centre (2009). Next-generation sequencing: Synergy with microarrays. Retrieved from http://www.microarrays.ca/info/2009Jan_NGS_SynergyWithArrays.pdf

[20] Valor, L.M. & Grant, N.G.S. (2007). Clustered Gene Expression Changes Flank Targeted Gene Loci in knockout Mice. PLoS ONE, 2(12); e1303. Retrieved from www.ploseone.org.

[21] Willenbrock, H., Salomon, J., Søkilde1, R., Barken, K.B., Hansen, T.N., Nielsen, F.C., Søren Møller, S., Litman, T. (2009). Quantitative miRNA expression analysis: Comparing microarrays with next-generation sequencing. RNA, 15, 2028-2034, doi: 10.1261/rna.1699809

